School of Economics and Finance

Credit Decision Support Methodology for Micro, Small and Medium Enterprises (MSMEs): Indonesian Cases

Novita Ikasari

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

June 2014

DECLARATION

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: 

Date: 23/6/2014

ACKNOWLEDGMENTS

Glory be to the Father, and to the Son, and to the Holy Ghost. As it was in the beginning, both now, and always, and to the ages of ages.

The path to thesis completion was dauntingly dark, but the way was never lost thanks to my present and past supervisors, Dr Tom Cronje, Dr Fedja Hadzic, and Professor Tharam S. Dillon; university members; and a collective of generous people whom I am honoured to have worked with and been befriended by.

Dr Fedja Hadzic has been instrumental to this research from the beginning. I am grateful for the opportunity to share his passion for data mining and his firm stance on academic integrity. On each collaboration, challenges were met and overcame. For the long hours of discussions, constructive feedback on many drafts and solutions to unexpected problems, I am indebted. An Indonesian proverb best expresses my appreciation: *hutang emas dapat dibayar, hutang budi dibawa mati*.

Becoming involved at a late stage of my research, Dr Tom Cronje imparted more knowledge than anyone should be capable of in such limited time. His systematic thinking and attention to detail, which this thesis has much benefited from, are things that I am thankful for. His “two-step-ahead” thinking was a delightful challenge in supervisory meetings.

I am fortunate to have been supervised by Professor Dillon in the early course of this research, where his professional experience contributed to its initial direction.

I could not thank enough each one of the following great individuals: Professor Kate Wright and Mrs Chris Kerin, for their unwavering support and for patiently clearing every barrier that existed. Associate Professor Fay Rola-Rubzen, Dr Tadayuki Mayamoto and Dr John Fielder for their time and encouraging words. I have learned to smile at the many facets of life from Lainey Weiser. Thanks to Annette Watkins, Chris Bright, and CIB100 tutors batch 2010–2013 for allowing me to be a part of such a professional team. To all members of the night club at EU4, thank you for your comradeship. To Farida, Sasipa, Ilham and Wilda, thank you for patiently listening to my life emergencies and showing what simple faith can do.

This research would have not been completed without the financial and moral support of DIKTI, University of Indonesia, and Curtin University. I am grateful for Dr Tafsir Nurchamid, Ak., M.Si, who went beyond his call of duty to support my data collection and to all loan staff who gave their time and effort for this research.

To my family, who dedicates their life to being human in the eyes of God and man, this thesis is a small token for their determination. Thank you to everyone: to the rock who keeps everyone on the ground, my mother, Juliana Sembiring; to my brothers Ichwan and Salomo, my sister Hana, who believes there is a way for everything when the mind seeks it, and my nephew Castra, with the hope that he stays gold. To my late father, Gustaf Adolf Panggabean, always.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Actionable Knowledge Discovery	AKD
Analytic Hierarchy Process	AHP
Artificial Intelligence	AI
Artificial Neural Network	ANN
Asian Financial Crisis	AFC
Back-propagation Neural Network	BPNN
Back-propagation Neural Networks	BPNN
Bank for International Settlements	BIS
Bank Indonesia	BI
Bank Negara Indonesia	BNI
Bank Rakyat Indonesia	BRI
Bank Tabungan Negara	BTN
Bantuan Likuiditas Bank Indonesia	BLBI
Capital Adequacy Ratios	CAR
Case-Based Reasoning	CBR
Classification and Regression Tree	CART
Credit Risk Assessment	CRA
Data Envelopment Analysis – Discriminant Analysis	DEA-DA
Database Structure Model	DSM
Debtor Information System	DIS
Decision Tree	DT
Diagonal Linear Discrimination Analysis	DLDA
Diagonal Quadratic Discrimination Analysis	DQDA
Discriminant Analysis	DA
Domain-Driven Data Mining	D^3M
Earnings Before Interest, Tax, Depreciation and Amortization	EBITDA
eXtensible Markup Language	XML
Factor Analysis	FA
Fuzzy Adaptive Resonance	FAR
Fuzzy-Support Vector Machine	F-SVM
Genetic Algorithm	GA
Genetic Programming	GP
Global Financial Crisis	GFC
Gross Domestic Product	GDP
Indonesian Rupiah	IDR
International Monetary Funds	IMF
k-Nearest Neighbour	k-NN
Knowledge Discovery in Databases	KDD
Kredit Ketahanan Pangan dan Energi	KKPE
Kredit Pengembangan Energi Nabati dan Revitalisasi Perkebunan	KPEN-RP
Kredit Usaha Pembibitan Sapi	KUPS
Kredit Usaha Rakyat	KUR
Learning Vector Quantization	LVQ

Least Square Support Vector Machine	LS-SVM
Lembaga Penjaminan Simpanan	LPS
Linear Discriminant Analysis	LDA
Linear Programming	LP
Linear Regression	LR
Loan Deposit Ratio	LDR
Logistic Regression	Logit
Micro, Small and Medium Enterprise(s)	MSME(s)
Mixture of Experts	MOE
Multilayer Perceptron	MLP
Multiple Criteria Liner Programming	MCLP
Multiple Criteria Non-Linear Programming	MCNLP
Multiple Discriminant Analysis	MDA
Multiple Discriminant Analysis	MDA
Multivariate Adaptive Regression Splines	MARS
Natural Language Processing	NLP
Neural Network	NN
Non-Performing Loan(s)	NPL(s)
Ordinal Logistic Regression	OLR
Ordinary Least Squares	OLS
Principal Component Analysis	PCA
Probabilistic NN	PNN
Probit Regression	PR
Quadratic Discriminant Analysis	QDA
Radial Basis Function	RBF
Random Survival Forest	RSF
Recursive Feature Elimination SVM	SVM-RFE
Recursive Partitioning Algorithm	RPA
Repeated Incremental Pruning to Produce Error Reduction	RIPPER
Return on Assets	ROA
Robert Morris Associates	RMA
Rough Sets	RS
Small and Medium Enterprise(s)	SME(s)
Small Business Administration	SBA
Strengths, Weaknesses, Opportunities, Threats	SWOT
Support Vector Machine	SVM
Two-stage Genetic Programming	2GP
United States of America	USA
University of California, Irvine	UCI
US Dollar	USD
Weight of Evidence	WoE
World Wide Web	WWW

LIST OF PUBLICATIONS

Book Chapters:

Ikasari, N., F. Hadzic and T. S. Dillon. 2011. Incorporating qualitative information for credit risk assessment through frequent subtree mining for XML. In XML Data mining: Models, method, and applications, ed. A. Tagarelli, 467–503. Hershey, PA: IGI Global.

Ikasari, N. and F. Hadzic. 2013. Structured Data Mining for Micro Loan Performance Prediction: The Case of Indonesian Rural Bank. In IAENG Transactions on Engineering Technologies, eds. G-C. Yang, S-L. Ao and L. Gelman, 641-653. Berlin: Springer-Verlag.

Conference Papers:

Ikasari, N. and F. Hadzic. 2012. Assessment of Micro Loan Payment Using Structured Data Mining Techniques: The Case of Indonesian People's Credit Bank. In *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2012*, vol.1, *WCE 2012, 4-6 July 2012, London, U.K.* (Best Paper Award)

Ikasari, N. and F. Hadzic. 2012. An assessment on loan performance from combined quantitative and qualitative data in XML. In *Discovery Science, Lecture Notes in Artificial Intelligence vol*. 7569, eds. J-G. Ganascia, P. Lenca and J-M. Petit, 268–283. Berlin: Springer-Verlag.

ABSTRACT

This research is aimed at constructing an objective and more accurate credit risk assessment method for MSMEs than is applied by banks in Indonesia at present. In providing finance to Micro, Small and Medium Enterprises (MSMEs), Indonesian banks apply different lending methods and collate quantitative (structured) and descriptive qualitative (unstructured text) data for assessing credit risk of all loan applicants. Their credit data are applied by way of manual discretionary credit assessment processes and/or credit scoring based on the 5Cs lending principles. A limited number of studies in this field demonstrate that the assessment can be enhanced by the use of statistical and machine learning techniques to distinguish good and bad loan applications; however, research about the development of comprehensive models that include text-based qualitative information to improve credit assessment for MSME has not been conducted. The current credit assessment methodology applied in Indonesia and the credit risk emanating from it serves as the principal reason for this research.

Three Indonesian banks provided credit risk data sets pertaining to existing performing and non-performing loans with them for this research. The eXtensible Markup Language (XML) was applied to these data sets to present quantitative and qualitative credit data in tandem and optimize the underlying information for analysis purposes. Based on this tree-structured credit data, a general structure (DSM) was extracted to create flat data; a flat format allows a more comprehensive use of data mining techniques. This research applied Decision Tree (DT) and Rule Induction (RI) to predict performing and non-performing loans, and extracted useful patterns from the credit data sets. The Forward Feature Selection technique was applied to financial and qualitative credit data to construct prominent credit parameters for MSMEs.

The results show that the prediction accuracy of DT and RI is influenced by the proportion of credit data sets with prediction accuracy improving when selected credit attributes are used. Qualitative attributes are predominantly selected across credit data sets. From a lending method perspective, this confirms that qualitative information adds value to credit risk assessment for MSMEs compared to the exclusive use of a scoring system. The knowledge discovery from the application of DT and RI provides a reduction in the misclassification of performing and non-performing loans. With a dynamic credit assessment system in place, banks are equipped with a more reliable, informed credit-granting decision process.

CHAPTER 1 - INTRODUCTION

1.1 Background

In every economy, the banking system plays an important role as financial intermediary, stimulating local business economic activities and driving the country's monetary stability. The Global Financial Crisis (GFC) that began in 2007 is testament to the significant need for a vigilant banking industry, particularly in a free enterprise type of market such as that of the United States of America (USA). The lack of rigorous governance over the shadow banking system (Gorton et al. 2010) launched the country's economy into depression. The crisis triggered a wave of recession across European countries, with Iceland the first to experience difficulty in meeting its short-term financial commitments (Sherwood 2008), followed by Greece and Ireland (Adrick 2010). The domino effect of the GFC has been more obvious in developed than in developing countries. This is in part the result of the crisis being triggered by widely-connected developed economy, as developing countries had improved their monetary backbone after the lessons learnt from the 1996 Asian Financial Crisis (AFC) (Naudé 2009). With this in mind, this research is focused on the Indonesian banking system: in particular, on how Indonesian banks perform their intermediary role with Micro, Small and Medium Enterprises (MSMEs).

Situated at the heart of international trade and business, Indonesia benefits from its strategic position and its rich natural resources; nevertheless, the economy has been unstable for more than three decades. Propelled by the oil boom in the 1970s, Indonesia became one of the forces behind Asia's meteoric economic ascent in the early 1980s. This era was typified by a steep rise in economic growth, a healthy Gross Domestic Product (GDP) surplus, relatively moderate inflation rates and high saving rates, in addition to investment and strong foreign direct investment (Berry, Rodriguez, and Sandee 2002; Dowling and Yap 2008; Karunaratne 1999; Kenward 1999). By the late 1980s, the Indonesian Government was very confident about the role of private capital inflow (Kenward 1999) and reformed the banking industry through various pro-business legal products, providing leeway for the establishment of banks, particularly privately owned ones. Unfortunately, this increase in the number of banks was not accompanied by the necessary supervision and control. In less than ten years the banking industry plummeted, with the high number of Non-Performing Loans (NPLs) and negative Capital Adequacy Ratio (CAR) (Dowling and Yap 2008) contributing to the economic downfall of 1996. The defaulted loans in general belonged to Indonesian large-sized and high profile companies with significant foreign exchange risk exposure.

Following the 1996 AFC the Indonesian Government concentrated on ways to strengthen the economic fundamentals, including revitalization of the banking industry and

promotion of MSMEs. On the banking side, a rationalization of the number of operating banks took place, prudent banking practice was enforced and a new and a stronger authoritative role for the central bank were mandated. For MSMEs, the Government established a ministry responsible for ensuring the sustainability of MSMEs in Indonesia, developed a credit scheme sensitive to MSMEs' needs and initiated development policies geared toward improvement of MSME business agility.

1.2 Types of banks operating in Indonesia

The Central Bank of Indonesia (Bank Indonesia, BI) categorized banks according to their operations as commercial banks and rural banks (Bank Perkreditan Rakyat, BPR) (Undang-Undang Republik Indonesia 1998). Aside from categorization specified by law, BI grouped commercial banks according to their ownership and activities for reporting purposes. These groupings are state-owned banks, private forex commercial banks, private non-forex commercial banks, regional development banks, joint venture banks, foreign banks and sharia commercial banks.

Commercial banks have a wider scope of banking activities than BPRs, performing conventional and sharia-based banking activities and providing payment transaction services such as interbank payment transfers. According to the Banking Law Number 10 (Undang-Undang Republik Indonesia 1998), commercial banks are entitled to execute any of the following activities: fund mobilization; providing loans; notes issuance; securities purchase-selling-guarantees; money transfers; funds placements; providing safe deposit boxes; custodial activities; factoring; credit card and trusteeship services; financing; other business that is aligned with prevailing laws; foreign exchange transactions; equity participation; and pension fund management. Sharia-based commercial banks carry out these activities under their own stipulations. Commercial banks are forbidden from carrying out insurance business. In terms of ownership, commercial banks can be owned by Indonesian citizens or legal entities and a joint operation between Indonesian citizens or legal entities and foreign legal entities, with respect to the 2012 legislation on capital requirement. A recent ruling from BI requires that commercial banks incapable of maintaining a minimum capital reserve of IDR 100 billion (ca. USD 10,397,171 with 1 USD = 9,618 IDR) converted into a BPR, with associated changes in business operations (Surat Edaran Bank Indonesia 2010b).

State-owned commercial banks are those in which the shares are fully or partially owned by the government of the Republic of Indonesia; they operate in a similar way to private commercial banks with the right to carry out foreign exchange transactions. There are four state-owned banks operating in Indonesia, P.T. Bank Negara Indonesia (Persero) Tbk (BNI), P.T. Bank Rakyat Indonesia (Persero) Tbk (BRI), P.T. Bank Tabungan Negara (Persero) Tbk (BTN) and P.T. Bank Mandiri (Persero) Tbk (Mandiri). All four banks are

strong performers in the banking industry and are in fierce competition with each other (Loan staff, Bank B, pers.comm).

Private forex commercial banks are owned by private Indonesian citizens/legal entities with the right to perform foreign exchange transactions directly such as payments, deposits, credits and other financial services in foreign currency. Essentially, these banks have international scope and are more exposed to exchange rate risk.

Private non-forex commercial banks are owned by private Indonesian citizens/legal entities, and do not have the right to perform foreign exchange transactions. This means the banks are operating within the domestic scope of business and are not directly exposed to exchange rate risk. Non-forex banks are able to upgrade their status to forex commercial banks subject to BI's assessment and approval.

Regional development banks are those whose shares are fully or partially owned by provincial government. These banks provide commercial banking services without the right to execute foreign exchange transactions and mainly operate within the national boundary.

Joint venture banks are established and owned by one or more Indonesian-based or -owned commercial banks and one or more foreign-based banks.

Foreign banks operate as branch offices of banks registered and located outside Indonesia. These banks operate commercially and are allowed to open branches in all regions across Indonesia.

Sharia commercial banks provide banking services based on sharia principles throughout their chain of business. These banks, and sharia units, abide by separate banking regulations and accounting practice. In its monthly statistics, BI reports sharia bank business performance separately.

BPRs are banks with banking activities limited mainly to taking saving deposits and providing finance (loans). These banks perform conventional and sharia-based banking activities but do not provide payment transactions services such as interbank payment transfers. According to Banking Law Number 10 (Undang-Undang Republik Indonesia 1998), BPRs are entitled to execute any of the following activities: fund mobilization, providing loans and fund placements; they are forbidden from carrying out any transactions related to demand deposits, foreign exchange transactions or business, equity participation or insurance business. They can only be established and owned by Indonesian citizens or legal entities, and are fully owned by Indonesian citizens or regional government or in joint ownership among these parties. Paid-up capital for establishment of a BPR is based on different requirements for four Indonesian regions, with the highest requirement of IDR 5 billion (ca. USD 521,648) for Jakarta.

1.3 An overview of the Indonesian banking industry

The Indonesian banking industry has gone through several development stages, the most critical during the 1988–1998 periods. Two studies performed in 2003, by Hamada (2003) and Enoch et al. (2003), suggest different categorizations of these stages, based on the distinct purpose of each study.

Hamada (2003) aimed at providing a comprehensive analysis of the development of Indonesia's financial sector from 1966 to 2003. The study discussed the role of banks as the country's financial instrument, and the manner in which the sector responded to economic challenges. In this study the development of the banking industry was divided into five stages: 1966–1972 (Formative Period), 1973–1982 (Policy Based Finance), 1983–1991 (Financial Reform), 1992–1997 (Period of Expansion) and 1998–2003 (Period of Financial Restructuring).

Enoch et al. (2003) aimed at providing a comprehensive analysis of how the industry reacted to the 1996 AFC. Development was categorized into seven stages covering the period 1988–1999: 1988–August 1997 (Unbalanced Liberization), October–November 1997 (Contained Banking Difficulties), December 1997 (Losing Control), January–February 1998 (Laying the Ground for Stabilization), March–May 1998 (First Initiatives and New Shock), June–September 1998 (Design of a Comprehensive Strategy) and October 1998–December 1999 (Indecisiveness Increases Costs).

In this thesis, an overview of the banking industry focuses on the period 1966–2011, providing a general overview of the prominence of different bank types, their intermediary activities, and important occurrences that resulted in changes to the banking industry within the context of public finance policy. This serves to contextualize the research applicable to this thesis. The stages used for the overview of the banking sector are: 1966–1981 (Supremacy of State-owned Banks), 1982–1997 (Supremacy of Privately–owned Banks) and 1998–present (Supremacy of Prudential Banking).

1.3.1 Stage 1: Supremacy of state-owned banks (1966 to 1981)

In the 16 years from 1966 to 1981, the lending market was dominated by a small number of state-owned commercial banks. Table 1.1 shows the lending market share of different commercial banks in Indonesia during this period. The difference between summation of the lending ratios and the total of 100% indicates the amount of lending provided by BI (Hamada 2003). This was channelled to various government institutions and infrastructural projects across the country, escalating the value of the real economy. Funds for development were sourced via both international borrowing from the International Monetary Fund (IMF) and domestic capital inflow. The surge of domestic capital inflow resulted from an increase in nominal deposit interest rates applied by state-owned banks. It

reached the highest rate of 72% per annum for time deposits in 1968 (Cole and Slade 1998b). The same trend did not apply to nominal lending interest rates: in 1968 lending interest rates were capped at 60% per annum, and the negative spread between deposit and lending interest rates were borne by BI.

Midway during this period, in 1973, the world oil price soared and accelerated the economy further. In order to bring down domestic capital inflow and expansion of credit, the government enforced credit ceilings and decreased deposit nominal interest rates until they were reduced to 9% per annum for time deposits in 1981 (Cole and Slade 1998b). In 1975, the state-owned banks showed a positive response to the credit ceiling when the lending ratio decreased from 72.2% to 58.2%. Since this policy was not applied to privately-owned banks, capital inflow was not aggressive, and credit outflows from these banks were stable although the number of banks decreased significantly, by around 40%: from 126 in 1972 to 75 in 1981. These banks were mainly involved in regional and international lending, supported by Indonesia's open capital economic policy account during this period (Cole and Slade 1998a). Simultaneously, businesses established accounts with banks in foreign countries and used them as part of their business financing scheme. The 1970 change from a floating to a fixed exchange rate system supported the international banking and lending. By the late 1970s, small businesses mostly dealt with informal and non-formal financial institutions providing them with the required financial support (Patten and Rosengard 1990); big and small businesses were thus served by different types of financial institutions.

Table 1.1 Total lending market share based on ownership of banks, 1966–1981

Year	State-owned Banks		Privately-owned Banks		Regional Government Banks		Foreign/Joint Venture Banks		Number of Banks
	Lending ratio (%)	Number of Banks	Lending ratio (%)	Number of Banks	Lending ratio (%)	Number of Banks	Lending ratio (%)	Number of Banks	
1966	57.50	n.a.	18.7	n.a.	0.00	n.a.	0.00	n.a.	n.a.
1967	45.80	7	15.4	121	0.00	23	0.00	n.a.	n.a.
1968	56.20	7	7.00	122	0.00	23	1.00	8	160
1969	56.20	7	7.00	122	0.00	23	1.00	11	163
1970	64.30	7	6.80	126	0.00	25	2.30	11	169
1971	69.30	7	6.60	129	0.00	25	3.20	11	172
1972	70.00	7	6.60	126	0.00	26	4.00	11	170
1973	72.80	7	6.70	114	0.00	26	5.10	11	158
1974	72.20	7	5.70	107	0.00	26	7.40	11	151
1975	58.20	7	4.80	97	0.00	26	4.40	11	141
1976	56.30	7	5.50	91	0.00	26	4.20	11	135
1977	56.80	7	6.40	85	1.30	26	4.60	11	129
1978	52.50	7	5.50	83	1.20	26	4.90	11	127
1979	52.20	7	6.50	78	1.40	26	5.50	11	122
1980	54.60	7	7.20	76	1.80	26	5.30	11	120
1981	57.90	7	8.20	75	2.40	26	5.40	11	119

Source: Miki Hamada, Transformation of the Financial Sector in Indonesia (Institute of Developing Economies [IDE-JETRO]), p. 7.

Note: The difference between the total lending ratio and 100% represents the portion of the loans provided by BI.

1.3.2 Stage 2: Supremacy of privately owned banks (1982–1997)

The fall in the world oil price in 1982 forced the government to find other sources to maintain Indonesia's balanced budget system. In order to stimulate the economy to overcome the dismal performance of the oil price, the government turned to the banking sector. The first banking reform package was enacted in June 1983. It entailed the abolishment of credit ceilings for all banks and of interest rate ceilings for state-owned banks. The deposit nominal interest rates for state-owned banks were around 17% to 18% per annum during 1983–1988. It is evident from Table 1.2 that the reform successfully encouraged privately-owned banks to take an active part in lending activities, resulting in a consistent increase in their lending market share. Over this same period, state-owned banks, regional government banks and foreign banks experienced a relatively stable lending market share.

In 1988 another banking reform package was issued to encourage the open market system and credit expansion. The most notable directive of the package was the elimination of restrictions on the establishment of new privately owned banks and branches. Privately-owned banks were given licences to operate with IDR 10 billion (USD 104,866, with 1 USD = 9,535.99 IDR) paid-up capital. This resulted in a 38% increase in the number of banks in just over one year. Up to the time of the AFC in 1996, the number of banks constantly increased despite the comparative modest rise in their lending ratio. A similar trend occurred for foreign banks, since a comparable general policy applied to their establishment. The government added two other preconditions for new joint venture banks: firstly, that a foreign bank had to be in partnership with an Indonesian domestic privately owned bank before a license to operate was issued; and secondly, that the joint-venture banks had to provide 50% of their total loans (amount of loans) to businesses involved in exports. This reform package did not affect the number of state-owned banks and regional development banks.

During this stage, banks focused on providing loans to either government projects or to businesses that were connected to powerful and influential figures (Grenville 2004). In addition, lending to acquaintances or families of shareholders was common practice. Realizing the peril of unsupervised banking practices, the government enacted a package of banking regulations in 1991 covering CAR (minimum of 8% by 1993), Loan Deposit Ratio (LDR), net open position for foreign banks, group lending limits and a score system to assess banks' health (Cole and Slade 1998b). In 1992, government also encouraged banks to provide 20% of total loans to Small and Medium Enterprises (SMEs), and made this a bonus point in the score system (Hamada 2003). According to Takeda-Hamada in (Hamada 2003), this presumably led banks to make superficial attempts to realize this requirement. This portion of loan allocation increased to 22.5% in 1997 (Peraturan Bank Indonesia 2001).

Table 1.2 Total lending market share based on ownership of banks, 1982–1997

Year	State-owned Banks		Privately-owned Banks		Regional Government Banks		Foreign/Joint Venture Banks		Number of Banks
	Lending ratio (%)	Number of Banks	Lending ratio (%)	Number of Banks	Lending ratio (%)	Number of Banks	Lending ratio (%)	Number of Banks	
1982	61.70	7	9.20	71	2.70	26	5.10	11	115
1983	64.00	7	12.30	70	2.70	27	5.60	11	115
1984	70.90	7	16.20	69	2.70	27	5.60	11	114
1985	69.40	7	18.50	69	2.90	27	4.80	11	114
1986	67.40	7	20.90	68	2.90	27	4.60	11	114
1987	66.00	7	22.70	67	2.90	27	4.30	11	112
1988	65.10	7	24.30	66	2.70	27	4.30	11	111
1989	62.20	7	29.20	91	2.60	27	4.90	23	148
1990	54.80	7	35.80	109	2.40	27	6.30	28	171
1991	52.70	7	36.80	129	2.30	27	7.50	29	192
1992	55.20	7	34.20	144	2.40	27	7.50	30	208
1993	47.60	7	40.20	161	2.40	27	9.80	39	234
1994	35.30	7	38.00	166	2.40	27	8.10	40	240
1995	39.80	7	47.60	165	2.20	27	10.30	41	240
1996	37.20	7	51.20	164	2.20	27	9.40	41	239
1997	40.50	7	44.60	144	20.00	27	12.90	44	222

Source: Miki Hamada, *Transformation of the Financial Sector in Indonesia* (Institute of Developing Economies [IDE-JETRO]), p. 7.

Note: The difference between the total lending ratio and 100% represents the portion of the loans provided by BI.

As in the previous stage, the major portion of small business finance was provided by local microfinance institutions. Of the operating banks, the Indonesian People's Bank (Bank Rakyat Indonesia, BRI) was the one that targeted the providing of finance to rural households and small businesses. With a long history of being the people's bank, BRI was fully owned by government and in later years gained an international recognition for its business performance (Patten, Rosengard, and Johnston 2001).

1.3.3 Stage 3: Supremacy of prudential banking (1998 to present)

The repercussions of the AFC became apparent in the banking industry when in one year (from 1996 to 1997) the number of privately-owned banks went down by 12% (shown in Table 1.2). The number went down again by approximately the same percentage in 1998, accompanied by a decrease in assets and liabilities (similar to deposits and loans), as highlighted in Table 1.3. All types of banks experienced liquidity problems; debtors of many banks failed to meet their loan obligations because large businesses were severely affected by the plunging IDR-to-USD conversion rate. After decades of a stable fixed-exchange rate at around IDR 2,000–2,500/USD, the USD rate rapidly doubled in 1996–1997. The worst situation occurred in 1998 when the IDR depreciated from 4,600/USD to 14,000/USD (Enoch, Frécaut, and Kovanen 2003). This affected BI, which was tasked to be the lender of last resort and to maintain the currency exchange rate. The volatility was highly uncontrollable and BI had inadequate reserves to maintain the fixed rate.

Table 1.3 The Indonesian banking sector at end of 1997 and at December 1998

Descriptions	State-owned Banks*		Privately-owned Banks		Joint Venture Banks		Total	
	1997	1998	1997	1998	1997	1998	1997	1998
Number of Banks	34	39	144	127	44	43	222	209
Assets (% of GDP)	47.90	55.00	46.00	22.80	11.80	9.90	105.70	87.70
Assets (% of Market Share)	45.30	62.70	43.50	26.00	11.20	11.30	100.00	100.00
Liabilities (% of GDP)	45.30	63.80	42.20	24.00	10.90	9.80	98.40	97.60
Liabilities (% of Market Share)	46.00	65.40	42.90	25.60	11.10	10.00	100.00	100.00
Equity (% of GDP)	2.60	(8.80)	3.80	1.20	0.90	0.10	7.30	(9.90)
Assets/Liabilities Ratio	105.8	86.2	109.1	95	108.2	101	107.5	89.9

Source: I Putu Gede Ary Suta and Musa Soebowo, Indonesian Banking Crisis: The Anatomy of Crisis and Bank Restructuring (Jakarta: Sad Satria Bhakti Foundation, 2004), pp. 230, 232.

* Including Regional Government Banks

BI's liquidity support program (Bantuan Likuiditas Bank Indonesia, BLBI), used to support banks with high NPLs and low CARs, reached 5% of GDP in December 1997 (Enoch, Frécaut, and Kovanen 2003). Table 1.4 indicates the escalating seriousness of the

situation that jeopardized the Indonesian economy as a whole. Post crisis analyses suggested that "connected lending" (Lindgren et al. 2000) and breach of lending limits (Enoch, Frécaut, and Kovanen 2003) steered banks into non-prudent lending practise. This also implied the weak supervision and control from BI. As a result, CAR position of these banks dropped significantly with an extreme case of 17 banks that incurred CAR worse than –2.5% (Lindgren et al. 2000) and eventually were closed down. BI was forced to accommodate the condition of banks and allowed them to have a minimum CAR of 4%.

During this turmoil, the government established an ad hoc agency, IBRA, with the ultimate goal of preventing the collapse of the banking industry. Based on audits by independent international auditors, these struggling banks were classified as A (CAR > 4%), B (CAR –2.5 – 4%) and C (CAR < –2.5%) (Lindgren et al. 2000). B and C classes were transferred to IBRA to be resolved. Owners of these banks were given time to improve their solvency and sort out their NPLs. If their efforts failed, the banks were closed or merged. Many of the problematic banks, previously considered sound and safe, experienced panicked behaviour from depositors, whose actions caused the banks more liquidity issues when news about their poor performance became known. Of the seven state-owned banks, four were merged and formed a new bank, while the others were included in the government recapitalization plan (Lindgren et al. 2000). BRI emerged as the most successful state-owned bank, because of the type of customer it served.

Table 1.4 NPL at end of each third quarter period, 1996–1999

Types of Bank	**NPL (%)**			
	1996	**1997**	**1998**	**1999**
All Commercial Banks	10.60	9.30	19.80	58.70
State Owned Banks	16.60	14.20	24.20	47.50
Private Foreign Exchange (Forex) Banks	4.00	4.40	12.80	76.90
Private Non- Foreign Exchange (Forex) Banks	14.70	16.50	19.90	38.90
Joint Banks	7.40	7.70	25.30	64.60
Foreign Banks	2.80	2.70	24.40	49.90
Regional Development Banks	18.50	13.90	15.80	17.00

Source: Miki Hamada, Transformation of the Financial Sector in Indonesia (Institute of Developing Economies [IDE-JETRO]), p. 18.

To improve the weak structure of the banking industry, from 1998 on the government issued regulations aimed at underpinning prudential banking practice and safeguarding deposits by establishing the Deposit Guarantee Institution (Lembaga Penjaminan Simpanan, LPS) (Undang-Undang Republik Indonesia 2004). In particular, regulations on loans and NPLs included the classification of loans, loan provisioning and debt restructuring (Lindgren et al. 2000). Loans had to be categorized as pass (or performing) loans, special mention

loans, substandard loans, doubtful loans and bad loans. Banks had to set aside respective loan provisions of 1%, 5%, 15%, 50% and 100% of net loan values for each category (Peraturan Bank Indonesia 2005). Based on this regulation, all loans included in the special mention, substandard and doubtful categories were classified as non-performing. In addition, BI re-enforced the minimum 8% CAR requirement and increased supervision on connected lending. The stringent implementation of prudent banking by BI was been rewarded as commercial banks have showed stronger performances ever since. This positive achievement did not necessarily apply to rural banks, also called the People's Credit Banks (BPR) as indicated in Table 1.6.

Tables 1.5 and 1.6 reflect the performance of the two groups of banks in recent years. Both BPRs and BRI targeted micro and small businesses; BPRs provided only a limited scope of financial services (saving and lending products), but performed competitively in terms of return on assets (Miyahara et al.). Initially BPRs also managed to channel their loan funds better, although commercial banks have levelled this in subsequent years with their LDR reaching almost 80%. In 2010, BI increased the minimum LDR level to 78% and statutory reserves to 8% of third-party funds for commercial banks, to be applied from March 2011 onwards (Peraturan Bank Indonesia 2010a). In this context, BI exercised its regulatory role to manage liquidity of banks.

Table 1.5 Selected performance indicators of commercial banks, 2000–2011

Indicators	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
	%	%	%	%	%	%	%	%	%	%	%	%
CAR	12.46	19.93	22.44	19.43	19.42	19.30	21.27	19.30	16.76	17.42	17.18*	16.06*
ROA	1.56	1.45	1.96	2.63	3.46	2.55	2.64	2.78	2.33	2.60	2.86	3.03
BOPO+	98.12	98.41	94.76	88.10	76.64	89.50	86.98	84.05	88.59	86.63	86.14	85.42
LDR	33.41	33.01	38.24	43.52	49.95	59.66	61.56	66.32	74.58	72.88	75.21	78.77
NPL	20.09	12.23	7.50	6.78	4.50	7.56	6.07	4.07	3.20	3.31	2.56	2.17

Source: Bank Indonesia, Indonesian Banking Statistics.

* From August 2010, CAR includes operational risk.

+ BOPO = Operations Expenses/Operations Income (%)

Table 1.6 Selected performance indicators of BPRs, 2000–2011

Indicators	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
	%	%	%	%	%	%	%	%	%	%	%	%
LDR	85.78	80.87	70.00	74.50	80.73	82.00	87.37	80.03	82.54	79.61	79.02	78.54
NPL	15.97	11.83	8.65	7.96	7.59	7.97	9.73	7.98	9.88	6.90	6.12	5.22
ROA	2.45	3.44	3.72	3.40	3.22	2.96	2.21	2.39	2.61	3.08	3.16	3.32
ROE	13.55	19.95	24.65	24.97	25.44	25.23	19.25	20.98	22.67	25.08	26.71	29.46

Source: Bank Indonesia, Indonesian Banking Statistics

.

During the 2000 to 2011 period, BI maintained a maximum 5% NPL limit, but BPRs NPL remained above this limit for the total period (Table 1.6). Commercial banks that had a significant share in the lending market, as shown in Table 1.7, were able to improve their loan market share performance significantly by 2007. Over the years, the five (2000–2008) and four (since 2009) state-owned banks dominated an average of over 30% of the market.

Table 1.8 provides the number of participating banks in selected years over the period 2000 to 2011. In 2011, there were 120 conventional commercial banks, 11 sharia commercial banks and 1,669 BPRs. Past experience had demonstrated the damage caused by domineering shareholders (domestic or foreign) in terms of bad corporate governance (Nasution 2012). In 2012, BI introduced legislation regarding maximum shareholder concentration groups for commercial banks: 40% ownership by a financial group; 30% ownership by a non-financial group; 20% individual ownership in the case of conventional banks; and 25% individual ownership for sharia commercial banks (Peraturan Bank Indonesia 2012). Banks owned by central government and those that were on the LPS watch list were exempted from this requirement. Banks were expected to comply with the regulation by December 2013.

Table 1.7 Loans of commercial banks and BPRs, 2000–2011 (in IDR billion)

Year	Commercial Bank Loans						BPRs	Total Loans	Lending ratio, commercial banks*	Lending ratio, State-owned banks+
	State-owned Banks	Forex Banks	Non-Forex Banks	Regional Development Banks	Joint Venture	Foreign-owned banks				
2000	68,677	45,732	10,596	10,015	3,714	14,291	3,619	156,644	97.68%	44.87%
2001	83,288	72,254	9,748	15,373	4,990	17,598	4,860	208,111	97.66%	40.97%
2002	109,704	106,902	11,574	21,486	5,273	17,750	6,683	279,372	97.60%	40.23%
2003	134,104	137,791	14,526	28,332	6,519	19,075	8,985	349,422	97.42%	39.39%
2004	171,427	181,832	15,101	37,209	9,904	23,172	12,149	450,794	97.30%	39.08%
2005	204,534	248,996	16,842	44,899	16,088	34,497	14,654	580,500	97.47%	36.14%
2006	231,582	276,768	19,114	55,919	18,166	36,951	16,948	655,448	97.41%	36.26%
2007	282,055	351,351	23,863	71,529	23,634	39,171	20,540	812,145	97.47%	35.63%
2008	396,024	452,613	27,122	95,751	33,360	49,420	25,472	1,079,761	97.64%	37.56%
2009	486,859	492,045	35,700	120,244	43,962	49,846	28,001	1,256,657	97.77%	39.62%
2010	569,041	630,777	48,757	143,174	50,311	50,348	33,844	1,799,689	98.11%	32.22%
2011	684,620	801,969	68,143	174,668	57,276	52,277	41,100	1,880,053	97.81%	37.22%

Source: Bank Indonesia, Indonesian Banking Statistics.

*Lending ratio commercial banks is the proportion of commercial bank loans to total loans.

+Lending ratio State-owned Banks is the proportion of State-owned Banks loans to commercial bank loans

In addition, the Bank for International Settlements (BIS) announced requirements for higher liquidity, capital requirement and a buffer, and revised the leverage ratio standard known as Basel III ("Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems" 2011). BI shows full commitment to adopt the Basel III accord by 2019, driven by the need to foster a resilient banking industry.

Table 1.8 Number of Indonesian banks and their offices, 2000–2011 (selected years)

Group of Banks	2000	2002	2004	2006	2007	2008	2009	2010	2011
State-owned Banks									
Total Banks	5	5	5	5	5	5	4	4	4
Total Offices	1,739	1,885	2,112	2,548	2,765	3,134	3,854	4,189	4,362
Forex Commercial Banks									
Total Banks	38	36	34	35	35	32	34	36	36
Total Offices	3,339	3,565	3,947	4,395	4,694	5,196	6,181	6,608	7,209
Non-Forex Commercial Banks									
Total Banks	43	40	38	36	36	36	31	31	30
Total Offices	533	528	688	759	778	875	976	1,131	1,288
Regional Development Banks									
Total Banks	26	26	26	26	26	26	26	26	26
Total Offices	825	909	1,064	1,217	1,205	1,310	1,358	1,413	1,472
Joint Venture Banks									
Total Banks	29	24	19	17	17	15	16	15	14
Total Offices	58	53	59	77	96	168	238	263	260
Foreign Banks									
Total Banks	10	10	11	11	11	10	10	10	10
Total Offices	53	61	69	114	142	185	230	233	206
Sharia Commercial Banks	2	2	2	3	3	5	6	11	11
Total Banks	55	113	263	346	398	576	711	1,215	1,390
Total Offices									
Rural Banks (BPR)									
Total Banks	2,419	2,141	2,158	1,880	1,817	1,772	1,733	1,706	1,669
Total Offices	2,482	2,747	3,507	3,173	3,250	3,367	3,644	3,910	4,172

Source: Bank Indonesia, *Indonesian Bank Statistics*.

1.4 Micro, small and medium enterprises (MSMEs)

Small businesses form an integral part of a country's economy, contributing not only to the number of enterprises managing resources to meet the needs and demands of society for products, but also to socioeconomic performance. Previously, small firms have been regarded as complementary to big firm business. They were to serve niche markets where big companies experienced difficulty achieving economies of scale, or they provided

components for their large-scale counterparts' operations in particular industries (Waite 1973). In recent times, small businesses have established their standing as equal contributors to the economy. Small-scale enterprises are also considered a factor in social development, since these enterprises are labour intensive in nature (You 1995).

In general, small businesses are associated with small-scale economic operations, very limited market share and low power distance between owners and customers. Recent studies have focused on the presence of this type of business and its relationship with the economic growth of countries. MSMEs form 99% of the total firms operating in European Union countries ("European SMEs under Pressure " 2010), in the USA (Bloodgood et al. 2010) and in Asia–Pacific countries ("SME Market Access and Internationalization: Medium-Term KPIs for the SMEWG Strategic Plan" 2010). They contribute to job creation, productivity and business innovation. Their numbers slightly vary over time but remain high in these countries. In Africa, SMEs constitute 90% of the economy and contribute to 50% of the GDP (Beck 2008; "European SMEs under Pressure " 2010; Isern et al. 2007; Klapper, Sarria-Allende, and Zaidi 2006). In Asia–Pacific countries such as Indonesia, Malaysia and Thailand, SMEs contribute 55.6%, 32%, and 37.8% to GDP, respectively ("SME Market Access and Internationalization: Medium-Term KPIs for the SMEWG Strategic Plan" 2010). In the US and Canada SMEs' contribution to GDP is 50.7% and 54.2% (Leung and Rispoli 2011), while it is 58.1% in European countries (Wymenga et al. 2012).

Different countries have different definitions for small businesses, usually based on the number of employees and on turnover. The European Union, for example, defines micro businesses as those that employ fewer than ten people and have ≤ €2 million annual turnover OR ≤ €2 million total assets. Small businesses are those that employ fewer than 50 people and have ≤ €10 million annual turnover OR ≤ €10 million total assets. Medium businesses employ fewer than 250 people and have ≤ €50 million annual turnover OR ≤ €43 million total assets ("The New SME Definition: User Guide and Model Declaration" 2004). In Australia, both the Australian Bureau of Statistics (Alhabshi, Khalid, and Bardai) and Fair Work use the number of employees as their criterion, while the Australian Taxation Office uses annual turnover to define MSMEs. In the USA and South Africa, the sector that businesses operate in is also considered, and MSMEs are defined using numbers of employees and annual turnover for different sectors. In South Africa the gross value of non-fixed assets also serves as an indicator. These different definitions have impeded direct comparisons across countries.

1.4.1 Definition of MSMEs in Indonesia

MSMEs in Indonesia are defined differently by three institutions according to their number of employees, net asset value (excluding land and buildings for business) and sales.

The Indonesian Bureau of Statistics and the State Ministry of Cooperative and SMEs each employ different criteria for SME classification, while the central government's Law on MSMEs covers micro businesses as well. Table 1.9 displays these classifications. In this study, the definition used is that of Law Number 20 Year 2008, as followed by BI.

Table 1.9 Definitions of Indonesian MSMEs

Institution	Micro	Small	Medium
Bureau of Statistics	Not available	5 to 19 employees	20 to 99 employees
State Ministry of Cooperatives and SMEs	Not available	Less than IDR 200 million (USD 20,874) of net assets, AND less than IDR 1 trillion (USD 104,373) of annual sales	Net assets value between IDR 200 million (USD 20,874) and IDR 10 trillion (USD 1,043,387)
The Government of Indonesia	Less than IDR 50 million (USD 5,218) of net assets, OR less than IDR 300 million (USD 31,312) of annual sales	Net assets value between IDR 50 million (USD 5,218) and IDR 500 million (USD 52,186), OR annual sales value between IDR 300 million (USD 31,312) and IDR 2.5 trillion (USD 260,933)	Net assets value between IDR 500 million (USD 52,186) and IDR 10 trillion (USD 1,043,387), OR annual sales value between IDR 2.5 trillion (USD 260,933) and IDR 50 trillion (USD 5,218,661)

Source: www.bps.go.id, www.depkop.go.id, Law Number 20 Year 2008 on Micro, Small and Medium Enterprises.

1.4.2 An overview of MSME contribution to the Indonesian economy

Indonesian MSMEs have been the backbone of the country's economy. Table 1.10 displays several indicators that confirm positive contributions of MSME to the whole economy in the periods before and after the GFC.

Table 1.10 Selected indicators of MSME development, 2005–2009

Indicator	2005	2006	2007	2008*	2009*
Number of MSMEs	47,017,062	49,021,803	50,145,800	51,409,612	52,764,603
Number of Non-MSMEs	5,022	4,577	4,463	4,650	4,677
Number of employee-MSMEs	83,586,616	87,909,598	90,491,930	94,024,278	96,211,332
Number of employee-Non MSMEs	2,719,209	2,441,181	2,535,411	2,756,205	2,674,671
% of MSME contribution to GDP	53.87%	56.23%	56.28%	55.67%	56.53%

Source: The Ministry for Cooperatives and Small Medium Enterprises, Perkembangan Data Usaha Mikro, Kecil, Menengah (www.bi.go.id) dan Usaha Besar (UB) Tahun 2005–2009, www.depkop.go.id, accessed 27 September 2012.

*Preliminary figures.

During 2005–2009, the number of MSMEs remained stable at 99% of the total number of firms operating in the country. This includes informal businesses that are not presented in

any formal statistical records. Similarly, MSMEs contributed to more than 90% of employment. In terms of productivity, MSMEs contribute slightly more than 55% of the GDP, which is not comparable to its proportion of the total number of businesses; nonetheless it is reasonable when economy of scale is considered.

1.4.3 Problems with small business sustainability and growth

In performing their business, MSMEs depend heavily on cash and often face issues with cash flow. In order to overcome cash shortage, MSMEs look for external financing. In general they have less access to external financing provided by formal financial institutions such as banks (Beck and Demirguc-Kunt 2006), regardless of the positive result of a recent study about banks' active involvement in financing MSMEs in South America (de la Torre, Martínez Pería, and Schmukler 2010). This limited access to external finance is the main reason for MSMEs' lack of growth (Schiffer and Weder 2001). A major issue for Indonesian MSMEs trying to maintain sustainable businesses relates to the availability of financing (Godau, Hiemann, and Jansen 2004; Rudjito 2003; Tambunan 2008).

On the managerial side, MSME owners and staff lack management and business skills (Bhasin 2010). When an individual plays a central role in a business, like the owner or manager of a small business, knowledge and skills about related business issues are required. In addition, MSMEs have limited access to information, markets, infrastructure and business exemptions (Tambunan, 2008b). Because of their small scale of activities, MSMEs incur high costs when they strive to expand their business networking and market, initiate or improve technology in their operations, apply for business-related licenses or attempt to acquire facilities such as tax exemptions on export goods.

1.4.4 Development policy for MSMEs

Cooperatives are considered one of the three economic development pillars in Indonesia outside the state and private sector (Liddle 1991). Introduced in the 1950s and nurtured by the first vice president of Indonesia, cooperatives were abandoned when oil prices increased in 1973 but revived shortly after the prices fell in 1983. The government assigned a ministry to foster cooperatives but never provided an integrated approach for this. After several years of the status quo, in 1996 small business development was added to the title of the ministry; in 2001 the ministry became The Ministry of Cooperatives and Small Medium Enterprises.

According to the Ministry's five-year strategic plan (Kementerian Koperasi dan Usaha Kecil dan Menengah 2010), development policies for MSMEs are aimed at revitalizing and increasing the number of training centres to increase entrepreneurial business skills and capacity; increasing MSME competitiveness by endorsing and promoting the use of local products countrywide; enhancing access to finance through funding schemes; and issuance of pro-MSME regulations. With regard to human resource and finance issues, BI also

provides various loan schemes (detailed in the next section), and provides subsidies for the training of bank staff, who in turn conduct training of MSME owners and staff.

1.5 Bank credit risk assessment (CRA)

The market for different commercial bank types depends on the services they provide. The majority of banks have retail and corporate banking sections. A small number of joint ventures and foreign banks provide banking services limited to businesses. Banks have actively extended their services to business clientele, particularly financial services, to MSMEs, in response to the 1996 and 2006 financial crises as well as government's decision to involve formal financial institutions in nurturing small businesses. BRI and BPRs have been in the forefront of MSME lending as they have dedicated all their loan portfolios to MSMEs. The scope of regional development banks and BPRs is mainly limited to small geographical areas; other commercial banks have wider geographical coverage.

Based on past experience, BI underlines the importance of credit governance, covering issues about concentrated borrower(s), legal lending limits and allowance for loan losses. A summary of credit regulations for commercial and rural banks is provided in Table 1.11.

Table 1.11 Comparison of credit regulations for commercial banks and BPRs

Items	Commercial Banks	BPRs
Lending guidelines[1]	Guidelines on related party and large exposures lending are mandatory. Items to be clearly outlined include criteria to assess borrowers' trustworthiness, legal lending limits, information and monitoring systems for provision of funds, and hedge plan. These guidelines have to be reviewed on annual basis.	Not available.
Legal lending limit[2]	Maximum of 10% of bank capital to related parties. Maximum of 20% of bank capital to an individual who is not a related party. Maximum of 25% of bank capital to a group that is not a related party.	Maximum of 10% of bank capital to related parties. Maximum of 20% of bank capital to an individual who is not a related party. Maximum of 30% of bank capital to a group that is not a related party.
Prescribed allowance for loan losses [3]	Minimum of 1% for pass loans. Minimum of 5% for net special mention loans. Minimum of 15% for net substandard loans. Minimum of 50% for net doubtful loans. Minimum of 100% of net bad loans.	Minimum of 0.5% for pass loans. Minimum of 10% for net substandard loans. Minimum of 50% for net doubtful loans. Minimum of 100% of net bad loans.

1 (Peraturan Bank Indonesia 2006)

2 (Peraturan Bank Indonesia 2009)

3 (Peraturan Bank Indonesia 2006)

Source: Bank Indonesia.

BI obligates commercial banks and BPRs to be attentive to any lending concentration to individual or single group borrowers and to carry out loan portfolio diversification to mitigate credit risks. BI has taken an aggressive approach by setting higher legal lending limits and lower weighted risks on MSME loans for commercial banks. From January 2012, the MSME lending limit has been IDR 1 trillion (ca. USD 104,493), an increase from IDR 500 million (ca. USD 52,246), and 75% weighted risk; a weight risk of 85% was applied previously.

1.5.1 Regulated assessment methodology for credit applications

BI stipulated assessment factors for loan quality measurement (Peraturan Bank Indonesia 2005) and enforced the use of Five Cs (5Cs) for good lending principles (Undang-Undang Republik Indonesia 1998), providing guidelines but not elaborating on methodologies. This implies that BI did not implement this particular issue as rigorously as it did legal lending limit percentages for loans. Based on the 2005 stipulation, credit assessment factors include borrower business potential, performance and ability to repay the loan. In order to assess loans based on business potential, banks need to determine the business profile of a borrower, focusing on business growth, market share, competitiveness, management, issues with employees, business networks and corporate responsibility, with reference to sustainability. The second feature, business performance, captures borrower financial capacity and is derived from financial statements such as profit and loss accounts, change of capital statements, cash flow and market sensitivity reports. The last feature, ability to repay, requires banks to review the historical loan performance including continuous loan repayment, availability and reliability of borrower financial information, a complete loan application submission, compliance with loan contractual agreements, appropriate use of funds and sound sources of funds for loan repayment.

The 5Cs for credit assessment purpose are: Capacity, Character, Condition of Economy, Collateral and Capital, relevant to the prospective borrower's credit risk (Weaver and Kingsley 2001). This credit risk parameter, as mandated by the Central Bank, is outlined on the Central Bank website. "Character" entails prospective borrower profiles. "Capacity" represents the applicant's managerial skills, in particular for managing firm resources and directing them into income generating resources. "Capital" exemplifies a prospective borrower's dependency on external financing. "Collateral" demonstrates a debtor's ability to repay the loan. The last "C", "Condition of Economy", is a set of possible economic events and the impact it may have on a borrower's business. Each of these Cs will be discussed in Chapter 2.

1.5.2 An overview of problems with credit risk assessment

Commercial banks, except for BRI, mostly transact with large businesses, and have implemented credit risk assessment methods suitable for such businesses. Large businesses systematically document their activities and communicate their business performance to a wide range of stakeholders such as investors, lenders, suppliers, consumers and government. This information enables stakeholders to understand the company's business risks. MSMEs – in particular micro and small businesses – do not have the same requirement since they are not exposed to as many stakeholders; and information on MSME business performance is hardly available. In the case of micro businesses, business information essentially resides with the owners who maintain and manage them, without documentary support.

Large companies and MSMEs pose different business risks. Small businesses are more responsive in adapting to economic changes and challenges (Berry, Rodriguez, and Sandee 2001; Rudjito 2003; You 1995). This means they are less sensitive to changes in economic factors such as price, tax or government policy, resulting in less impact on the whole business operation than is experienced by big firms.

The government's urging of banks to allocate part of their loan portfolios to MSMEs was not supplemented with relevant risk assessment instruments. Banks were left to develop their own assessment methods, taking into account the different nature of MSMEs, and were faced with a major problem regarding credit risk assessments, due to the unavailability of sufficient clear and verifiable information. Banks need a proven set of credit risk indicators to assess loan applications, and that enable them to establish good benchmarks based on common features indicative of good and bad credit risks.

1.5.3 Credit risk assessment for MSMEs

Cash shortages and limited access to formal financial institutions referred to in Section 1.4.3 are some of the obstacles to business longevity that MSMEs encounter. In Indonesia, the government has mandated banks to channel funds to MSMEs to overcome such obstacles. However, banks encounter asymmetric information issues caused by the unavailability of relevant information on which to make informed lending decisions. The lack of systematic bookkeeping stems from the fact that, in particular in emerging markets, many entrepreneurs have established informal businesses (Nichter and Goldmark 2009). Since these are not registered at tax offices or elsewhere, they have little incentive to perform meticulous documentation of daily operations. Furthermore, the non-separation of professional and personal activities confuses the use of assets or income generated from economic undertakings, adding to the complexity of preserving relevant documentation. From a banks' perspective, the lack of financial data that can be considered reliable

undermines prudential practice. The asymmetric information available increases bank credit risk during the loan assessment process.

Based on the type of data that is predominantly used to decide on the loan disbursement, and how this data is collated, lending methods for MSMEs are categorized as transaction and relationship lending (Berger and Udell 2006). Transaction lending refers to credit decisions based primarily on quantitative data, such as can be derived from financial statements. If some forms of qualitative data are also used to arrive at the decision, e.g. financial information and ratios, it is still called transaction lending. Examples of transaction lending are financial statement lending, small business credit scoring, asset-based lending, factoring, fixed asset lending, and leasing (Berger and Udell 2006). Relationship lending exists when credit is granted on the basis of soft information collected through years of business. When a decision is based on the trust built over time, then it is relationship lending. BRI is a perfect example of this lending methodology. Having been in banking since 1895, and nationalized in 1946, it has branched out across Indonesia providing financial services to generations of rural businesses.

Both types of lending present advantages and disadvantages. As transaction lending is predominantly characterized by quantitative data, as is the case with credit scoring, it produces more consistent decisions. Credit scoring techniques provide benefits such as low information cost (Frame, Srinivasan, and Woosley 2001), faster decision making (Dinh and Kleimeier 2007), consistency (Abdou, Pointon, and El-Masry 2008; Yu, Wang, Lai, et al. 2008) and objectivity (Chye, Chin, and Peng 2004; Yu, Wang, Lai, et al. 2008). One vital drawback is its limited applicability to information-opaque businesses such as MSMEs. Relationship lending, on the other hand, works well with this type of businesses since it decreases the information asymmetry (Behr 2011), loan rates and credit rationing (Berger and Udell 1995; Elsas and Krahnen 1998), and decreases monitoring costs (Lehmann and Neuberger 2001). However, the disadvantage is its dependency on bank staff experience, which results in a certain degree of subjectivity (Abdou, Pointon, and El-Masry 2008; Dinh and Kleimeier 2007).

1.5.3.1 Microfinance institutions and microcredit

Microcredit is a service provided by financial institutions to small businesses and households, and incorporates social missions such as poverty reduction and social capital (Hulme 2000; Kah, Olds, and Kah 2005). The most prominent example of a microcredit program is that provided by the Grameen Bank, an institution established in 1976 by 2006 Nobel Prize winner Muhammad Yunus. Grameen relies on social networks and strong kinships between credit providers and recipients to address the economic welfare of borrowers. A microfinance institution is different from a commercial bank in terms of funds

mobilization and interest (Dowla and Alamgir 2003). Where banks are legalized to accept public savings, microfinance institutions are only allowed to collect savings from entities to whom they provide finance. Savings in microfinance institutions are a safety net for lenders in case of delayed loan payments (Hulme 2000). With the growing need to be financially sustainable, microfinance institutions are moving toward commercial types of financial institution (Copestake 2007). These are not considered microfinance institutions since the financial transactions between lender and borrower are for purely economic purposes.

1.5.3.2 An overview of types of finance available to MSMEs

Banks are entitled to set their own credit schemes for MSMEs, although the central government through BI also offers four schemes (www.bi.go.id) by which funds are made available by the executing banks, and BI subsidizes the interest rates. Three of these credit schemes are applicable to specific industries; the last one covers every industry. These schemes are Food and Energy Sustainability Finance, Bio Energy Development and Plantation Revitalization Credit, Credit for Cattle Breeding Business and People's Business Credit.

Food and Energy Sustainability Finance (Kredit Ketahanan Pangan dan Energi, KKPE)

KKPE is an investment type of finance that is channelled to groups of farmers or cooperatives. This credit scheme includes the financing for plantations of, among others, rice, corn, soybeans, chilli, for livestock and fisheries, and for the procurement or maintenance of equipment and machinery to support these enterprises. The maximum loan is IDR 50 million (USD 5,218) for plantations, livestock, fisheries and cooperatives, and IDR 500 million (USD 52,186) for procurement or maintenance of equipment and machinery. Interest rates are set at 5% above the LPS rate for sugar cane plantations and 6% for other businesses. The loan period cannot be more than five years. Selected regions where this credit scheme is offered are Sumatra, Java, Bali, South Sulawesi, South Kalimantan, Papua and Riau. The executing banks are regional development banks, three of the four state owned banks and a very limited number of privately owned banks. Issues surrounding this credit scheme include MSME failure to provide collateral; insufficient credit risk assessment methods; limited outreach of the credit scheme; and limited loan purposes.

Bio Energy Development and Plantation Revitalization Credit (Kredit Pengembangan Energi Nabati dan Revitalisasi Perkebunan, KPEN-RP)

This scheme is offered to support the development of feedstock crops for bio-fuels, and also for agriculture revitalization programs. Financing is available for expansion, rehabilitation and replanting of palm oil, rubber and cocoa; at the moment the scheme is under review for extension until 2014. The maximum loan amount is fixed by the Directorate

General of Plantations, and credit interest is set at a maximum of 5% above the LPS rate. The maximum loan period is 13 years for palm oil and cacao and 15 years for rubber. The scheme is offered in Sumatra, Riau, Jambi, Bengkulu, West Java, Lampung, Maluku, Sulawesi, Kalimantan, Papua and West Papua. The funds are provided by regional development banks, three of the four state-owned banks and a small number of privately owned banks. Issues that impede KPEN-RP are mostly administrative and bureaucratic. The many local government offices involved in providing documents for credit assessment have resulted in much red tape and complicated procedures that discourage MSMEs.

Credit for Cattle Breeding Business (Kredit Usaha Pembibitan Sapi, KUPS)

KUPS is a specific financing scheme for cattle breeding activities monitored through a microchip identification system. The scheme is for groups of breeders and will end in 2014. The maximum loan is IDR 66,315 million (USD 6,921) for each group of breeders. The interest rate is set at a maximum of 6% above the LPS rate, with the longest loan duration of six years and a grace period of 24 months. The regions eligible for this scheme are limited to East Java, West Papua, Jogyakarta and Central Java. Participating are some regional development banks, two of the four state-owned banks and two privately owned banks. The problems surrounding KUPS are related to complicated documents and procedures required by the banks.

People's Business Credit (Kredit Usaha Rakyat, KUR)

In 2007 the Government initiated this loan scheme specifically for MSMEs and cooperatives that did not qualify for any existing bank or government financing schemes at the time. Although its introduction was welcomed by MSMEs and banks, the channelled funds did not meet expectations. The problem arose from a misconception about the promotion of "non-collateral loans" made by the government. The government was not specific about which type of collateral (principal or additional) MSMEs were exempt from, and the banks assumed that MSMEs had to provide principal collateral but were released from additional collateral. This confusion was eventually addressed by the government. Within three years, policies related to administrative requirements and collateral were relaxed to increase the amount of loan disbursement. The government added more money to guarantee MSME loans, provide latitude on access to loan applications, allow lower demand for collateral and waive the requirement to obtain credit references from the Central Bank's Debtor Information System (DIS). These were sufficient to overcome periodic financing difficulties for MSMEs, but they jeopardize the banks' approach to business, forcing them to compromise the 5Cs Good Lending Concept, the primary tool for credit assessment.

The scheme targets the "non-bankable" group of MSMEs, the majority of which are micro enterprises that are unable to secure loans because they have no proof of collateral to

banks. The maximum loan amount is IDR 5 million (USD 522) for micro loans and IDR 500 million (USD 52,186) for retail loans. Interest rates are set at 22% per annum for micro and 14% per annum for retail loans, and the maximum duration is three years for the former and five years for the latter. The lending period for both types of loan can be extended to six and ten years. The scheme is offered nationally and channelled through all four state-owned banks, two privately owned banks, 13 regional development banks and one sharia bank. BI has noted issues surrounding this scheme, including lack of promotion, high interest rates, late payment by LPS, difficulties in assessing debtors, and disputes over some schemes.

Since its inception, banks have reported notable achievements in channelling their funds through the KUR scheme (Barth, Lin, and Yost 2011). Table 1.12 shows the achievements of six commercial banks and thirteen regional banks in only four years. The staggering number of loans provided to non-bankable firms testifies to the success of MSME financing. The only notable deficiency is the lack of published information about NPLs amounts, which would provide a more balanced perspective of KUR performance.

Table 1.12 Loan amounts of KUR, 2008–2011 (in IDR billion)

Bank	2008–2009	2010	2011	Total
BNI	1,527.9	1,630.6	3,348.4	6,506.9
BRI	12,841.1	9,879.8	16,796.2	39,517.1
Mandiri	1,505.7	2,100	3,396.4	7,002.1
BTN	263.3	710.1	933.5	1,907
Cooperative Bank (Bukopin)	669.3	245	170.3	1,084.6
Syari'ah Mandiri	382	452.2	660.3	1,494.5
Regional Development Banks	Not participating	2,211	3,697.5	5,908.5
Total	17,189.3	17,228.6	29,002.6	63,420.6

Source: The Ministry for Cooperatives and Small Medium Enterprises, www.depkop.go.id, accessed 23 January 2012.

1.5.3.3 The need for reliable and appropriate credit risk indicators for MSMEs

As this discussion has indicated, Indonesian banks are not equipped with a set of good credit assessment indicators for MSMEs. With the exception of BRI and BPRs, other commercial banks have developed their own assessment methods based on traditional methods to assess large firms' credit risk (Loan Staff, Bank B, pers.comm). In addition, banks have to cope with MSMEs' generally unsystematic bookkeeping and lack of valid and reliable financial information (Berger and Udell 1995; Berger, Klapper, and Udell 2001). In order to overcome the incompleteness of information, bank staff have to collect the required information through interviews and observations; this method requires considerable personal

judgment, which raises questions about objectivity and consistency. Although a scoring system is used to enhance objectivity, there exists an issue of transfer of knowledge and skill from one loan officer to another when staffing changes.

Within the last six years, MSME (BI has combined micro businesses with SMEs in these reports) financing has shown an unstable tendency as shown in Table 1.13. Increased net expansion indicates that higher loan disbursements occurred, while a decrease suggests loan repayments outpaced loans granted. Considering KUR and government's efforts to drive banks into taking active involvement in MSME lending, the numbers indicate that banks have been cautious in increasing the MSME portion of their loan portfolios.

Table 1.13 Net expansion of loans, 2006–2011 (in IDR billion)

Recipients of Credits	2006	2007	2008	2009	2010	2011
MSMEs	58,017.6	96,178.2	136,270.8	106,270.8	194,807.3	230,150.3
Non-MSMEs*	37,928.7	112,639.5	169,020.6	20,300.4	142,951.5	227,521.8
Total	99,056.2	213,614.8	311,027	133,100.4	334,673.1	457,672.1
% MSMEs	58.57	45.02	43.81	79.84	58.20	50.28

Source: Bank Indonesia, Net Expansion of Micro Small and Medium Enterprises Credits, various papers.

*Non-SMEs credits include credit card.

Net expansions are the subtraction of outstanding MSME loans in one period by outstanding MSME loans at the end of previous year. The information presented in Tables 1.13 and 1.14 provides insight into the quality of MSME loans. While the portion of SME net loans decreased every year from 2006 to 2008, the portion of MSME NPL increased every year in the same period. In 2009, both expansion and NPL percentages increased, but with different variance: the increase in NPL was smaller than that of the expansion. In the years 2010 and 2011, the portion of MSME NPL was stable while the percentage of MSME loan expansion decreased. This MSME NPL movement shows that loan quality is an ongoing issue. In Table 1.13, the preference of banks to provide loans to large businesses is evident. With the exception of 2009, the portion of large business lending increased, including the last two years (2007 and 2008), possibly because large businesses offer more certainty as they can provide reliable information about the business and collateral.

Table 1.14 NPLs of loans, 2006–2011 (in IDR billion)

Type of Firms	2006	2007	2008	2009	2010	2011
MSMEs	18,766.9	19,295	21,244.7	24,744.4	16,473.1	17,443.3
Non-MSMEs	29,271.8	20,346.3	20,044.6	20,939.6	30,935.2	32,960.9
Total	49,705.9	42,480.6	44,493.5	45,714	47,408.3	50,404.2
% MSMEs	37.75	45.42	47.74	54.13	34.75	34.60

Source: Bank Indonesia, NPL of Micro Small and Medium Enterprises Credits, various papers.

1.6 Problem statement

Banks should be equipped with computational intelligence method that enables them to assess the credit worthiness of MSMEs with reasonable accuracy, notwithstanding imperfect financial data and the distinct characteristics of entities in the MSME business sector. The method allows the incorporation of qualitative credit information to achieve a more informed credit decision.

1.7 Research objectives and significance

1.7.1 Research objectives

The aim of this research is to construct a methodology for the development of a set of financial assessment tools that will enable automatic extraction of novel and useful knowledge patterns that can provide banks with evidence-based support for making strategic decisions.

With the achievement of the aim, the objectives below are fulfilled by the work of this research:

a. To develop a template to capture quantitative and qualitative types of data in a domain-specific way and effectively organize (contextualize) the available information
b. To identify an appropriate data mining technique and a suitable way of applying it to the combined data so that derived knowledge patterns encompass a comprehensive picture about MSME risk profiles
c. To develop a conceptual framework for the construction of comprehensive credit parameters for MSMEs that integrates quantitative and qualitative information
d. To develop an approach for constructing credit parameters for MSMEs in different business sectors
e. To develop a decision support methodology for MSME credit risk assessment based on knowledge patterns derived from the application of data mining.

1.7.2 Research significance

The significance of this research can be identified in both scientific and socioeconomic fields.

1.7.2.1 Scientific significance

This research will:

a. Contribute to the research about credit parameters for Indonesian MSMEs, which is scarce at present
b. Extend the currently limited research on MSMEs' credit risk assessment methods

c. Enable the application of data mining techniques to quantitative and qualitative information to provide an informative analysis of data
d. Support banks by providing a more reliable decision making method for assessing MSME loan applications. Automation of the process will decrease the implication of subjectivity in the decision making process
e. Provide a decision support methodology that can be applied to other developing countries that practise judgmental credit assessment of MSMEs based on the study of Indonesian banks.

1.7.2.2 Socioeconomic significance

The outcome of this research will:

a. Support banks with a more standardized and comprehensive method of assessment. This research will present a systematic framework for performing credit risk assessment of MSMEs
b. Support banks to fulfil their obligations as mandated by the Government of Indonesia, which is to strengthen MSME business capacity, in particular by providing financial backing to MSMEs
c. Provide MSMEs with limited financial and non-financial information with better access to financial support from banks, by equipping banks with tools that provide faster, more prudent and standardized loan decisions
d. Provide feedback to banks on relevant data to be collected that will contribute to a sound loan decision.

1.8 Research methodology

This research consists of a literature review that investigates the existing body of work in credit risk assessment, covering statistical and machine learning methods. The review is conducted to indicate the existing gaps relating to credit risk parameters of MSMEs where quantitative and qualitative information are available in tandem. Furthermore, data mining as part of a knowledge discovery process is explored.

Three banks with particular geographical locations have been selected. Two are located in Jakarta and one in Tangerang, a municipality in the vicinity of Jakarta. The reasons for the selection of these banks are their specific micro and small loan exposures and relative lending methods and criteria. With empirical research, a set of activities to collate and organize information on existing credit parameters for MSMEs is performed. Credit data are taken from the loan application forms of the banks. These forms contain quantitative and descriptive qualitative data and are presented in a structure that is crucial for credit-granting decisions. The data is presented using eXtensible Markup Language (XML) to capture domain-specific needs and effectively organize the available information according to its

context. Before any data mining technique is applied to the dataset, data pre-processing is carried out and a method for effectively representing XML (tree-structured data) in a structure preserving flat format as proposed by (Hadzic 2012) is used. The next step is to apply appropriate data mining techniques to extract novel and useful patterns, followed by a validation of these patterns in terms of their operability. The proof-of-concept of the decision support tool is finally carried out.

1.9 Thesis structure

This thesis is structured as follows:

Chapter 2 presents credit assessment principles applied to MSMEs in general, followed by existing studies on credit risk assessment methods. This body of literature covers traditional statistical methods and machine learning methods. In addition, an interface of techniques between these two fields or within each field, termed hybrid methods, is covered. This chapter presents a comparison between BI policy on the 5Cs and the application of this credit risk assessment by the three Indonesian banks whose data are being used for the empirical analysis. The chapter will conclude with a summary of existing bank practice and insights about its implications for credit risk assessments of MSMEs.

Chapter 3 introduces and describes data mining as the method used in this research. The chapter starts with an introduction of knowledge discovery in databases and data mining, and continues with a description of the common steps in the knowledge discovery process. These include data selection, data pre-processing, application of data mining tasks and post-processing of the discovered knowledge. Each step is explained and elucidated, using its application in various domains. To ensure the practicability of the discovered knowledge, quality and interestingness (understandability, validity and usefulness) measures commonly used to evaluate the discovered knowledge are explored.

Chapter 4 elucidates the methodology applied in this research. Methods, scope of data collection and the data collected from the banks are presented. In this chapter, the data are utilized following the steps used in data mining. Pre-processing techniques used, applications of selected data mining techniques and evaluation of the discovered knowledge are also described.

Chapter 5 explains the findings of the research. The selected patterns generated by applications of data mining techniques are discussed, construed and analyzed in accordance with the lending policy of each bank. The role of descriptive qualitative credit information is also substantiated. Practical implications for the banking industry in general and each bank in particular are discussed.

Chapter 6 concludes the thesis with a summary of the study and directions for future research. In this chapter, significant findings, as well as academic and managerial contributions of these findings and the limitations of the study, will be highlighted.

CHAPTER 2 – CREDIT RISK ASSESSMENT PRINCIPLES AND METHODS FOR MSMEs

2.1 Introduction

The rationale for the development of credit risk assessment methods for Indonesian MSMEs addressed in Chapter 1 reveals that MSMEs carry higher credit risks than their larger businesses counterparts. These stem from unsystematic risks regarding financial positions and performance, so that banks are forced into making uninformed credit decisions. The 5Cs credit risk assessment principles imposed by BI provide guidance for the assessment of loans irrespective of the practical problems faced by banks dealing with MSMEs. As banks with the major lending market share in Indonesia have historically served large businesses, the established overarching principles of general credit assessment leaves a practical gap in that there are no equivalent methods of assessing MSMEs for credit purposes within the boundary of current practice.

This chapter opens by enumerating different applications of credit risk assessment principles and the corresponding credit risk assessment methods. These methods are classified according to their mode of application: namely traditional discretionary, statistical and machine learning methods, with reference to MSME credit risk assessment. The weaknesses of each method are summarized to justify the objective of this research. The rest of the chapter is dedicated to Indonesian credit risk assessment principles and methods. The 5Cs principles set by BI are discussed in more detail, as is the way that the principles are applied by each bank on which this research focuses.

2.2 General credit risk assessment principles

Lending is an integral major part of a bank's business. The credit risk associated with lending is defined by the Bank for International Settlements (BIS) as "the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms" (*Principles for the Management of Credit Risk* 2000). This potential for failure starts when the credit is approved, and prevails over the period of borrowing time. Consequently, banks perform rigorous credit risk assessment procedures that are aimed at achieving reliable and informed credit decisions, applying the well-known 5Cs lending principles: "character", "capacity to pay", "capital", "collateral" and "conditions of economy".

The first C, "character", refers to the applicant's intention to repay the loan and honour the loan contract. This includes being accessible, providing accurate and relevant information and displaying unwavering commitment to meet repayment obligations. The

main focus is to determine moral and integrity. Banks collate information related to the applicant's social and professional lifestyle, and historical loan performance with the bank, with other banks and with credit bureau reports (Cronje 2013).

The second C, "capacity", refers to the applicant's financial repayment ability. Banks are required to utilize the financial information to determine the applicant's financial ability to fulfil his/her repayment obligation over a specified period of time. The focus of this assessment is to analyze the income and expenses of the applicant.

The third C, "capital", is concerned with the applicant's financial strength. While an assessment of the applicant's capacity shows the repayment ability, an assessment of capital measures the net worth of the applicant, i.e. assets minus liabilities. It also entails the analysis of the cash flow capability of a business (Weaver and Kingsley 2001).

The fourth C, "collateral", is focused on assets which the bank can use as security in case of default. It entails valuing assets that might serve as security, as well as assessing the time and cost constraints to liquidate them.

The fifth C, "conditions of economy", is targeted at the external factors that affect the applicant's capacity to repay the loan (Weaver and Kingsley 2001). Local, national and international trends or changes in the economy are used to analyze the applicant's future financial repayment ability.

2.3 Credit risk assessment methods

These principles form the basis of all credit risk assessments conducted by banks, which adapt the principles to the different market segments they provide finance to. Each bank develops its own methods to analyze information based on the basic credit principles, and to do so effectively and efficiently. Each bank employs traditional discretionary methods and automated methods. Traditional discretionary credit risk assessment is based on policies and procedures used by the bank and utilizing its accumulated knowledge and experience. Automated credit assessment may be either statistical or machine learning-based.

2.3.1 Traditional discretionary manual assessment

Traditional discretionary assessment is performed using credit-related information collated from credit application forms, face-to-face interviews, observations, emails and other bank records. In cases where an existing banking relationship between a customer and the bank already exists, a less formal procedural assessment, termed relationship lending, may be performed.

2.3.1.1 General application of the 5Cs by banks

The character of the prospective borrower is assessed with respect to integrity and business competencies. Honesty and openness are two components that are addressed by

way of data verification of the completed loan application form, including the applicant's willingness to provide relevant documents and provide easy access to required information. The data may be verified using information from independent agencies such as credit bureaus (Cronje 2013; Weaver and Kingsley 2001). Since many small businesses are sole proprietorships, the business skills and personal qualities of the owner are related to the viability and success of the business. There is a positive relationship between an owner's characteristics such as age, educational background, personal values, personal traits and the like on technological adoption (Hairuddin, Noor, and Malik 2012; Thong and Yap 1995), innovativeness (Khan and Manopichetwattana 1989; Lefebvre and Lefebvre 1992; Miller and Toulouse 1986; Stewart et al. 1999) and firm performance (Chandler and Jansen 1992; Cooper, Gimeno-Gascon, and Woo 1994; Gray and Mabey 2005; Haber and Reichel 2007; Kotey and Meredith 1997; Man, Lau, and Chan 2002; Richbell, Watts, and Wardle 2006; Wiklund and Shepherd 2003). A lavish lifestyle is not regarded as a strong point since it does not positively correspond with individual and financial stability; and stability is necessary to ensure loan repayment during the borrowing period (Weaver and Kingsley 2001). Banks also need to establish a prospective borrower's banking credibility. This information is easy to collate if the borrower has been a customer of the bank for a period of time. The banking relationship in terms of historical loan performance (if any) acts as a "letter of reference". In cases where the prospective borrower is a new customer, the bank has to acquire information about his/her banking credibility from other banks where loan history may exist. External credit bureau information is normally also obtained to verify previous loan performance of a customer using other institutions. Essentially, this information is used to establish the credit reputation and trustworthiness of the prospective borrower.

The applicant's capacity to pay is assessed in terms of income and expenses and, in the case of a business, of the financial ability to repay the loan with cognisance of possible environmental changes over a specified period. The primary documents that banks use to derive this knowledge are proof of income for individuals or financial statements of business. Financial ratio analysis is a common and well established tool to assess the historical and present financial performance of businesses. Edminster (1971) and Edmister (1972) indicate that pre-selected ratios of different categories are useful in predicting business failure, which may compromise loan repayments by small businesses. Danos, Holt and Imhoff (1989) present a more contemporary result showing that accounting information can be used as lending decision determinant for medium-sized borrowers with sales between USD 20 and 40 million.

Many times, especially in micro lending cases, bank staff have to construct a prospective borrower's financial statements based on interviews. This situation arises when the applicant does not keep well-constructed accounting records. When the prospective

borrower is a customer of the bank, account transactions (deposits and withdrawals) provide good information from which to deduce the customer's cash management and business state of affairs. Dan (2011) proposes that banks should use the information on hand about SME borrowers, such as existing credit lines and account and supply-chain credit, to detect early signals about their ability to repay loans. If the applicant has another source of income, then additional supporting documents are required as proof of income. With all these documents and information, banks attain an understanding of the prospective borrower's existing repayment capacity. Since loans have to be repaid over a period of time, banks perform sensitivity analysis using factors of economy to determine whether the ability to repay will prevail for the total period during which repayment is to be conducted.

Capital is assessed to gain an understanding of the net worth of the applicant (business or individuals). The equity to debt ratio of applicants is regarded as a risk indicator. Banks also prefer a certain level of internal equity or capital to be provided by customers for loan transactions (Ou and Haynes 2006). In addition, banks use the available financial statements of businesses to analyze the ability to generate cash, since the flow of cash is regarded by banks and other creditors as an indicator of business health (Jarvis 2000; Mramor and Valentincic 2003). In this context, banks analyze the applicant's capital cash flow based on the turnover of the payables, receivables and inventory, as well as the liquidity of other business assets.

Conditions of economy can be seen as those conditions relating to the loan itself as well as those external factors that may affect the repayment capacity of the borrower (Weaver and Kingsley 2001). The loan conditions inter alia include the interest rate, fees, and the loan term. Changes to any of these will have an impact on the borrower since they will affect the repayment sizes and affordability. External factors that affect the prospective borrower's repayment capacity include economic conditions, political stability, and business cycles; the effects of these factors on small businesses in general and with respect to their sectors are well documented (Audretsch and Acs 1994; Beaver 2004; Berger and Udell 1998; Dobbs 2007; Everett and Watson 1998; Gertler and Gilchrist 1994; Kangasharju 2000).

Collateral consists of back-up assets and/or sources of repayment that the bank has a claim to in order to reduce the implications of loan repayment defaults. Banks have different thresholds for collateral to loan ratios. When assessing the suitability of assets for collateral, they require information about the realizable liquidation value of the collateral, and the time and effort it will take to conduct the liquidation. Collateral has been researched by many academics, covering a variety of issues. Plaut (1985) explored how interest rate, loan term, repayment scheme, principal and collateral of a loan are theoretically interrelated. Studies done by Stiglitz and Weiss (1981), Wette (1983) and Bester (1985, 1987) put special emphasis on imperfect credit markets and the value-at-risk when considering collateral.

More straightforward relationship studies with regard to collateral and loan riskiness were conducted by Berger and Udell (1990), Inderst and Mueller (2007), Jiménez, Salas, and Saurina (2006) and Ortiz-Molina and Penas (2008). However, in small business lending, studies of the associations between collateral and relationship lending are inconclusive. While a considerable number of studies find that longer credit relationships, for instance reduced asymmetric information, induce lower collateral requirements (Berger and Udell 1995; Chakraborty and Hu 2006; Degryse and Cayseele 2000; Steijvers, Voordeckers, and Vanhoof 2010; Voordeckers and Steijvers 2006), a small number of studies negates this result (Menkhoff, Neuberger, and Suwanaporn 2006; Ono and Uesugi 2005).

2.3.1.2 The practice of relationship lending

Relationship lending refers to a nurtured banking relationship that allows the bank to access a customer's privileged information (Berger and Udell 1995; Boot 2000). Berger and Udell (2006) compared relationship lending to transaction lending, which they describe as a type of lending mechanism that relies on soft or qualitative information as the primary source of information, collected through direct contact over a period of time. The scope of banking rapport includes undisclosed information communicated by the customer during the course of banking as well as the insight that the loan staff develop about the customer by way of continuous interaction. This allows the loan staff to make an informed credit decision in the absence of the usual supporting documents. In keeping with prudential banking practice, loan staff subsequently complete the relevant information and documentation, although these actions are not officially recognized in the standardized credit assessment procedures of banks.

The effect of relationship lending to MSMEs on loan price has been inconclusive, with some studies finding that loan interest rates increase in the case of longer relationships (Angelini, Di Salvo, and Ferri 1998; Baas and Schrooten 2006; Degryse and Cayseele 2000) while others find the opposite (Berger and Udell 1995; Boot 2000) or no significant associations (Lehmann and Neuberger 2001; Petersen and Rajan 1994). The result from Degryse and Cayseele (2000) is interesting since it shows that loan rates are affected by the range of products the customer utilizes: the wider the range, the lower the loan interest rate. Some studies also find that relationship lending shows a positive correlation with the availability of finance (Berger and Udell 1995; Cole 1998; Petersen and Rajan 1994).

2.3.2 Statistical methods

Credit risk assessment can be performed in an automated way. Techniques to assess credit risk include statistical methods, machine learning, Artificial Intelligence (AI) and data mining. These techniques are interrelated since machine learning, AI and data mining use statistical computations. Statistics and data mining are particularly similar as they answer

rationally constructed questions by "learning from the data" (Glymour et al. 1997, p.11). Nevertheless, a distinction between statistics and data mining remains, determined by the perspective applied to the data analysis. Statisticians perform analysis based on hypotheses constructed from research questions (Hand 1998). This hypothesis-based analysis emphasizes the reliability and validity of results. Data miners perform data analysis using the population data in a retrospective manner (Glymour et al. 1997); in other words, data analysis is based on sets of questions, and the information resides in the data itself. This results in the necessity to derive representative models and to understand that knowledge must be extracted from within the data.

Before credit scoring gained its renowned position in the credit risk assessment field, Bierman and Hausman (1970) proposed the use of dynamic programming and Bayesian probability to address the problems of credit-granting decisions. The focus of their study was to integrate the previous and future probability of loan repayment into the system. Although the study is intended for personal loans and assumes one loan price for all, the inclusion of multi-period analysis is a notable contribution. Based on this work, Dirickx and Wakeman (1976) propose an extended model that allows subsequent analysis to be performed on a borrower who has defaulted in the past. While Bierman and Hausman's model discontinues analysis when it predicts default, the extended model continues to calculate potential profit should the loan be granted. The Bierman-Hausman model assumes all loan instalments and bank expenditure occur at exactly the same time, and has been improved by Srinivasan and Kim (1987b) to reflect a real-world situation in which cash goes in and out at different times.

When effectiveness becomes a subject matter, credit scoring is regarded as a well-established method. It is "the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit" (Thomas, Edelman, and Crook 2002, p.1). Two steps are required to develop credit scoring methods: in the first, relevant attributes from historical events of loan performance are selected; and in the second, weights are applied to these attributes to form the decision models. Notwithstanding an early study by Harter (1974) and a particular case of microcredit discussed by Schreiner (2002), a lending decision based on credit scoring is perceived to be more efficient and reliable than subjective, judgmental lending (Asch 1995; Chalos 1985; Chandler and Coffman 1979; Eisenbeis 1996; Hand and Henley 1997; Rosenberg and Gleit 1994). However, some studies find judgmental lending to be equally effective, leading to similar lending decisions as those reached using statistical scoring methods (Dinh and Kleimeier 2007; Edmister 1988; Emel et al. 2003; Thomas 2000). It should be noted that these studies are confined to financial ratios as credit risk determinants.

Scholars have explored credit scoring techniques with the intent to improve the techniques and achieve better scoring accuracy (Dong, Lai, and Yen 2012; Huang, Tzeng,

and Ong 2006; Lee et al. 2002; Pang, Wang, and Bai 2002; West 2000) and to find models to improve the accuracy of scoring (Blöchlinger and Leippold 2006; Finlay 2008; Finlay 2010; Thomas 2000). Two statistical techniques commonly used are discriminant analysis (DA) and logistic regression (logit). In the forefront of the initiative to develop credit scoring was Durand (1941) who pioneered the use of quadratic discriminant analysis (QDA) to formalize credit scoring. Some later studies showed that DA provides good credit classification (Hand and Henley 1997; Lee et al. 2002; Myers and Forgy 1963; Orgler 1970) although some researchers criticize the normality assumptions and sample bias that exist in the application of the technique (Eisenbeis 1977, 1978; Reichert, Cho, and Wagner 1983; Wiginton 1980). A number of studies (Baesens, Gestel, et al. 2003; Desai, Crook, and Overstreet 1996; Tsai et al. 2009; Yobas, Crook, and Ross 2000) affirm the competitive edge of this technique, while other studies find DA inferior than techniques such as neural network (NN) (Abdou, Pointon, and El-Masry 2008; Lee et al. 2002; Li, Pang, and Xu 2002; Malhotra and Malhotra 2003; Tam and Kiang 1992) and Support Vector Machine (SVM) (Bellotti and Crook 2009; Lee 2007; Yao and Lu 2011).

Logit is used to develop credit scoring where normality assumptions of the independent variables and linearity or variance distributions are not required (Tabachnick and Fidell 2007). Orgler (1970) was a pioneer in the application of logit with several notable features. He used the judgment of bank staff to perform loan quality classifications; some of the independent variables were non-financial ratios and were presented as dummy variables; the technique was subject to bias due to the judgment of staff. The good performance of logit itself, (West 2000) compared to other techniques used for the development of credit scoring models, is supported by a number of studies (Baesens, Gestel, et al. 2003; Crone and Finlay 2012; Desai, Crook, and Overstreet 1996; Dinh and Kleimeier 2007; Dong, Lai, and Yen 2012; Lee and Chen 2005; Limsombunchai, Gan, and Lee 2005; Thomas 2000; West 2000; Wiginton 1980).

Another statistical technique used for credit scoring is the k-Nearest Neighbour (k-NN) with modest performance (Henley and Hand 1996; West 2000). The objective of k-NN is to improve classification performance by establishing a metric to define the distance between two discriminating points. According to Henley and Hand (1996), the issues related to credit assessment models such as irregularities over feature space, multidimensionality, and practicality can be addressed by k-NN to produce the lowest bad risk rate.

2.3.2.1 Discriminant analysis (DA)

Discriminant analysis (DA) is a technique used to sort a data set into two groups by using a sub-set of variables as predictors (Tabachnick and Fidell 2007). The benefits of applying DA include its practical application and clarity of assessment, which means that

when a loan is rejected, bank staff are able to provide the applicant with the reason why. However, DA has practical problems that lead to its limited use in credit risk assessment. These problems relate to normal distribution assumption, group definition, population probability, unequal covariance matrices, dimension reduction, interpretation of individual variables, and estimation of classification error rates. These criticisms and counter criticisms about DA are discussed in detail by Eisenbeis (1977, 1978), Rosenberg and Gleit (1994) and Hand and Henley (1997).

The first attempt to apply DA in small lending risk assessment was by Durand (1941). Using a credit scoring and efficiency index, his study aimed at formulating credit risk attributes from consumer lending cases and a risk assessment tool. Although his study was not focused on MSMEs, Durand's was the first to construct systematically a credit risk assessment method. The sample consisted of 7,200 consumers with small loan principals (< USD 500) and very little collateral, which generated higher interest rates. The sample of 7,200 loans was made up of 4,000 loans provided by banking companies and 3,200 by personal finance, automobile finance and appliance finance companies. The attributes and their weightings were constructed from a survey involving banking companies and financial agencies. It was found that banks assigned higher weightings to moral characteristics than to other financial, employment, personal and loan characteristics, while the financial companies considered employment and past payment records more important than loan characteristics. In the final model, Durand eliminated moral characteristics (including past payment records) due to the unavailability of data. The complete and final attributes used were divided into two categories, as shown in Table 2.1. The study resulted in three rating formulae, encompassing the importance of the industry, the locations of the lenders, and the change of the economic cycle in developing indicators for assessing consumer credit risk. Durand noted three limitations to the study. Firstly, it lacked generalization power since the formulae were constructed from pre-calculated risk analysis. Secondly, the failure to collect data about payment history and moral character implicated reliability issues. This is particularly important: further research (Hassler, Myers, and Seldin 1963) shows that payment history is a significant factor in distinguishing between performing and NPLs. Thirdly, the formulae lacked explanation power and could not easily be explained by a non-statistician or non-mathematician. Durand's major contribution was the deductive way in which credit risk indicators were constructed.

Table 2.1 Credit risk attributes applied to Durand's study

Category	Sub-Category	Name of Attributes
Financial factors	Income	Monthly income
	Amount of loan	Amount of loan
	Length of loan	Number of equal monthly payments
	Security of loan	Security of loan
	Cash price	Cash price of article purchased
	Down payment	Amount of down payment
		Repossession of assets
	Borrower assets and liabilities	Availability of bank account
		Availability of life insurance
		Availability of real estate
Non-financial factors	Stability of occupation	Duration of borrower's present employment
	Stability of residence	Duration of residence at borrower's present address
	Occupation and industry	Occupation
	Personal characteristics	Age of borrower
		Marital status and sex of borrower
		Gender
	Purpose of loan	Intended use of funds

The second study was carried out by Robert Edminster (1971), who associated business failures with loan performance. Edminster classified small businesses as bankrupt when their loans were written off by the lender in the study, the Small Business Administration. SBA is a US state-owned organization providing services ranging from finance to business counselling to small businesses (http://www.sba.gov/about-sba-services/our-history). Referring to Altman's (1970) z-score computation for corporate bankruptcy, Edminster used selected financial ratios to validate their role in predicting bankruptcy and their application in the credit-granting decision. The sample consisted of 42 loans (21 good loan cases and 21 bad), and a synthetic set of data was generated for the hold-out sample. Edminster selected 19 financial ratios based on the existing body of literature about corporate bankruptcy and collected data of these ratios for three years. The ratios were calculated relative to their respective industry ratios to find their "industry relative" ratio performance. Any upward or downward trend of the ratio was included as part of the constraint. The industry ratio was provided by either SBA or Robert Morris Associates (RMA) Statement Studies. Using DA, Edminster concluded the study with seven financial ratios (Table 2.2) and a model that showed 92% accuracy based on the 42-loan sample. The performance of the model on the hold-out data could only be projected with 80% accuracy. From this result, financial ratios as determinant for loan collection prediction are confirmed. Edminster emphasizes the importance of a diverse group of financial ratios rather than diverse ratios in single financial ratio groups. In addition, the use of industry relative ratio provides a more objective perspective of a borrower's business performance. The synthetic data used to validate the

model reduces the generalization power of the model and can be regarded as a weakness of the study.

Table 2.2 Credit risk attributes applied to Edminster's study

Category	Name of Attributes
Financial ratios	Cash flow/Current liabilities Equity/Sales Net working capital/Sales Current liabilities /Equity Inventory/Sales RMA (upward trend) Quick ratio/RMA (downward trend) Quick ratio/RMA (upward trend)

The function developed for this study is as follows:

$z = 95 - 42X_1 - 29X_2 - 48X_3 + 28X_4 - 45X_5 - 35X_6 - 92X_7$

where

$X_1 = 1$ if Cash flow/Current liabilities ratio < 0.05, otherwise $X_1 = 0$

$X_2 = 1$ if Equity/Sales ratio < 0.07, otherwise $X_2 = 0$

$X_3 = 1$ if Net working capital/Sales ratio divided by RMA ratio < - 0.02, otherwise $X_3 = 0$

$X_4 = 1$ if Current liabilities/Equity divided by SBA ratio is averaged at < 0.48, otherwise $X_4 = 0$

$X_5 = 1$ if Inventory/Sales RMA ratio has shown upward trend and is < 0.04, otherwise $X_5 = 0$

$X_6 = 1$ if Quick ratio/RMA ratio has shown downward trend and is < 0.34, otherwise $X_6 = 0$

$X_7 = 1$ if Quick ratio/RMA ratio has shown upward trend, otherwise $X_7 = 0$

z = z-score from 0 to 1, where 0 means the borrower is likely to fail to make the loan payment, and closer to 1 means the borrower is more likely to make the loan payment.

The third documented study in this field was carried out by Viganò (1993) for the development of a credit scoring system by African development banks. Although a credit scoring model is constructed, the study focuses on the performance of the selected attributes more than the methodology. This work needs to be carefully considered since it relates to development banks in emerging economies: that is, banks partially financed by grants and donors. Using 100 loans from one development bank (The Caisse Nationale de Crédit Agricole, CNCA, of Burkina Faso), Viganò partitioned them into 69 sample loans and 31 hold-out sample loans. Good loans were defined as those that were not written off and bad

loans were those loans that were. Attribute selection and values were not clearly explained, although they were the underlying focus of the study. Questionnaires and face-to-face interviews were carried out to collect borrower credit data from bank staff. The qualitative attributes were transformed into binary (dummy) variable values. Viganò stressed the importance of these qualitative attributes, although the unavailability of financial data for the borrowers might have forced this. Attributes applied to the in sample and hold-out samples (Table 2.3) were the same, although the categorization (or "factor") of each attribute was slightly different based on the factor analysis performed for each sample. The author indicates that he used factor analysis to reduce the number of attributes and avoid over-fitting. The statistical techniques used were the quadratic discriminant analysis (QDA) and the Jackknife methods. Jackknife is used to reduce bias by resampling, which in this study might have occurred as a result of the small sample size. The accuracy of the models constructed with the techniques is contained in Table 2.4 for both the in-sample and hold-out sample loans.

Table 2.3 Credit risk attributes applied to Viganò's study – in-sample

Category	**Name of Attributes**
Borrower's attitude toward the financial transactions	Public or foreign projects Foreign donors initiative Place first meeting Donor intervention in first meeting
Profitability and revenue stability	Agriculture Trade Technological degree Structural flexibility Weather dependence Monetary revenue Access to markets Price taker Loan destination: agriculture Loan destination: trade Amount determination criteria Mortgage
Customer's financial exposure and ability to manage several borrowing contracts	Access to markets Links with other bank loans Other bank loans/project amount Self-financing Automatic repayment on sales Total borrower actual balance/initial balance
Profitability and variability on livestock sector	Breeding Loan destination: breeding
Stability of the customer's personal and economic situation	Marital status

Category	Name of Attributes
	Retired Source of revenue Telephone Automatic repayment on salary
Customer general economic conditions	Currently expiring loans Total borrowing Loan amount Mortgage
Customer's financial situation and wealth	Telephone Repaid loans Total borrowing
Customer's attitude toward the obligation	Sex Belonging to a production organization
Bank's ability to control credit risk	Amount granted/amount demanded Compound interest rate Savings account
Quality of information and information adequacy (dummy variables)	Proximity to the bank Loan destination: whole enterprise Savings account Quantity of information available New enterprise
Quality of the customer's banking behaviour	Past defaults

From Table 2.4, it is apparent that QDA and the use of the Jackknife method are useful in classifying bad loans; however both are only moderately capable of classifying good loans.

Table 2.4 Credit risk attributes applied to Viganò's study

Actual Classification	Model Classification (n = 69)		Model Classification (n = 31)	
	Good	Bad	Good	Bad
QDA				
Good	62.75%	37.25%	68.42%	31.58%
Bad	8.16%	91.84%	8.33%	91.67%
Jackknife Method				
Good	62.75%	37.25%	63.16%	38.84%
Bad	20.41%	79.59%	8.33%	91.67%

The author provided explanations of the attributes and their categorization, the categories and attributes; and the study includes domain experts in the attribute identification process, which increases the utility of the attributes.

2.3.2.2 Logit

In comparison with DA, logit allows a wide range of data formats for independent variables and does not have requirements such as normality, linearity predictors or equal variance of the predictors (Tabachnick and Fidell 2007). The data can be in discrete, continuous, or dichotomous formats, or a mix of these. Three studies of MSME credit risk assessment by means of logit have been carried out by Longenecker, Moore and Petty (1997), Altman and Sabato (2007) and Gool et al. (2009).

The work of Longenecker, Moore and Petty (1997), as well as that of Asch (1995) and Eisenbeis (1996), is focused on describing the development of RMA/Fair Isaac's credit scoring for MSME. Since its development in 1995, the score has been widely used by around 75 large US banking institutions, prompting the development of SME credit scoring for Japan, Mexico, Belgium, Italy, Portugal and Germany by RMA/Fair Isaac (Asch 2000). The US model was developed using extensive loan data with respect to geography and loan amounts. The input data consisted of more than 5,000 loans from various banks across the country. A good loan was defined as one that had not been categorized as 30-day delinquent more than twice within the four-year loan duration. A bad loan was defined as one that had been delinquent for 60 days or more. The preliminary attributes that were applied amounted to more than 134, itemized as follows: 15 from application documents submitted to banks; some from balance sheets and income statements that were not disclosed in the study, 27 from business credit reports provided by credit bureaus; 16 related to the financial credibility of the principal borrower[1]; 36 related to historical loan information of the principal borrower provided by credit bureaus; 30 from financial ratios developed with the inclusion of RMA; eleven generated from information on principals. The final applied attributes and model function were not disclosed, but it was implied that the attributes for the final credit scoring models were not mutually exclusive. Throughout the development process, lenders acted as domain experts and were involved in weighting the attributes. In order to derive an effective predictive model, the data was segmented into seven clusters: gross sales, net sales, type of business, region, loan type, industry, and requested loan amount. It was found that the requested loan amount provides the best predictive ability with USD 35K as the cut-off point. Then the reject inference technique is applied to develop a scorecard to predict how rejected loans would have performed had they been accepted. In consultation with the domain experts, three credit scorings were finally developed, designated for loans (1) greater than USD 35K, (2) less than USD 35K with financial profile, and (3) less than USD 35K without financial profile (Eisenbeis 1996). These models were tested empirically by

[1] Person/s owning the business for which the loan is requested

extensive use across the country and updated every two years to incorporate any changes in the related attributes. The model updates and validations were based on practical findings.

The study of MSME credit risk assessment carried out by Altman and Sabato (2007) was aimed at developing a distress prediction model for SMEs with respect to loan repayment ability. The sample consisted of 1,890 non-default loans and 120 default (declared bankrupt) loans for borrowers with sales of less than USD 65 million. Data used to develop the model were obtained from a public database, WRDS COMPUSTAT, for the period 1994 to 2002. Altman and Sabato selected 17 financial ratios that were grouped into Leverage, Liquidity, Profitability, Coverage and Activity ratios. Upon application of a forward stepwise selection technique and GINI index, five attributes were found to be good predictors of default loans (Table 2.5).

Table 2.5 Credit risk attributes applied to Altman and Sabato's study

Category	Sub-category	Name of Attributes
Financial ratios	Profitability	EBITDA*/Total assets
		Retained Earnings/Total assets
	Leverage	Short-term debt/Equity book value
	Liquidity	Cash/Total assets
	Coverage	EBITDA*/Interest expenses

* Earnings before interest, tax, depreciation and amortization

The following model with original predictors and with log predictors was developed:

$KPG = 4.28 + 0.18X_1 - 0.01X_2 + 0.08X_3 + 0.02X_4 + 0.19X_5$ (original predictors)

$KPG = 53.48 + 4.09 \text{ -}LN(1 - X_1) - 1.13LN(X_2) + 4.32 \text{ -}LN(1 - X_3) +$

$1.84LN(X_4) + 1.97LN(X_5)$ (log predictors)

where

X_1 = EBITDA/Total assets

X_2 = Short-term debt/Equity book value

X_3 = Retained earnings/Total assets

X_4 = Cash/Total assets

X_5 = EBITDA/Interest expenses

KPG = Known Probability of Being Good, i.e. the score for non-default, 0 if default, 1 if non-default.

Although the log-likelihood test showed significant results for the function with original predictors, Altman and Sabato transformed the predictors into logarithmic values to increase the accuracy of the unlogged function by allowing the variations of attribute values to be reduced. The accuracy level of the developed functions tested on the in-sample data

increased from 75% for original predictors to 87% for log predictors. The models were then validated using 432 samples (406 non-default loans and 26 default loans), and compared with Z''-Score and Multiple Discriminant Analysis (MDA). These results are presented in Table 2.6. The performance is measured by the level of accuracy and misclassification errors. The Type I error refers to a misclassification of defaulted firms as non-defaulted, while the Type II error refers to a misclassification of non-defaulted firms as defaulted. In the last column, the average accuracy based on the combination of both Type I and Type II error, is provided.

Table 2.6 Results of Altman and Sabato's study

Models	Accuracy	Type I error	Type II error	1 – Average Error
Original Predictors	75.43%	21.63%	29.56%	74.41%
Logged Predictors	87.22%	11.76%	27.92%	80.16%
Z''-Score	68.79%	25.81%	29.77%	72.21%
MDA	59.87%	30.12%	29.84%	71.52%

From these results it is evident that the model with the log structure outperformed the other three models. MDA showed the worst performance. The authors mentioned that the use of only quantitative attributes was inevitable since the data set was taken from publicly available information. They therefore recommended the use of qualitative information to improve the accuracy of the model with variables such as the number of employees, legal form of the business, locality, industry, etc. Another suggestion made by the authors was to make use of loan amounts as the basis for sample selection as they might show different results.

As opposed to the work of Altman and Sabato, Gool et al. (2009) used non-financial attributes to develop a credit scoring model for MSME. This study, similar to the RMA/Fair Isaac model, incorporated domain experts in attribute selection. The contribution of domain experts was optimal to the extent that they provided not only the attributes but also the categorization of the continuous attributes. Essentially, this study shed light on the conflict of choice between technical and traditional lending processes for microfinancial organizations in Bosnia-Herzegovina; the study indicates that credit scoring is a good substitute for the manual assessment process. The data consisted of 6,722 personal loans provided by Bosnian micro lenders from June 2001 to November 2008. Around 78.3% were good loans and 21.7% were bad. The total number of 6,722 loans was partitioned to 70% in-sample and 30% hold-out sample loans. There is no indication whether the authors maintained the proportion of good and bad loans consistent for the in-sample and the hold-out data sample for classifying loans as good or bad. The authors defined good loans as those with average delay

of ≤ 2 days per instalment, while bad loans were those with an average delay of > 2 days per instalment. Sixteen predictor attributes were used to develop two models (Table 2.6). In the first model, all categorical attribute values were transformed into binary (dummy) variables including those with continuous values. The second model used a statistical approach where discretization was performed using the weight of evidence (WoE) coding technique to avoid over-fitting that may arise from a significantly large coefficient. With WoE, categories were grouped on their similarities and then coded in accordance with the distributed good and bad loans. This simplified the model, as can be seen in Model 2. The difference between the two models is that the near-singularity problem occurred with the usage of the attributes "Branches over Bosnia-Herzegovina" and "Name of loan officer" in Model 1; these attributes were discarded for this model.

Table 2.7 Credit risk attributes applied to Gool, Baesens, Sercu and Verbeke's study

Category	Name of Attributes
Borrower Characteristics	Age Job Experience (in years) Net Earnings of Business Business Capital Business Register Net Earnings of Household Household Capital Other Debt
Loan Characteristics	Purpose Amount Requested Duration Cycles Beginning Month (of loan) Year of Initiation
Lender Characteristics	Branches over Bosnia-Herzegovina Name of loan officer (proxy for experience)

The two functions developed are:

Model 1: General Binary Logit Model

$$\pi_i = E(Y_i = 1) = \frac{1}{1 + e^{-logit_i}},$$

$$logit_i = \beta_0 + \beta_{11}x_{i11} + \beta_{12}x_{i12} + \cdots + \beta_{148}x_{i148}$$

where

Y_i = binary dependent variable, 0 if good, 1 if bad

β_0 = intercept, β_0 = regression coefficient

x_i = dummy coded explanatory variable

Model 2: Binary Logit Model with WoE Coding

$$\pi_i = E(Y_i = 1) = \frac{1}{1 + e^{-logit_i}},$$

$$logit_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{16} x_{i16}$$

where

Y_i = binary dependent variable, 0 if good, 1 if bad

β_0 = intercept, β_0 = regression coefficient

x_i = weight of evidence coded explanatory variable

Validation of the models was carried out in terms of stability, interpretability and discriminatory power. Model 1 showed superior performance on stability and discriminatory power compared to Model 2, and both models performed well in terms of interpretability. Since the attribute selection was done by the bank staff, this study showed the role of domain experts with their "intuitive" approach about credit scoring. Taking into account the better performance of Model 1 where the contribution of bank staff was incorporated, it was concluded that credit scoring should be considered a supplementary tool rather than a substitute for the manual credit-granting process, and that the inclusion of domain experts is mandatory in credit risk assessment in micro lending.

2.3.3 Machine learning

The development of machine learning as a tool to solve practical problems started with the idea to "educate" machines to attain human-like inference ability to classify events (Solomonoff 1957). At the core of this classification task are statistical techniques, so the term "machine learning" refers to the development of learning technologies covering the computational theories of learning and construction of learning systems (Michalski 1986). Despite its close interface with AI, Michalski underlines their differences. AI is limited to the information programmed into the system to make deductive inferences, but machine learning is able to create new knowledge through inductive inferences. Essentially, machine learning techniques are used to generate predictions in an automated iterative manner based on a given set of historical categories for a given sample set (Michalski 1986; Kantardzic 2011; Solomonoff 1957). The rapid development of machine learning generates interest because it offers efficient and accurate data processing. In addition, machine learning is not limited by statistical assumptions. Since the initial start-off with Knowledge Discovery in Databases (KDD) and data mining, machine learning has been a field of study that is aimed at enabling the mining of hidden patterns from large databases and representing the extracted knowledge. Machine learning and data mining also interface with statistics, artificial intelligence, database technology and information retrieval (Han, Kamber, and Pei 2012). In the field of credit risk assessment, machine learning and other soft computing-based methods

enable efficient, objective, equal and reliable credit decisions. This is beneficial for banks that are exposed to high numbers of loan applications. Of the many machine learning techniques extant, the NN and SVM are prominent in methodological studies of credit risk assessment and show overall good performance.

In NN, the causal relationships between credit attributes and credit performance are represented by artificial neurons that reflect the information processing functions of the human brain (Hecht-Nielsen 1988). The neurons are positioned in input, hidden, and output layers. The numbers of neurons for the input and hidden layers are predetermined by the operator; the numbers in the output layer are determined in accordance with the desired target class. The neurons are connected by "a signal transmission pathway" (Hecht-Nielsen 1988, p. 36). Figure 2.1 depicts a multilayer NN with one hidden layer. Each input neuron represents one variable (or attribute). These variables are processed through the weighted connections between the neurons in each layer, allowing the knowledge to be stored in the weight of each neural interconnection. Learning rules are used to modify these weights based on the knowledge acquired from the learning. This is considered an advantage, since it reduces any issues arise from applying different discretization techniques and is free of normal distribution requirements (Bahrammirzaee 2010). Despite these positive points, NN poses an over-fitting problem and weak explanatory power notoriously known as "black box". The lack of explanatory power is the major obstruction to the practical use of NN in the credit domain since the domain requires descriptive rationalization for any negative decision.

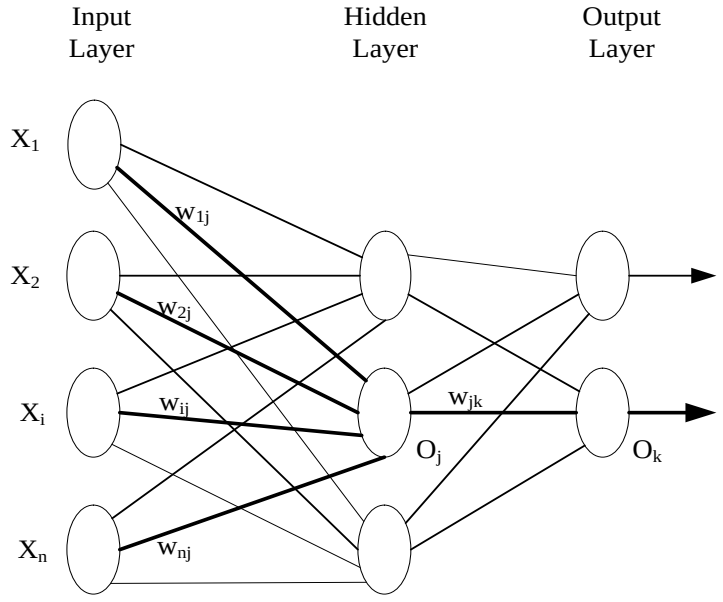


Figure 2.1 Three-layer neural network

Source: Han, Kamber, and Pei (2012, p. 399)

Jensen's (1992) early work on credit risk assessment using NN applied Back-propagation Neural Networks (BPNN) to develop credit scoring. These allow the network to learn from errors and improve the weight of the connectors, providing stronger connections between layers and resulting in better performance. The study used credit data for 100 loans, partitioned into a 75 training[2] data set and a 25 test[3] data set. Jensen found NN to be a promising method of credit scoring, despite the modest performance of 80% accuracy on the training data and 76% accuracy on the test data. The BPNN showed much better performance in correctly classifying good loans (85.7%) than bad loans (25%). The performance of NN has been ambiguous in subsequent studies, with the majority showing NN superiority in classifying bad loans (Angelini, di Tollo, and Roli 2008; Desai, Crook, and Overstreet 1996; Lai, Yu, et al. 2006b; Malhotra and Malhotra 2003; Yu, Wang, and Lai 2008). Other studies found consistent results similar to Jensen's (Abdou, Pointon, and El-Masry 2008; Tsai et al. 2009; West 2000). When NN is compared with traditional statistical methods or other machine learning techniques, the results are in some cases inconclusive. Table 2.8 shows the performance of NN in comparison with other methods in the field. In order to overcome the black-box limitation, neuro-fuzzy logic (Piramuthu 1999) or Rule Extraction (Baesens, Setiono, et al. 2003) is proposed.

SVM was introduced by Vapnik (1999) as a classification technique aimed at maximising the margin of the hyperplane separating classes (Yu, Wang, Lai, et al. 2008). The input data is mapped into a high-dimensional feature space separated by a hyperplane that creates a distance between classes. The training points that are closest to the hyperplane are called the support vector. In Figure 2.2, the circled white and black rings are the support vectors. SVM finds the maximum classifier margin using constrained optimization quadratic programming techniques (Yu, Wang, Lai, et al. 2008). In Figure 2.2, the dotted lines on the left and the right of the firm line are hyperplanes 1 and 2, which represent the boundaries of the binary classes, constructed by finding the support vectors and calculating the parameters of the gradient. In the credit domain, SVM emerged as a technique preferable to NN since it moderates the over-fitting problems that typically occur in NN (Lee 2007; Li, Shiue, and Huang 2006). Overall, SVM is known for its ability to handle linear and non-linear cases (Schebesch and Stecking 2005) as well as its strong generalization power (Burges 1998; Lahsasna, Ainon, and Wah 2010). However, particularly in the credit domain, SVM presents

[2] Training (in-sample) data refers to a portion of the dataset that is used to build a model. A dataset is partitioned into training and test data using a certain ratio. If the ratio is 70:30, 70% of the entire dataset is randomly selected to build the model and the other 30% is used to validate it.

[3] Test (hold-out) data refers to the portion of the dataset that is used to test the accuracy of (or validate) the model. Test data also refers to new records or unseen data. In this thesis, the terms in-sample/hold-out data and training/test data will be used interchangeably.

limitations related to efficiency (Burges 1998; Huang, Chen, and Wang 2007; Wei, Li, and Chen 2007; Yu et al. 2007), lack of explanatory power (Huang, Chen, and Wang 2007; Lahsasna, Ainon, and Wah 2010) and choice of input features and parameter settings (Zhou, Lai, and Yu 2010). Hens and Tiwari (2012) propose the use of F-score and stratified sampling to reduce the computational time of SVM. In order to overcome the explanatory deficiency, Martens et al. (2007) suggest the use of the rule extraction technique. With regard to the input feature and parameter setting issues, Wang, Wang, and Lai (2005) and Hao et al. (2010) propose the use of fuzzy membership. Yu et al. (2007) advocate a least squares fuzzy approach, while Peng et al. (2008) suggest the use of a Multi-criteria Convex Quadratic programming model to overcome the complexity issue.

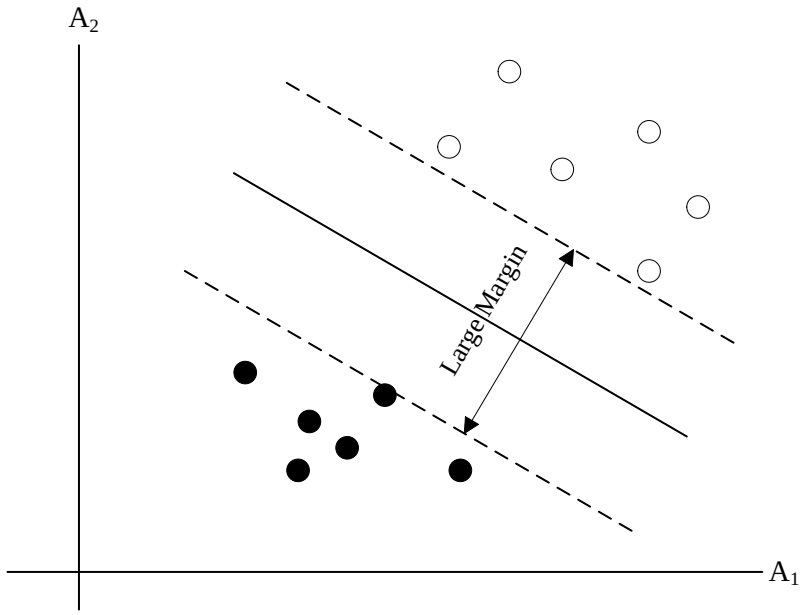


Figure 2.2 Basic principle of SVM

Source: Han, Kamber, and Pei (2012, p. 410)

The intention behind using SVM in credit risk assessment is to overcome limitations experienced by applying NN. Studies in the credit domain have compared SVM with other statistical and machine learning techniques. These studies also combined statistical or machine learning techniques with SVM. The combination is viable since SVM is essentially a constrained optimization of discrimination problems. Table 2.8 shows the satisfactory performance of SVM and SVM-based techniques. The use of SVM and its variations on corporation and consumer credit are discussed by Yu, Wang, Lai, et al. (2008).

Other machine learning techniques applied to corporate credit risk assessment are the Decision Tree (DT), evolutionary algorithms (Genetic Algorithm (GA) and Genetic Programming (GP)) and fuzzy-classifier techniques. DT is considered effective decision-

making tools as it is highly explanatory through the attributes in the nodes. DT has demonstrated overall good classification ability in the credit domain (Table 2.8). Regardless of the usefulness of DT, studies of its use for MSME credit risk assessment are scarce.

GA, introduced by Holland (1975), is inspired by Darwin's theory of the survival of the fittest. The technique automatically and continuously configures potential solutions to an optimization problem through a constant mutation of the sets of variables until it finds the best solution (Goldberg 1989). Following the principle of natural selection, only the most promising candidates (i.e. sets of variables) survive, according to an evaluation function. The iterations to find the best solution are stopped when improvement of the potential solution reaches a point below the predetermined threshold (Finlay 2009). GA has been applied to test the achievement of the objectives of the credit scoring model (Finlay 2009) and to include rejected instance[4] analysis (Chen and Huang 2003). In his study Finlay (2009) found that the model and the business objective of such model were correlated, and that the credit scoring model developed using GA showed similar accuracy with those developed using LR and logit. When GA was used to assign rejected loan applications into their possible class (where they would be if they had been accepted), Chen and Huang (2003) found that it performed the classification task efficiently. Its limitation is that it is substantially subject to mathematical conventions that increase its complexity (Koza 1994). GP, introduced by Koza (1994), addresses this issue by providing a tool that automatically writes the program for the task that GA was set to do. The advantage of GP is the comprehensibility of the outcome, since it applies a tree-based structure (Yu, Wang, Lai, et al. 2008). Table 2.8 shows the two studies in the credit risk assessment field where their performance is compared with other techniques. Fuzzy-based techniques, another technique of machine learning, also achieve good performance in the credit domain in particular for their use in enhancing the explanatory power of the extracted rules (Hoffmann et al. 2007).

To consider the advantages and limitations of statistical and machine learning techniques, several studies have been conducted in the field of credit risk assessment using a combination of techniques. Hybrid techniques have been used with emphasis on finding the perfect combination of techniques to produce the best result. The results of these studies are presented in Table 2.8. The majority of hybrid technique studies show good accuracy compared with other respective techniques. Although hybrid application has been conducted to improve the classification power of the relevant techniques, they have not been tested in cases involving MSMEs. In the next sub-section, the few studies pertaining to MSME credit risk assessment by way of machine learning are discussed in detail.

[4] The term refers to the analysis of how rejected loans would have performed had they been accepted.

Table 2.8 Selected relevant studies using computational intelligence and other techniques

Author(s)	Classification Algorithm	Compared with	Data	Findings
NEURAL NETWORK (NN)				
West (2000)	NN (MLP, MOE, RBF, LVQ, FAR)	LDA, Logit, k-NN, Kernel Density, CART	Australian and German credit data from UCI	1. MOE, RBF, MLP and Logit were superior for both data sets in comparison with other techniques. 2. VVQ, Kernel Density, FAR and CART were inferior for both data sets in comparison with other techniques.
Baesens, Gestel, et al. (2003)	Logit, LDA, QDA, LP, LS-SVM, NN, DT, naïve Bayes, k-NN classifiers		German and Australian Credit data from UCI, two major Benelux financial institutions, four UK financial institutions	1. LS-SVM and NN showed good accuracy performance. 2. Simple and linear classifiers (eg. LDA and Logit) performed closely to LS-SVM and NN. 3. QDA, naïve Bayes and DT showed weak performance.
Baesens, Setiono, et al. (2003)	NN (MLP), Neurorule	C4.5, C4.5 rules, Logit	German Credit data from UCI, and two major Benelux financial institutions	1. C4.5 outperformed other method across data set for accuracy on training data; however provide significantly excessive rules, thus increasing complexity. 2. When complexity is taken into account, Neurorule was the best technique for all data sets.
Lai, Yu, et al. (2006b)	NN-based Metalearning	LDA, Logit, Single ANN, Single SVM, Majority-voting based metamodel	Japanese credit data from UCI	NN-based Metalearning outperformed other techniques in predicting good and bad loans.
Lai, Yu, et al. (2006a)	NN ensemble (Reliability-based)	Logit, ANN, SVM	Japanese credit data from UCI	Reliability-based NN outperformed other techniques in Type I, Type II and overall accuracy.
Tsai and Wu (2008)	ANN (single and multiple classifiers)		Australian, German, and Japanese credit data from UCI	Single multiplier was more suitable for bankruptcy prediction and credit scoring development.

Author(s)	Classification Algorithm	Compared with	Data	Findings
Šušteršič, Mramor, and Zupan (2009)	BPNN	Logit	Credit data from a Slovenian bank	1. BPN outperformed Logit in overall accuracy, Type I error and Type II error. 2. Efficiency of BPNN increased with reduction of attributes.
Tsai et al. (2009)	DA, Logit, NN, DEA-DA		Credit data from a Taiwan financial institution	1. DEA-DA outperformed other techniques in overall and bad loans accuracy, followed by NN. 2. All techniques had good predictive capability, with DEA-DA and NN being the optimal ones.
SUPPORT VECTOR MACHINE (SVM)				
Van Gestel et al. (2003)	LS-SVM	OLS, OLR, NN (MLP)	Financial institutions in BankScope database	LS-SVM outperformed other techniques in 4 out of the 5 credit ratings.
Dikkers and Rothkrantz (2005)	LR, Logit, SVM (tree decoding and logit decoding)		Credit data from ABN AMRO	1. All techniques showed compatible accuracy performance on test data. 2. SVM-tree decoding showed superior accuracy performance on train data reflecting over-fitting problem.
Wang, Wang, and Lai (2005)	F-SVM	LR, Logit, NN, SVM, unilateral-weighted F-SVM, Bilateral weighted F-SVM	UK corporations from public database, Japanese credit card application from UCI, credit data set from CD-ROM of another study.	F-SVM showed promising results compared to other techniques, but not consistent across data set.
Lai, Yu, Zhou, et al. (2006)	LS-SVM	LR, Logit, ANN, SVM	Credit data from Thomas, Edelman and Crook (2002)	LS-SVM outperformed other techniques in Type I, Type II and overall accuracy.
Lee (2007)	SVM	MDA, CBR, BPNN	Not clarified	SVM outperformed other methods
Wei, Li, and Chen (2007)	SVM-MK (with Mixture of Kernel)	MCLP, MCNLP, DT, NN	Credit data from major US commercial bank	SVM-MK outperformed other techniques in overall accuracy, Type I error and Type II error.

Author(s)	Classification Algorithm	Compared with	Data	Findings
Yu, Wang, Wen, et al. (2008)	RS-SVM	Logit, ANN, SVM, RS, Fuzzy-SVM, Neuro-RS,	UK credit data from FAME database and Australian credit data from UCI	1. For the single method, SVM outperformed other techniques on overall, Type I and Type II accuracy across data set. 2. For the hybrid method, RS-SVM outperformed other hybrid techniques on overall, Type I and Type II accuracy across data sets.
Zhang and Hui (2009)	SVM	BPNN	Australian and German credit data from UCI	SVM and BPNN showed comparable accuracy performance for both data sets.
Zhou, Lai, and Yu (2010)	LS-SVM of two ensembles based on (1) reliability of the decision and (2) weight assignment strategies.	LDA, DLDA, QDA, DQDA, Logit, PR, DT C4.5, DT ID3, CART, BPNN with tangent sigmoid transfer function in hidden layer, BPNN with symmetric saturating linear transfer, Probabilistic NN, Bayesian classifier, k-NN, Adaboost algorithm, LS-SVM with RBF Kernel and with Linear Kernel, 1-norm SVM with RBF Kernel and with Linear Kernel	German Credit data from UCI and England financial service company	1. SVM with weight-based outperformed other techniques for German data 2. K-NN50 was superior for England data. 3. All SVM ensembles poorly classified bad loans for England data.
Hens and Tiwari (2012)	SVM	SVM-GA, BPNN, GP	Australian and German credit data from UCI	Given the reduced computational time, SVM showed competitive accuracy results with other techniques for both credit data.

Author(s)	Classification Algorithm	Compared with	Data	Findings
DECISION TREE (DT)				
Srinivasan and Kim (1987a)	RPA, MDA, Logit, GP, AHP		214 corporate credit data	RPA outperformed other techniques on overall accuracy performance. RPA was slightly superior in replicating expert's judgment.
Joos et al. (1998)	Logit, DT		Credit records from a Belgium bank	1. Logit performed better than DT on financial ratio attributes on overall accuracy. 2. DT performed better than logit on non-financial ratio attributes on overall accuracy. 3. DT had lower Type I error than logit on credits with longer maturity date.
GENETIC PROGRAMMING (GP)				
Ong, Huang, and Tzeng (2005)	GP	NN (MLP), C4.5, RS, Logit	Australian and German credit data from UCI	GP outperformed other techniques, but NN and logit showed satisfactory performance for alternative techniques.
Huang, Tzeng, and Ong (2006)	2GP	NN (MLP), CART, C4.5, RS, Logit	Australian and German credit data from UCI	1. 2GP outperformed other techniques, but GP, NN and logit showed satisfactory performance for alternative techniques. 2. The model was more understandable for users since they are presented in IF-THEN rules.
FUZZY CLASSIFIERS				
Hoffmann et al. (2007)	Approximate Fuzzy, Descriptive Fuzzy	DA, Bayes, ANN, C4.5	Australian and German credit data from UCI, credit data from major Benelux financial company	1. The fuzzy classifiers showed comparable results as other techniques. 2. There was a trade-off between smaller number of complex rules and a larger rulebase composed of more intuitive linguistic rules.

Author(s)	Classification Algorithm	Compared with	Data	Findings
HYBRID METHODS				
Hoffmann et al. (2002)	Genetic Fuzzy, Neuro Fuzzy	C4.5 and C4.5 rules	Credit data from a Benelux financial institution.	1. Genetic Fuzzy outperformed other techniques but the rules are less understandable for users. 2. Neuro Fuzzy showed inferior performance but the rules are understandable for users.
Huang, Chen, and Wang (2007)	Hybrid SVM (with GA, Grid, Grid+F-score)	BPNN, GP, DT	Australian and German credit data from UCI	SVM-based techniques performed identical to BPNN and GP.
Yu, Wang, Wen, et al. (2008)	SVM-RS	Logit, ANN, SVM, RS, Fuzzy-SVM, Neuro-RS	UK corporations credit data from Financial Analysis Made Easy (FAME) CD-ROM	SVM-RS outperformed other techniques in overall, Type I and Type II accuracy rate.
Zhou, Zhang, and Jiang (2008)	RS-C4.5 (RSC)	Single C4.5, BPNN, GP, SVM-GA	Australian and German credit data from UCI	Using 10-fold cross validation, RSC outperformed other techniques in both data sets.
Yao (2009)	Hybrid SVM (neighbourhood RS for input features, Grid search for kernel parameters, 10-fold cross validation)	LDA, Logit, NN	Australian and German credit data from UCI	1. Neighbourhood RS and SVM-based classifiers outperformed other techniques for overall classification rate on both credit data. 2. Features that were selected using the neighbourhood RS showed the best accuracy compared to other feature selection techniques for both credit data.
Chen and Li (2010)	Hybrid SVM (Classifiers: LDA, DT, RS, F-score)	(Original) SVM	Australian and German credit data from UCI	1. All hybrid models achieved similar classification accuracy for the Australian credit data. LDA-SVM presents fewest input features. 2. F-score+SVM outperformed other models and have fewer input features for German credit data.

Author(s)	Classification Algorithm	Compared with	Data	Findings
Yao and Lu (2011)	Hybrid SVM (neighbourhood RS for input features, Grid for kernel parameters, 10-fold cross validation)	On input features: t-test, correlations matrix, stepwise regression, RS, CART, MARS. On classification accuracy: LDA, Logit, NN	Australian and German credit data from UCI	1. SVM-based classifiers and neighbourhood RS outperformed other techniques for overall classification rate on both credit data. 2. Features that were selected using the neighbourhood RS shows the best accuracy compared to other feature selection techniques for credit data.

2.3.3.1 Neural network (NN)

Wu and Wang (2000), Bensic, Sarlija, and Zekic-Susac (2005), Angelini, di Tollo, and Roli (2008), Derelioğlu, Gürgen, and Okay (2009) and Dima and Vasilache (2009) conducted studies on credit risk assessment for small businesses using NN or its variants as the main technique and compared the performance of it with other selected techniques.

Wu and Wang (2000) applied NN to take advantage of its fast processing time and good handling of noise and missing data, and found it particularly useful for small business lending. Their study aimed at replicating prior credit-granting decisions (acceptance or rejection) of loan officers. LDA, QDA and k-NN were used for comparative purposes. The data set was obtained from a bank in New York, and consisted of 182 small business loan applications with principal loan amounts of less than USD 4 million. Of the 182 cases, 28 were rejected. The inclusion of rejected applications presented an advantage as these data are often exceptionally difficult to attain. There were 19 attributes used for the classification task (Table 2.9), obtained from documents submitted during the loan application and from credit agencies. During the experiment stage, cases selected for training and test were rotated between both sets; thus the training and test data sets consist of every available case. In the pre-processing step, the data were normalized using the zero-mean technique. Afterwards, the back-propagation technique was applied to the whole sample to extract the discriminating variables.

Table 2.9 Credit risk attributes applied to Wu and Wang's study

Category	Name of Attributes
Company's profile	Business' age
Financial Report Information	Combined debt service ratio Business net cash flow Debt to net worth ratio Current ratio Return on total assets Sales to receivables ratio Receivables turnover Sales to total assets
Credit Rating	Salary and other income Credit Bureau score Total amount of existing credit obligations The Dun & Bradstreet Credit Bureau score Number of major derogatory ratings Number of inquiries in last six months Percentage of tradelines never delinquent Maximum number of delinquency ever
Loan Structure	Total loan amount requested Term of loan

The results from classification training in the experiment stage were then compared to actual occurrences, and presented in the form of the confusion matrix (Table 2.10). NN had 100% accuracy in true positive cases, and other methods failed to match this performance. Although NN provided low accuracy for the rejected cases, the results remained better than those of other techniques. The researchers argue that the failure in identifying rejected cases, based on discussions with knowledgeable and experienced bank staff, was due to the conservative outlook of the bank. Another reason for the bad performance was the omission of qualitative information in the study although it was used in the actual decision making process. The exclusion of descriptive qualitative information was inevitable since NN works only with numerical values.

Table 2.10 Results of Wu and Wang's study

Actual Classification	**Model Prediction**	
	Accept	**Reject**
NN		
Accepted	100.00%	0.00%
Rejected	57.10%	42.90%
LDA		
Accepted	92.90%	7.10%
Rejected	67.90%	32.10%
QDA		
Accepted	94.40%	4.60%
Rejected	78.60%	21.40%
k-NN		
k=1 Accepted	94.10%	5.80%
Rejected	78.60%	21.40%
k=3 Accepted	98.00%	1.90%
Rejected	85.70%	12.30%
k=5 Accepted	98.00%	1.90%
Rejected	92.90%	7.10%

The study by Bensic, Sarlija, and Zekic-Susac (2005) focused on the extraction of discriminant credit risk factors for small business lending in a transitional economy. The authors applied logit, BPNN, RBFN, Probabilistic Neural Network (PNN) and LVQ, and CART decision tree algorithms to classify good and bad loans. Good loans were indicated as those that had not experienced delinquency for 45 days or more, while bad loans were those that had experienced at least one overdue payment in 46 days or more. Data, collected from a Croatian saving and loan association, consisted of 144 accepted (66 good and 78 bad loans)

and 16 rejected small businesses loans. Experiments were performed on two levels. In the first level, the authors used only the accepted loans and NN algorithms. The rejected loans were used in the second level experiment where all techniques were applied. The best algorithm (BPNN) and logit classified the rejected loans differently, with BPNN dividing the rejected loans equally as good and bad loans while logit classified 30% as good and 70% as bad. The authors combined the rejected loans with the 144 accepted samples rather than performing a separate analysis on them. These data were partitioned into 75% for training and test data and the other 25% kept for validation purposes. Oversampling was used to maintain equal numbers of good and bad loans in the training and test data sets. Bad loans formed 65.79% of the validation data set. The 31 initial attributes were reduced to 19 after the feature selection process. All categorical data were transformed into binary (dummy) variables.

Results from this study are contained in Tables 2.11 and 2.12. Table 2.11 shows the attributes for credit risk assessment extracted by the different techniques applied. These discriminating attributes were derived from the data set that included rejected loans.

Table 2.11 Discriminant credit risk attributes in Bensic, Sarlija and Zekic-Susac's study

Category	Name of Attributes	Logit	NN				CART
			BPNN	RBF	PNN	LVQ	
Small business characteristics	Main activity (industry) of the small business			Y	Y		
	Starting a new business undertaking			Y			
Personal characteristics of entrepreneurs	Entrepeneur's occupation		Y	Y	Y		
	Entrepeneur's age		Y				
	Business location			Y			
Credit programme characteristics	Method of interest repayment	Y	Y	Y	Y		
	Method of principal repayment				Y		
	Length in months of the credit repayment period				Y		
	Grace period in credit repayment				Y		
	Interest rate			Y	Y		
	Amount of credit		Y				Y
Growth plan	Planned value of the profit reinvested (percentage)		Y	Y	Y	Y	Y
Entrepreneurial idea	Clear vision of the business	Y	Y	Y	Y	Y	Y
	Sale of goods/services		Y				
Marketing plan	Advertising goods/services	Y					Y
	Awareness of competition			Y	Y		

Table 2.12 shows the overall classification performance of the model developed with each technique. PNN was presented under NN as this algorithm outperformed the other three NN algorithms. From the table it can be seen that NN has the lowest Type I error while LR has the lowest Type II error. The authors found no statistically significant difference between the models generated by the techniques.

Table 2.12 Results of Bensic, Sarlija and Zekic-Susac's study

Actual Classification	**Model Prediction**	
	Good	**Bad**
LR		
Good	88.00%	12.00%
Bad	46.15%	53.85%
NN		
Good	84.00%	16.00%
Bad	15.38%	84.62%
CART		
Good	68.00%	32.00%
Bad	38.46%	61.54%

Angelini, di Tollo, and Roli (2008) applied classical and ad-hoc feed-forward NN in an effort to overcome weaknesses originating from human factors. These included subjectivity, lack of sectoral knowledge of 5Cs, and imprecision and technical limitations stemming from statistical assumptions. The study aimed at classifying accepted loans as good and bad. Good loans were those that were repaid in full at the end of the analysis period, while bad loans were those that were not. As many as 76 SMEs from a wide range of industries were used, based on three-year loan data collected from an Italian bank. The data were partitioned into two groups: 53 loan applications used as training data and 23 as test data, with a relatively equal ratio of good and bad loans in each set. A total of 11 ratios (Table 2.13) were selected as attributes after removing missing and wrong values.

The authors found that the classical feed-forward NN is a better technique in classifying bad loans cases, with 13.3% of Type II error in the test data only. In comparison, the ad-hoc feed-forward NN had Type II errors in both the training and test data sets. The researchers propose that both techniques should be applied complementarily to assess SME credit risk. The problems in this study relate to data pre-processing. The value of some of the financial ratios used differed due to the different industries of the SMEs, therefore showing outlier values. The authors used arithmetic means to smooth the range and replace missing values. The effectiveness of this normalization technique remains unclear because of possible information loss.

Table 2.13 Credit risk attributes in Angelini, di Tollo and Roli's study

Category	Name of Attributes
Financial Ratios	Cash flow/Total debt Turnover/Inventory Current liability/Turnover Equity/Total assets Financial costs/Total debts Net working capital/Total assets Trade account receivables/turnover Value added/Total assets
Other Ratios	Actual credit used/maximum credit threshold Transpassing medium-long term/maximum credit threshold medium-long term Actual credit used medium-long term/maximum credit threshold medium-long term

Derelioğlu, Gürgen, and Okay (2009) added to these studies by targeting dimension reduction through feature selection. Their objective was to classify MSME loans as good and bad. With MSMEs forming 95% of the economy in Turkey, there was a pressing need to find proper discriminating credit risk attributes. In this study, DT, Recursive Feature Elimination SVM (SVM-RFE), Factor Analysis (FA) and Principal Component Analysis (PCA) were used to extract attributes, and then a classification of loans into good and bad was performed using MLP. The performance of MLP was compared with k-NN and SVM. The data set contained 512 loans, 144 good and 368 bad. There were 27 attributes covering aspects of demography, financial, general risk, and delinquency.

The study was conducted in two stages. The first stage was to find a reduced data set. There were six sets of data with attributes extracted by different feature selection techniques. Several accuracy measurements were applied to each set to find the most robust one based on the confusion matrix, accuracy rate, Matthews Correlation Coefficient and misclassification rate. In the second stage three techniques were applied to find the best classifier. Although k-NN outperformed the other two techniques, it was found to be too sensitive to economic fluctuations; the authors regard MLP the best classifier as it shows good accuracy performance and outperforms SVM. The authors also acknowledge the limited generalizability of the results due to random sampling.

The work of Dima and Vasilache (2009) was aimed at identifying businesses (micro or small) that were likely to be bad payers. The data used were of 3,000 businesses in Bucharest, 2,000 of which had delayed loan payments more than one month (bad payers) and the other 1,000 which had delayed loan payments for less than one month (good payers). Using probit regression, it was found that an increased loan amount caused greater delay in payments for both types of business, and that the delay occurred more among small

businesses than micro businesses. The authors used MLP to classify the businesses into good and bad payers. Table 2.14 shows the classification performance of MLP for training (classification) and test (prediction) data.

Table 2.14 Results of Dima and Vasilache's study

Actual Classification	Model Classification		Model Prediction	
	Good payers	**Bad payers**	**Good payers**	**Bad payers**
Good payers	82.20%	17.80%	81.30%	18.70%
Bad payers	24.60%	75.40%	17.70%	82.30%

2.3.3.2 Support vector machine (SVM)

A limited number of studies of MSME credit risk assessment use SVM. Other than the aforementioned study by Derelioğlu, Gürgen, and Okay (2009), only two other published studies could be retrieved. Chen and Li (2009) applied SVM to small businesses in China, and Kim and Sohn (2010) used this technique for government-funded small businesses in Korea.

The work by Chen and Li (2009) was to identify discriminant credit risk attributes for small businesses and produce a credit risk classification model using SVM. The authors compared the technique with BPNN. The small data set of 32 SMEs limited the generalization power of the developed model. Of the 32 cases, 24 were good loans and the remainder were bad. They used six attributes for the classification task (Table 2.15), obtained using factor analysis and SVM. Table 2.16 shows the better performance of SVM in comparison with BPNN. The authors argue that SVM is more robust than BPNN as the decrease in the SVM's accuracy performance on test data was smaller than that of BPNN.

Table 2.15 Credit risk attributes applied to Chen and Li's study

Category	Name of Attributes
Operating efficiency	Credit ratio The rate of supply contract compliance Promotion cost rate Product sales rate
Customer profitability	The relative market share The personal target completion rate

Table 2.16 Results of Chen and Li's study

Technique	Model Classification		Model Prediction	
	Accuracy rate	Misclassification rate	Accuracy rate	Misclassification rate
BPNN	75.00%	25.00%	79.20%	20.80%
SVM	87.50%	12.50%	91.67%	8.30%

The study by Kim and Sohn (2010) was driven by the high default rate of SMEs that had received Korean government loans. The applications had been assessed based on financial statements, general characteristics of the company and a technology scorecard. The authors argued that the scorecard had flaws, and this study was aimed at finding a credit risk assessment model specific for SMEs applying for the technology-fund loan grants. Technology-fund is scheme provided by Korean government for SMEs to stimulate the use of technology in business. In this study, SVM was used and compared with BPNN and logit. The data set consisted of 3,827 loans with 3,103 well performing loans and 724 defaulting loans. Defaulting loans were those with any one of the following incidents within three years after receipt of the funds: delayed payment, bounced check, failed product commercialization, bad credibility of managers, business closure or corporate reorganization.

Table 2.17 Credit risk attributes applied to Kim and Sohn's study

Category	Name of Attributes
SME characteristics	Listed in the stock market
Financial ratios	Net income/Stockholders' equity Total assets turnover Total assets growth rate Debt ratio
Technology scorecard	Technology knowledge Fund supply Output of technology development Technology commercialization potential Market potential Market characteristics Business progress (for new company) or amount of sales (for old company)
Economic indicators	Economic situations index of SMEs Economic preceding index

Initially there were 43 attributes comprising eight SME characteristics, nine financial ratios, 16 technology evaluation scores and ten economic indicators; after using stepwise selection, 14 attributes were selected, shown in Table 2.17. The authors used oversampling

to maintain equal numbers of performing and defaulted loans in the training set. The data was partitioned into 80% training and 20% test sets, except for BPNN where the data was partitioned into 60% training, 20% validation and 20% test sets. The experiments showed that SVM has the highest prediction accuracy of 66.16% in comparison with the 64.16% of logit of and 64.23% of BPNN.

2.3.3.3 Rough set (RS)

Rough set (RS), a technique introduced by Pawlak (1991), typically is used to handle vague and imprecise data by finding the lower and upper approximations. All data that cannot be clearly distinguished using the available information is placed on the upper approximation and is roughly sorted into classes. A study by Daubie, Levecq, and Meskens (2002) applying RS to MSME credit risk assessment used RS to select the attributes and perform the classification of the data set. The results were compared to those of DT. The data set was provided by a Belgian bank and consisted of 1,102 loans with 633 low loan risks and 469 high. Initially there were 21 attributes, eleven describing characteristics of the company and ten related to the company's financial condition. After the attribute selection process was completed by applying RS, eight attributes were used (Table 2.18). Since RS works only with categorical data, the values of these attributes were discretized in the pre-processing stage.

Table 2.18 Credit risk attributes applied to Daubie, Levecq and Mesken's study

Category	Name of Attributes
SME characteristics	Quality of management
Financial ratios	Stockholders' equity/Total assets Net worth/(Net worth + Long Term Debt) Quick assets/Current liabilities Current assets/Current liabilities Net income/Total assets Value-added/Salary expenses Pre-tax funds flow/Interest expenses

From Table 2.19 it is evident that the pruned DT outperformed RS in terms of overall accuracy. For misclassification, RS showed better performance than DT in classifying bad loans, but DT classified good loans better for this data set. Since type I errors, i.e. bad loans classified as good loans, are more costly for banks, DT is regarded the better technique. The authors mentioned an over-fitting issue that occurred when DT was used.

Table 2.19 Results of Daubie, Levecq and Mesken's study

Technique	Accuracy rate	Type I error rate	Type II error rate
RS	76.68%	15.05%	9.81%
DT	87.50%	8.90%	17.40%

2.3.3.4 Random survival forest (RSF)

The work of Fantazzini and Figini (2009) was the first application of the Random Survival Forest (RSF) in the credit domain. It was introduced by Ishwaran et al. (2008) as an extension of the Random Forest method proposed by Breiman (2001). The idea of this extended technique is to produce trees with branches that take into consideration all attributes, providing lower variance and bias. Since RSF is a non-parametric technique, the authors compared the performance with logit, a parametric technique. The data set comprised credit data of 1,003 SMEs provided by Creditform, a rating agency for SMEs in Germany. The attributes used in this study comprised 16 financial ratios (Table 2.20). RSF used all attributes to generate the trees and calculate the importance of values for these attributes. After this calculation, four attributes that are marked with an asterisk in Table 2.20, were found to be significant.

Table 2.20 Credit risk attributes applied to Fantazzini and Figini's study

Category	Name of Attributes
Financial ratios	Supplier's target days (days of debts and other payables)* Outside capital structure (capability of the firm to receive non-bank loans for financing) Industrial rights (degree of innovation in relation to its size) Liquidity ratio Debt ratio Equity ratio Capital tied up (gross long-term debt/total assets) Short-term/ long-term debt Tax/Sales Provisions/Sales Personnel expenses/Sales* Depreciation/Sales* Net income/Total assets* Equity/Debt Short-term debts/Overall debt Interest income/Total assets

Table 2.21 shows that the RSF credit risk model performed much better with the training data than with the test data. It was argued that the high degree of bias that could be associated with logit produced low variance estimates and low boundary errors. In terms of

input attributes, the authors recommended the inclusion of qualitative attributes to obtain a more complete understanding of MSME credit risks.

Table 2.21 Results of Fantazzini and Figini's study

Technique	Model Classification	Model Prediction
RSF	93.17%	76.69%
Logit	85.13%	84.14%

2.3.4 Identification of the research gap

It is evident from the literature survey that statistical and machine learning techniques provide good credit risk assessment models for MSME. With the increasing need for efficient credit assessment models (Ince and Aktan 2009), machine learning techniques have been applied more than statistical techniques in studies to date. Structured credit data used in the statistical and machine learning techniques were either in numeric or categorical format; categorical data were transformed into binary variables. The quality of credit data used in these studies can be regarded as reliable since they were obtained from actual loan documents provided by financial institutions or credit bureaus. Studies performed in developing countries were more subject to a lack of relevant credit risk information than studies in developed countries. Information obtained from local credit bureaus was applied in some studies to overcome this problem. Table 2.22 on the next page shows the type of attributes used in the various studies and the countries in which they were conducted.

The research for this thesis focuses on Indonesian MSMEs. Credit risk assessment for MSMEs in Indonesia is typified by a tedious process of data collection, verification and processing to ensure sufficient information is available for prudent credit assessment. The lack of reliable independent credit bureaus makes it hard to obtain reliable credit information. In this regard, the research presents a different setting. Credit risk assessment for MSMEs as conducted by the three banks that are sampled for this research is discussed in the next section. The credit data about Indonesian MSMEs is a combination of structured and unstructured types. The structured data comes from financial information, while the unstructured data comes from text that describes business management and performance in a non-numerical way. Hence, the challenge is to develop a technique to present and analyze these two types of data in tandem. The unstructured data creates complications, since the well developed and explored statistical and machine learning/data mining techniques are applicable to relational data with well-defined structure. This can be seen in the studies discussed above, where categorical data has been transformed into binary variables.

Table 2.22 Highlights of studies of MSMEs' CRA

Authors and Year of Study	Type of Attribute Data		Country of Study
	Numeric	Binary	
Durand (1941)	Y		United States of America
Edminster (1971)	Y		United States of America
Viganò (1993)	Y	Y	Africa
Longenecker, Moore, and Petty (1997)	Y		United States of America
Altman and Sabato (2007)	Y		United States of America
Gool et al. (2009)	Y	Y	Bosnia-Herzegovina
Wu and Wang (2000)	Y		United States of America
Bensic, Sarlija, and Zekic-Susac (2005)	Y	Y	Croatia
Angelini, di Tollo, and Roli (2008)	Y		Italy
Derelioğlu, Gürgen, and Okay (2009)	Y	Y	Turkey
Dima and Vasilache (2009)	Y		Romania
Chen and Li (2009)	Y		China
Kim and Sohn (2010)	Y		Korea
Daubie, Levecq, and Meskens (2002)	Y		Belgium
Fantazzini and Figini (2009)	Y		Germany

2.4 Credit risk assessment principles and methods in Indonesia

Banks in Indonesia are required to apply the 5Cs to assess all loan applications as per Banking Law Number 10. This law does not go into detail about the application of principles and assessment methods; it stipulates only that collateral take the form of the goods, financed project, or assets financed, or accounts receivable. Banks are prohibited from taking land with traditional ownership as collateral. Banks are not obliged to demand additional collateral in a form of assets not directly linked to the financed project. More detailed guidance is provided in BI Regulation Number 7/2/PBI/2005, Quality Assessment of Commercial Bank Assets. The aspects of 5Cs covered in this regulation are business prospects, financial performance and capacity to pay the loan. These are broadly addressed with a focus on basic guidelines. With this level of directive from the central bank, operating banks in Indonesia remain autonomous in their practice of the 5Cs principles and their application of assessment methods. Banks are supported in their assessments by the Debtor Information System (DIS) by BI, which serves as an interbank credit bureau providing information on historical credit performance of debtors. Banks can obtain information about poor credit performance of debtors such as unpaid loans overdue, defaults or bounced check issuance. As mentioned in Section 1.8, the three sample banks are selected based on their

micro and small loan exposures. Aside from the three selected banks, no other banks comply with this requirement. The three banks also apply different lending methods and criteria. In the next sub-sections, the application of credit risk assessment principles and methods for MSMEs as practiced by each bank used in this research is discussed.

2.4.1 Credit risk assessment application – Bank A

Bank A is a rural bank or BPR, owned by an Indonesian citizen and daily managed by an appointed director. It is located in an area that is highly populated and renowned for its high number of small businesses. The market segment of Bank A consists of formal and informal micro businesses. Many of Bank A's customers are recurring debtors who have over time established professional and personal rapport with the loan officers.

The credit risk assessment process starts with the collection of customer information required for a credit decision. The loan officer interviews the applicants and performs on-site visits using a survey guideline. The collated financial information covers cash inflows and outflows for one year of business operation. Cash outflows include both business and personal expenses, since these are difficult to separate and are normally paid from the business income. This case is typical for customers who run businesses from their home. Taking into account their unsystematic (non-formal) financial documents, loan officers have to infer much of this financial information from the available documents and their on-site visits. The non-financial information covers the managerial skills of customers (applicants), business competition faced and the repayment plan. The repayment plans of the customers are stated as an expression of intent[5] to comply with the bank's credit instalment payment policy.

After the data collection is completed, it is verified by another banker. In the verification process, a request for historical loan performance is sent to the BI. If the customer is listed on DIS, the application is rejected. Each loan application involves two staff members: one is in charge of the data collection and the other is responsible for data verification. A credit report based on this verified information consists of three sections: "Loan Information", "Objective Analysis" and "Subjective Analysis". Loan Information details the proposed loan and its repayment plan. Objective Analysis presents the constructed financial information of the customer. Subjective Analysis provides a descriptive computation of each amount written in the Objective Analysis and also contains the non-financial information. The determinant of the credit decision is the proportion of the daily instalment payment to daily net income.

[5] This is part of the qualitative analysis whereby loan staff are required to obtain customers' expressed compliance with the bank's loan payment policy.

Once the report is finished, the loan officer discusses the proposed loan and repayment plan with the customer. The proposed loan amount can be different from the amount requested by the customer as it is calculated based on the constructed financial information and projected daily repayment capacity. The bank has a policy to provide loans in accordance with what the bank calculates as the required finance for the business. Many times, the customers request higher amounts than the bank is willing to advance; the difference, presumably, arises from the fact that business and personal expenditures are not clearly separate in micro businesses. When an agreement cannot be reached, the loan application is rejected. Applicants are always provided with reasons for rejection. When an agreement is reached, the loan officer recommends the loan application for approval.

The bank conducts daily meetings where all aspects, such as collection of instalment payments and data collection, are discussed. In this way, credit risk assessment is performed collectively by bank staff, but the final decision is made by either the loan officer or the senior manager. The director and the owner oversee the reliability and validity of the credit data on loan applications. The authority of the staff to accept or reject an application is based on the amount of the loan principal. The loan staff are authorized to accept or reject loans up to IDR 10 million (USD 103).[6] The head of credit in the bank is authorized to make the decision about loans between IDR 10 million and IDR 25 million (USD 2,562), and the director about loans between IDR 25 million and IDR 100 million (USD 10,247). Loans of more than IDR 100 million are authorized by the director and the owner as a joint decision. Two persons are required to sign off all applications. Bank A keeps all loan applications in hard and soft copy. Computer files about applicants are kept separately, which means a database containing information about all loan applications does not exist. It is evident, from the assessment methodology applied by Bank A, that the experience and knowledge of loan staff are the primary decision-making tools.

Bank A requires daily payment on both loan principal and interest. The payment is collected each day (including Saturdays and Sundays) and is collected in person by bank staff. On public holidays the payments are put off; customers have the choice to pay before or after the holidays, and this is included in the expression of intent.

2.4.2 Credit risk assessment application – Bank B

Bank B is a state-owned commercial bank with more than 9,000 offices across the country. This bank originated from village banks providing finance to farmers and traders. The bank maintains its existence at grassroots level, and focuses on micro, small and

[6] The conversion rate used in this chapter and subsequent chapters is USD 1 = IDR 9,7087.37864, rounded to the nearest dollar.

medium businesses. The type of commercial finance this bank provides is revolving credit. Although the maximum duration of working capital finance is three years, most of the finance used for this research is for twelve months. As with Bank A, many of the customers are recurring debtors who have good rapport with the loan officers. The process of credit risk assessment in Bank B is carried out in three stages based on the screening sections on the credit scoring form.

The first screening section relates to the business sector of the applicants, and their credit trustworthiness as recorded by the central bank as well as by Bank B. In this section, the account officer marks a "Yes" or "No" box to indicate the answer. Bank B regularly updates a list containing business sectors that are not recommended (high risk) and those that are not approved in financing. An application is rejected when the business sector is confirmed (by a "Yes") as in one of the not approved sectors. An applicant whose business sector is one that is not recommended for financing can still advance to the next stage if two other conditions are satisfied: cash collateral is provided and special permission is obtained from the senior manager. Information about the credit trustworthiness of the applicant is obtained from BI and from the Bank B system. An applicant cannot be listed as a default debtor on DIS and Bank B must have no negative credit records of the applicant.

When an applicant satisfies these requirements, the account officer interviews the applicant and performs on-site visits to collect data for the second stage of assessment. The data is used to complete the second section of the scoring form, which relates to financial information. If the applicant has good documentation of financial transactions, in the last three years, this is used. If the applicant does not have systematic documentation, the account officer works with the applicant to construct elementary financial statements. The credit risk assessment system, which was developed by an independent agency for Bank B, generates selected financial ratios and other growth indicators from these financial statements, which are then compared with the required conditions set by Bank B for each ratio and growth. For example, Bank B requires a current ratio above 1.4, so the system compares the calculated current ratio with this pre-set condition. Based on this, the system provides a score to each ratio and growth indicator that shows if they satisfy the bank's condition. The loan application progresses to the next step if the account officer concludes from the score that it is acceptable to the bank.

In the third stage of assessment, the account officer collects non-financial information through a series of more comprehensive interviews and on-site visits than the ones conducted for the previous stage. While interviews are conducted with the owner of the business, who then still has to fill in the financial section of the scoring form, they are extended to business associates and family members to complete the non-financial section of the form. The collated information is detailed in another report that comprises information about the

business, the 5Cs credit risk assessment principles, and risk mitigations. Essentially, the main part of this report consists of narration of the credit scoring form. On the report, as part of the application of 5Cs principles, the financial standing of the applicant is described and analyzed using more ratios than the ones that were scored. The application of 5Cs credit risk assessment principles depends on the loan amount: a loan application up to IDR 500 million (USD 51,235) is assessed based on repayment capacity and character only. On completion of the report, all information is verified by another account officer, including obtaining legal opinions from relevant legal officers on content where needed; for example, an opinion from a notary on the validity of the legal documentation on land that is submitted as collateral. The account officer scores the non-financial performance of the applicant according to the pre-set credit scoring categories of Bank B. For each category, a weight is assigned and a total score is acquired. The final credit score is not a numerical figure: the form has three blocks, black, grey and white. One of these has to be marked. If the applicant does not completely comply with any of the three assessment categories, then black is marked and the application is rejected. If the application is not totally rejected because of definite non-compliance, the perceived risk may be indicated by marking the grey block, if the opinion is that the loan application should still be considered. In the case of total compliance, the white block is marked. This recommendation and the compiled documents (scoring form, report, invoices, legal proofs, etc.) together with the proposed loan covenant that is agreed to the applicant are forwarded to the head of the branch office for approval. At any assessment stage, the bank staff have to provide reasons why rejection of an application is recommended.

If the head of the branch office grants a credit that was originally recommended to be rejected, the responsibility for that particular applicant shifts from the account officers to the head. The authority of the different level in the bank to accept or reject a credit application is based on the amount of the loan principal. The head of a branch office is authorized to make decisions about loans up to IDR 2 trillion (USD 204,939). Loans between IDR 2 trillion to IDR 40 trillion (USD 4,098,781) are forwarded to the head of the regional office. Loans above IDR 40 trillion require a decision from a committee formed at the head office of Bank B. Two persons are required to sign all recommended loan applications and the signature of the head of the branch office is required to sign off these applications. Bank B keeps all applications in hard and soft copy. Computer files about applicants are kept separately, which means a database containing information about all loan applications does not exist. It is evident, from the assessment methodology applied by Bank B, that the scoring system is the primary decision-making tool. When the colour score is grey, the loan decision has to be made at a higher level than the normal authorized level. For example, a loan amount of IDR

2 trillion can be granted by the head of the branch office. If this loan is scored grey, the decision is to be made by the head of the regional office.

Bank B requires monthly payments of interest, although there are cases where payments are scheduled quarterly or bi-annually. The instalments can be transferred through automatic teller machines (ATM) or personally deposited at any office of the bank.

2.4.3 Credit risk assessment application – Bank C

Bank C is a private forex commercial bank with more than 400 offices across the country. Initially established as a bank that finances co-operatives, Bank C expanded its operations to lending to MSMEs. Information regarding the credit risk assessment process is not made available by Bank C, and has been inferred from the data on the credit application form and from various personal communications with the responsible loan officer.

The credit risk assessment method depends on whether the requested financing is below or above IDR 500 million (USD 51,235). The account officer collects customer information through interviews and on-site visits. Each item related to the credit assessment is given a score. A ratio (grade) is assigned to each category of assessment. For loan applications below IDR 500 million, the account officer performs a seven-step risk assessment. The first step is an assessment of the applicant's financial capacity, where three groups of financial ratios are calculated. The individual ratios in these groups and their corresponding industry benchmarks are presented in a separate section of the credit proposal and a score is given to each. An evaluation of the quality of the financial information and banking activities are assessed in the second step. The third and fourth steps are based on the quality of management and the business environment. These four categories together form the basis for the allocation of an assessment rating of the borrower. The next assessment steps are credit facility analysis (step 5), collateral analysis (step 6) and analysis of collateral (step 7). Upon the completion of the seven steps, the account officer forwards the loan application to a credit committee for a decision. For loans above IDR 500 million, the assessment is performed by focusing on qualitative information first, followed by quantitative information. The qualitative information is collected by the account officer and concerns the management and competition of the business as well as the applicant's banking history with this and other banks. The quantitative section is also constructed by the account officer, and contains the financial ratios, analysis of the required financing, business prospects, repayment capacity, collateral and risk analysis.

This proposal does not contain any scores or ratings. As with applications below IDR 500 million, the proposal is forwarded to a credit committee for approval. It is evident from the assessment methodology applied by Bank C that there are two decision-making tools used. These are a scoring and rating system for loans below IDR 500 million and a

discretionary method for loans above IDR 500 million. As in Bank B, computer files about applicants are stored on the laptop of the account officer, which means a database on loan applications does not exist.

2.5 Summary

In general, the 5Cs credit assessment principles are designed to assist lenders to analyze the credit risk profile of prospective borrowers. In previous studies, the 5Cs credit assessment principles have been analyzed by way of statistical and machine learning methods with good accuracy of loan decisions. In the case of MSME loan applications, the use of techniques from both methods has given arbitrary results. These studies were conducted on reliable and complete MSME credit data provided by banks and credit bureaus in a number of countries where they were not subject to the problems introduced by unreliable and incomplete credit data, such as exists in Indonesia. The way that the three sample banks in Indonesia apply the 5Cs principles depend on their market segment, since the BI provides only a broad directive on the use of the principles in Banking Law Number 10 and BI Regulation Number 7/2/PBI/2005. All Indonesian banks included in this research collect quantitative and qualitative credit data and practise a form of relationship lending after initial lending has been conducted. As found in the overview of studies relating to credit risk assessment methods, the use of quantitative credit data is prominent in the use of credit scoring models for MSMEs. Within an academic and practical context, the inclusion of qualitative credit data has not been explored to its maximum potential. This thesis provides an approach that will allow the extraction of valuable and novel knowledge from both types of credit data using data mining techniques.

In the next chapter, data mining as a process of knowledge discovery is discussed.

CHAPTER 3 – KNOWLEDGE DISCOVERY IN DATABASES AND DATA MINING

3.1 Introduction

In order to process the credit information the banks have gathered into a credit decision, they apply either traditional discretionary manual assessment or statistical or machine learning methods. Processing the information into a credit decision offers different challenges, particularly given the often opaque information that is available from Indonesian MSMEs. One approach to process it is to use data mining.

Data mining is concerned with developing methods to acquire patterns that express relationships of a causative nature among the many variables available in large databases. The mined patterns reflect instance behaviour, and have to be verified for *actionability*. Instances (also known as examples, cases or records) are "objects from which a model will be learned, or on which a model will be used" (Kohavi and Provost 1998, p. 273). Actionability refers to "the ability of a pattern to suggest to a user that they take some concrete actions to his/her advantage in the real world" (Cao et al. 2010, p. 96). To represent knowledge, the extracted patterns are converted to rules. When users are able to make a decision based on these rules, the discovered knowledge is said to be actionable. Therefore, data mining is part of a knowledge discovery methodology designed to exploit information registered in a large repository.

This chapter presents the role and functionality of data mining in a general context, with deeper focus on data mining applications in specific domains. Sections 3.2 and 3.3 provide a brief explanation on the sequential steps of knowledge discovery in databases (KDD) as well as of data mining. Subsequent sections are dedicated to the elaboration of the main concepts in KDD, namely types of data, common data mining tasks, types of knowledge representation, and the evaluation of the discovered knowledge.

3.2 General description of KDD and data mining

In order to maintain their competitive edge, firms are driven to provide business excellence by being customer-oriented. With the growth of business, customer orientation is extended from positioning the customers as the end target by ensuring customer satisfaction (Dean and Bowen 1994) to maximizing the involvement of customers throughout the firm production chain (Lengnick-Hall 1996). This leads to the extensive compilation of data by a firm beyond customer-related information to include internal business activities such as employee recruitment and retention, financial transactions and reports, and product

innovations that are recorded and stored in databases. These expanding databases do not exist only in business domains. Fields such as health, science and telecommunications are notable for their dependence on an influx of data to deduce a conclusive finding. For example, in the health sector, a study of early detection of adverse drug events demonstrated the existence of large databases such as that of the Food and Drug Administration and prescription event monitoring containing more than two million and one million reports respectively (Wilson, Thabane, and Holbrook 2004). With the growing size of databases, firms need to have an efficient way to manage the data so that useful information is generated and fed back to the executives for strategic decision-making. The process to achieve this is termed knowledge discovery in databases (KDD), defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, Piatetsky-Shapiro, and Smyth 1996, p. 40-41).

KDD was launched at the International Joint Conference on Artificial Intelligence workshop in 1989, where topics included the role of domain knowledge, the development of efficient, incremental and effective algorithms, and the advancement of discovery tools (Piatetsky-Shapiro 1991a). Large databases store uncountable data, some of which is likely to be irrelevant. This triggers the need to develop effective methods to extract novel and valid patterns. To reduce the size of the search space to be covered, domain knowledge imposes constraints on a search (Piatetsky-Shapiro 1991a). The inclusion of domain experts' input at the earliest stage of the mining process implied that the extracted patterns would already incorporate some degree of domain knowledge. In subsequent developments, domain expertise in KDD was given wider recognition and became a separate subject known as actionable knowledge discovery (AKD) (Cao et al. 2010); this is discussed in more detail in Section 3.7.2. The development of algorithms and discovery tools to extract hidden patterns from large databases is the substance of data mining. It is an automated method to extract such patterns, and is a multidisciplinary approach in which statistics, machine learning, database technology and information retrieval are related (Han, Kamber, and Pei 2012).

The term data mining emerged in the advent of computer-based analysis around the 1960s (Fayyad, Piatetsky-Shapiro, and Smyth 1996) and is often used interchangeably with other terminologies such as knowledge mining, knowledge extraction, pattern analysis data archaeology and data dredging (Chen, Han, and Yu 1996). There are various ways to describe data mining; all point to the same meaning: the way of making sense of sizeable data. In terms of scope, a slight variation can be found as can be seen from two definitions. Fayyad (1996b) refers to data mining as the act of extracting patterns out of large databases, while Han, Kamber and Pei (2012) call it a process of discovering interesting knowledge from large quantities of data. Fayyad provides a definition from the narrow perspective of an action; Han, Kamber and Pei take a wider perspective. The results of applying data mining

techniques to a database are patterns. Since these patterns are extracted from enormous quantities of data, they come in excessive amounts and present a practical challenge for users. In other words, the knowledge residing in the patterns is difficult to act upon directly. The patterns need to be evaluated and transformed before they become useful knowledge. Such evaluation includes the application of quality and interestingness measures, while a transformation of patterns means to present the knowledge in a form that is meaningful for the users and to incorporate this discovered knowledge with domain knowledge. Thus data mining is not a standalone project; it is part of the KDD process.

Data mining is particularly well applied in the science domain. Databases in fields such as astronomy, earth sciences, medical imaging, security and surveillance are not only large, but also host data that are complex in terms of format and dimensions (Kamath 2003). Research in these fields by Kamath (2003) provides examples of complexity and an overview of the data mining application suitable for various purposes. For example, a classification algorithm was used on the data containing more than 30 thousand 1024 x 1024 pixel images to map and list volcanoes in Venus.

In the next section, the steps in KDD are described.

3.3 KDD as a process

KDD provides a systematic passage to convert massive and autonomous data into meaningful knowledge (Fayyad, Piatetsky-Shapiro, and Smyth 1996). It encompasses steps from the initial stage of data collation to the final stage of knowledge extraction, as illustrated in Figure 3.1.

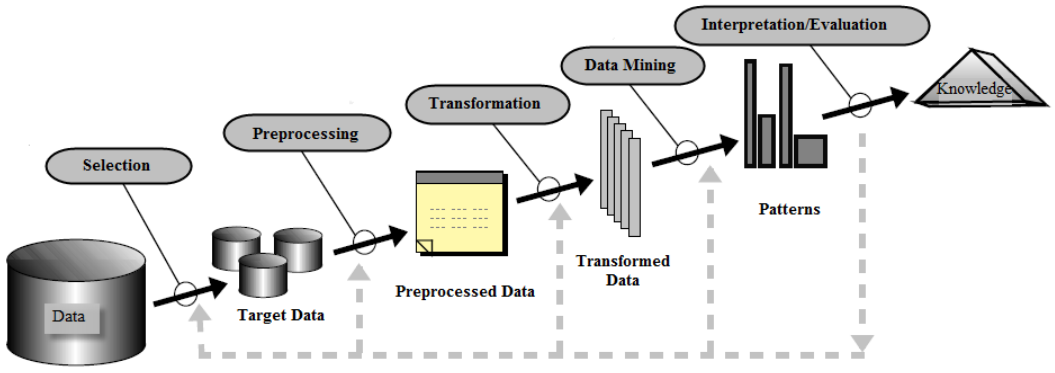


Figure 3.1 An overview of the steps that compose the KDD process

Source: Fayyad, Piatetsky-Shapiro and Smyth, 1996, p. 41.

Since a database often consists of multi-purpose data, the first step begins with data selection, whereby relevant data are obtained for a particular purpose based on the discretion of domain experts. The step to transform the raw data into readily-mined data is called pre-processing. Methods for data pre-processing include data cleaning, data integration, data transformation and data reduction (Han, Kamber, and Pei 2012). The necessity of a pre-processing step emerges from the fact that data are compiled by domain experts, and real-world data may not be in a format that mining tasks can be applied to. Data cleaning is performed to clear the data from incompleteness, noise and inconsistencies. Missing value problems can be overcome by discarding the record(s) or inputting the value manually, or by entering a number to complete missing values such as using a global constant, a central tendency of the attribute, a central tendency of the same target class or the most possible value. The problem of noise is handled by applying techniques such as binning, regression and outlier analysis, while inconsistencies can commonly be solved by manual correction. Data integration takes place when the data exist in multiple databases from different sources and must be merged into a single data warehouse. Important concerns here include entity identification, attribute naming inconsistencies and resulting correlation of the data, duplication and the use of different measurement scales or representation for attribute values. Data transformation is aimed at attaining a more efficient mining process with techniques such as smoothing, attribute construction, aggregation, normalization, discretization, and concept hierarchy generation. The final method of pre-processing, data reduction, also aims at a more efficient mining process by presenting a reduced data size. Reduction can be achieved through dimensionality reduction (feature subset selection), numerosity reduction or data compression.

Once the data are in an appropriate format, data mining techniques are applied depending on the intended purpose, namely descriptive or predictive (Han, Kamber, and Pei 2012). Data mining is the step in KDD that entails automated extraction of patterns. Although the term KDD and data mining are frequently used in the same context (Fayyad 1996a; Han, Kamber, and Pei 2012), the distinction is clear in that KDD is the entire process of knowledge discovery and data mining is a component of this process.

Mattheus (1993) notes that one challenge in KDD and data mining comes from the constant changing, and imperfect nature, of real-world databases. The successful application of KDD and data mining is documented in several fields of business as well as in domains such as health (Dinu, Zhao, and Miller 2007; Li et al. 2004), and sport (Delen, Cogdell, and Kasap 2012). In the business domain, the application of data mining includes fields of marketing (Ngai, Xiu, and Chau 2009; Shaw 2001), finance (Kirkos, Spathis, and Manolopoulos 2007; Zhou and Kapoor 2011) and banking (Hormozi 2004). A more detailed explanation of these tasks is presented in Section 3.5.

The output of data mining tasks is patterns that represent the model[7] of a particular subset of the data (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Extracted patterns are not easily understood or of value to all domain experts. In order to ensure that patterns are understandable for users, they may be transformed into a form of predictive rules or decision trees (see Section 3.6). The utility of these rules is attained through a post-processing step performed to measure the quality and interestingness of the rules. Although conformity about the meaning of interestingness does not exist, a general understanding of the term entails conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability (Geng and Hamilton 2006). Types of interestingness measures are discussed in more detail in Section 3.7. Once quality and interestingness measures are applied to the extracted patterns, these interpretations become dynamic knowledge for the particular domain. Notwithstanding the limited generalization power of the patterns that is confined to the similar context of the data subset, the overall ability to extract novel and understandable knowledge from massive databases remains a positive contribution of KDD.

3.4 Types of data

The process of uncovering hidden knowledge begins with the selection of relevant data in a database. A database is a collection of systematically organized data that enables easy management of information (Date 1977). An enterprise can have databases located in one site or dispersed. More important issues than locality are the form of data made available to be mined and the purpose of the knowledge discovery process. The different forms of data are described in this section, and the purpose of knowledge discovery is addressed in the next. According to Cao et al. (2010), the traditional practice of KDD and data mining is to use a data-centred process with the dominant technical feature to extract patterns. Thus, the selection of the right mining technique to mine the potential knowledge effectively is crucial, and dependent on the type of data. The types (from the most simple to the most complex data to mine) are relational, sequential, semi-structured and unstructured data (Hadzic, Tan, and Dillon 2011).

3.4.1 Relational data

The relational database system, introduced by Codd (1970, 1979), is identified by its relational structure, update rules on the primary key, and relational mathematical expressions. The same principles apply to the data itself. Relational-type data are collected and stored in numerous tables that represent domain-specific relationships. Each relation table consists of records or tuples (in rows) and a set of relevant attributes (in columns) that

[7] A model is "a structure and corresponding interpretation that summarizes or partially summarizes a set of data, for description or prediction" (Kohavi and Provost 1998, p. 273).

describes an entity (Han, Kamber, and Pei 2012). Furthermore, a relation table is also used to present associations between or among entities. With such clear structure, a relational database is developed based on a well-defined relational *schema*. A schema describes all entities, attributes and the prevailing entity-relationships (Kohavi and Provost 1998). Relational data are commonly used to capture information in business and are expected to remain a preferred format for documented business data due to their reliability, scalability, tools and performance (Shanmugasundaram et al. 2001).

3.4.2 Sequential data

In a sequence database, items or events are stored according to their order of occurrence and may or may not be associated with a time stamp (Han, Kamber, and Pei 2012). Examples of these types of data include customer shopping sequences, biological sequences, and business process sequences. Han, Kamber and Pei (2012) distinguish sequential data with characteristics as follows: *time series data* that are numerical and recorded at equally chosen time intervals; *symbolic sequence data* that consist of sequential events or nominal data, where time stamps are not considered relevant or observations are made at non-equal time intervals, and *biological sequences* that are long and incorporate hidden semantic information. Each data sequence records the transactions and item (or set of items) that appear in particular order on each transaction. A simple illustration involves data-sequencing the behaviour of a customer who purchases a book: suppose A purchases *The Hobbit* on 3 January and *To Kill a Mockingbird* on 15 January. The data sequence of A is displayed as "(The Hobbit) (To Kill a Mockingbird)". If A bought *And Then There Were None* together with *The Hobbit* on the first transaction, the data-sequence becomes "(The Hobbit And Then There Were None) (To Kill A Mockingbird)", and the elements of the first transaction are referred to as an item set.

Early studies using sequential data for input include that of Agrawal and Srikant (1995) on basket data, Mannila (1997) on the sequence of alarms in a telecommunication network, and Agrawal, Gunopulos and Leymann (1998) on business process. The mining techniques developed for sequential data analysis discover patterns where the chronological and optionally temporal aspects of the input data are captured (as in Agrawal and Srikant 1995; Pei et al. 2004; Yan, Han, and Afshar 2003; Zaki 2001). The subsequent development of the sequence mining techniques have been reported as successful in domains such as bioinformatics (Jones and Pevzner 2004), health (Jie et al. 2010), and hydrography (Salas et al. 2012), and in impact-targeted activity such as debt-related activities (Cao, Zhao, and Zhang 2008).

3.4.3 Unstructured data

Unstructured data types or raw data do not have a pre-defined schema with precise structure. Unstructured data are generally found in free-format objects such as text-based items (e.g. emails, blogs, documents, and social media) and images (e.g. audio and video). Since most commonly considered data in this unstructured data category is in textual form, considerable effort has been devoted to discover hidden patterns from documents by way of text mining (Feldman and Dagan 1995; Lent, Agrawal, and Srikant 1997; Mei and Zhai 2005; Nahm and Mooney 2001; Nasukawa 2001). Although recent initiatives on metadata mining have led to the production of schema-type information from unstructured data (Han, Kamber, and Pei 2012), the lack of a precise schema to guide the pattern search of unstructured data creates a difficult task: Han, Kamber and Pei (2012) note the increasing complexity of unstructured and semi-structured data. These data, often collected in real time, include dynamic data streams, spatial data (e.g. geospatial), spatiotemporal data (e.g. moving-object), cyber-physical system data, multimedia data and web data. There has been a significant amount of work done on mining complex data, but there is still a considerable gap in the development of efficient and effective mining techniques for unstructured and semi-structured data.

3.4.4 Semi-structured data

Semi-structured data types emerged with the development of the World Wide Web (WWW) technology and the need to integrate data from multiple sources with different structures (Abiteboul 1997; Buneman 1997). In a detailed construction of this particular data type, Abiteboul (1997) describes the characteristics of semi-structured data structure as being irregular, implicit, partial and indicative. Irregularity refers to the various types of elements, while implicit structure means that additional tasks are required to develop the structure. Although a schema with strict indications of attributes and relationships as in relational data types is not achievable with all data, a predefined schema in the form of a graph (Buneman et al. 1997; Calvanese, De Giacomo, and Lenzerini 1999; Wang, Yu, and Wong 2000) or some non-graph form (Beeri and Milo 1999; Goldman and Widom 1997) is attainable. The schema captures the relationships among attributes in semi-structured types of data that need to be preserved and appropriately presented, especially in domains where contexts are important. As a general rule, a graph or a tree is the underlying structure of semi-structured documents used to capture structural relationships (Buneman 1997). Data mining techniques are developed to extract patterns from graph and tree-structured data that incorporate the structural aspect of the input data. The early works on algorithms for graph mining came from Cook and Holder (1994) and Yoshida, Motoda and Indurkhya (1994). Some notable

works in algorithm development followed in subsequent years (Chakrabarti and Faloutsos 2006; Inokuchi, Washio, and Motoda 2003). The application of graph mining algorithms include domains such as chemical analysis (Cook and Holder 1994; Geibel and Wysotzki 1996; Inokuchi, Washio, and Motoda 2003; Yoshida, Motoda, and Indurkhya 1994) and web mining (Broder et al. 2000; Mendelzon and Wood 1995). With regard to tree-structured data, there are many developed algorithms over the years including those by Abe et al. (2002), Asai et al. (2002), Miyahara et al. (2001), Taniguchi et al. (2001) and Zaki (2005b) with applications in domains such as bioinformatics (Hashimoto et al. 2008; Shapiro and Zhang 1990; Zaki 2005b). Several semi-structured databases exist, such as Resource Description Framework (RDF) and XML databases, which are standards for web data exchange as set by The World Wide Web Consortium (W3C). This research uses XML to present the semi-structured nature of credit data. The structural aspects of XML and common techniques for XML mining will be elaborated in Chapter 4.

3.5 Common data mining tasks

The choice of mining tasks to be applied to data depends on the use to which the obtained knowledge will be put. Firms engaged in various types of business, science and government have collected and stored multipurpose data. It is a customary practice to fill the repositories with an abundance of data without a clear initial understanding of their subsequent use. In the case of fields for which activities are closely related to complex types of data, the inflow of a massive amount of data is generally so automatic and continuous that firms have no option other than to retain it all. Data mining conveniently sorts the data and extracts interesting patterns from them. However, the effectiveness of the tool depends on the utility of the extracted patterns, and it is important for the end user to specify the purpose the mined data will be used for before the mining techniques are chosen.

There are two main purposes for the application of data mining algorithms on a dataset: descriptive and predictive (Han, Kamber, and Pei 2012). Descriptive data mining tasks are performed to generate patterns or models applicable to subsets of the data; they explain the properties of the data and can easily be comprehended by users to achieve an understanding of the data and the relationships within them. This means describing a concept and the associations of its features in accordance with the corresponding domain. Since all attributes are treated equally, with none identified as target attributes or class(es), the mining task is known as unsupervised learning. Output examples of descriptive data mining tasks are customer profiles, natural object catalogues, faulty product management, and policy development processes. Predictive mining tasks are performed to predict missing values or the behaviour of a new dataset, using patterns extracted from the training data set. In generating models from the training data, attributes are categorized as predictors

(independent) and targets (dependent). The mining process for this is called supervised learning. Application examples of predictive data mining tasks are credit assessment, genomic medicine, and fraud identification. Common data mining tasks applicable to such end purposes are association rules mining, classifications, clustering, and outlier detection and analysis.

3.5.1 Association rule mining

Association rule mining (Han, Kamber, and Pei 2012) is a task to mine patterns that express related occurrences of items in the database, indicating correlations among the items. This task is of a descriptive nature and is typically performed to non-labelled instances, so is categorized as unsupervised learning. Nevertheless, after the introduction of an associative classification framework (Li, Shen, and Topor 2002), the use of association rules to build models for classification and prediction tasks is increasing in popularity (Cheng et al. 2007; Cheng et al. 2008; Veloso, Meira, and Zaki 2006; Yin and Han 2003). One example of association rule mining application in a business domain is the well-known market basket analysis problem. The knowledge from market basket analysis may be transformed into, among other possibilities, product promotion strategy development, since the rules provide insights about customer purchase behaviour. For example, an association rule of "motorcycle => helmet" describes a concurrent purchase of these particular goods by a customer. Depending on the frequent patterns found, a firm could design a pricing strategy and discount schemes that would attract combined purchase behaviour.

Association rules mining was introduced by Agrawal, Imielinski and Swami (1993) in a study in which two problems of this task are identified and addressed. The first is the problem of finding large item sets within the database. An item set consists of a number of items in one transaction. The example of "motorcycle => helmet" is a two-item set since two goods were acquired at the same time. Item sets are indicated by frequent occurrence that satisfies a minimum support threshold (minsupport). The support level indicates the percentage of transactions found in the data that contain a particular item set. The second problem arises once frequent item sets have been identified, and the task then becomes generating association rules that satisfy a specified minimum confidence threshold. The confidence level indicates the probability that the presented relationship between the items will actually occur. From the market basket example, for instance buy(Z,"motorcycle") AND buy(Z,"helmet") → [support = 2%, confidence = 70%], 2% support means that this behaviour exists in at least 2% of the records in the database. The 70% confidence means that there is a 70% possibility of a helmet being bought by a customer at the same time a motorcycle is purchased. A strong association rule has to provide both minimum support and confidence levels (Han, Kamber, and Pei 2012).

The Apriori algorithm pioneered by Agrawal and Srikant (1994) is a well-known and commonly used technique for association rule mining (Tufféry 2011). Its performance and functionality were improved through the works of Brin et al. (1997), Park, Chen, and Yu (1995), and Savasere, Omiecinski, and Navathe (1995). Other algorithms developed for the pre-requisite task of frequent item set generation include frequent pattern tree growth (FP-growth) algorithms (Han, Pei, and Yin 2000), hyper-structure mining of frequent patterns (H-Mine) (Pei et al. 2001), and those for mining maximal[8] and closed[9] frequent item sets (Grahne and Zhu 2003).

3.5.2 Clustering

Clustering is a descriptive mining task that is applied to segregate cases or records into groups or clusters based on existing logical relationships in the data, without any predefined labels. It aims to maximize the intra-cluster similarity and minimize the inter-cluster similarity (Han, Kamber, and Pei 2012). Since the groupings are not determined by any target class, clustering is an unsupervised learning approach. The task of partitioning the data depends on the chosen algorithms, and these choices affect the results (Han, Kamber, and Pei 2012). This presents a cluster validity issue, which is determining how robust the cluster outputs are (Jain, Murty, and Flynn 1999). Addressing this problem, Han, Kamber and Pei (2012) outline prerequisites for clustering tasks, covering scalability, algorithms for complex data types, noise handling, involvement of domain experts, and responsive but insensitive clustering techniques. Some practical examples of clustering in different domains include market segmentation in marketing (Huang 1998; Hsu and Chen 2007; Liu and Ong 2008), image recognition (Coleman and Andrews 1979; Jain and Flynn 1996; Lowe 2001) and web search (Cai et al. 2004; Joshi and Jiang 2002; Zamir and Etzioni 1999). Clustering is a well-established subject and has close interconnections with other disciplines such as statistics and machine learning. One of the first clustering algorithms published was *k*-means by Lloyd in 1957 (Han, Kamber, and Pei 2012). Others to name only a few, include partitioning around medoids (Joshi and Jiang), clustering large applications (CLARA), agglomerative nesting (AGNES), divisive analysis (DIANA) and monothetic analysis (Batista, Prati, and Monard) by Kaufman and Rousseeuw (1990), balanced iterative reducing and clustering (BIRCH) by Zhang, Ramakrishnan and Livny (1996), and clustering using representatives (CURE) by Guha, Rastogi and Shim (2001).

[8] Maximal frequent item set refers to an item set that is not a subset of any other frequent item set (Gouda and Zaki 2001).

[9] Closed frequent item set refers to an item set that has no superset that has the same support count as the original item set (Wang, Han, and Pei 2003).

3.5.3 Classification

A classification task is concerned with the construction of models from the training data and the application of these models to unseen or new test data (Han, Kamber, and Pei 2012). The classification algorithm, applied to the training data, is aimed at building models by capturing the relationship between the predictors and the target classes.[10] Since the models are generated from a labelled data set, the classification task is known as supervised learning. These models can be applied to previously unseen data sets for which the target class is unknown. The purpose is to predict the target class that each instance of unseen data can be classified into (also known as prediction task). Applying the models to unseen data (also referred to as test data) is a way to evaluate their predictive accuracy and generalization power. Although classification and clustering are similar in the way that both segregate the data into particular groups, these two mining tasks are different in the way the target class is designated. In classification, labels are defined for the training data prior to the application of the mining task, with the objective of acquiring characteristics or classifiers of a particular class. On the other hand, in clustering, labels are not defined in advance, but are generated subsequently. A notable example of classification with real-world data is credit risk assessment, where loan applications are categorized as risky or non-risky and lead to rejection or approval of an application. This type of mining task is also useful to detect fraudulent behaviour (Kirkos, Spathis, and Manolopoulos 2007) and for health diagnosis and treatment (Li et al. 2004).

Before any steps of classification are carried out, a training data set must be prepared. The data consist of attributes with their corresponding values as predictors or independent variables, and the class label as the target or dependent variables. When the training data is applied to compile the model, the issues of imbalanced target classes and over-fitting should be taken into consideration since these conditions affect the performance of the model. An imbalanced class occurs when the number of instances for each target class is noticeably disproportional, distressing the performance of the learning algorithms and ultimately increasing misclassification costs (He and Garcia 2009). Over-fitting arises when the number of instances is considerably small relative to the number of attributes. This causes the model to be overtrained. Such cases are commonly found when decision tree and neural network algorithms are used (Tufféry 2011). When the training data is set, the classification task involves the learning and classification step (Han, Kamber, and Pei 2012). In the learning step, a model is constructed by classification algorithms applied to the

[10] Target classes are the dependent variables/attributes in supervised learning. For example, in the credit domain the target classes can be “accepted” and “rejected”, or classified as “performing loan” or “non-performing loans”, etc.

training data. This model, or classifier, captures the attributes' values/constraints associated with a particular target class. To assess the predictive ability of the model, a classification step is carried out whereby the model is applied to previously unseen/test data to determine its accuracy rate: that is, the percentage of correctly classified instances derived from the data. When the model shows good performance based on the accuracy rate, it is used to classify a new dataset for which the target class is not known.

There are several basic methods for conducting the classification task, such as decision tree induction, naïve Bayesian classification, and rule-based classifiers (Han, Kamber, and Pei 2012). Quinlan's iterative dichotomizer (ID3) (Quinlan 1979, 1983) and C4.5 (Quinlan 1993) were among the first decision tree induction algorithms developed. They were followed by others such as CART (Breiman et al. 1984; Crawford 1989), FACT (Loh and Vanichsetakul 1988), BOAT (Gehrke et al. 1999) and SLIQ (Mehta, Agrawal, and Rissanen 1996). Naïve Bayesian is a statistical technique based on Thomas Bayes' theory of probability (Han, Kamber, and Pei 2012), and thereafter explored and extended by, inter alia, Weiss and Kulikowski (1990), Mitchell (1997), and Duda, Hart, and Stork (2001).

Rule induction algorithms are used to extract rules in a sequential manner, and include AQ15 (Hong, Mozetic, and Michalski 1986), CN2 (Clark and Niblett 1989; Clark and Boswell 1991) and RIPPER (Cohen 1995). More advanced methods for the classification task include the Bayesian belief network (Heckerman 1996), neural network (Avner 1995; Gallant 1993; Rumelhart, Hintont, and McClelland 1986), support vector machine (Boser, Guyon, and Vapnik 1992; Vapnik and Chervonenkis 1971), frequent patterns (Dong and Li 1999; Liu, Hsu, and Yiming 1998), lazy learners (Dasarathy 1991), genetic algorithms (Skowron and Rauszer 1992), rough set approach (Pawlak 1995) and fuzzy set approach (Zadeh 1965; Zadeh 1983).

3.5.4 Outlier detection and analysis

An outlier is "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1980, in Knorr and Ng 1998, p. 393). The definition suggests data values that are of irregular nature, or exceptions from the rest of the data. Such irregularities found in a real-world database, and in many cases can act as a red flag to fraudulent or irregular transactions; outlier detection is therefore important. Although closely related to clustering, outlier or anomaly detection emphasizes the finding of out-of-the-ordinary cases in a data set rather than of clusters in the data (Han, Kamber, and Pei 2012). Because of this, outlier detection methods are categorized as statistical, proximity-based and cluster-based methods. In statistical methods, outliers are considered a by-product of normal data, with the application of detection methods such as maximum likelihood, Mahalanobis distance and Grubbs' Test (Han, Kamber, and Pei 2012).

The proximity-based methods use disparity in distance as an indicator of outlier existence; several algorithms have been pioneered and further developed by Knorr and Ng (1998). Lastly, with the clustering-based method, outliers are detected as small clusters or non-clustered objects with methods such as bootstrap (Barbará et al. 2003) and FindCBOLF algorithms (He, Xu, and Deng 2003).

3.6 Types of knowledge representation

The ultimate purpose of KDD is to provide enterprises with useful and interesting knowledge derived from their substantial databases. With the use of data mining techniques or any other data analysis tools, inferences about a firm's operations can be efficiently derived, and can enhance strategic planning and decision-making. Due to the automated process, products of mined data are focused on providing users who are not experts in data mining or other information technology to take advantage of this new knowledge; therefore, it is necessary to transform the outputs into readily-interpreted knowledge from which direct action may be taken. The two common types of knowledge representation are production rules and decision trees.

3.6.1 Productive rules

Various techniques, in particular those with classification and association rule mining tasks such as rule induction, decision table, or Apriori, extract and present patterns in the form of a rule. Depending on the task it is produced from, a rule expresses deductive reasoning or relationships of attributes in words. Where a rule presents knowledge about logical arguments, it is formed as condition-action (de Maindreville and Simon 1988) with antecedent-consequence connected by "AND" for a conjunctive rule or "OR" for a disjunctive rule. The condition or antecedent consists of the value of the attributes that precedes a particular outcome, while the action/consequence indicated by the value of the target class, and is typically shown by the "IF-THEN" rule (Han, Kamber, and Pei 2012).

3.6.2 Decision tree

A decision tree presents the rules in a graphic format, easy to comprehend by non-expert users. A tree consists of nodes (or vertices) and connecting edges. Figure 3.2 displays a decision tree showing simplified hypothetical loan-granting rules. The root node ("duration with bank") is the starting point from which the case is built, followed by interior nodes ("info from credit bureau", "avg monthly deposit balance", "loan historical performance", "collateral type") reflecting the relevant attributes; the leaf node ("accept", "reject") at the end of each branch consists of the target class. The outgoing edges of a node display the possible outcome of that particular node. Interior nodes have both outgoing and incoming edges; while a root node and a leaf node only have outgoing and incoming edges. A rule path

consists of a series of interior nodes and their connecting edges, starts from the root node and ends at a particular leaf node. For example, one rule path is “duration with bank (new)” AND “info from credit bureau (negative)” → “reject”. The “IF-THEN” rule format of the path is formed where the traverse along the interior nodes forms the conditions of “IF” and the leaf node is the conclusion following the “THEN”.

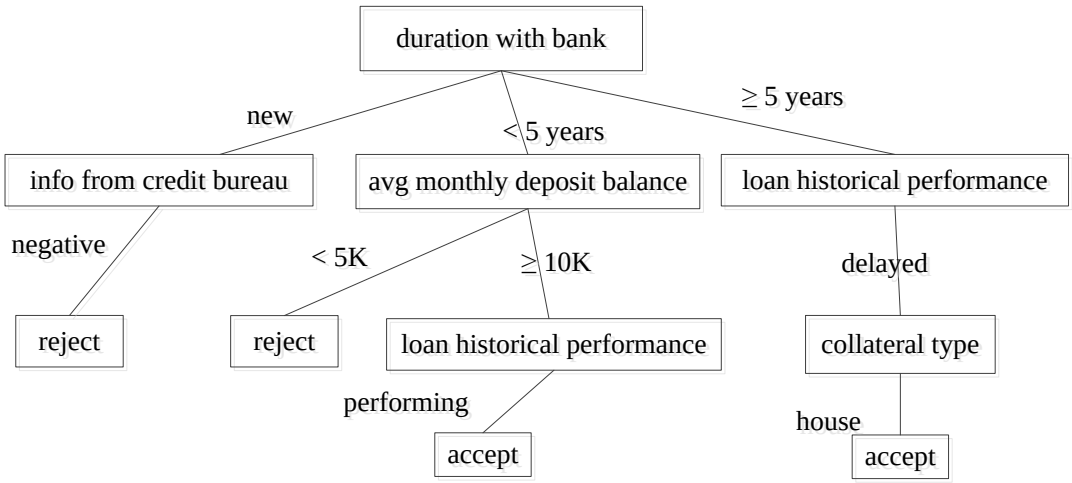


Figure 3.2 Illustration of credit-granting rules

Each path represents one rule, and a decision tree can become complex and troublesome when the mining task generates excessive rules.

3.7 Post-processing of discovered knowledge

With the automation of pattern extraction, data mining is deemed to provide a time-effective approach to retrieve hidden patterns from a database. However, it can be ineffective if users have to make sense of an excessive number of patterns. This emphasizes the issue of pattern usefulness or interestingness in data mining, which extends beyond minimum support and confidence discussed in Section 3.5.1. Criteria and measures of interestingness have been proposed in various studies (Cao et al. 2010; Geng and Hamilton 2006; Han, Kamber, and Pei 2012; Lavrač, Flach, and Zupan 1999). Geng and Hamilton (2006) compiled nine criteria of interestingness: conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability. In Table 3.1 the definition of each criterion is provided with reference to the studies from which they were retrieved. The last feature, actionability, is part of an emerging field in data mining, which highlights the increasingly critical roles of human or domain experts in the knowledge discovery process (Cao et al. 2010).

Table 3.1 Criteria for interestingness

Criterion	Description	References
Conciseness	A pattern contains few attribute-value pairs. A pattern set contains few patterns.	Bastide et al. (2000), Padmanabhan and Tuzhilin (2000)
Coverage	A pattern covers a relatively large subset of a dataset.	Agrawal and Srikant (1994), Webb and Brain (2002)
Reliability	The relationship described by a pattern occurs in a high percentage of applicable cases.	Ohsaki et al. (2004), Tan et al. (2002)
Peculiarity	A pattern is far away from the other discovered patterns according to some distance measure.	Barnett and Lewis (1984), Knorr et al. (2000), Zhong et al. (2003)
Diversity	The elements of the pattern differ significantly from each other or the patterns in a set differ significantly from each other.	Hilderman and Hamilton (2001)
Novelty	A person did not know the pattern before and is not able to infer it from other known patterns.	Sahar (1999)
Surprisingness	A pattern contradicts a person's existing knowledge or expectations.	Bay and Pazzani (1999), Carvalho and Freitas (2000), Liu et al. (1997; 1999), Silberschatz and Tuzhilin (1995; 1996)
Utility	A pattern is of use to a person in reaching a goal.	Chan et al. (2003), Lu et al. (2001), Shen et al. (2002), Yao et al. (2004)
Actionability	A pattern enables decision making about future actions in the domain.	Ling et al. (2002), Wang et al. (2002)

Source: Geng and Hamilton, Choosing the right lens: Finding what is interesting in data mining, *Studies in Computational Intelligence* (SCI)), p.4.

Patterns extracted from a database represent models applicable to the partitioned data. These models are converted to rules, with conditions and outcomes clearly presented. In this section, the last step of the KDD process relating to measurement of the interestingness rule is discussed, with a note that a number of techniques for quality measurement can be used simultaneously to assess the discovered knowledge models. According to Freitas (1999), there are three frequently used quality measures: coverage, completeness and confidence. Coverage level, stated as a percentage, shows the scope of applicability of the rules within the database. For example, a set of rules that has 90% coverage level means that 90% of the database is covered by these rules. Completeness, also stated as a percentage, shows the scope of the extracted patterns in the corresponding target class within the database. There is an expectation from users to obtain all rules that exist in the data. However, completeness is difficult to achieve, especially when support and confidence levels are set by the user. In cases where completeness is achieved, users are still left with an overabundance of rules, including rules that only partially cover the data. Confidence level, as described in Section 3.5.1, is a conditional probability of the occurrence

of the presented relationship between the items. Excessive rules can be reduced by removing the contradictive (Padmanabhan and Tuzhilin 1998; Zhang and Zhang 2001) and redundant (Bayardo, Agrawal, and Gunopulos 2000; Yan et al. 2008) rules. However, this decreases completeness and creates oversimplification of the model due to the lower coverage rate of the test data. Limited generalization is a trade-off between accuracy and coverage rate.

3.7.1 Technical interestingness measures

Technical or objective interestingness measures are aimed at assessing the structure of the generated model (Han, Kamber, and Pei 2012). There are so many technical measures available from probability, statistical and information theories that the issue of selecting the right measures becomes significant. Several studies conducted in this area include Lallich, Teytaud, and Prudhomme (2007), Lenca et al. (2007) and Tan, Kumar, and Srivastava (2002), each using a list of different – though not mutually exclusive – measures for association rules. The criteria used for selection in these studies are diverse, but there is tacit conformity among the studies about the role of domain experts. A comprehensive study by Geng and Hamilton (2007) groups 38 measures according to the data mining tasks of association, classification and summary. Aside from support and confidence described in Section 3.5.1, and coverage and accuracy in Section 3.7.1, the list of measures for association and classification rules include, among others, Loevinger (Hilderman and Hamilton 2000), Lift (Brin, Motwani, and Silverstein 1997), Laplace (Good 2003), Gini Index, Goodman and Kruskal's Predictive Association, Piatetsky-Shapiro (Piatetsky-Shapiro 1991b), Information Gain (Church and Hanks 1990), Sebag and Schoenauer (1988) and Zhang (2000). The criteria used for selection are peculiarity, surprisingness and conciseness; the subjective features used are unexpectedness, novelty and actionability. The interestingness measures for the summaries task is denoted by class distribution and number of classes (Hilderman and Hamilton 2001); in other words, by diversity measurements. These include variance (Rosenkrantz 1997), Simpson Index (Simpson 1949), Shannon Index (Shannon and Weaver 1949), Berger Index (Berger and Parker 1970), Schutz Index (Schutz 1951), Theil Index (Theil 1967) and Atkinson Index (Atkinson 1970).

3.7.2 Domain-specific interestingness measures

Domain-specific interestingness measures are part of Domain-driven data mining (D^3M). D^3M refers to a logical framework that allows incorporation of domain expertise and contexts into the discovered knowledge to attain actionable discovered knowledge (Cao et al. 2010). It is an emerging field with domain-specific or subjective interestingness as one of its focal points. The rationale for the subjective interestingness measure is the actionability of rules for the users, as noted in Section 3.7. The current practice of KDD is a proven effective way to discover hidden knowledge, but the process is not yet developed to allow full

integration of the domain working environment and users find it difficult to utilize the discovered knowledge for decision-making purposes. The domain working environment comprises constraints of domain, data, interestingness and deliverables (Cao et al. 2010). Domain constraints include the elements related to the functioning of the business such as terminology used, business processes, policies enforced and nature of the problems. Data constraints pertain to the dynamics and challenges of real-world data such as large volume, bad structure, type (e.g. multimedia, high frequency, density and distribution), high dimensionality, and data privacy. Interestingness constraints, also termed business interestingness, have regard to the socio-economic aspects that determine the usefulness of the extracted knowledge. Deliverables constraints refer to the sustainability of the discovered knowledge, which requires infrastructure establishment and support to ensure that the knowledge discovery process is an integral part of the company's information systems. In order to attain user-oriented knowledge, interaction with domain experts throughout the knowledge discovery steps is imperative. Cao (2010) suggests some techniques to integrate domain knowledge into the discovery process, such as dynamic user modelling, different types of focus group discussions, human-centred data mining, and online user interactions.

Given the significant portion of human involvement, Cao (2010) proposes four interestingness measures: objective and subjective technical interestingness, and objective and technical business interestingness. If the discovered knowledge satisfies all four measures, then it achieves knowledge actionability interestingness. Objective technical interestingness is measured by its minimum support and confidence, while subjective technical interestingness is measured by a probability-based belief, as suggested by Padmanabhan and Tuzhilin (1998). Objective and technical business interestingness measures are formulated with specific regard to the domain application (Luo et al. 2008). D^3M methodology, to different extents, is applied to several fields such as finance (Cao and Zhang 2006; Ou et al. 2008), web mining (Chau et al. 2003), marketing (Piton et al. 2009) and customer relationship management (Yang et al. 2003).

3.8 Summary

In this chapter the general concept of KDD and data mining is discussed. Each step in the knowledge discovery process is explained, showing the role of data mining as a tool to extract hidden, novel and interesting patterns from a large database. The mining task to be applied to data depends on the type of data and the purpose of its application. Each of these determinants, and the development of data mining techniques as well as their applications in various domains, is explained. In order to enhance the usefulness of the extracted patterns for users, the patterns need to be presented in forms that are easily understood. Finally, interestingness measures used to evaluate the mined patterns are discussed.

CHAPTER 4 – RESEARCH METHODOLOGY APPLIED

4.1 Introduction

In the previous chapter data mining was discussed in the context of the knowledge discovery process applicable to databases. The application of data mining techniques enables the extraction of patterns representing causative relationships among attributes in large repositories. Such patterns should be comprehensible to the users, making decision trees and productive rules the most common types of knowledge representation. In this regard, common data mining techniques used for predictive and descriptive purposes were discussed. The application of data mining techniques in various domains was provided.

In this chapter, the methodology used to develop a set of credit risk assessment tools is discussed. An overview of the methodology is presented in Section 4.2. Section 4.3 explains the structure and content of the credit data sets[11] that were obtained from the three sample banks. In Section 4.4, the steps taken to transform the credit data sets into the format appropriate for data analysis are explained. Section 4.5 focuses on the data mining techniques applied on the credit data sets. The chapter concludes with a summary in Section 4.6.

4.2 An overview of the research methodology

The methodology to achieve the objectives set out in Chapter 1 comprises four steps, illustrated in Figure 4.1.

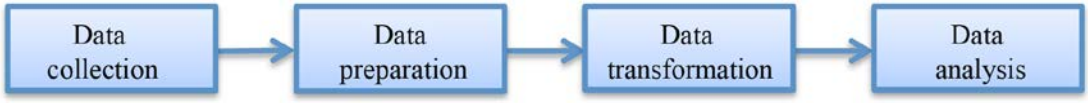


Figure 4.1 Research methodology steps

The credit data sets used for this research were obtained from three Indonesian banks. The customer types of the banks and the way that they conduct credit assessments were described in Chapter 2. The non-formatted credit data sets are prepared for processing by changing them into a more structured format. During this step, the credit data sets are

[11] In this thesis, the term "credit data set" refers to the collated biographic, financial and other data of a single borrower that is utilized for credit risk assessment by a bank. The phrase "credit data sets of a bank" refers to the total number of individual collated biographic, financial and other data of single borrowers of a bank. The term "credit data sets" has the same meaning as "sample" in the finance domain and "instances or records" in the data mining domain.

allocated tree-structures using XML and then transformed into a flat data format. This format allows the application of a wide range of data mining techniques while preserving the structural information in the extracted patterns. The aforementioned steps are discussed in detail in the ensuing sections, and a recapitulation of the steps is provided in the summary section of this chapter.

4.3 MSME credit data sets

The three sample banks were selected as providers of credit data sets based on their target market similarity, namely that the financing activities of all three are primarily targeted at micro- and small-scale businesses. The credit data sets obtained from the banks also provide a good geographical spread, as Bank A operates only in rural areas, Bank B operates in both metropolitan and rural areas, and Bank C operates only in metropolitan areas. The operational credit assessment methodologies of the three banks (discussed in Chapter 2) allow comparison between the data mining results of this research, with manual discretionary assessment and the presently utilized credit scoring assessment of the banks. In this regard, Bank A applies the discretionary lending method and Bank C applies the scoring method, while Bank B applies both.

4.3.1 Credit data sets collection method

Based on Banking Law Number 10, banks are obliged to maintain confidentiality of borrowers' credit data. Therefore, a stringent confidentiality protocol was followed in the collection of the data sets used in this study. This included signing confidentiality agreements with each bank. The credit data of loans granted within a period of four years were collected from banks over the period April—November 2011, and collated manually for the purpose of this research. Interviews were conducted with loan staff of the banks to determine how, when and where they obtained credit data and to verify the methodologies employed to utilize the data in their credit assessment processes. Based on the aforementioned confidentiality protocols and business competitiveness, no statistical descriptive information about the actual performing and non-performing loans are provided.

4.3.2 Credit data sets content and structure

The credit data sets collected from the sample banks differ in terms of the number of sets, the business sectors of the borrowers, geographical areas, and the data format and content. Each bank was asked to provide 300 data sets consisting of equal numbers of performing and non-performing loans. None was able to do this due to the different operational collation processes followed by each banks, and because the researcher depended on the goodwill of the banks to provide data and had no direct control over what they offered. This resulted in a smaller numbers of credit data sets available for this research

(Table 4.1). Moreover, the banks could not provide equivalent numbers of performing and non-performing loan credit data sets, resulting in extremely unbalanced cases, especially in the cases of Banks B and C. This will be addressed in the detailed description of the credit data sets of each sample bank in Sections 4.3.2.1 to 4.3.2.3.

Table 4.1 Summary of the credit data sets of the sample banks

Description	**Bank A**	**Bank B**	**Bank C**
Number of credit data sets	Relatively balanced	Unbalanced	Significantly unbalanced
Diversity of borrowers	Relatively homogeneous	Relatively segmented	Segmented
Service area	Rural	Rural and metropolitan	Metropolitan
Credit data sets Financial data	Customer financial information collected by bank.	Customer financial information collected by bank and derived financial ratios.	Financial ratios constructed by bank.
Qualitative data	Record of information obtained during interview with customer.	Record of information obtained during interview with customer; interpretation of financial information; additional collateral information; and SWOT analysis conducted by bank.	None.
Core loan data	Loan amount, collateral negotiated, loan term and business sector.	Loan amount, collateral negotiated, loan term and business sector.	Loan amount, collateral negotiated, loan term and business sector.
Scores	None.	27 scores applied for credit risk assessment by bank.	117 scores applied for credit risk assessment by bank.

4.3.2.1 Bank A

Bank A operates in a rural area close to Jakarta, renowned for its density of micro businesses. The majority of businesses in the credit data sets supplied by Bank A, 77.09% of the total borrowers, are in retail trading, as indicated in Table 4.2. It is normal for micro business owners to operate more than one business, and this is reflected in the credit data sets: 36.46% of the borrowers have multiple businesses. The additional ventures can be very similar (e.g. trading vegetables and refilled domestic gas bottles) or involve different activities (e.g. trading vegetables and operating a car rental business).

Table 4.2 Borrower profile based on business sectors - Bank A

Sector	Number of borrowers	Percentage (%)
Trade	74	77.09
Manufacturing	6	6.25
Agribusiness	2	2.08
Services	14	14.58
Total	96	100.00

Bank A can be regarded as a small bank since it had only 300 customers to whom finance was provided at the time of data collection; 96 credit data sets were obtained from this bank. The performing loans represent 60.42% (58 credit data sets), and non-performing loans 39.58% (38 credit data sets). Although not equivalent in numbers, this is regarded as relatively balanced. A considerable portion of the loans (70.83%) are concentrated in the low to middle range, between IDR 1 million (USD 10.3) and IDR 50 million (USD 515), based on the standard loan size categorization applied by Bank A (Table 4.3).

Table 4.3 Borrower profile based on sizes of loans – Bank A

Size of Loans (x)	Number of loans	Percentage (%)
IDR 1 million < x < IDR 6 million	23	23.96
IDR 6 million < x < IDR 25 million	31	32.29
IDR 25 million < x < IDR 50 million	14	14.58
IDR 50 million < x < IDR 100 million	13	13.54
IDR 100 million < x < IDR 250 million	3	3.13
IDR 250 million < x < IDR 500 million	12	12.50
Total	96	100.00

Bank A's credit data sets consist of financial, qualitative and core loan data. The financial data include numerical figures like income and expenses, and bank loan instalments to be paid. The qualitative data are descriptive and unstructured. They entail aspects such as borrower characteristics, borrower business practices, and aspects related to the management of the business. The core loan data consist of a mix of descriptive unstructured data (e.g. business sector and type of collateral) and structured/quantified data (e.g. size of loan and value of collateral).

4.3.2.2 Bank B

Bank B is the largest of the three sample banks, with more than 9,000 branches across Indonesia. For this research, credit data sets were provided by four branches. Branch 1 is

located in a rural area in Jakarta surrounded by traditional markets and micro businesses. Branches 2, 3 and 4 operate in the metropolitan Jakarta area, serving customers from diverse sectors.

The credit data sets provided by Bank B represent customers from a more diverse business environment than that of Bank A. Table 4.4 contains the business sector classification of the credit data sets. Hajj Pilgrim (Religious traveling service to the Holy Land of Mecca for Muslims) is separately classified, because this type of service is heavily controlled and closely monitored by the government and therefore different from the general service sector. The credit data sets from the trade sector represent the majority (50.91%), followed by the service sector; these sectors are also the largest for Bank A (77.09% trade sector and 14.58% service sector). There are 16.36% (27 credit data sets) where the borrowers operate multiple businesses; in all of these cases, the additional ventures are in different sectors from the main business.

Table 4.4 Borrower profile based on business sectors – Bank B

Sector	**Number of borrowers**					**Percentage (%)**
	Branch 1	**Branch 2**	**Branch 3**	**Branch 4**	**Total**	
Trade	26	29	15	14	84	50.91
Construction	10	4	5	2	21	12.73
Agribusiness	0	1	1	0	2	1.21
Services	16	12	5	6	39	23.64
Hajj Pilgrim	7	2	0	0	9	5.45
Production	4	3	0	1	8	4.85
Not available	1	1	0	0	2	1.21
Total	64	52	26	23	165	100.00

Bank B provided a total of 165 credit data sets with a very unbalanced proportion of performing loans (149 credit data sets) and non-performing loans (16 credit data sets). All the loans are short-term revolving credit. Table 4.5 shows the proportion of performing and non-performing loan credit data sets received from each branch. The very limited number of non-performing credit data sets are all loans that were restructured due to repayment difficulties on the part of the borrowers (Loan staff, Bank B, pers. comm).

Table 4.5 Proportion of performing and non–performing loans – Bank B

Credit Data Sets	Number received					Percentage (%)
	Branch 1	Branch 2	Branch 3	Branch 4	Total	
Performing Loans	58	50	23	18	149	90.30
Non-Performing Loans	6	2	3	5	16	9.70
Total	64	52	26	23	165	100.00

From Table 4.6 it is apparent that the loan sizes of the credit data sets differ, and loans between IDR 500 million and 2 trillion form the majority (58.79%) of the credit data sets.

Table 4.6 Borrower profile based on size of loans – Bank B

Size of Loans (x)	Number of loans					Percentage (%)
	Branch 1	Branch 2	Branch 3	Branch 4	Total	
$x \leq$ IDR 500 million	20	12	7	5	44	26.67
IDR 500 million $< x \leq$ IDR 2 trillion	41	30	15	11	97	58.78
IDR 2 trillion $< x \leq$ IDR 4 trillion	3	9	4	7	23	13.94
$x >$ IDR 4 trillion	0	1	0	0	1	0.61
Total	64	52	26	23	165	100.00

As with Bank A, Bank B collects financial, qualitative and core loan data in accordance with the 5Cs lending principles. However, due to differences in lending methods, Bank B collects more credit data for credit risk assessment. The financial data consist of financial ratios derived from the financial statements of borrowers (structured type of credit data). The qualitative data are descriptive and unstructured. They entail such aspects as borrower characteristics, borrower business practices, interpretations of financial information, additional information about collateral, and SWOT analysis for each credit data set. The core loan data consist of a mix of descriptive unstructured data (e.g. business sector and type of collateral) and structured/quantified data (e.g. size of loan and value of collateral). Bank B applies a scoring system whereby particular financial and qualitative data are numerically scored to reflect the acceptability of loan applicants. In the case of qualitative data, the scores are applied to different qualitative credit attributes retrieved from the unstructured descriptive qualitative data; the final scores are in structured numerical form to represent the financial and qualitative information.

4.3.2.3 Bank C

Bank C provides financing to MSMEs throughout Indonesia. A total of 140 credit data sets were obtained from loan staff working in branches in the metropolitan area of Jakarta. The bank applies a more detailed sector classification than the other two sample banks, as it specifies sub-sectors, allocating different credit risk ratings to the various sub-sectors (Table 4.7).

Table 4.7 Borrower profile based on business sectors – Bank C

Sector and Sub-Sector	Number of borrowers		Percentage (%)
	Sector	Sub-Sector	
Home Industry: Handicraft	18	 18	12.86
Services: Construction Transportation Car Service Catering Health Service Printing Heavy Equipment Communication Head Hunter	44	 27 6 4 2 1 1 1 1 1	31.43
Finance: Pawnshop	1	 1	0.71
Trade: Convenient Store Food and Beverages Automobile Distribution Goods and Services Garment (Production) Handicraft Other Retail Heavy Equipment	61	 21 16 10 5 2 2 2 2 1	43.57
Agribusiness: Poultry Farming	15	 15	10.72
Manufacturing: Pharmacy	1	 1	0.71
Total	140	140	100.00

As with Bank A and B, the majority of the credit data sets for Bank C are from the trade (43.58%) and service (31.44%) sectors. Considering the further sub-sector classification of Bank C, the credit data sets from the construction sub-sector (forming part of the services sector) and convenience store sub-sector (forming part of the trade sector) are the most prominent.

All credit data sets of Bank C represent loans below IDR 500 million (USD 5,150). The proportions of performing and non-performing loans are extremely unbalanced, consisting of 95% performing loans (133 credit data sets) and 5% non-performing loans (7 credit data sets).

Most loans (55%) are between IDR 300 million and IDR 500 million. The classification of the loans into different group sizes (Table 4.8) was done by the researcher in a discretionary manner since information about standard loans size categorization in Bank C was not available.

Table 4.8 Borrower profile based on sizes of loans – Bank C

Loan Principal (x)	Number of loans	Percentage (%)
IDR 50 < x ≤ IDR 100 million	8	5.71
IDR 100 million < x ≤ IDR 200 million	24	17.14
IDR 200 million < x ≤ IDR 300 million	31	22.15
IDR 300 million < x ≤ IDR 400 million	21	15.00
IDR 400 million < x ≤ IDR 500 million	56	40.00
Total	140	100.00

Bank C provided credit data sets consisting of financial data in the form of financial ratios derived from the financial statements of borrowers, combined with credit scores allocated to each of these ratios from their credit assessment. The scores allocated to financial ratios serve as an indication of a borrower's financial performance and ability to comply with the repayment of a loan. These figures are structured credit data. The core loan data provided by Bank C are a mix of descriptive structured text (e.g. business sector, type of collateral) and quantitative structured (e.g. size of loan, value of collateral) types of credit data. Credit scores are also allocated to this data by Bank C. No descriptive qualitative (unstructured) data was provided. The Bank applies the combined financial ratio scoring and the core loan data scoring to determine the credit risk acceptability of loan applicants.

4.3.3 Preparation of raw credit data sets

In order to ensure compliance with the confidentiality protocol, personal data like borrower identity and other information like business names and addresses that may result in

the identification of such persons or entities, have been removed from the credit data sets. Total anonymity exists. The research includes the use of descriptive qualitative information that has been provided in Bahasa (an Indonesian language). In order to ensure the correct interpretation thereof and to preserve the data in the original format, aligned with the context utilized by the bank, no translation was conducted.

The data sets of the banks are not all the same. As such, the basic preparation of the data sets differed to a certain extent, as indicated below.

4.3.3.1 Bank A

The credit data sets of Bank A were in the form of audio recordings. The recorded financial, qualitative and core loan data of all credit data sets were transcribed using a standard uniform format and structure and verified by the researcher. The data were then integrated into a single data source.

4.3.3.2 Bank B

Financial data were made available in hard copy, but qualitative and core loan data were in audio recorded format. Credit scores allocated to the data in the original credit assessment process of the bank were copied into a template for each credit data set. Due to inconsistencies, the financial statements for all credit data sets were reproduced to ensure identical financial ratios were available in each credit data set. As with Bank A, the qualitative and core loan data were transcribed and verified. Subsequently, the complete credit data sets were integrated into a single data source.

4.3.3.3 Bank C

The credit data sets of Bank C were obtained in soft copy (Word document) and subsequently integrated into a single data source.

4.4 Transformation of credit data sets for data analysis

The transformation of the flat format credit data sets for analysis required discretization [12] of numerical attribute values and categorization of text-format attribute values. The discretized and categorized data requires less processing time and memory usage as well as providing more comprehensible mining results (Han, Kamber, and Pei 2012). The number of distinct continuous values was limited to group categories, and the unstructured descriptive text-based data were organized according to concept significance. The details are provided in the ensuing sections.

[12] Discretization is a data transformation strategy applicable to numerical attributes. It transforms numerical values into categorical values by replacing the raw figures with interval or conceptual labels (Han, Kamber, and Pei 2012).

4.4.1 Discretization of the structured types of data

The numerical attributes of the credit data sets were manually and automatically discretized. Manual discretization was performed by selecting the ranges and cut-points of categories for the attributes on a discretionary basis by applying domain knowledge. In order to attain comprehensive results, automated discretization was carried out by applying supervised [13] and unsupervised [14] discretization techniques using data mining software RapidMiner (Mierswa et al. 2006). The supervised discretization technique applied credit performance as the target class. The entropy-based technique is one of the supervised discretization techniques, was selected for its accuracy of performance in credit data (Dougherty, Kohavi, and Sahami 1995; Kohavi and Sahami 1996). It was based on information[15] content in relation to the target classes (performing or non-performing loans). The cut-points for each range were determined by applying a recursive process of selecting values with the least information requirements. However, the imbalanced target class proportion within the credit data sets of the banks also warranted the use of an unsupervised discretization technique. Equal-width binning is such a technique, and was selected because of its promising results and its simplicity and accuracy when applied in a credit domain (Dougherty, Kohavi, and Sahami 1995; Ikasari and Hadzic 2012). The automatic equal-width binning discretized the attribute values into a predefined number of bins (or ranges). This resulted in a uniform number of bins for all attributes. All credit risk assessment scores allocated to credit attributes by the banks were exempted from the discretization process since the score values already represent defined score categories that differ from distinct continuous values like financial figures and financial ratios.

4.4.1.1 Bank A

As stated in the previous section, financial data collected by Bank A consists of the income and expenses of the customers as well as bank loan information such as calculated loan instalments, size of loans and value of collateral. Manual discretization was performed by the researcher applying the standard categorization employed by Bank A as well as a discretionary approach: the size and duration of loans were discretized based on the bank's internal categorization, and other numerical attributes were discretized by dividing the spread of minimum to maximum attribute values into discretionary numbers of categories. The cut-points or the boundaries of the different categories for each attribute were

[13] Discretization is supervised when it is implemented with use of target class information (Han, Kamber, and Pei 2012).

[14] Discretization is unsupervised when it is implemented without any use of target class information (Han, Kamber, and Pei 2012).

[15] The term "information" reflects the data relevance considering target classes within a statistical context.

determined with cognizance of multiple attribute interdependency. For example, the categories containing values of vehicle maintenance expenses have direct relationships with categories that contain values of vehicle registration expenses.

4.4.1.2 Bank B

The numerical attributes in Bank B's credit data sets consist of financial ratios and loan sizes. Manual discretization was performed by applying domain knowledge. Categories allocated to loan sizes were aligned with the level of approval required for credit assessment approval of different loan sizes by Bank B. The raw values of financial ratios were discretized by dividing the spread of minimum to maximum attribute values into discretionary numbers of categories reflecting comparative different levels of performance. The category cut-points for each of the financial ratios were designed to capture the underlying financial risk. As with Bank A, entropy-based and equal-width binning discretization techniques were applied to all other numerical attributes (non-financial ratios). In addition to financial ratios, Bank B assigned numerical scores to particular financial data. These scores represent a standard categorization employed by Bank B for financial figures; therefore these scores were not discretized.

4.4.1.3 Bank C

Bank C provided credit risk assessment scores for all financial data. As mentioned in Section 4.4.1, such scores are not discretized. The only additional numerical attribute provided by this bank represented loan sizes. The categorization of loan sizes was done in a discretionary way.

4.4.2 Structuring the unstructured types of data

The categorization of the descriptive unstructured text data was performed manually, notwithstanding the development in the Natural Language Processing (NLP) field of study. The primary reason for this lies in the original data being produced in Bahasa, for which no parser or corpus is readily available. All unstructured descriptive text in the credit data sets was manually categorized by applying the process shown in Figure 4.2.

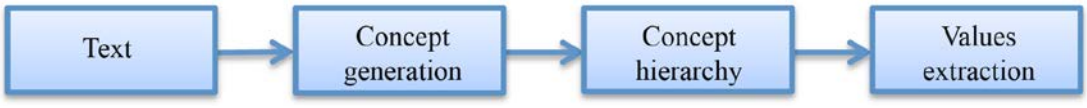


Figure 4.2 Text structuring steps

Text structuring begins by generating concepts from text. This text can be a paragraph or a section. Next, each of the generated concepts is registered into a hierarchy, presenting contextual knowledge of the relevant attribute. These iterative steps are carried out until all

concepts are contained in the hierarchy. Text structuring ends with enumeration of the attribute values.

The application of these steps to Bank A and B credit data sets is discussed in the following sections. The credit data sets of Bank C contain no text-formatted data except core loan data (e.g. name of bank, sector) and assigns credit assessment scores with a brief textual annotation (e.g. good turnover, good reputation). Since the text can be regarded as pre-defined following the respective numerical score values, it can be regarded as structured; categorization is therefore not required.

4.4.2.1 Bank A

The concept generation and hierarchy derived from qualitative data in the credit data sets are based on the implicit structure found in the raw text. Although loan staff do not use a prescribed list of questions, Bank A has a credit report template; this provides an inherent structure, according to which the extracted information was sorted.

4.4.2.2 Bank B

Bank B collected extensive qualitative data comprising the 5Cs lending principles and core loan data, and did a business risk analysis (SWOT) for each credit data set. Although the credit report has a systematic layout with headings and sub-headings for the related sections, the relative content is scattered among various sections. This is due to completion of the forms is undertaken by different loan staff and branches, requiring a rigorous iterative process of concept generation and hierarchy in completing the text structuring.

4.4.3 Structuring the format of credit data sets

Credit risk assessment is a systematic process to evaluate the credit worthiness of prospective borrowers. This entails a structure relating to important aspects (5Cs) indicated in the credit reports. To preserve this structure and present structured and unstructured data in tandem, the credit data sets were automatically extracted into an XML document. The quantitative attributes such as financial ratios are similar across banks. However, qualitative attributes are different for each bank. This means the general structure of the template could be applied with a top level split of quantitative/qualitative attributes for all banks. Similar quantitative attributes are retained and specific individual bank quantitative/qualitative attributes are added. The XML template for credit data therefore contains common and specialised content which makes the template to a certain extent generic.

The structural aspects of XML are explained using partial credit data taken from Bank B data sets. The researcher developed the XML template using XMLSpy (Icon Information Systems GmbH) software. Figure 4.3 shows a fragment of the XML template, including selected attributes and their respective values. The actual XML document for Bank B is

extensive in terms of the quantity of attributes. There are 48 attributes under quantitative analysis, altogether comprising three-year time series of financial ratios.

```
<?xmlversion="1.0"encoding="utf-8"?>
<CreditApplication>
    <loanapplication>
        <sector1>hajjpilgrim</sector1>
        <sector2>NA</sector2>
        <debtorstatus>new</debtorstatus>
        <qualitativeanalysis>
            <character>
                <characterqual>reliableandgood</characterqual>
                <accmngmt>NA</accmngmt>
                <businessrep>excellent^goodbusinessrelations^wellknownintheindustry^acti
                veinvolvementinassociation</businessrep>
                <debtorattitude>NA</debtorattitude>
            </character>
        </qualitativeanalysis>
        <quantitativeanalysis>
            <cr-3>[>200]</cr-3>
            <cr-2>[>200]</cr-2>
            <cr-1>[>200]</cr-1>
        </quantitativeanalysis>
        <scoreanalysis>
            <financial>
                <crmorethan140percent>0</crmorethan140percent>
                <quickratiomorethan135percent>0</quickratiomorethan135percent>
                <ttlscorefin>6</ttlscorefin >
            </financial>
            <nonfinancial>
                <charactersc>
                    <leveloftrustsc>0</leveloftrustsc>
                    <accmngmtsc>0</accmngmtsc>
                    <businessrepsc>0</businessrepsc>
                    <debtorattitudesc>0</debtorattitudesc>
                    <ttlchar>0</ttlchar>
                </charactersc>
                <marketsc>
                    <qualprodserv>0</qualprodserv>
                    <stratmrktsc>1</stratmrktsc>
                    <ttlmarketsc>1</ttlmarketsc>
                </marketsc>
                <ttlscorenf >5.5</ttlscorenf>
            </nonfinancial>
            <ttlscore>11.5</ttlscore>
        </scoreanalysis>
        <swotanalysis>
            <strengthdet>management^skilledemployee^systematicbookkeeping</strengthdet>
            <weakness>management^onemanshow^systematicbookkeeping</weakness>
            <opportunity>demandedservice^muslimmajority</opportunity>
            <threat>production^financial_baddebt_conversion^business_pricecomp</threat>
        </swotanalysis>
    </loanapplication>
    <CPerf>P</CPerf>
</CreditApplication>
...
```

Figure 4.3 Fragmented XML document – Bank B

The above attributes are termed XML elements. XML elements in Figure 4.3 encapsulate the hierarchy of the elements within the credit data sets, representing "parent-child" and "ancestor-descendant" relationships. Based on this structure, the XML document is modelled in the form of a rooted, ordered, labelled tree. As described in Chapter 3, a tree consists of nodes (or vertices) and connecting edges. An element (e.g. sector1) or a node can have a value (e.g. "hajjpilgrim"). A node without any value in a tree-structured data format can serve the purpose of providing context to the nodes emanating from its outgoing edges. In this thesis, such a node is referred to a "contextual node". Some examples of contextual nodes taken from Figure 4.3 are "loanapplication", "qualitativeanalysis", and "character". Figure 4.4 shows a tree structured format of a partial segment of the XML document that was presented in Figure 4.3. The parent-child relationship constitutes a top-to-bottom relationship (e.g. "loanapplication-sector1", "loanapplication-sector2", and "character-characterqual"). The "ancestor-descendant" relationship represents the top-to-bottom relationship within the same path (e.g. "qualitativeanalysis" is ancestor of "characterqual"). In this regard, the hierarchy existing in the XML document is secured in the paths of the tree. An element (or attribute) and its value represent one node in the mapping.

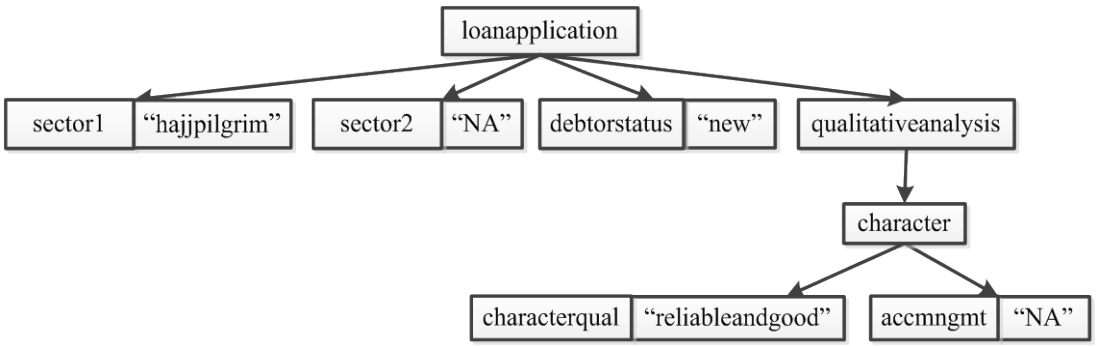


Figure 4.4 Tree-structure of fragmented XML document – Bank B

In order to increase efficiency of processing, elements (e.g. "sector1") and their values (e.g. "hajjpilgrim") in the tree-structured data format are generally converted into a single string (e.g. 'sector1["hajjpilgrim"]') and subsequently into an integer-based format. For the string representation, the data mining approach adopted for this research (as discussed in Hadzic 2012) applies the pre-order (depth-first) string encoding of Zaki (2005a).

In order to describe the conversion of tree-structured data to an integer-mapped format and subsequently to a flat representation of the credit data, two credit data sets with different credit performance outcomes are illustrated in XML document format in Figure 4.5 and integer-mapped format in Figure 4.6.

```
<?xmlversion="1.0"encoding="utf-8"?>
<CreditApplication>
    <loanapplication>
        <sector1>hajjpilgrim</sector1>
        <sector2>NA</sector2>
        <debtorstatus>new</debtorstatus>
        <qualitativeanalysis>
            <character>
                <characterqual>reliableandgood</characterqual>
                <accmngmt>NA</accmngmt>
                <businessrep>excellent^goodbusinessrelations^wellknownintheindustry^acti
                veinvolvementinassociation</businessrep>
                <debtorattitude>NA</debtorattitude>
            </character>
        </qualitativeanalysis>
        <quantitativeanalysis>
            <cr-3>[>200]</cr-3>
            <cr-2>[>200]</cr-2>
            <cr-1>[>200]</cr-1>
        </quantitativeanalysis>
    </loanapplication>
    <CPerf>P</CPerf>
</CreditApplication>
<CreditApplication>
    <loanapplication>
        <sector1>trade</sector1>
        <sector2>NA</sector2>
        <debtorstatus>new</debtorstatus>
        <qualitativeanalysis>
            <character>
                <characterqual>goodwill</characterqual>
                <accmngmt>NA</accmngmt>
                <businessrep>experienced</businessrep>
                <debtorattitude>NA</debtorattitude>
            </character>
        </qualitativeanalysis>
        <quantitativeanalysis>
            <cr-3>[>200]</cr-3>
            <cr-2>[0.01-50]</cr-2>
            <cr-1>[0.01-50]</cr-1>
        </quantitativeanalysis>
    </loanapplication>
    <CPerf>NP</CPerf>
</CreditApplication>
```

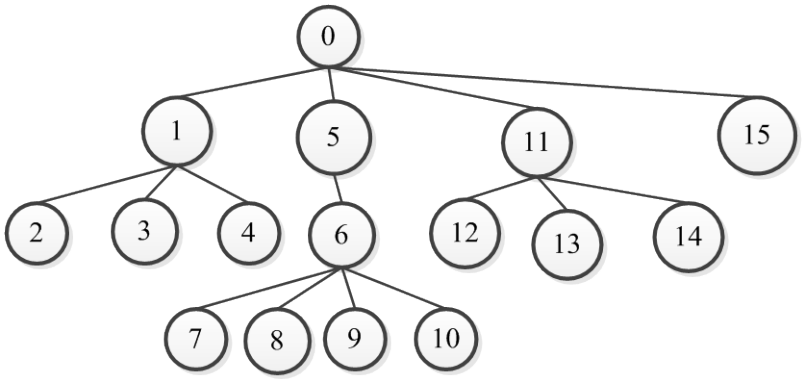
Figure 4.5 Selected credit data of two borrowers in XML document – Bank B

The credit data sets for Borrower 1 and Borrower 2 in the XML document start and end with the "CreditApplication" element. Figure 4.6 shows the integer mapping of these credit data sets done in sequential manner.

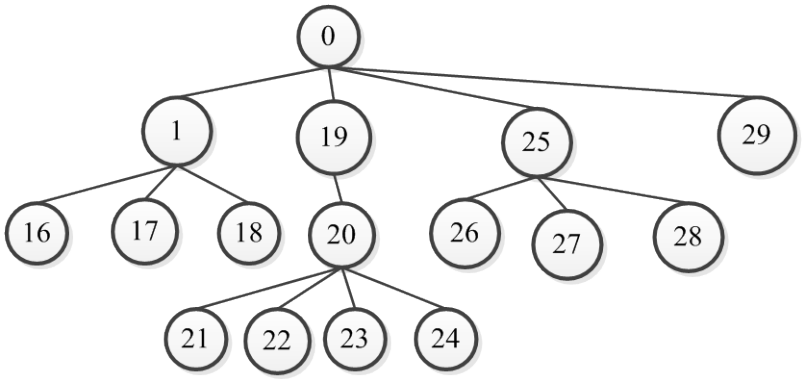
CreditApplication	0
loanapplication	1
sector1[“hajjpilgrim”]	2
sector2[“NA”]	3
debtorstatus[“new”]	4
qualitativeanalysis	5
character	6
characterqual[“reliableandgood”]	7
accmngmt[“NA”]	8
businessrep[“excellent^goodbusinessrelations^wellknown-intheindustry^activeinvolvementinassociation”]	9
debtorattitude[“NA”]	10
quantitativeanalysis	11
cr-3[“>200”]	12
cr-2[“>200”]	13
cr-1[“>200”]	14
CPerf[“P”]	15
sector1[“trade”]	16
sector2[“NA”]	17
debtorstatus[“new”]	18
qualitativeanalysis	19
character	20
characterqual[“goodwil”]	21
accmngmt[“NA”]	22
businessrep[“experienced”]	23
debtorattitude[“NA”]	24
quantitativeanalysis	25
cr-3[“>200”]	26
cr-2[“0.01-50”]	27
cr-1[“0.01-50”]	28
CPerf[“NP”]	29

Figure 4.6 Example of string to integer mapping of Figure 4.5

Given the integer-mapping of elements and their values, the pre-order string encoded representation of the underlying hierarchy of the fragmented credit data set of the first borrower (Figure 4.5) was transformed into “0 1 2 -1 3 -1 4 -1 -1 5 6 7 -1 8 -1 9 -1 10 -1 -1 -1 11 12 -1 13 -1 14 -1 -1 15”. The second borrower (Figure 4.5) credit data set was transformed into “0 1 16 -1 17 -1 18 -1 -1 19 20 21 -1 22 -1 23 -1 24 -1 -1 -1 25 26 -1 27 -1 28 -1 -1 29”. Figure 4.7 shows the tree presentation of the integer-mapped string encoding for Borrowers 1 and 2.



Borrower 1



Borrower 2

Figure 4.7 Tree presentation of integer-map of partial XML document

Techniques that can be applied to mine association rules from credit data sets represented in a tree-structured data format are limited to the application of frequent sub-tree mining algorithms (Chi et al. 2005; Hadzic 2012; Hadzic, Tan, and Dillon 2011; Zaki 2003). In this regard, XML was used to present structured and unstructured types of synthetic credit data in tandem. The occurrence of contextual nodes escalated the combinatorial complexity of the tree-structured credit data sets. The excessive number of extracted patterns obstructed the analysis and interpretation of the results as experienced by Ikasari, Hadzic, and Dillon (2011). It was necessary to transform the tree-structured credit data sets into a flat format to enable direct application of a wider range of data mining techniques. The technique to convert the credit data sets from string encoding into a flat data format (denoted as a table), as detailed by Hadzic (2012), has been adopted in this research.

The tree-structured database containing the string encoding representation of the two credit data sets were as follows:

1	0 1 2 -1 3 -1 4 -1 -1 5 6 7 -1 8 -1 9 -1 10 -1 -1 -1 11 12 -1 13 -1 14 -1 -1 15
2	0 1 16 -1 17 -1 18 -1 -1 19 20 21 -1 22 -1 23 -1 24 -1 -1 -1 25 26 -1 27 -1 28 -1 -1 29

Figure 4.8 Example of segments of the tree-structured database of Bank B

The conversion started with extraction of the structural properties that exist in every credit data set in the database. This structure is referred to as the Database Structure Model (DSM). The algorithm for DSM extraction developed by Hadzic (2012) traverses every credit data set and continuously improves the DSM. The DSM is complete when all credit data sets can be matched against it. Since the credit data sets of Bank A and Bank B were formatted using the XML template, their structural properties were identical. The string encoding of x_i and b_j was then used to represent the DSM. The x_i notation was used for the attribute names (e.g. sector1), where i corresponds to the pre-order position of the node in the tree. The backtracks were noted by using b_j, with j corresponding to the backtrack number in the string encoding. Thus, for the partial extraction of Bank B credit data sets presented in Figure 4.6, the DSM became "$x_0\ x_1\ x_2\ b_0\ x_3\ b_1\ x_4\ b_2\ b_3\ x_5\ x_6\ x_7\ b_4\ x_8\ b_5\ x_9\ b_6\ x_{10}\ b_7\ b_8\ b_9\ x_{11}\ x_{12}\ b_{10}\ x_{13}\ b_{11}\ x_{14}\ b_{12}\ b_{13}\ x_{15}$". The string encoding of the DSM then became the first row of the table. The matching columns were filled by element names (for contextual nodes) or element names and their corresponding values (for non-contextual nodes). The DSM was used to match the attributes with their positions in the tree-structured database. The structural information from the XML template is still preserved since the attributes are indexed/distinguished by their pre-order position in the tree. The backtracks (-1 in string encoding or b_j in the flat credit data sets) were omitted from the flat credit data sets, because in this application the backtrack attributes have no predictive power. Similarly, the contextual nodes were also excluded during the application of the data mining techniques because they have identical values for all entries. The flat credit of Figure 4.8 subsequent to the application of DSM was presented in Figure 4.9.

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	16	17	18	19	20	21	22	23	24	25	26	27	28	29

Figure 4.9 Flat representation of the tree-structured database

The aforementioned step was performed to the credit data sets of Bank A and Bank B. Bank C credit data sets are structured with each attribute explicitly contextualized (Section

4.4.2), therefore the use of XML was not applied in this case. The XML data of Bank A credit data sets consists of 701 attributes and attribute values; 78 nodes in each credit data set tree; a maximum tree height of four; and a maximum fan-out of 12 for each credit data set tree. After the application of DSM to the XML data, the flat credit data sets of Bank A consist of 78 attributes comprising 62 attributes with values and 16 contextual nodes. The XML data structure of Bank B credit data sets consists of 2,290 attributes and attributes values; 213 nodes in each credit data set tree; a maximum tree height of five and a maximum fan-out of ten for each credit data set tree. The flat credit data sets contain 213 attributes of which165 have values and 48 are contextual nodes.

4.5 Application of the selected data mining techniques

The application of various machine learning techniques to increase the accuracy of credit risk assessment for MSMEs was reviewed in Chapter 2. The applied techniques in these studies employed structured data in the form of numerical and binary credit data, omitting the use of descriptive qualitative (unstructured) data to assess MSME credit risk. In order to overcome this shortcoming, this research employed both quantitative and descriptive qualitative credit data. In the initial stage of this research, a total of 50,000 credit data sets were automatically created based on a limited number of samples provided by an Indonesian bank. They were presented in a tree-structured format using XML, and frequent tree mining algorithms were applied to this synthetic credit data (Ikasari, Hadzic, and Dillon 2011). The algorithms that were used include TreeMiner (Zaki 2005a) for mining frequent embedded sub-trees, IMB3-miner (Tan et al. 2006) for mining frequent induced/embedded sub-trees, and CMTreeMiner (Chi et al. 2004) which mines closed/maximal sub-trees. Both TreeMiner and IMB3-miner algorithms failed to extract frequent patterns even at relatively high support thresholds. The closed/maximal sub-tree algorithm CMTreeMiner performed better as it only enumerated those frequent sub-trees of which no superset was frequent (maximal) or for which no superset had the same frequency (closed). However, it extracted an excessive number of patterns at a low support threshold that deterred timely analysis. At the lower support threshold required for better class discrimination, the sub-tree enumeration task became unfeasible. At feasible support thresholds, the frequent sub-trees were too general, with poor class-discriminating power. XRules algorithm (Zaki 2003), which is an XML classifier based on association rules formed from frequent sub-trees, was also used. However, it demonstrated similar limitations because of the excessive number of frequent sub-trees that needed to be extracted to form class-discriminating associations. This warranted the application of more general data mining techniques to extract class discriminating attributes without the need to generate frequent sub-trees. It prompted the transformation of the credit data from a tree-structured to a structure-preserving flat data

format to allow the application of more comprehensive data mining techniques such as Decision Tree (DT) and Rule Induction (RI). In addition, the Forward Feature Selection technique was selected to determine credit risk parameters for MSMEs.

4.5.1 Final set-up of credit data sets

The data used for this research consist of credit data sets with different formats and different target class proportions for each of the sample banks (Table 4.9). The data sets of each sample bank were randomized and partitioned into 70% training and 30% testing sets. The testing data was kept separately (unseen) throughout the process in order to verify the accuracy of the models developed with the training data. In addition to the manual and automatic discretized data sets, one set of data containing the raw attribute values was also prepared. The very low portion of the non-performing credit data sets provided by the sample banks (Table 4.9) resulted in imbalanced data sets.

Table 4.9 Composition of credit data sets of the sample banks

Credit Data Sets	Bank A	Bank B	Bank C
Number and type of financial data	30 attributes (structured)	48 attributes (structured)	10 attributes (structured)
Number and type of descriptive qualitative data	13 attributes (unstructured)	33 attributes (unstructured)	None
Number and type of core loan data	18 attributes (12 structured, 6 unstructured)	56 attributes (15 structured, 41 unstructured)	34 attributes (20 structured, 14 unstructured)
Number and type of scores	None	27 attributes (structured)	117 attributes (68 structured, 49 unstructured)
Total data sets provided	96 credit data sets	165 credit data sets	140 credit data sets
Number of performing loans	58 credit data sets	149 credit data sets	133 credit data sets
Number of non-performing loans	38 credit data sets	16 credit data sets	7 credit data sets

As previously indicated, this imbalance of performing and non-performing loans can be ascribed to the low number of non-performing loans existing in the branches/geographical areas where the sample sets were sourced. The effect of training data set size and class imbalance[16] in DT has been addressed by Catlett (1991); accordingly, in this research,

[16] Imbalanced two-class data problems occur where the main target class of interest is represented by a minority of instances compared to the other target class (Han, Kamber, and Pei 2012). In this research, the main target class of interest is the non-performing loan credit data sets.

oversampling[17] was performed by way of applying replication to the training data sets of the sample banks. Essentially, oversampling was carried out to reduce the skewed proportion between performing and non-performing loan credit data sets and to increase the number of non-performing loan credit data sets. The same replication procedures were applied to the credit data sets of all the sample banks to ensure that the results achieved from each bank could be compared. Specific final set-up actions performed for each of the sample banks are provided in the sections below.

4.5.1.1 Bank A

The 96 credit data sets were partitioned into 67 training data sets and 29 testing data sets. The original training data sets consist of 44 performing and 23 non-performing loans while the original testing data sets contain 14 performing and 15 non-performing loans. The training data sets increased to 134 with the first application of oversampling when performing and non-performing loan credit data sets were all replicated once. With the second oversampling process, performing loan credit data sets were each replicated once and non-performing loan credit data sets were replicated three times. This replication increased the training data sets to 180. The unequal replication applied in this case was to achieve a fairly equal number of performing and non-performing loan credit data sets. Table 4.10 summarizes the different training data sets that existed after replication was performed. It is important to note that the replication was applied to all formats in which the training data exists, namely original (raw) format, manually discretized and automatically discretized.

Table 4.10 Final training data sets – Bank A

Data Sets	Attribute Values	Number of Data Sets in the Training Data Set		
		Performing	Non-performing	Total
Data Set AM1	Manually discretized	44	23	67
Data Set AM2	Manually discretized	88	46	134
Data Set AM3	Manually discretized	88	92	180
Data Set AR4	Original	44	23	67
Data Set AR5	Original	88	46	134
Data Set AR6	Original	88	92	180
Data Set AA7	Automatically discretized	44	23	67
Data Set AA8	Automatically discretized	88	46	134
Data Set AA9	Automatically discretized	88	92	180

[17] Oversampling is one of the sampling strategies whereby resampling is performed on the minor target class to achieve more equal representation of both target classes in the training data set.

4.5.1.2 Bank B

The 165 credit data sets were partitioned into 115 training data sets and 50 testing data sets. The original training data set consisted of 104 performing and 11 non-performing loan credit data sets. The disproportion also existed in the original testing data set of 45 performing and five non-performing credit data sets. Due to the large difference in the numbers of these sets, the number of performing loan credit data sets was not replicated. The replication was only applied to the non-performing sets, with the highest level of replication resulting in 66 non-performing sets of a total of 170 training data sets. Compared with the replication conducted for Bank A, attaining a fairly equal number of performing and non-performing loan credit data sets was not attempted since the original skewness of the proportion was regarded as too much, and oversampling with replication was discreetly carried out for Bank B to avoid extensive distortion of the results from excessive exact duplication of small credit data sets. Table 4.11 summarizes the different training data sets that existed after replication was performed.

Table 4.11 Final training data sets – Bank B

Data Sets	Attribute Values	Number of Data Sets in the Training Data Set		
		Performing	Non-Performing	Total
Data Set BM1	Manually discretized	104	11	115
Data Set BM2	Manually discretized	104	22	126
Data Set BM3	Manually discretized	104	44	148
Data Set BM4	Manually discretized	104	66	170
Data Set BR5	Original	104	11	115
Data Set BR6	Original	104	22	126
Data Set BR7	Original	104	44	148
Data Set BR8	Original	104	66	170
Data Set BA9	Automatically discretized	104	11	115
Data Set BA10	Automatically discretized	104	22	126
Data Set BA11	Automatically discretized	104	44	148
Data Set BA12	Automatically discretized	104	66	170

4.5.1.3 Bank C

The 140 credit data sets were partitioned into 98 training data sets and 42 testing data sets. Bank C data sets presented the most extreme imbalance between performing and non-performing loan credit data sets. The original training data set consisted of 94 performing and only four non-performing loan credit data sets, while the original testing data set contained 39 performing and three non-performing loan credit data sets. Although

oversampling was prudently conducted, it was necessary to reduce the extreme skewness in the number of performing and non-performing loan credit data sets. The number of performing data sets was not replicated; replication only applied to the non-performing credit data sets, with the highest level of replication resulting in 48 non-performing data sets of a total of 142 training data sets. As with bank B, no attempt was made to attain a fairly equal number of performing and non-performing loan credit data sets due to the original skewness of the proportions. Table 4.12 summarizes the different training data sets that existed after replication was performed.

Table 4.12 Final training data sets – Bank C

Data Sets	Attribute Values	Number of Data Sets in the Training Data Set		
		Performing	Non-Performing	Total
Data Set CM1	Manually discretised (by bank)	94	4	98
Data Set CM2	Manually discretised (by bank)	94	16	110
Data Set CM3	Manually discretised (by bank)	94	32	126
Data Set CM4	Manually discretised (by bank)	94	48	142

4.5.2 Decision tree (DT)

The DT algorithms used from RapidMiner include a tree induction implementation based on learning mechanisms of C4.5 (Quinlan 1993) and CART (Breiman et al. 1984) algorithms, and J4.8, an implementation of Quinlan's C4.5 algorithm (Quinlan 1993) from the Weka machine learning workbench (Holmes, Donkin, and Witten 1994). A top-down decision tree consists of nodes representing the discriminating attributes that are selected with respect to the number of instances. In this research, the parameter used to determine the selection of the discriminating attributes (or splitting nodes) was gain ratio. Gain ratio is known to smooth the bias toward attributes with an excessive number of values that occur with information gain[18], and therefore produces higher accuracy performance (Quinlan 1993). As previously mentioned, the issue of class imbalance in DT is a well-known problem. Catlett (1991) indicates that the problem can be overcome by way of pruning, while other scholars propose oversampling and undersampling strategies (Chawla 2003; Drummond and Holte 2003). Although Drummond and Holte (2003) find that

[18] Information gain is an attribute selection criterion based on the minimum entropy. An entropy is a measure of information that emanates from the probability of a value occurrence (Shannon 1997). If the same sequence of values exists across attributes, the entropy is said to be minimum, resulting in high information gain.

undersampling outperforms oversampling, it could not be applied to this research because of the extremely small number of non-performing loan credit data sets.

Over-fitting is another issue commonly known to DT (Breiman et al. 1984; Quinlan 1986). Pruning[19] of the tree can be applied if over-fitting occurs. The model generated by the training data set captures specific characteristics that do not exist in the testing data set. This means the model is lacking generalizability. Removal of branches that are considered not to improve the performance of the tree to a significant extent increases the generalizability of the model. Pruning of a tree can be performed during its construction of the tree (pre-pruning) or after the tree is constructed (post-pruning). Pre-pruning concerns stopping the growth of the tree based on a pre-determined number of splits, to keep the tree from "perfectly fitting" the training data set. Post-pruning is performed by removing nodes, based on the misclassification rate.

4.5.3 Rule induction (RI)

RapidMiner utilizes an RI algorithm similar to that of Repeated Incremental Pruning to Produce Error Reduction (RIPPER), introduced by Cohen (1995). This algorithm reduces the over-fitting issue, also known to exist in the rule induction learners. The learning algorithm in RapidMiner and in Weka uses the less dominant classes as a starting point and prunes the rules exhaustively. Information gain is used as the criterion to select the splitting node. The use of the rule induction learner is highly relevant when practicality of the outcome becomes an issue. With the "IF-THEN" presentation of the rules, RI provides even more comprehensible rules compared to the decision tree, especially when the tree size is big (Han, Kamber, and Pei 2012).

4.5.4 Forward feature selection

Feature selection is a data reduction technique, performed to address the over-fitting problem prompted by the substantial number of attributes in the data sets. The benefits of feature selection include improvement of performance accuracy and increased knowledge comprehension (Dash and Liu 1997; Guyon et al. 2006). Forward feature selection is a method of finding the optimal subset by adding one attribute at a time to an initially empty subset (Schaffernicht et al. 2009). Attributes that increase performance are added to the subset. The iterative process ends when the addition of attributes does not improve the performance, based on a predetermined minimum threshold.

Feature selection is commonly applied before any performance analysis takes place. In this research, performance analysis was initially based on all the credit data set attributes

[19] Pruning of a tree involves removing the sub-tree that originated from a node. As a result the node becomes a leaf node (Shawkat Ali and Wasimi 2007).

provided by the sample banks. The forward feature selection technique was then applied to the sets to provide the sample banks with discriminative attributes useful for the refinement of present credit risk assessment. In this regard, the forward feature selection was applied to the financial and qualitative data attributes of Bank A and Bank B using DT and RI algorithms measured by 10-fold cross-validation. The forward feature selection was not applied to Bank C credit data sets as the scores represented a pre-selection of attributes by the bank. Finally, the accuracy performance of the subset was tested in two different ways. First the selected attributes were used as selection criteria to determine their performance as credit risk predictors with the testing data sets. This meant only the selected attributes were used to predict the credit performance of loans contained in the testing data sets. Second, the selected attributes were tested in combination with the core loan attributes and scores of the testing data set. It excluded all non-selected financial and qualitative data attributes to which the feature selection was applied.

4.6 Summary

The methodology used to achieve the research objectives of this study consisted of four steps, with each step comprising extensive and different processes. Figure 4.10 provides a comprehensive summary of these steps, and shows how the methodology relates to the research objectives of this study. The incorporation of structured (quantitative) and descriptive qualitative (unstructured) data for credit risk assessment required a series of data format conversions. The initial credit data provided by the sample banks was structured and unstructured, and the composition of such data differed for each bank. It was necessary to synchronize the data format by structuring all quantitative and qualitative credit data in a similar way. Once in a structured format, the data was presented in tandem in one XML document with the hierarchical and contextual information of the data preserved. The first research objective was achieved by structuring all qualitative and quantitative credit data of the sample banks in a similar format and presenting them in one document; the use of XML enabled such a presentation. A selection of XML mining techniques was applied, such as XRules and frequent closed/maximal sub-tree algorithms. However, the structure of the tree-structured data became too complex and these techniques failed to produce optimum results. The complexity issue was resolved by applying a conversion process to flatten the data. The extracted general structure (DSM) was used to create the flat-format data while preserving the existing inherent structure. This process made the direct use of well-established data mining techniques possible. The second objective was attained with appropriate data mining techniques selected and applied to the credit data, based on the converted format. The completion of the processes in research objectives one and two was required for the achievement of research objectives three, four and five.

Research objective three was to determine a conceptual framework for the construction of comprehensive credit parameters for MSMEs that integrates quantitative and qualitative information. The use of all credit data supplied by the sample banks (after the structuring of all such data for data mining technique application) resulted in wide-ranging credit data inclusion.

Research objective four was focused to develop an approach to construct credit parameters for MSMEs. The predictive attribute selection and the comparison thereof with other non-selected credit attributes made this possible. These were achieved and validated using the testing credit data sets.

Finally, research objective five was to develop a Decision Support Methodology for MSME credit risk assessment based on knowledge patterns derived from the application of Data Mining. The spheres of importance highlighted by the predictive attributes served as important criteria in highlighting credit-related data that could be used in decision-making about new loans in banks.

In the next chapter, the research results are presented and the research findings discussed.

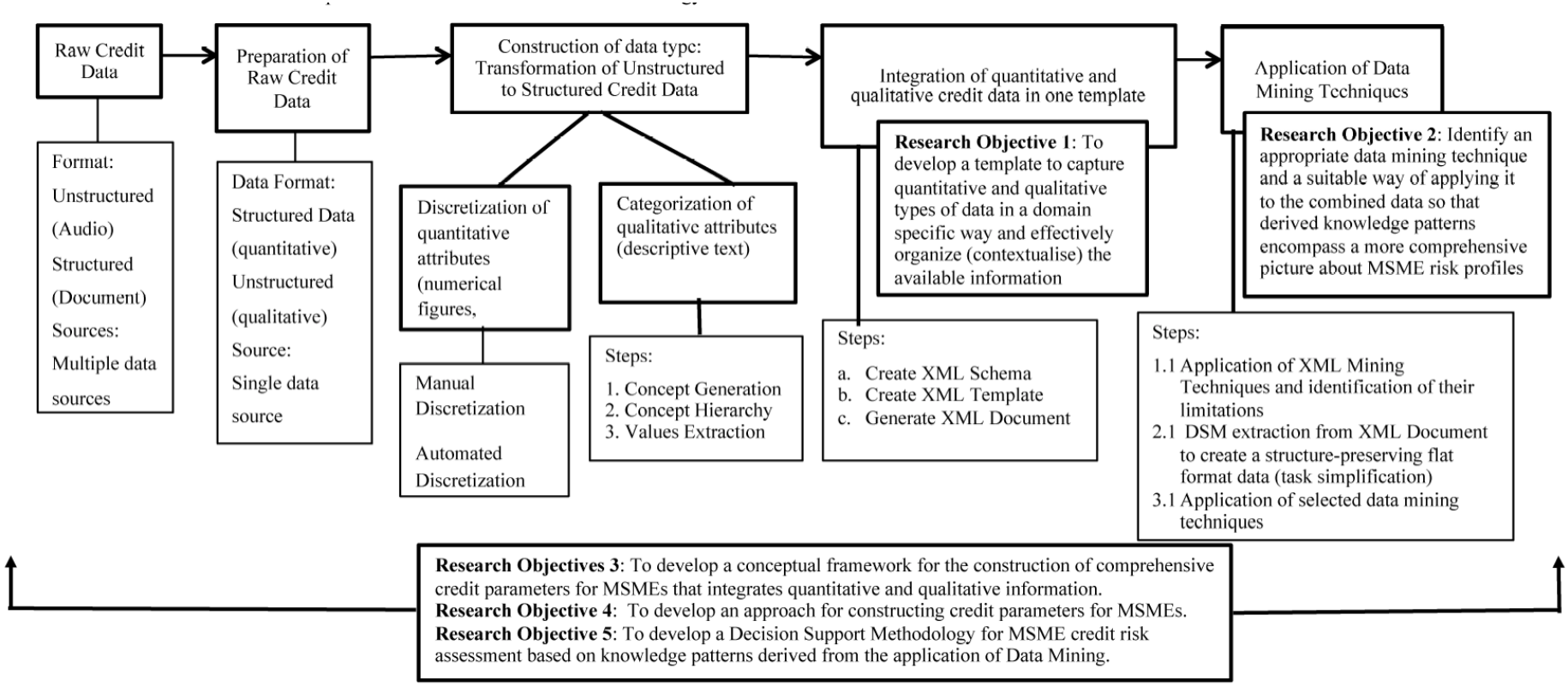


Figure 4.10 A comprehensive view of research methodology

CHAPTER 5 – RESEARCH FINDINGS

5.1 Introduction

In the previous chapter, the methodology to develop reliable credit risk assessment for MSMEs was discussed. The steps in the methodology included a rigorous conversion and transformation process for the credit data sets obtained from the sample banks. The quantitative and qualitative credit data collected from the banks were structured into one document for analysis purposes. The application of the DSM technique transformed the data set into a format that allowed the application of the data mining techniques to all credit data.

In this chapter, the findings derived from the application of data mining techniques to the credit data sets are discussed. Section 5.2 contains the results for the data sets of the different sample banks. In this section, the implications of discretization applied to the attribute values and the oversampling performed to the data set samples are presented, and the credit risk parameters derived from the data sets as well as their accuracy performance are discussed. The derived level of additional knowledge and a comparison of the level of contribution of this study for the different sample banks are presented in Section 5.3. This chapter is then concluded with a summary.

5.2 Results

DT and RI algorithms were applied to the 70% training and 30% testing credit data sets obtained from each sample bank. Table 5.1 lists the different data sets, the applied discretization modes and the extent of oversampling applied. For each sample bank, the performance accuracy attained from the application of DT and RI are presented and analyzed in the following sections.

Table 5.1 Summary of training data sets for the sample banks

Data Sets	**Discretization**	**Oversampling**	
		None	**Number of Times Replicated (PL*;NPL**)**
AM1-AM3	Manual	AM1	AM2 (1;1), AM3 (1;3)
AR4- AR6	None (original)	AR4	AR5 (1;1), AR6 (1;3)
AA7- AA9 AA7*-AA9*	Automatic	AA7 AA7*	AA8 and AA8* (1;1), AA9 and AA9* (1;3)
BM1-BM4	Manual	BM1	BM2 (None;1), BM3 (None;3), BM4 (None;5)
BR5-BR8	None (original)	BR5	BR6 (None;1), BR7 (None;3), BR8 (None;5)
BA9-BA12 BA9*-BA12*	Automatic	BA9 BA9*	BA10 and BA10* (None;1), BA11 and BA11*(None;3), BA12 and BA12* (None;5)
CM1-CM4	Manual (by Bank)	CM1	CM2 (None;3), CM3 (None;7), CM4 (None;11)

*PL = Performing loan credit data sets

**NPL = Non-Performing loan credit data sets

5.2.1 General interpretation of the results

This research focuses on the prediction accuracy of different data mining techniques with regard to performing and non-performing loan classes. Banks are exposed to monetary loss with each misclassification of loan applications. Approval of a loan that is predicted as a performing loan when it actually turns out to be a non-performing loan (false positive error), or the rejection of a loan that is predicted to be non-performing when it actually turns out to be performing (false negative error), are loss-making (unprofitable) decisions for banks. The extent of potential loss from incorrect classification is beyond the scope of this research, but results are interpreted in the context of the extent to which they might assist banks in enhancing the ratio of performing to non-performing loans in comparison with the existing performing versus non-performing loan ratio.

Classification accuracy, prediction accuracy and coverage rate of each data set are presented in this chapter. Coverage rate refers to the percentage of the individual credit data sets within the total data set covered by the model, or by individual rules that make up the model. The coverage rates of the models are provided separately for classification (training data sets) and prediction (testing data sets). The coverage rates of the individual rules that make up the models are discussed based on the percentage of cases in the testing data set that they cover. Although the results are affected by the proportion of the credit data sets, they still serve as good indicators of the applicability and comparative performance of the different data mining techniques that are applied within the context of credit risk prediction.

5.2.2 Bank A

The content of the individual credit data sets of Bank A can be summarized as 30 numerical attributes of customer financial position, 13 textual attributes of customer business and non-business particulars, and 18 attributes of core loan information (a mixture of numerical and textual). Bank A is a small rural bank with 77.09% of the customers in its credit data sets operating as traders. The major portion of the financing (70.83%) provided to the customers sampled for this research is between IDR 1 million to IDR 50 million.

5.2.2.1 Decision tree (DT)

Table 5.2 shows that in all training data sets, the DT algorithm produced good accuracy in classifying the performing and non-performing loans with a maximum of one second of processing time. However, complete misclassification of non-performing loans occurred when it was applied to the testing data sets AM1, AM2, AM3, AA7 and AA8*.

Table 5.2 Performance accuracy of DT – Bank A

Data Sets	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
AM1	8	11	75.47%	71.43%	100.00%	48.28%	0.00%	100.00%
AM2	38	41	90.48%	76.00%	100.00%	48.28%	0.00%	48.28%
AM3	39	40	100.00%	89.32%	98.89%	48.28%	0.00%	55.17%
AR4	10	17	81.08%	53.33%	100.00%	57.89%	70.00%	100.00%
AR5	**15**	**24**	**95.35%**	**87.50%**	**100.00%**	**73.33%**	**78.57%**	**89.66%**
AR6	29	33	100.00%	92.00%	100.00%	61.90%	87.50%	75.86%
AA7	4	4	64.52%	20.00%	100.00%	48.28%	0.00%	100.00%
AA7*	6	8	73.47%	55.56%	100.00%	60.87%	100.00%	96.55%
AA8	39	40	83.51%	81.08%	99.00%	60.87%	100.00%	62.07%
AA8*	42	47	90.80%	80.85%	100.00%	48.28%	0.00%	100.00%
AA9	39	40	100.00%	82.88%	100.00%	60.87%	100.00%	62.07%
AA9*	**39**	**40**	**100.00%**	**88.46%**	**100.00%**	**70.00%**	**100.00%**	**58.62%**

* denotes the application of entropy discretization technique.

The application of DT on Data Sets AA9* and AR5 produced superior average prediction accuracy performance (85% and 75.95%) compared with the rest of the data sets. Data Set AR5 generated a smaller tree than AA9*. That indicates better generalization of the model from more accurately selected nodes; thus, the accuracy performance of Data Set AR5 is selected as the more ideal model. The training data set AR5 contains replicated performing and non-performing loans (each one time replicated) with numerical attribute values kept in original format. Figure 5.1 displays the tree generated from Data Set AR5. The root node of the DT is the attribute "ratio of collateral to the loan principal (X25_rat_nvalcoll_loan)". It branches out to 15 leaf nodes.

The model built by training data set AR5 has an 89.66% testing data set coverage rate consisting of 15 extracted rules for performing and non-performing loans. An exemplary rule for performing loans is "X25_rat_nvalcoll_loan= rat_nvalcoll_loan (>1.014) AND X25_rat_nvalcoll_loan = rat_nvalcoll_loan (≤3.807) AND X42_othopinstexp = othopinstexp (≤208500) AND X60_netincpm = netincpm (>1801750) AND X61_ratinstdeppd = ratinstdeppd (≤45.575) AND X71_salessust = salessust (high) AND X78_ comprofrisk = comprofrisk (low)". This rule covers seven credit data sets with 85.71% prediction accuracy. The credit parameters for predicting a performing loan comprise of a ratio of collateral to the loan principal between 1.014 and 3.807; other operating instalment expenses of less than IDR 208,500; earned income of more than IDR 1,801,750 per month; a maximum of 45% of

the daily income disbursed for loan repayment; high sustainability of sales; and a low compliance risk profile. The compliance risk profile comprises the origin of customers (locale or non-locale), historical loan payment (continuous or disrupted) and continuity of payment (in the absence of person in charge). A low compliance risk profile reflects continuous previous loan repayments, notwithstanding the origin of customers and the availability of a guarantee on daily instalment payments in the absence of the person in charge. This extracted rule is supported by 60 out of 88 performing loans cases in the training data set that is replicated once (i.e. 30 out of 44 actual performing loan cases in the sample provided). This implies its prominence and therefore can be used to refine the credit information collection process.

There are six rules for non-performing loans. These have distinct coverage rates in the training data set, resulting in low coverage rates in the testing data set because of the limited number of credit data sets. The highest testing data set coverage is 10.35% for any non-performing loan rules. An exemplary rule for non-performing loans is "X25_rat_nvalcoll_loan = rat_nvalcoll_loan (>1.014) AND X25_rat_nvalcoll_loan = rat_nvalcoll_loan (≤3.807) AND X42_othopinstexp = othopinstexp (>208500) AND X25_rat_nvalcoll_loan = rat_nvalcoll_loan (>1.153)". This rule has a coverage rate of 6.89% and predicts two non-performing loans with 100% accurately. The credit parameters for predicting a non-performing loan comprise a ratio of collateral to the loan principal between 1.014 and 3.807 but higher than 1.53; and other operating instalment expenses of more than IDR 208,500.

The over-fitting problem for performing credit data sets prevails in all data sets, while in the case of non-performing loans it occurs in Data Sets AM1, AM2, AM3, AR5, AR6, AA7, AA8*. The models generated by the training data sets to classify performing and non-performing loans contain too-specific parameters (attributes and attribute values) as the variety of attribute values that could be related to performing and non-performing loans are not necessarily broadly reflected in the testing data sets. This can be ascribed to the limited number of performing and non-performing credit data sets used in the training and testing data sets (Table 4.10, page 111).

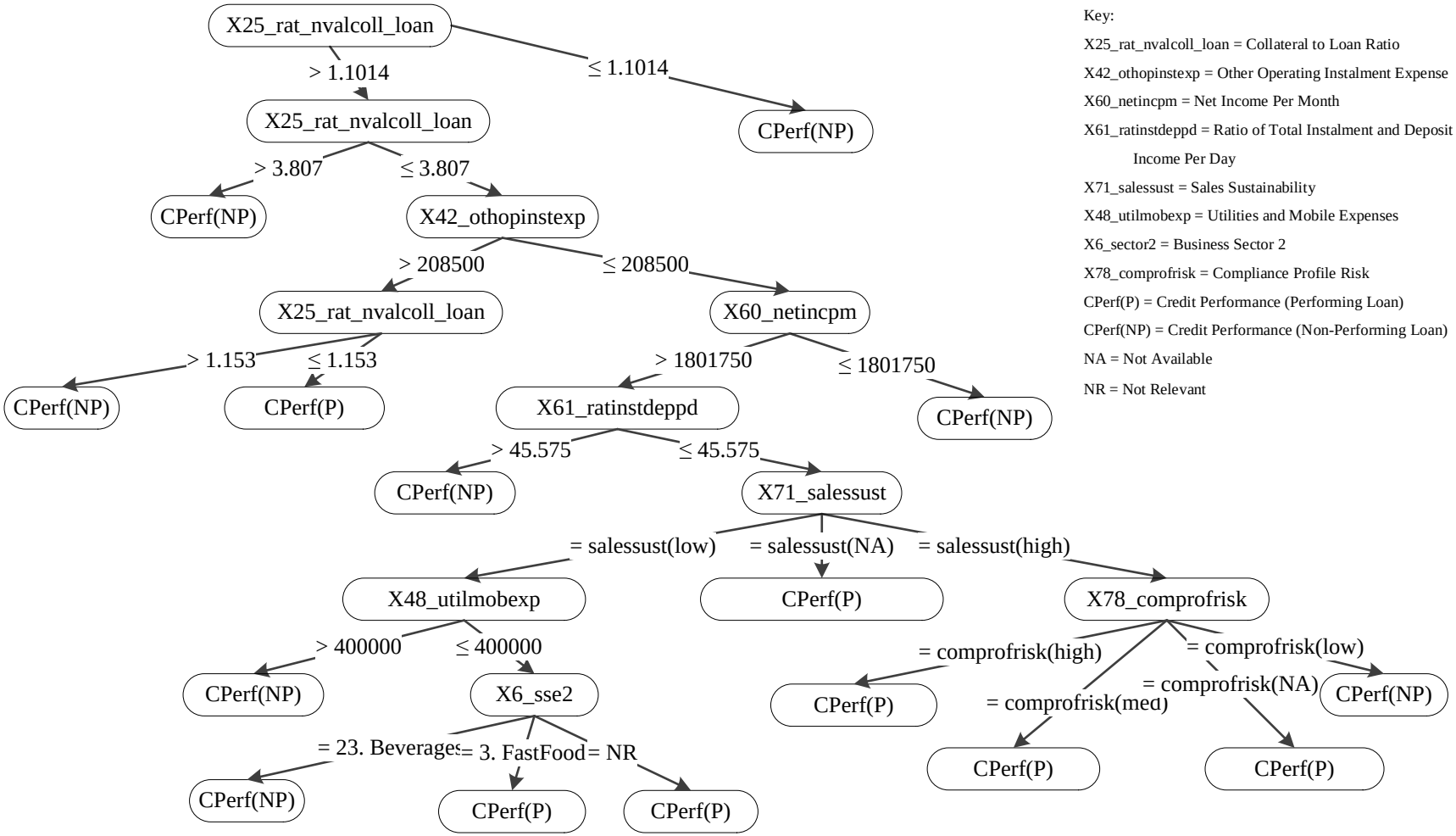


Figure 5.1 Decision tree from Data Set AR5 – Bank A

5.2.2.2 Rule induction (RI)

Table 5.3 shows the performance of the RI application with all data sets. The application of the RI algorithm to the automatically discretized Data Sets AA7, AA8, AA8*, AA9, and AA9* provides on average more performance accuracy than its application to the other data sets. Although findings in previous studies are that the use of entropy-based discretization provides better accuracy performance in comparison with equal-width binning discretization (Dougherty, Kohavi, and Sahami 1995; Kohavi and Sahami 1996), it is not conclusively supported with regard to these credit data sets. Data Set AA8, where numerical attributes were discretized with equal-width binning in one second's processing time, provides the best average prediction accuracy for performing and non-performing loans in the testing data sets (84.07%). The over-fitting problem is more apparent in the case of performing loans, since over-fitting of such loans occurs in all data sets except Data Set AA7. In the case of non-performing loans it occurs in fewer data sets, namely AM1, AM2, AM3, AR5 and AA7*. The models generated by the training data sets to classify performing and non-performing loans contain too-specific credit parameters (attributes and attribute values) because of the limited number of performing and non-performing credit data sets used in the training and testing (14 performing and 15 non-performing credit data sets) data sets (Table 4.10, page 111).

Table 5.3 Performance accuracy of RI – Bank A

Data Sets	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
AM1	3	74.14%	88.89%	94.03%	48.28%	0.00%	93.10%
AM2	5	75.00%	66.67%	99.25%	48.28%	0.00%	93.10%
AM3	8	85.23%	85.87%	99.44%	51.72%	0.00%	100.00%
AR4	7	72.55%	56.25%	98.51%	56.25%	61.54%	96.55%
AR5	3	86.52%	75.56%	97.01%	54.55%	71.43%	86.21%
AR6	3	97.50%	90.00%	96.11%	68.42%	90.00%	79.31%
AA7	1	66.04%	35.71%	94.02%	100.00%	58.33%	86.21%
AA7*	3	72.73%	66.67%	91.04%	48.28%	0.00%	51.72%
AA8	**4**	**81.44%**	**75.68%**	**99.25%**	**76.47%**	**91.67%**	**96.55%**
AA8*	3	72.38%	58.62%	95.52%	70.59%	83.33%	89.66%
AA9	4	88.37%	87.23%	98.89%	60.87%	100.00%	100.00%
AA9*	7	95.29%	92.63%	98.33%	66.67%	100.00%	89.66%

* denotes the application of entropy discretization technique.

The performing and non-performing loans in the AA8 training data set were replicated once. The model built by training data set AA8 has a 96.55% testing data set coverage rate. There are four extracted rules for performing and non-performing loans. One exemplary rule for performing loans is "if X18_ratioofdailyinstalmentanddeposittoloan = ratioofdailyinstalmentanddeposittoloan (range2 [1.592 - 1.734])". This rule covers 14 credit data sets with 57.14% prediction accuracy. The rule consists of a single credit parameter that is a ratio of the daily bank deposits by a customer to the daily loan instalment of between 1.592 and 1.734. Another rule with higher prediction accuracy for performing loans is "if X78_ comprofrisk= comprofrisk (low) AND X6_sector2= sector2 (NR)". This rule covers eight credit data sets with 75% prediction accuracy. The credit parameters are low compliance risk profile; and single business venture. Taking into account the Bank A borrower profile (Chapter 4, Section 4.3.2.1) where 36.46% of the borrowers have multiple businesses, this rule suggests that careful consideration is required regarding the number of businesses the applicant possesses when assessing his/her credit worthiness. This extracted rule is supported by 34 out of 88 performing loan cases in the training data set that is replicated once (i.e. 17 out of the 44 actual performing loan cases in the provided sample). This implies its prominence.

The rule with the highest prediction accuracy for non-performing loans in the testing data set is "if X21_purposeofloan = purposeofloan(additionalworkingcapital)". This rule covers five credit data sets with 100% prediction accuracy. This rule comprises one credit parameter: that the loan is intended for additional working capital. This parameter on its own does not seem to be realistic criterion, since working capital is an important form of finance for any business. However, it should be interpreted from the point of view that the RI analysis identified it as a prominent non-performing loan predictor based on the training data sets obtained from Bank A. It is supported by 16 out of 46 non-performing loans in the training data set that is replicated once (i.e. eight out of the 23 actual non-performing loan cases in the provided sample). Although no over-fitting is evident in this data set, it is evident that the limited size of the total data sets may have contributed to this rule.

5.2.2.3 Forward feature selection

The 30 numerical attributes of customer financial position and the 13 attributes of customer business and non-business particulars in the individual credit data sets serve as the base data for the selection of attributes with the Forward Feature Selection technique. Core loan data attributes are excluded from the Forward Feature Selection since they are all regarded as important attributes that should be utilized for credit risk assessment in combination with the selected attributes. Figures 5.2 (on DT) and 5.3 (on RI) show the selected attributes for each of the data sets.

Data Sets	1. Total sales	2. Ratio of total sales to total COGS	3. Maintenance expense	4. Holiday allowance expense	5. Other operating instalment expense	6. Household expense	7. Miscellaneous expense	8. Education and children exp	9. Total expense	10. Ratio of non-operating to operating expense	11. Total instalment and deposit per day	12. Net income per day	13. Total COGS	14. Business conditions	15. Business experience	16. Payment of sales	17. Sales Sustainability	18. Payment of purchases	19. Purchases Sustainability	20. Employee	21. Bookkeeping	22. Consumers	23. Business Turnover	24. Character	25. Suppliers	26. Compliance Profile Risk
AM1					X																	X				X
AM2					X		X	X			X			X												X
AM3					X	X				X																X
AR4	X		X				X																			X
AR5							X					X		X			X	X								
AR6				X						X											X					X
AA7		X												X								X				X
AA*																				X						X
AA8				X										X					X							
AA*		X			X																					X
AA9									X					X	X	X				X		X				X
AA*					X								X	X	X				X	X			X	X	X	X
Total	1	2	1	2	5	1	3	1	1	2	1	1	1	5	2	1	1	1	2	3	1	3	1	1	1	10

Figure 5.2 Forward feature selection technique attributes (DT) – Bank A

5.2.2.3.1 Test of DT on selected attributes

A total of 26 attributes, comprising 13 income and expense and 13 qualitative credit data attributes, are selected by the technique. The most frequently selected attribute is "compliance risk profile", which is extracted from ten of the 12 data sets. This is followed by "business conditions" (five data sets) and "other operating instalment expenses" (five data sets). These attributes also form part of the DT decision rules in Figure 5.1. The processing times in brackets and performance accuracy of DT applied the selected attributes in the different data sets are provided in Tables 5.4 and 5.5. Table 5.4 shows that over-fitting of performing loans occurs in all data sets except AM1, while for non-performing loans it occurs in all data sets except AA7, AA7* and AA8*. Data Set AM1, with manually discretized numerical attribute values, provides the best average prediction accuracy of 83.34%. The performing and non-performing loans in this training data set were not replicated.

Table 5.4 Performance accuracy when applying forward feature selection technique attributes (DT) – Bank A

Data Sets (in seconds)	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
AM1 (2)	**8**	**10**	**80.39%**	**81.25%**	**100.00%**	**100.00%**	**66.67%**	**96.55%**
AM2 (1)	32	44	85.42%	84.21%	96.27%	60.00%	77.78%	82.76%
AM3 (1)	38	50	98.78%	92.86%	100.00%	52.63%	60.00%	44.83%
AR4 (1)	6	8	80.77%	86.67%	100.00%	70.59%	83.33%	96.55%
AR5 (1)	24	40	92.05%	84.78%	96.27%	57.89%	70.00%	82.76%
AR6 (1)	23	39	100.00%	95.83%	100.00%	71.43%	73.33%	44.83%
AA7 (0)	6	7	77.78%	84.62%	100.00%	60.87%	100.00%	100.00%
AA7* (0)	7	8	76.36%	83.33%	100.00%	61.90%	87.50%	93.10%
AA8 (1)	11	13	78.85%	80.00%	100.00%	48.28%	0.00%	100.00%
AA8* (0)	4	4	87.23%	85.00%	100.00%	60.87%	100.00%	100.00%
AA9 (1)	32	43	100.00%	95.83%	100.00%	65.00%	88.89%	96.55%
AA9* (0)	30	41	100.00%	97.87%	100.00%	48.28%	0.00%	89.66%

* denotes the application of entropy discretization technique.

The model built by training data set AM1 has 96.55% testing data set coverage rate. There are eight extracted rules for performing and non-performing loans, of which four rules exist for performing loans. Their coverage rate ranges between 3.44% and 34.48%. An exemplary rule for performing loans that has the best coverage rate is "X42_ othopinstexp = othopinstexp (NR) AND X78_ comprofrisk = comprofrisk (low)". This rule, with a coverage

ratio of 34.48%, covers ten credit data sets with 76.92% prediction accuracy. The credit parameters for predicting performing loans comprise non-existing other operating instalment expenses and a low compliance risk profile.

The rule from AM1 that provides the best prediction accuracy for non-performing loans is "X42_ othopinstexp = othopinstexp (NR) AND X78_ comprofrisk = comprofrisk (high)". This rule covers six credit data sets with 100% prediction accuracy. The credit parameters for predicting non-performing loans comprise non-existing other operating instalment expenses combined with a high compliance risk profile.

Table 5.5 shows the performance accuracy of each data set, with processing time in brackets. The over-fitting of performing loans occurs in all except data sets AM1, AM2 and AR4 when combining the selected attributes with the core loan attributes, and in four sets (AM1, AM2, AR6 and AA8*) for non-performing loans. There is only one set (AR4) without any over-fitting problems for performing and non-performing credit data sets, and DT failed to generate a decision tree for one data set (AA7). Data Set AA7*, with numerical attribute values that are automatically discretized using the entropy-based discretization technique, provides the best prediction accuracy (86.64%) when combining the selected attributes with the core loan attributes (Table 5.5). No replication of performing and non-performing credit data sets is applied to this data set.

Table 5.5 Performance accuracy when applying forward feature selection technique attributes combined with core loan attributes (DT) – Bank A

Data Sets (in seconds)	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
AM1 (1)	6	7	75.93%	76.92%	100.00%	100.00%	63.64%	96.55%
AM2 (1)	38	41	84.09%	69.57%	100.00%	100.00%	60.87%	44.83%
AM3 (1)	39	40	100.00%	90.20%	100.00%	70.00%	100.00%	51.72%
AR4 (1)	10	16	68.00%	41.18%	100.00%	72.73%	66.67%	100.00%
AR5 (1)	38	43	86.67%	77.27%	100.00%	80.00%	85.71%	44.83%
AR6 (1)	26	32	100.00%	92.93%	100.00%	70.59%	83.33%	51.72%
AA7 (1)	1	1	65.67%	0.00%	100.00%	48.28%	0.00%	100.00%
AA7* (1)	**5**	**6**	**77.34%**	**50.00%**	**100.00%**	**73.68%**	**100.00%**	**100.00%**
AA8 (1)	39	40	81.32%	67.44%	98.51%	60.87%	100.00%	44.83%
AA8* (1)	45	49	84.27%	71.11%	92.54%	48.28%	0.00%	55.17%
AA9 (1)	39	40	100.00%	86.79%	97.78%	60.87%	100.00%	75.86%
AA9* (0)	39	40	98.70%	88.35%	100.00%	70.00%	100.00%	58.62%

* denotes the application of entropy discretization technique.

The model built by training data set AA7* has a 100% testing data set coverage rate. There are five extracted rules for performing and non-performing loans, of which three rules are for performing loans with a coverage rate between 3.44% and 34.48%. An exemplary rule for performing loans with the best coverage rate is "X78_ comprofrisk = comprofrisk (low)". The credit parameter for predicting a performing loan is low compliance risk profile. This rule has a coverage rate of 10.34%. It covers ten credit data sets with 71.43% prediction accuracy. The rule for predicting non-performing loans is "X78_ comprofrisk = comprofrisk (high)". This rule covers six credit data sets with 100% prediction accuracy. The credit parameter for predicting a non-performing loan is high compliance risk profile. Both the abovementioned rules are based on the same attribute but applying different parameters.

In comparison with the performance accuracy, where all credit data attributes are used, the prediction accuracy improves by applying the Forward Feature Selection Technique. This is particularly evident in the data sets where the credit attribute values are manually discretized. When all credit data is used, all three data sets (AM1, AM2 and AM3) have identical average prediction performance of 24.14% (Table 5.2), while the average performance increases to more than 24.14% when a reduced number of attributes is used, showing the more relevant importance of the fewer attributes to credit risk assessment (Tables 5.4 and 5.5). The effect of the reduced number of credit attributes on the predictability of the extracted rules is less significant where raw attribute values and automatically discretized attribute values are used in data sets. For example, the average prediction performance of Data Set AR4 is 63.95% when all credit data is used; it increases to 76.96% when selected attributes are used but decreases to 69.70% when the selected attributes are used in combination with core loan attributes.

5.2.2.3.2 Test of RI on selected attributes

When RI is applied, the forward feature selection technique identifies 25 attributes consisting of 14 numerical attributes of customer financial position and eleven textual attributes of customer business and non-business particulars. The most frequently selected attribute is the "compliance risk profile", extracted from eight of the 12 data sets, followed by "other operating instalment expenses", extracted from five sets.

Data Sets	1. Total COGS	2. Ratio of total sales to total COGS	3. Security and cleaning expense	4. Rego allowance expense	5. Depreciation expense	6. Miscellaneous expense	7. Total operating expense	8. Household expense	9. Education and children expense	10. Other operating instalment expense	11. Ratio of non-operating to operating expense	12. Net income per day	13. Net income per month	14. Other instalment expense	15. Sales Sustainability	16. Purchase Sustainability	17. Business conditions	18. Bookkeeping	19. Consumers	20. Employees	21. Payment of purchase	22. Compliance Profile Risk	23. Suppliers	24. Payment of sales	25. Business Turnover
AM1		X								X			X												
AM2				X						X									X			X			
AM3			X		X							X		X					X			X			
AR4								X		X					X										
AR5	X				X						X									X		X			
AR6	X											X						X							
AA7						X				X						X									
AA7*																	X					X			
AA8									X										X			X			
AA8*		X								X												X	X	X	
AA9							X										X	X			X	X			
AA9*	X													X								X			X
Total	3	2	1	1	2	1	1	1	1	5	1	2	1	2	1	1	2	2	3	1	1	8	1	1	1

Figure 5.3 Forward feature selection technique attributes (RI) – Bank A

The processing times and performance accuracy of RI applied to the selected attributes in the different data sets is provided in Tables 5.6 and 5.7. Compared with the DT results, over-fitting of performing loans occurs to 12 data sets for RI and 11 for DT (Tables 5.6 and 5.4). On the other hand, in the case of non-performing loans, over-fitting occurs with seven data sets for RI and nine for DT (Tables 5.6 and 5.4). When compared with performance where all credit attributes are used for credit risk assessment, the number of data sets affected by over-fitting problems for non-performing loans is similar to that when all attributes are used for credit risk assessment. This is evident in Table 5.3, where over-fitting implicates only five data sets when RI is applied to all credit attributes.

Table 5.6 shows that the best average prediction accuracy is achieved with Data Set AA9 (81.82%). The performing loans are replicated once, and non-performing loans are three times replicated in the AA9 training data set, while the numerical attribute values are discretized using the equal-width binning technique. Classification and prediction accuracy based on the entropy-based discretization technique for Data Set AA7* is not available as the selected attributes from Data Set AA7* are descriptive qualitative text ("business condition" and "compliance profile risk"). There are more variations in the rules discovered by RI in comparison to those of DT as a result of the greater variety in the attributes selected.

Table 5.6 Performance accuracy when applying forward feature selection technique attributes – RI

Data Sets (in seconds)	**# of Rules**	**Classification Accuracy with Training Credit Data Sets**		**Coverage Rate**	**Prediction Accuracy with Testing Credit Data Sets**		**Coverage Rate**
		Performing Loans	**Non-Performing Loans**		**Performing Loans**	**Non-Performing Loans**	
AM1 (0)	2	81.63%	77.78%	95.52%	53.85%	100.00%	89.66%
AM2 (1)	5	85.26%	82.05%	97.01%	54.17%	80.00%	65.52%
AM3 (1)	12	80.68%	81.52%	97.78%	63.16%	80.00%	79.31%
AR4 (0)	5	74.00%	58.82%	98.51%	54.55%	71.43%	100.00%
AR5 (1)	2	83.50%	93.55%	94.78%	48.28%	0.00%	48.28%
AR6 (1)	2	81.11%	83.33%	99.44%	66.67%	81.82%	100.00%
AA7 (0)	3	78.72%	65.00%	97.01%	0.00%	51.72%	75.86%
AA7*	-	-	-	-	-	-	-
AA8 (1)	7	72.65%	82.35%	98.51%	64.71%	75.00%	96.55%
AA8* (1)	3	83.50%	63.04%	94.78%	0.00%	51.72%	51.72%
AA9 (1)	**5**	**89.53%**	**88.30%**	**95.00%**	**63.64%**	**100.00%**	**79.31%**
AA9* (1)	5	79.79%	84.88%	100.00%	65.00%	88.89%	100.00%

* denotes the application of entropy discretization technique.

The model built by training data set AA9 has a 79.31% testing data set coverage rate. There are five extracted rules for performing and non-performing loans. An exemplary rule with the highest prediction accuracy for performing loans is "if X78_comprofrisk = comprofrisk (low)". This rule covers five credit data sets with 80% prediction accuracy. The rule consists of a single credit parameter namely a low compliance risk profile.

An exemplary rule with the highest prediction accuracy for non-performing loans is "if X78_comprofrisk = comprofrisk (high) AND X63_busscon = busscon (stable)". This rule covers six credit data sets with 100% prediction accuracy. The rule comprises the following credit parameters: a high compliance risk profile and stable business operations.

When the selected attributes are tested together with the core loan attributes, over-fitting of performing loans for RI and DT are similar (ten data sets for RI and nine for DT – Tables 5.7 and 5.5). It affects fewer numbers of data sets (ten data sets), compared with 12 sets when RI is applied to Forward Feature Selection Technique attributes without core loan attributes (Table 5.6). The over-fitting for non-performing loans (four data sets – Table 5.7) is less than that when RI is applied to Forward Feature Selection Technique attributes without core loan attributes (seven data sets – Table 5.6), and similar to that of DT applied to the same attributes (four data sets for RI and DT– Tables 5.7 and 5.5).

Data Set AM3 provides the best average predictive accuracy (86.78%) for the testing data set. The performing loans in this dataset are replicated once and the non-performing loans three times. The credit data sets in AM3 contain manually discretized numerical attribute values.

The model built by training data set AM3 has a 96.55% testing data set coverage rate. There are four extracted rules for performing and non-performing loans. An exemplary rule with the highest prediction accuracy for performing loans is "if X52_othinstexp = othinstexp (NR)". This rule covers eleven credit data sets with 90.91% prediction accuracy. The rule consists of a single credit parameter, namely non-existence of other instalment expense.

There are three rules for non-performing loans. The rule with the highest prediction accuracy for non-performing loans is "if X21_ purposeofloan = purposeofloan (additionalworkingcapital)". This rule covers five credit data sets with 100% prediction accuracy and is composed of one credit parameter: the loan is intended for additional working capital of the business. This is identical with the rule extracted from Data Set AA8 in Table 5.3 when RI is applied to all attributes. As previously mentioned, this credit parameter on its own does not seem a realistic criterion since working capital is an important form of finance for any business. However, it should be interpreted from the point of view that the RI analysis identified it as a prominent non-performing loan predictor based on the training data sets obtained from Bank A. It is supported by 40 out of 92 cases of non-

performing credit data sets in the training data set that is replicated three times (i.e. ten out of 23 actual non-performing loan cases in the provided sample). It is evident that the limited size of the total data sets may have resulted in this rule.

Table 5.7 Performance accuracy when applying forward feature selection technique attributes combined with – RI

Data Sets (in seconds)	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
AM1 (1)	1	73.33%	100.00%	100.00%	58.33%	100.00%	100.00%
AM2 (1)	2	80.00%	76.47%	95.52%	66.67%	100.00%	55.17%
AM3 (1)	**4**	**90.43%**	**96.51%**	**99.44%**	**92.31%**	**81.25%**	**96.55%**
AR4 (1)	4	66.67%	37.50%	95.52%	48.28%	0.00%	48.28%
AR5 (0)	6	81.25%	73.68%	97.76%	66.67%	100.00%	93.10%
AR6 (2)	9	87.95%	84.54%	99.44%	65.00%	88.89%	96.55%
AA7 (1)	3	69.35%	80.00%	94.03%	48.28%	0.00%	65.52%
AA7* (1)	2	72.22%	61.54%	94.03%	90.91%	72.22%	82.76%
AA8 (1)	5	84.54%	83.78%	99.25%	68.42%	90.00%	96.55%
AA8* (1)	2	74.51%	62.50%	97.01%	58.33%	100.00%	58.62%
AA9 (1)	7	88.75%	83.00%	99.44%	60.87%	100.00%	89.66%
AA9* (1)	2	90.79%	81.73%	95.56%	0.00%	51.72%	82.76%

* denotes the application of entropy discretization technique.

5.2.2.4 Summary of results – Bank A

Table 5.8 shows the data sets with the best performance accuracy for the application of DT and RI to the different attribute combinations, namely all credit attributes, numerical attributes of customer financial position, and textual attributes of customer business and non-business particulars identified with the Forward Feature Selection Technique; and a combination of core loan attributes with numerical attributes of customer financial position and textual attributes of customer business and non-business particular attributes identified with the Forward Feature Selection Technique. When all credit data attributes from Bank A are used, the application of RI provides the best average prediction performance with the equal-width binning discretized data set AA8 (84.07%).

When selected credit data attributes are used, the best average prediction performance is achieved with RI applied to manually discretized data set AM3 (86.78%). Both data sets consist of replicated performing and non-performing loans. The use of the Forward Feature

Selection Technique improves the performance accuracy of the data sets. The credit parameter most frequently selected with this technique is "compliance risk profile". The over-fitting problem of the performing loans is reduced by the application of Forward Feature Selection, while the effect of a reduced number of attributes is less evident in the case of non-performing loans.

Table 5.8 The best performance accuracy for Bank A credit data sets

Data Mining Techniques		**Data Sets**	**Classification Accuracy with Training Credit Data Sets**		**Prediction Accuracy with Testing Credit Data Sets**		**Average Prediction Accuracy**
			Performing Loans	**Non-Performing Loans**	**Performing Loans**	**Non-Performing Loans**	
DT		AR5	95.35%	87.50%	73.33%	78.57%	75.95%
RI		AA8	81.44%	75.68%	76.47%	91.67%	84.07%
Forward Feature Selection and DT	Selected Attributes	AM1	80.39%	81.25%	100.00%	66.67%	83.34%
	Combined Attributes	AA7*	77.34%	50.00%	73.68%	100.00%	86.64%
Forward Feature Selection and RI	Selected Attributes	AA9	89.53%	88.30%	63.64%	100.00%	81.82%
	Combined Attributes	**AM3**	**90.43%**	**96.51%**	**92.31%**	**81.25%**	**86.78%**

5.2.3 Bank B

The content of the individual credit data sets of Bank B can be summarized as 48 numerical financial position attributes of customers; 33 textual attributes related to customer business and non-business particulars including a SWOT analysis; a mixture of 56 numerical and textual attributes of core loan data; and 27 numerical score attributes. Bank B is one of the largest banks in Indonesia and provides financial services predominantly to micro and small businesses. Traders (50.91%) and service providers (23.64%) are the dominant borrowers in the credit data sets provided by Bank B. The majority of the loans (58.79%) are between IDR 500 million to IDR 2 trillion. Bank B collects core loan information, financial reports and business information (hereafter referred to as "QI" for "Qualitative Information") from loan applicants, to be used in the credit risk assessment process. In addition to the collected information for each applicant, Bank B calculates financial ratios (hereafter referred to as "FI" for "Financial Information") from the reports and assigns credit scores to selected financial ratios and business information. In this regard, the credit data sets contain "raw" (QI and FI) and "analyzed QI and FI" (credit scores) information. This combination of

information available from Bank B requires substantial analysis to determine the extent to which the compiled scoring can contribute to performance accuracy. The analysis should shed light on whether the scores add value to the existing raw FI and QI information or only duplicate FI and QI information in an analyzed format.

The imbalanced (skewed) proportion of performing and non-performing loans in the training and testing data sets affects the performance accuracy of the data mining techniques irrespective of the levels of replication applied to the training data sets. Consequently, the application of DT and RI to the data sets primarily provides an indication of the comparative significance of such data sets for constructing prediction methods and the potential contribution of the data mining techniques to the traditional credit risk methodology of Bank B. As such, the best prediction rules in this section are selected in terms of their usefulness in real-world scenarios, notwithstanding their prediction accuracy based on the individual data sets provided by Bank B.

5.2.3.1 Decision tree

The DT algorithm is applied to different combinations of attributes:

Combination1: Core loan attributes, QI, FI and compiled scores (all credit attributes);

Combination 2: Core loan, FI and QI attributes; and

Combination 3: Core loan attributes and scores.

5.2.3.1.1 Test of all credit attributes

The DT algorithm applied to core loan attributes, QI, FI and compiled scores reflects the class (performing and non-performing loan) skewness in the training and testing data sets. The prediction accuracy of 90% for performing loans and 0% for non-performing loans in Data Sets BM1, BM2, BM3, BM4, BR5, BA9, BA9* and BA10* are evidence of this skewness and the trees do not perform any splits, as applying constraints (or splits) does not improve the overall classification performance. The minimum size of the trees (one leaf node) is therefore retained. These single leaf node trees are known as default trees and do not represent valid attainable accuracy based on the data at hand. The non-performing loans in the BM1, BR5, BA9 and BA9* training data sets are not replicated, and constitute 11 of the 115 loans in the training data sets. The non-performing loans in the BM2 and BA10* training datasets are replicated once, and constitute 22 of the 126 loans in the training data sets. The non-performing loans in the BM3 training datasets are replicated three times and represent 44 of 148 total loans in the training data sets. The non-performing loans in BM4 training data sets are replicated five times and represent 66 of 170 total loans in the training data sets. The non-performing loans make up five of the 50 loans in the testing data sets.

As indicated in Chapter 4, different levels of replication are applied to non-performing loan credit data sets to smooth the effect of the large imbalance in the number of performing

and non-performing loan credit data sets. Table 5.9 shows the classification accuracy for eight data sets (BR6, BR7, BR8, BA10, BA11, BA11*, BA12 and BA12*) where non-performing loans replicated at different levels are on average generating more classification accuracy than data sets where replication have not been applied (BM1, BR5, BA9 and BA9*). Furthermore, the higher classification accuracy associated with replication does not necessarily result in high prediction accuracy, since in the cases of BA11, BA11*, BA12 and BA12*, no prediction accuracy exists for non-performing loans in the testing data sets.

Table 5.9 Performance accuracy of DT applied to all credit attributes – Bank B

Data Sets	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1	1	1	90.27%	0.00%	100.00%	90.00%	0.00%	100%
BM2	1	1	82.93%	33.33%	100.00%	90.00%	0.00%	100%
BM3	1	1	70.27%	0.00%	100.00%	90.00%	0.00%	100%
BM4	1	1	61.78%	46.15%	100.00%	90.00%	0.00%	100.00%
BR5	1	1	89.91%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6	10	18	95.79%	58.06%	100.00%	93.02%	28.57%	100.00%
BR7	11	19	100.00%	75.86%	100.00%	91.30%	28.00%	100.00%
BR8	11	19	100.00%	82.50%	100.00%	91.30%	28.00%	100.00%
BA9	1	1	90.35%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9*	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10	**13**	**22**	**87.85%**	**47.37%**	**100.00%**	**97.44%**	**36.36%**	**98.00%**
BA10*	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11	3	4	79.53%	85.71%	100.00%	89.80%	0.00%	100.00%
BA11*	4	6	77.27%	87.50%	100.00%	90.00%	0.00%	100.00%
BA12	3	4	72.22%	70.45%	100.00%	89.80%	0.00%	100.00%
BA12*	16	26	97.92%	86.49%	100.00%	90.00%	0.00%	100.00%

* denotes the application of entropy discretization technique.

The over-fitting problem for non-performing loans prevails in all data sets that provide results (BR6, BR7, BR8, BA10, BA11, BA11*, BA12 and BA12*). Therefore, the models generated by the training data sets for the non-performing loans contain too-specific credit parameters (attributes and attribute values) and show no or low predictive ability power when applied to the testing data sets. The over-fitting problem for performing loans occurs in four (BR6, BR7, BR8 and BA12*) of these eight data sets. This problem does not occur for performing loans in Data Set BA10, which provides the best average prediction accuracy

(66.90%) in zero second processing time. In this data set, numerical attribute values are automatically discretized using equal-width binning, and the non-performing loans are one-time replicated in the training data set. Figure 5.4 shows the tree generated from Data Set BA10. The root node of the decision tree starts with the attribute "applicant' account management (X14_accmngmtsc)" and branches out to 13 leaf nodes.

The model built by training data set BA10 has a 98.00% testing data set coverage rate. There are 13 extracted rules for performing and non-performing loans. An exemplary rule from BA10 that predicts the performing loans the most accurately is "X47_acctmngmtsc= acctmngmtsc (≤ 0.500)". This rule covers 17 credit data sets with 94.12% prediction accuracy. It consists of a single credit parameter, namely the applicant's standing with the bank. This credit parameter comprises a good standing of the applicant with the bank consisting of timely repayments of previous loans, submission of all required application documents and full compliance with previous credit contracts.

There are six rules for non-performing loans. These rules have distinct coverage rates in the training data set, resulting in low coverage rates in the testing data set because of the skewed proportion and limited number of credit data sets. The highest testing data set coverage is 20% for any non-performing loan rules. One exemplary rule for non-performing loans is "X144_acctmngmtsc= acctmngmtsc (>0.500) AND X158_ttlscorenf= ttlscorenf (> 7.875) AND X159_ttlscore=ttlscore (>24.250) AND X15_amt3=amt3 (range4 [4521919166.400 - 6029225555.200])". This rule has a coverage rate of 20% with a 33.33% prediction accuracy rate. It comprises more credit parameters than the previously mentioned rule for performing loans. The credit parameters are the applicant's somewhat poor standing with the bank, indicative of failure to make timely repayments of previous loans, and/or submission of all required application documents and/or full compliance with previous credit contracts; unstable non-financial standing; high overall credit risk; and existence of outstanding financial obligations with a third bank the applicant has conducted business with. These financial obligations to the third bank include acquired business finance (such as working capital) and personal finance (such as home loans, credit cards, and other consumer loans).

Another rule for non-performing loans is "X144_acctmngmtsc= acctmngmtsc (>0.500) AND X158_ttlscorenf= ttlscorenf (>7.875) AND X159_ttlscore=ttlscore (>24.250) AND X15_amt3=amt3 (range1 [-∞ - 1507306388.800]) AND X75_industryfut = industryfut (grow)". The credit parameters for this rule differ from the previous rule in two ways. First, the parameter refers to a smaller amount of outstanding financial obligation with a third bank the applicant has conducted business with. Second, there is an additional parameter, namely the growing future prospect of the industry. This rule also has a coverage rate of 20.00% with a 33.33% prediction accuracy rate. This extracted rule is supported by eight out of 22

cases in the training data set that is replicated once (i.e. four out of eleven actual performing loan cases in the provided sample). This implies its prominence and therefore can be used to refine the credit information process, particularly with regard to specific industry information more relevant to credit risk.

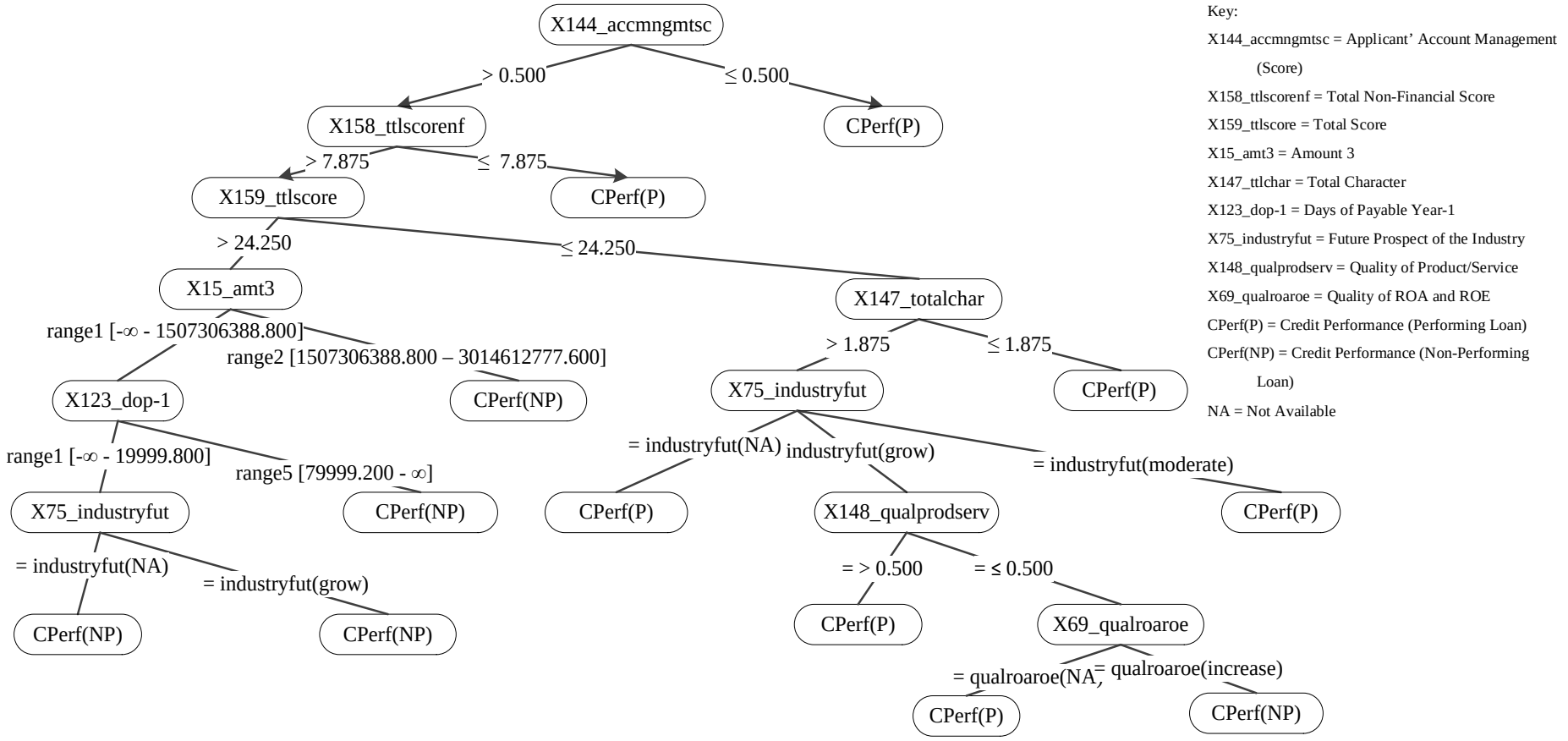


Figure 5.4 Decision tree from Data Set BA10 – Bank B

5.2.3.1.2 Test of core loan, FI and QI attributes

Table 5.10 shows the performance accuracy of using core loan, FI and QI attributes. Each test is completed in one second processing time. The DT algorithm produces ten default trees. Four of the six data sets that have multiple leaf nodes show prediction ability with regard to non-performing loans in the testing data sets. Two of these data sets have numerical attribute values that kept in original format (BR6 and BR7) and the other two have numerical attribute values that are entropy-based discretized (BA11* and BA12*).

Table 5.10 Performance accuracy of DT applied to core loan, FI and QI attributes – Bank B

Data Sets	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BM3	1	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4	1	1	61.18%	0.00%	100.00%	90.00%	0.00%	100.00%
BR5	3	4	89.72%	0.00%	100.00%	89.80%	0.00%	100.00%
BR6	**10**	**18**	**90.20%**	**50.00%**	**100.00%**	**92.86%**	**25.00%**	**100.00%**
BR7	11	20	84.62%	52.63%	100.00%	90.91%	16.67%	100.00%
BR8	20	31	100.00%	77.65%	100.00%	88.64%	0.00%	100.00%
BA9	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9*	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10*	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11	1	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11*	17	22	88.35%	71.11%	100.00%	95.83%	13.64%	88.00%
BA12	1	1	61.18%	0.00%	100.00%	90.00%	0.00%	100.00%
BA12*	101	106	97.50%	71.11%	100.00%	95.83%	13.64%	76.00%

* denotes the application of entropy discretization technique.

The best average prediction accuracy (58.93%) is achieved with BR6 with numerical attribute values kept in their original format. The non-performing loans are replicated once in the BR6 training data set. Figure 5.5 shows the tree generated from Data Set BR6. The root node of the decision tree is the attribute “profit margin ratio for the second year before loan application is lodged (X95_pm-2)” and branches out to ten leaf nodes.

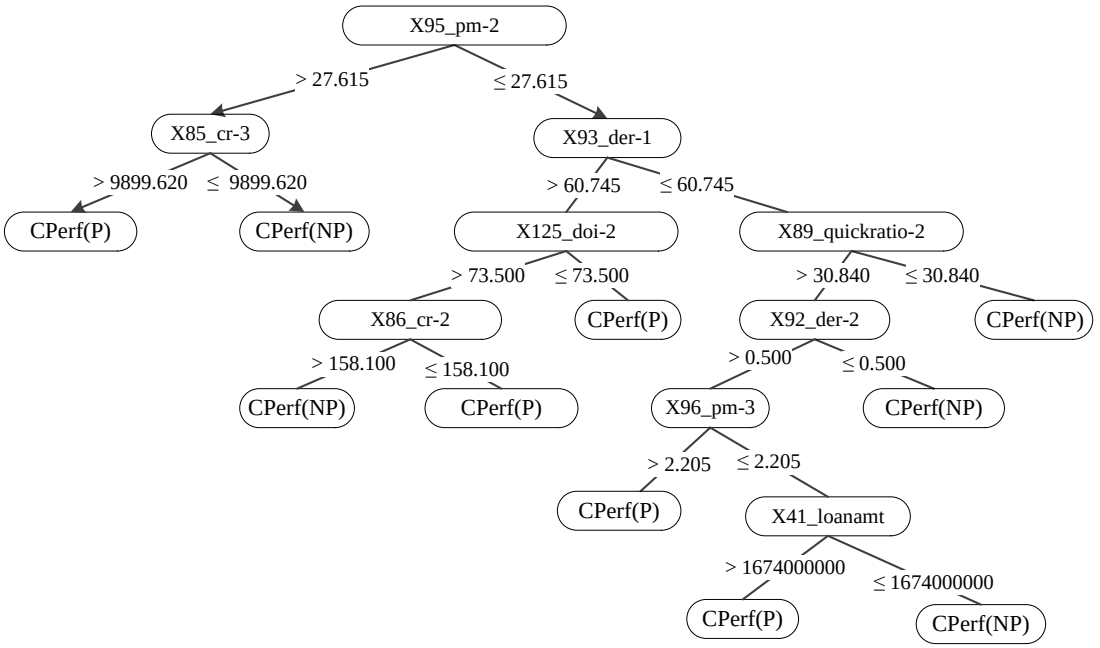


Figure 5.5 Decision tree from Data Set BR6 – Bank B

Key:

X95_pm-2 = Profit Margin Y-2

X85_cr-3 = Current Ratio Y-3

X93_der-1 = Debt to Equity Ratio Y-1

X125_doi-2 = Inventory Turn Over Y-2

X89_quickratio-2 = Quick Ratio Y-2

X86_cr-2 = Current Ratio Y-2

X92_ der-2 = Debt to Equity Ratio Y-2

X96_ pm-3 = Profit Margin Y-3

X41_loanamt = Loan Amount

CPerf(P) = Credit Performance (Performing Loan)

CPerf(NP) = Credit Performance (Non-Performing Loan)

The model built by training data set BR6 has 100% testing data set coverage rate. There are ten extracted rules for performing and non-performing loans, of which eight consist only of financial attributes as credit parameters. An exemplary rule for predicting performing loans is “X95_pm-2=pm-2 (>27.615) AND X85_cr-3=cr-3 (>9899.620)”.)”. This rule covers four credit data sets with 75.00% prediction accuracy. It comprises the profit margin and current ratio. There are five rules for non-performing loans, with distinct coverage rates in the training data set resulting in low coverage rates in the testing data set because of the skewed proportion and limited number of credit data sets. The highest testing data set coverage is 20% for any non-performing loan rules. An exemplary rule for non-performing loans is “X95_pm-2 = pm-2 (≤ 27.615) AND X93_der-1=der-1 (>60.745) AND X125_doi-2=doi-2 (>73.500) AND X86_cr-2=cr-2 (>158.100)”. This rule has a coverage rate of 20% with a 33.33% prediction accuracy rate. It comprises the profit margin ratio, debt to equity ratio, inventory turnover time period, and current ratio. These credit parameters are derived from three-year historical financial information and consist of attributes selected by DT in any of the three years.

The over-fitting problem of performing credit data sets occurs in two data sets, BR8 and BA12*. Over-fitting for non-performing loans prevails in four of the six data sets that

successfully generate decision trees; it is evident that the models generated by the training data sets to classify non-performing loans contain too-specific parameters (attributes and attribute values) and therefore have low prediction ability when applied to the testing data sets.

5.2.3.1.3 Test of core loan and scores attributes

When DT is applied to core loan and score attributes, more performance accuracy is achieved for training and testing data sets than when it is applied to the previous core loan, FI and QI attributes. Table 5.11 shows that the average classification accuracy increases for three data sets where numerical attribute values are kept in their original format (BR6, BR7, BR8); one manually discretized data set (BM4); and three of the automatically discretized data sets (BA10, BA11 and BA12).

Table 5.11 Accuracy performance of DT applied to core loan attributes and scores – Bank B

Data Sets	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1	1	1	90.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2	1	1	82.93%	33.33%	100.00%	90.00%	0.00%	100.00%
BM3	1	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4	76	81	88.54%	74.32%	100.00%	91.49%	33.33%	84.00%
BR5	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6	12	22	95.00%	65.38%	100.00%	91.30%	25.00%	100.00%
BR7	11	20	100.00%	73.33%	100.00%	90.48%	12.50%	100.00%
BR8	8	14	100.00%	80.49%	100.00%	89.74%	9.09%	100.00%
BA9	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9*	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10	**10**	**18**	**90.48%**	**57.14%**	**100.00%**	**97.44%**	**36.36%**	**100.00%**
BA10*	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11	3	4	76.34%	76.47%	100.00%	89.80%	0.00%	100.00%
BA11*	1	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BA12	3	4	71.77%	67.39%	100.00%	89.80%	0.00%	100.00%
BA12*	8	12	86.02%	68.83%	98.24%	90.00%	0.00%	94.00%

* denotes the application of entropy discretization technique.

The DT algorithm produces eight default trees. Of the eight data sets that have multiple leaf nodes, five data sets show prediction ability with regard to non-performing loans in the testing data sets. One data set has manually discretized numerical attribute values (BM4), three have numerical attribute values that are kept in original format (BR6, BR7 and BR8) and one has its numerical attribute values discretized using equal-width binning (BA10).

BA10, with numerical attribute values discretized by applying equal-width binning, provides the best average prediction accuracy (66.90%), identical to that achieved with DT applied to all attributes (Table 5.9). Non-performing loans are replicated once in the BA10 training data set. Figure 5.6 shows the tree generated from Data Set BA10. The root node of the decision tree is the assigned score for the attribute "applicant's account management (X14_accmngmtsc)" and branches out to ten leaf nodes. This decision tree is very similar to the tree generated when DT is applied to all credit data (Figure 5.4).

The model built by training data set BA10 has 100% testing data set coverage rate. There are ten extracted rules for performing and non-performing loans. The exemplary rule for performing loans that provides the highest prediction accuracy is "X144_acctmngmtsc= acctmngmtsc (≤0.500)". This rule covers 17 credit data sets with 94.12% prediction accuracy. It constitutes the applicant's good standing with the bank, consisting of timely repayments of previous loans, submission of all required application documents, and full compliance with previous credit contracts.

Another rule for performing loans is "X144_acctmngmtsc= acctmngmtsc (>0.500) AND X158_ttlscorenf = ttlscorenf (>7.875) AND X159_ttlscore = ttlscore (>24.250) AND X15_amt3 = amt3 (range1 [-∞ - 1507306388.800]) AND X157_ttlmngmnt = ttlmngmnt (>3.750) AND X24_loantype5 = loantype5 (NR)". This rule consists of credit parameters: applicant's somewhat poor standing with the indicative of failure to make timely repayments of previous loans, and/or submission of all required application documents and/or full compliance with previous credit contracts; unstable non-financial standing; high overall credit risk; outstanding financial obligations with a third and fifth bank the applicant has conducted business with; and substandard commercial and technical (entrepreneurship) qualifications. These are indicative of non-performing loans characteristics. Despite the apparent disqualifying credit risk assessment parameters, this rule covers four credit data sets with 75% accuracy in the testing data set. The rule covers seven out of 45 cases of performing loans in the training data set where no replication is applied. This implies its prominence and therefore can be used to refine the scoring system and its application by loan staff. The contradictory nature of these credit parameters and the predicted class may have been a result of the limited number of performing and non-performing loan data sets obtained from Bank B. Nevertheless, the extracted rule is correct based on the provided credit data sets.

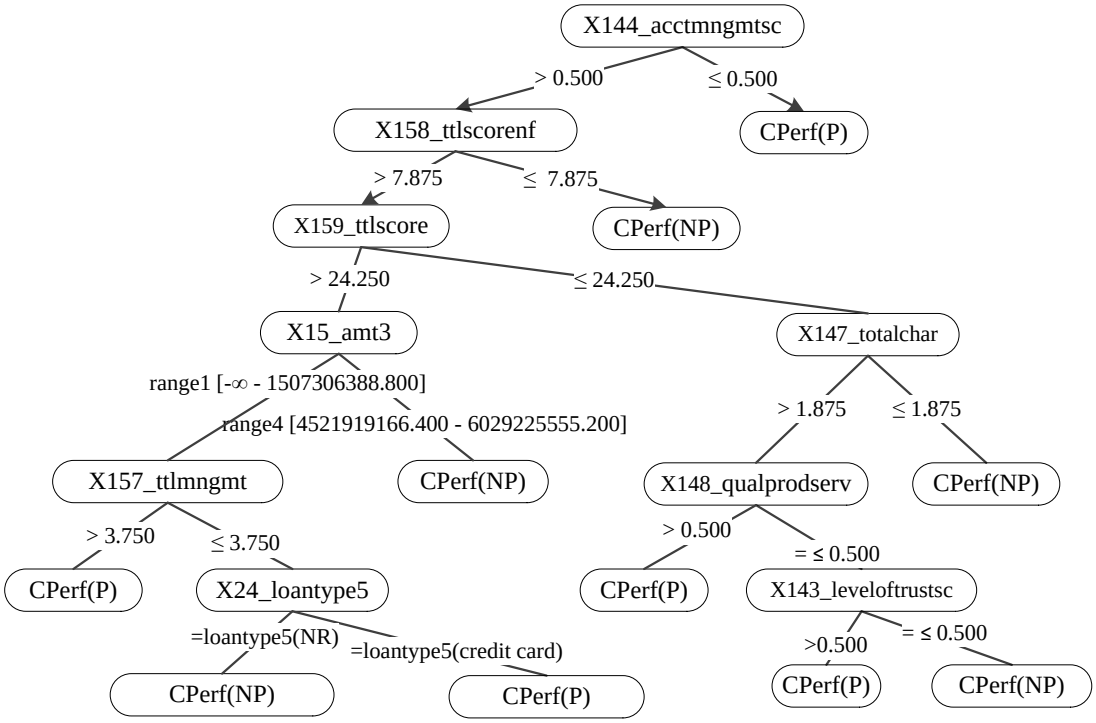


Figure 5.6 Decision tree from Data Set BA10 – Bank B

Key:

X144_accmngmtsc = Applicant' Account Management (Score)

X158_ttlscorenf = Total Non-Financial Score

X159_ttlscore = Total Score

X15_amt3 = Amount 3

X157_ttlmngmtsc= Total Management (Score)

X147_ttlchar = Total Character

X148_qualprodserv = Quality of Product/Service

X24_loantype5 = Loan Type 5

CPerf(P) = Credit Performance (Performing Loan)

CPerf(NP) = Credit Performance (Non-Performing Loan)

There are five rules for non-performing loans. These rules have distinct coverage rates in the training data set resulting in low coverage rates in the testing data set due to the skewed proportion and limited number of credit data sets. The highest testing data set coverage is 40% for any non-performing loan rules. An exemplary rule for non-performing loans is "X144_acctmngmtsc = acctmngmtsc (>0.500) AND X158_ttlscorenf = ttlscorenf (>7.875) AND X159_ttlscore = ttlscore (>24.250) AND X15_amt3 = amt3 (range1 [-∞ - 1507306388.800]) AND X157_ttlmngmnt = ttlmngmnt (≤3.750) AND X24_loantype5 = loantype5 (NR)". This rule has a coverage rate of 40% with 66.67% prediction accuracy rate. The credit parameters of this rule are: applicant's somewhat poor standing with the bank indicative of failure to make timely repayments of previous loans, and/or submission of all required application documents and/or full compliance with previous credit contracts; unstable non-financial standing; high overall credit risk; outstanding financial obligations with a third and fifth bank the applicant has conducted business with; and commercial (bookkeeping and financial management) and technical (entrepreneurship) qualifications that

the applicant possess. This rule has the same contradictive nature as the previous rule for performing loans. The credit parameters for this rule are conflicting since three are in traditional sense indicative of non-performing credit risk attributes while the other three are traditional performing/qualifying credit risk attributes. The fact that the conflicting credit parameters are combined in a single rule should be considered from the perspective of the interaction between the attributes forming part of the rule.

The over-fitting problem occurs in more data sets where DT is applied to core loan attributes and scores (Table 5.11) in comparison with the application of it to core loan, FI and QI attributes (Table 5.10). Over-fitting of non-performing loans prevails in all of the eight data sets from which decision trees are successfully generated. Over-fitting of performing loans occurs in three (BR6, BR7 and BR8) of these eight data sets. The testing Data Set BA10 that provides the best average prediction performance accuracy shows over-fitting with regard to non-performing loans but not to performing loans. The non-performing credit data sets that have been replicated once in the training BA10 data set may cause this result.

5.2.3.1.4 Comparison of DT test results

Table 5.12 shows the comparative prediction accuracy achieved with DT for the different combinations of Bank B data set attributes. Overall, the best average prediction accuracy is achieved with BA10 (66.90%). The prediction accuracy of BA10 is the same where all credit attribute are used as well as the core loan attribute and score combination. From Table 5.12 it is evident that the scores provide better prediction accuracy when used in combination with FI and QI compared to situations where FI and QI are used exclusively.

Table 5.12 A comparison of prediction performance accuracy with DT for different combinations of credit risk attributes – Bank B

Data Sets	All Credit Attributes		Core Loan, FI and QI Attributes		Core Loan Attributes and Scores	
	Performing Loans	**Non-Performing Loans**	**Performing Loans**	**Non-Performing Loans**	**Performing Loans**	**Non-Performing Loans**
BM1	Default Tree		Default Tree		Default Tree	
BM2	Default Tree		Default Tree		Default Tree	
BM3	Default Tree		Default Tree		Default Tree	
BM4	Default Tree		Default Tree		91.49%	33.33%
BR5	Default Tree		89.80%	0.00%	Default Tree	
BR6	93.02%	28.57%	**92.86%**	**25.00%**	91.30%	25.00%
BR7	91.30%	28.00%	90.91%	16.67%	90.48%	12.50%
BR8	91.30%	28.00%	88.64%	0.00%	89.74%	9.09%
BA9	Default Tree		Default Tree		Default Tree	
BA9*	Default Tree		Default Tree		Default Tree	
BA10	**97.44%**	**36.36%**	Default Tree		**97.44%**	**36.36%**
BA10*	Default Tree		Default Tree		Default Tree	
BA11	89.80%	0.00%	Default Tree		89.80%	0.00%
BA11*	90.00%	0.00%	95.83%	13.64%	Default Tree	
BA12	89.80%	0.00%	Default Tree		89.80%	0.00%
BA12*	90.00%	0.00%	95.83%	13.64%	90.00%	0.00%

* denotes the application of entropy discretization technique.

5.2.3.2 Rule induction

RI, similar to DT, is applied to different combinations of attributes:

Combination1: core loan attributes, QI, FI and compiled scores (all credit attributes);

Combination 2: core loan, FI and QI attributes; and

Combination 3: core loan attributes and scores.

Thus it is evident that the credit data sets contain “raw” and “interpreted/analyzed” information.

5.2.3.2.1 Test of all credit attributes

Table 5.13 shows the performance of the RI application with all data sets. In contrast to the DT results, the performance accuracy of RI is more affected by the skewed proportions of performing and non-performing loans than by the discretization technique applied. The prediction accuracy of 90% for performing loans and 0% for non-performing loans in Data Sets BM1, BM2, BM3, BR5, BR6, BR7, BA9, BA9*, BA10, BA10*, BA11 and BA11* are evidence of this skewness. The RI algorithm attains the best performance accuracy where

non-performing loans have been replicated five times (BM4, BR8, BA12 and BA12*). Only a single rule "CPerf(P)" is extracted for all other data sets, with 90% prediction accuracy for performing loans and 0% for non-performing loans; these do not provide valid attainable accuracy with the data at hand.

The non-performing loans in the BM1, BR5, BA9 and BA9* training data sets are not replicated and constitute 11 of the 115 loans in the training data sets. The non-performing loans in BM2 and BA10* are replicated one time and constitute 22 of the 126 loans in the training data sets. The non-performing loans in BM3 training datasets are replicated three times and represent 44 of the 148 total loans in the training data sets. The non-performing loans make up five of the 50 loans in the testing data sets. The over-fitting problem for non-performing loans appears in all data sets that generate classification results (BM4, BR8, BA12 and BA12*), while over-fitting for performing loans occurs in only one (BR8).

Table 5.13 Performance accuracy of RI applied to all credit attributes– Bank B

Data Sets	**# of Rules**	**Classification Accuracy with Training Credit Data Sets**		**Coverage Rate**	**Prediction Accuracy with Testing Credit Data Sets**		**Coverage Rate**
		Performing Loans	**Non-Performing Loans**		**Performing Loans**	**Non-Performing Loans**	
BM1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2	1	82.40%	0.00%	100.00%	90.00%	0.00%	100.00%
BM3	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4	6	84.54%	69.86%	99.41%	90.24%	11.11%	100.00%
BR5	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BR7	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BR8	6	94.19%	72.62%	99.41%	90.70%	14.29%	86.00%
BA9	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9*	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10*	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11*	1	69.44%	0.00%	100.00%	90.00%	0.00%	100.00%
BA12	2	91.30%	74.36%	93.53%	95.45%	50.00%	58.00%
BA12*	**6**	**87.00%**	**75.71%**	**99.41%**	**95.45%**	**50.00%**	**96.00%**

* denotes the application of entropy discretization technique.

Data Sets BA12 and BA12* both provide best average prediction accuracy (72.73%), although the average classification accuracy of BA12* is somewhat lower (81.36%) than that of BA12 (82.83%). BA12* has significantly higher training and testing data set coverage rates (99.41% of training and 96% of testing data set) compared to those of BA12 (93.53% of training and 58% of testing data set); therefore BA12* is selected as the best data set. Non-performing loans are replicated five times in the training data set and automatic discretization with the entropy-based technique is applied to BA12*.

There are six extracted rules for performing and non-performing loans from BA12*. An exemplary rule with the best prediction accuracy for performing loans is "if X158_ttlscore = ttlscore (range2 [10.250 - 24]) AND X43_intrate = intrate (range1 [-∞ - 14])". This covers 20 credit data sets with 100% prediction accuracy. The credit parameters contained in it are low overall credit risk combined with applied interest rates of below 14% per annum on existing loans. There are two rules for non-performing loans. These rules have distinct coverage rates in the training data set, resulting in low coverage rates in the testing data set because of the skewed proportion and limited number of credit data sets. The highest testing data set coverage is 20% for any non-performing loan rule. One rule for non-performing loans is "if X130_wcto-3 = wcto-3 (range1 [-∞ - 212.960])". This has a coverage rate of 20%, with a low 5.26% prediction accuracy rate that can be ascribed to the over-fitting problem that results in too-specific parameters. This rule consists of one credit parameter: a working capital turnover ratio of less than 212.96.

The aforementioned RI rules that predict performing and non-performing loans are correct, but the reality of parameters like "interest rate of less than 14% per annum" as indicators of performing loans and "working capital turnover ratio of less than 212.96" show that a bias result exists because of the limited number of performing and non-performing loan data sets obtained from Bank B. Even in a context where no over-fitting is evident for performing loans in any of the data sets, the parameters "instability related to future planning by local authority" and "a lack of information about non-primary business activities of applicant" are inconsistent with real-world interpretation of performing loan characteristics/attributes; this emphasizes that RI can be used well in finding rules, but that it depends on the validity of the data that it is applied to.

5.2.3.2.2 Test of core loan, FI and QI attributes

Table 5.14 shows the performance accuracy of RI applied to core loan, FI and QI credit attributes with a processing time of maximum one second. As in the findings where RI is applied to all credit attributes, it distinguishes the performing and non-performing loans in those data sets where non-performing loans are replicated five times (BM4, BR8, BA12 and BA12*) and in one data set where non-performing loans are replicated three times (BR7). A

single rule "CPerf(P)" is extracted for all other data sets and does not provide valid attainable accuracy with the data at hand.

The over-fitting problem occurs in three data sets (BM4, BR8 and BA12) for both performing and non-performing credit data sets of the five data sets from which prediction results for both performing and non-performing loans are successfully generated. BM4 is manually discretized and BA12 is automatically discretized using the equal-width binning technique. BA12*, which is entropy-based discretized, and BR7, with numerical attribute values kept in original format, show the over-fitting problem of non-performing credit data sets.

Table 5.14 Performance accuracy of RI applied to core loan, FI and QI attributes – Bank B

Data Sets	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2	1	82.40%	0.00%	100.00%	90.00%	0.00%	100.00%
BM3	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4	7	94.32%	74.39%	98.82%	86.36%	7.14%	84.00%
BR5	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BR7	**1**	**71.76%**	**41.18%**	**100.00%**	**93.94%**	**17.65%**	**66.00%**
BR8	5	94.68%	80.26%	99.41%	90.70%	14.29%	98.00%
BA9	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9*	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10*	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11*	1	69.44%	0.00%	100.00%	90.00%	0.00%	100.00%
BA12	9	92.22%	73.75%	99.41%	90.91%	10.71%	100.00%
BA12*	1	67.74%	56.52%	99.41%	92.86%	13.64%	38.00%

* denotes the application of entropy discretization technique.

The best average prediction accuracy is obtained from BR7 (55.80%). The model built by training data set BR7 has a moderate testing data set coverage rate of 66%. There is only one extracted rule and it classifies performing loans. The rule is "if X96_ pm-1 = pm-1 (<=

24.14) AND X129_ nwcto-1= nwcto-1 (>96.365) AND X15_ amt3 = amt3 (<= 3576112489) AND X130_ wcto-3 = wcto-3 (> 101.43)". This rule covers 33 credit data sets with 93.94% prediction accuracy. The credit parameters for this rule are a profit margin ratio less than or equal to 24.14, a net working capital turnover ratio of more 96.365, outstanding financial obligations with a third bank the applicant has conducted business with; and a working capital turnover ratio of more than 101.43.

5.2.3.2.3 Test of core loan and scores attributes

When RI is applied to core loan and score attributes, prediction results are generated for both performing and non-performing loans in three data sets, compared with four sets when all attributes (Table 5.13) and five when core loan, FI and QI attributes (Table 5.14) are considered. Table 5.15 shows these results are from the data sets that is manually discretized (BM4); one data set with numerical attribute values kept in their original format (BR8); and one automatically discretized data set (BA12*). RI distinguishes the performing and non-performing loans in data sets where non-performing loans are replicated five times, similar to the findings where RI is applied to all credit attributes. The exception is BA12 where it does not produce any prediction accuracy for non-performing loans, notwithstanding the extracted non-performing loan rules from the training data set. A single rule "CPerf(P)" is extracted for all other data sets and the results do not provide valid attainable accuracy with the data at hand.

The over-fitting problem of non-performing credit data sets occurs in all three sets (BM4, BR8 and BA12*) for which classification and prediction results are successfully generated. These data sets experience the same over-fitting problem when core loan, FI and QI attributes are used (Table 5.14). The over-fitting problem related to performing credit data sets does not occur in any of these. The best average prediction accuracy of 65.35% is obtained from BR8. Non-performing loans are replicated five times in the BR8 training data set, with numerical attributes values kept in their original format.

Table 5.15 Performance accuracy of RI applied to core loan attributes and scores – Bank B

Data Sets	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2	1	82.40%	0.00%	100.00%	90.00%	0.00%	100.00%
BM3	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4	9	88.68%	84.38%	99.41%	93.33%	15.00%	98.00%
BR5	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BR7	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BR8	**5**	**95.40%**	**74.70%**	**99.41%**	**97.37%**	**33.33%**	**92.00%**
BA9	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9*	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10*	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11*	1	69.44%	0.00%	100.00%	90.00%	0.00%	100.00%
BA12	7	84.68%	83.05%	99.41%	88.64%	0.00%	90.00%
BA12*	7	86.11%	82.26%	99.41%	91.30%	25.00%	100.00%

* denotes the application of entropy discretization technique.

The model built by training data set BR8 has a 92% testing data set coverage rate. There are five extracted rules for performing and non-performing loans. An exemplary rule that provides substantial prediction success for the performing loans is: "if X149_stratmrktsc = stratmrktsc (=≤ 1.500)". This rule covers 40 credit data sets with 100% prediction accuracy. It consists of one credit parameter: the applicant's exceptional marketing strategy.

Two rules for non-performing loans are extracted from the BR8. These are very distinct in both training and testing data sets, but are impractical for real-world interpretation of non-performing loan characteristics/attributes and are excluded from this discussion. Over-fitting applies to them, and they have low prediction ability when applied to the testing data set that contains five non-performing credit data sets, since best rule predicts only one of the five cases correctly.

5.2.3.2.4 Comparison of RI test results

Table 5.16 shows the comparative prediction accuracy achieved with RI for the different combinations of Bank B data set attributes in the sections above. Overall, the best

average prediction accuracy is achieved with BA12 (72.73%) where all credit attributes are used.

Table 5.16 A Comparison of the prediction accuracy for different combinations of credit risk attributes when RI is applied – Bank B

Data Sets	All Credit Attributes		Core Loan, FI and QI Attributes		Core Loan Attributes and Scores	
	Performing Loans	**Non-Performing Loans**	**Performing Loans**	**Non-Performing Loans**	**Performing Loans**	**Non-Performing Loans**
BM1	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BM2	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BM3	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BM4	90.24%	11.11%	86.36%	7.14%	93.33%	15.00%
BR5	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BR6	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BR7	Single Conjunctive Rule		**93.94%**	**17.65%**	Single Conjunctive Rule	
BR8	90.70%	14.29%	90.70%	14.29%	**97.37%**	**33.33%**
BA9	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BA9*	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BA10	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BA10*	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BA11	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BA11*	Single Conjunctive Rule		Single Conjunctive Rule		Single Conjunctive Rule	
BA12	95.45%	50.00%	90.91%	10.71%	88.64%	0.00%
BA12*	**95.45%**	**50.00%**	92.86%	13.64%	91.30%	25.00%

* denotes the application of entropy discretization technique.

5.2.3.3 Forward feature selection

The Forward Feature Selection technique is applied to the 48 FI and 33 QI attributes. It is not applied to the core loan attributes, since all of them are regarded as important. On the other hand, scores are excluded from this process since they already represent "interpreted/analyzed" information based on FI and QI credit data.

5.2.3.3.1 Test of DT on selected attributes

A total of 40 attributes, comprising 20 FI and 20 QI attributes, are selected by this technique. The six attributes that are the most prominent are contained in Figure 5.7. Each is selected recurrently by the technique in at least three of the data sets. The most frequently selected attribute is "going concern", extracted from nine of the 15 data sets. It is followed by "level of trust" (six data sets), "suppliers" and "target market" (each from four data sets)

and “working capital Y-1” and “commercial qualification” (each from three data sets). Five of the six attributes form part of the QI data and serve as indicator of the importance of QI as a credit risk assessment criterion, based on the data sets provided by Bank B. The rest of the selected attributes (34 of 40) from both FI and QI and are selected between one and two times. These are relevant attributes since they have been selected by the Forward Feature Selection Technique, but are not as important as the aforementioned more frequently selected attributes.

Data Sets	Going Concern	Level of Trust	Suppliers	Target Market	Working Capital Turnover Y-1	Commercial Qualification
BM1		X				
BM2	X	X				
BM3	X		X			
BR5		X				
BR6	X	X				
BR7	X					
BR8	X			X		
BA9*		X				
BA10	X					
BA10*	X	X	X			
BA11	X		X		X	X
BA11*	X			X	X	X
BA12				X	X	X
BA12*			X	X		
Total	9	6	4	4	3	3

Figure 5.7 Forward feature selection techniques attributes (DT) – Bank B

Table 5.17 shows the performance accuracy (with processing time indicated in brackets) of the selected attributes with DT applied to the different data sets. The DT algorithm produces prediction results for non-performing loans in five data sets: BM2, BM3, BM4, BR7, and BA11. In BM2, BM3 and BM4 the numerical attribute values are manually discretized; in BR7 the numerical attribute values are kept in their original format; and in BA11 the attribute values are automatically discretized with equal-width binning. The DT algorithm produces five default trees with training and testing data set coverage rates of 100%.

The best average prediction accuracy (66.67%) with the selected attributes is achieved with BM2, for which the numerical attribute values are manually discretized and non-performing loans are replicated once in the training data set. The model built by training data

set BM2 has a 90% testing data set coverage rate. There are 18 extracted rules for performing and non-performing loans. An exemplary rule from BM2 that provides the highest prediction accuracy for performing loans is "X57_gconcern = gconcern (NA) AND X94_pm-3 = pm-3 ([0.01-10])". This rule covers ten credit data sets with 90% prediction accuracy. The credit parameters for predicting performing loans consist of a condition where no information related to the sustainability of the business is obtained from the applicant, and a profit margin ratio between 0.01 and 10. The parameters in the abovementioned DT rule contradict real-world interpretation of performing loan characteristics/attributes. The rule is correct in the sense that it emanates from the data sets provided by Bank B, and in the context that no over-fitting is evident for performing loans in this data set. However, it shows that the limited number of non-performing loan data sets compared with the number of performing loan data sets obtained from Bank B biases the results; as is also indicated by previous findings.

Table 5.17 Performance accuracy of applying selected attributes (DT) – Bank B

Data Sets (in seconds)	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1 (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2 (1)	**18**	**22**	**85.96%**	**50.00%**	**100.00%**	**93.33%**	**40.00%**	**90.00%**
BM3 (1)	32	40	97.98%	85.71%	100.00%	91.49%	33.33%	98.00%
BM4 (1)	23	31	100.00%	85.71%	100.00%	90.24%	11.11%	96.00%
BR5 (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6 (1)	18	28	98.98%	75.00%	100.00%	89.58%	0.00%	100.00%
BR7 (1)	16	24	100.00%	86.27%	100.00%	93.02%	28.57%	100.00%
BR8 (1)	20	32	100.00%	92.96%	100.00%	89.58%	0.00%	96.00%
BA9 (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9* (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10 (1)	9	10	85.12%	80.00%	100.00%	89.80%	0.00%	100.00%
BA10* (1)	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11 (1)	24	34	94.29%	88.37%	100.00%	91.49%	33.33%	96.00%
BA11* (1)	16	21	85.58%	100.00%	100.00%	100.00%	0.00%	96.00%
BA12 (1)	22	32	93.94%	84.51%	100.00%	88.10%	0.00%	100.00%
BA12* (1)	16	22	100.00%	95.65%	100.00%	90.00%	0.00%	100.00%

* denotes the application of entropy discretization technique.

There are four rules for non-performing loans. These rules have distinct coverage rates in the training data set resulting in low coverage rates in the testing data set because of the skewed proportion and limited number of credit data sets. As such they are impractical for real-world interpretation of non-performing loan characteristics/attributes and are excluded from this discussion. Over-fitting prevails to non-performing loans for data set BM2, and it provides (as does the best rule for performing loans) a contradictory best rule for the prediction of non-performing loans.

Table 5.18 shows the performance accuracy of the data sets when selected attributes identified with the Forward Feature Selection Technique are combined with the core loan attributes. The DT algorithm produces seven default trees, of which the results are not valid representatives of attainable accuracy for the data at hand. The over-fitting of performing loans occurs in four (BR6, BR7, BR8 and BA12*) of the nine data sets that produce results, while the over-fitting of non-performing loans happens in all nine data sets.

Table 5.18 Performance accuracy of forward feature selection technique attributes combined with core loan attributes and scores (DT) – Bank B

Data Sets (in seconds)	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1 (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2 (1)	1	1	82.93%	33.33%	100.00%	90.00%	0.00%	100.00%
BM3 (1)	1	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4 (1)	76	81	87.37%	72.00%	100.00%	91.49%	33.33%	78.00%
BR5 (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6 (1)	13	24	92.59%	77.78%	100.00%	90.70%	14.29%	100.00%
BR7 (1)	12	22	100.00%	78.57%	100.00%	88.10%	0.00%	100.00%
BR8 (1)	8	14	100.00%	80.49%	100.00%	89.74%	9.09%	100.00%
BA9 (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9* (1)	1	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10 (1)	**11**	**20**	**90.29%**	**52.17%**	**100.00%**	**97.44%**	**36.36%**	**98.00%**
BA10* (1)	1	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11 (1)	3	4	76.34%	76.47%	100.00%	89.80%	0.00%	100.00%
BA11* (1)	4	6	97.12%	25.00%	100.00%	100.00%	0.00%	100.00%
BA12 (1)	3	4	71.77%	67.39%	100.00%	89.80%	0.00%	100.00%
BA12* (1)	11	18	93.33%	75.00%	100.00%	90.00%	0.00%	98.00%

* denotes the application of entropy discretization technique.

DT extracts patterns consisting of one attribute from the application of the Forward Feature Selection Technique, six score attributes and two core loan attributes. Data set BA10 provides the best average prediction accuracy (66.90%). This is a data set where numerical attribute values are automatically discretized using equal-width binning and non-performing loans are replicated once in the training data set. The decision tree generated from BA10 is similar to that of the same data set (BA10) that is tested for all attributes (Figure 5.4) and for core loan attributes and scores (Figure 5.6). The model built by training data set BA10 has a 98% testing data set coverage rate. There are 11 extracted rules for performing and non-performing loans. Two exemplary rules are selected to demonstrate their potential for credit assessment practice improvement. An exemplary rule for performing loans is "X144_acctmngmtsc = acctmngmtsc (≤ 0.500)". This rule covers 17 credit data sets with 94.12% prediction accuracy. The single credit parameter for the prediction is a good standing of the applicant with the bank, consisting of timely repayment of previous loans, submission of all required application documents, and full compliance with previous credit contracts.

As in previous discussions in this chapter (Section 5.2.3.1.3), a rule for performing loans with conflicting credit parameters occurs here: "X144_acctmngmtsc = acctmngmtsc (>0.500) AND X158_ttlscorenf = ttlscorenf (>7.875) AND X159_ttlscore = ttlscore (≤24.250) AND X147_ttlchar = ttlchar (>1.875) AND X148_qualprodserv = qualprodserv (>0.500)". This rule consists of credit parameters: the applicant's somewhat poor standing with the bank, indicative of failure to make timely repayments of previous loans and/or submission of all required application documents and/or full compliance with previous credit contracts; unstable non-financial standing; high overall credit risk; unstable personal character; and uncompetitive goods/service provision. The last four credit parameters are indicative of traditional non-performing loans characteristics. Despite the apparent disqualifying credit risk assessment parameters, this rule covers 19 credit data sets, with 100% accuracy in the testing data set. Furthermore, it covers 49 out of 104 cases of performing loans in the training data set where no replication is applied. This implies its prominence, and it therefore can be used to refine the scoring system and its application by loan staff.

There are five rules for non-performing loans. These are very distinct in both training and testing data sets. The highest testing data set coverage rate is 20% for any non-performing loan rule. An exemplary rule for non-performing loans that is potentially useful for improvement of credit assessment practice, and that provides the highest prediction accuracy, is "X144_acctmngmtsc = acctmngmtsc (>0.500) AND X158_ttlscorenf = ttlscorenf (>7.875) AND X159_ttlscore = ttlscore (>24.250) AND X15_amt3 = amt3 (range1 [-∞ - 1507306388.800]) AND X123_dop-1= dop-1 (range1 [-∞ - 19999.800]) AND X149_stratmrktsc = stratmrktsc (≤ 1.500) AND X24_loantype5= loantype5 (NR)". This rule

has a testing data set coverage rate of 20% with 50% prediction accuracy. It includes more credit parameters than the previous rule for performing loans. The parameters are: poor standing of the applicant with the bank, indicative of failure to make timely repayment of previous loans and/or submission of all required application documents and/or full compliance with previous credit contracts; unstable non-financial standing; high overall credit risk; existing outstanding financial obligations with a third bank the applicant has conducted business with; no information about the accounts payable turnover time; an exceptional marketing strategy; and no outstanding financial obligations to a fifth bank the applicant has conducted business with. The financial obligations to the third and fifth bank include acquired business finance (such as working capital) and personal finance (such as home loans, credit cards, and other consumer loans). An exceptional marketing strategy and no outstanding financial obligations to a fifth bank contradict real-world associations with nonperforming loan characteristics/attributes. The fact that the conflicting credit parameters are combined in a single rule should be considered from the perspective of the interaction between all the attributes forming part of the rule.

From a broad perspective, the combination of the Forward Feature Selection Technique attributes together with the core loan attributes and scores contributes to the significance of the rules, although the rules still show definite bias. The skewed proportion of the performance and non-performance loans in the training and testing data sets might have an effect on these results.

5.2.3.3.2 Test of RI on selected attributes

When the RI algorithm is applied, the Forward Feature Selection Technique identifies 31 attributes consisting of 19 FI and 12 QI. The four attributes that are the most prominent are contained in Figure 5.8. Each of these attributes is selected recurrently by the technique in at least three of the data sets. The most frequently selected attribute is “current ratio Y-3”, extracted from five of the 15 sets. This followed by “level of trust” (four sets), and “profit margin Y-3” as well as “profit margin Y-2” (each from three sets). Three of the four attributes are from the FI data, reflecting the importance of FI as credit risk assessment criterion based on the data sets provided by Bank B. The rest of the selected FI and QI attributes (27 of 31) are selected between one and two times. These are relevant attributes but not as important as the more frequently selected attributes.

Data Sets	Current Ratio Y-3	Profit Margin Y-3	Profit Margin Y-2	Level of Trust
BM1				X
BM2		X		
BM3		X		
BR5				X
BR6	X	X		
BA9	X			
BA9*				X
BA10	X			
BA10*				X
BA11			X	
BA11*			X	
BA12	X			
BA12*	X		X	
Total	5	3	3	4

Figure 5.8 Forward feature selection techniques attributes (RI) – Bank B

The performance accuracy of all data sets is provided in Tables 5.19 and 5.20, with processing time indicated in brackets next to each. Only six data sets (BM2, BR7, BR8, BA11*, BA12 and BA12*) produce non-performing loan classifications with the training data, while two data sets (BA11 and BA12) predict non-performing loans with the testing data. A single rule "CPerf(P)" is extracted from 11 data sets (BM1, BM2, BM3, BM4, BR5, BR6, BA9, BA9*, BA10, BA10*, BA11). The over-fitting problem for non-performing loans occurs in all data sets that produce results, while for performing loans it occurs in three data sets (BR7, BR8 and BA12*).

The best average prediction accuracy is achieved by Data Set BA11 (52.38%). Non-performing loans are replicated three times in the training data set of BA11, and consist of numerical attribute values that are kept in their original format. The model built by training data set BA11 has a 90% testing data set coverage rate. There is only one extracted rule and it is for performing loans: "if X95_pm-2 = pm-2 = ([-∞ - 27.430])". The rule has one credit parameter, a profit margin ratio of below 27.43.

Table 5.19 Performance accuracy of forward feature selection technique attributes (RI) – Bank B

Data Sets (in seconds)	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1 (2)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2 (1)	1	85.00%	66.67%	100.00%	90.00%	0.00%	100.00%
BM3 (2)	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4 (1)	1	61.18%	0.00%	100.00%	90.00%	0.00%	100.00%
BR5 (1)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6 (1)	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BR7 (1)	5	91.43%	81.40%	99.32%	90.00%	0.00%	100.00%
BR8 (1)	7	97.94%	87.67%	99.41%	89.13%	0.00%	100.00%
BA9 (2)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9* (1)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10 (1)	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10* (1)	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11 (1)	1	70.27%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11* (1)	**1**	**71.94%**	**55.56%**	**100.00%**	**93.33%**	**11.43%**	**90.00%**
BA12 (1)	64	70.55%	95.83%	97.65%	82.61%	3.70%	42.00%
BA12* (1)	6	94.29%	92.31%	99.41%	90.00%	0.00%	94.00%

* denotes the application of entropy discretization technique.

Table 5.20 shows the performance accuracy of the data sets when selected attributes that were identified with the Forward Feature Selection Technique are combined with the core loan attributes. The same number of data sets produces a single rule: “CPerf(P)”, as the test was conducted with selected attributes only (BM1, BM2, BM3, BR5, BR6, BR7, BA9, BA9*, BA10, BA10* and BA11). Over-fitting for performing loans occurs in only one of the five data sets that produced results, BR8, while over-fitting for non-performing loans occurs in four data sets.

Data Set BM4 provides the best average prediction accuracy (80.15%). This is a data set where numerical attribute values are manually discretized and non-performing loans are replicated three times in the training data set. The selected attributes used in Data Set BM4 are “target market (in detail)”, “economy, social and political conditions”, “net working capital turnover Y-3” and “EBITDA per future interest Y-2”. The extracted rules from BA12 comprise five score attributes, four core loan attributes, and only one Forward Feature Selection attribute.

Table 5.20 Performance accuracy of applying selected attributes combined with core loan attributes and scores (RI) – Bank B

Data Sets (in seconds)	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
BM1 (1)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BM2 (1)	1	82.93%	33.33%	100.00%	90.00%	0.00%	100.00%
BM3 (1)	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BM4 (1)	**9**	**62.03%**	**50.00%**	**99.41%**	**93.62%**	**66.67%**	**100.00%**
BR5 (0)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BR6 (0)	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BR7 (0)	1	70.59%	33.33%	100.00%	90.00%	0.00%	100.00%
BR8 (2)	5	98.91%	83.33%	99.41%	89.74%	9.09%	100.00%
BA9 (1)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA9* (1)	1	90.43%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10 (1)	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA10* (1)	1	82.54%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11 (1)	1	69.86%	0.00%	100.00%	90.00%	0.00%	100.00%
BA11* (1)	11	82.24%	60.98%	99.32%	90.00%	0.00%	96.00%
BA12 (1)	3	84.68%	83.05%	99.41%	97.22%	28.57%	76.00%
BA12* (1)	7	86.61%	87.93%	99.41%	90.00%	0.00%	96.00%

* denotes the application of entropy discretization technique.

The model built by training data set BM4 has a 100% testing data set coverage rate. There are nine extracted rules for performing and non-performing loans. An exemplary rule that provides the highest prediction accuracy is "if X144_acctmngmtsc = acctmngmtsc (0)". This rule covers 36 credit data sets with 100% prediction accuracy. The single credit parameter for the prediction is the good standing of the applicant with the bank, consisting of timely repayments of previous loans, submission of all required application documents and full compliance with previous credit contracts.

Although the over-fitting problem with regard to non-performing loans does not occur in BM4, the prediction accuracy of the extracted rules is low, and they contradict real-world interpretation of performing loan characteristics/attributes. As with the results from applying DT, the skewed proportion of the performance and non-performance loans in the training and testing data sets might have an effect on these results. The extracted rules for non-performing loans are impractical for real-world scenarios and are excluded from this discussion.

When the RI algorithm is applied to data sets with only selected attributes (Table 5.19) and to selected attributes combined with core loan attributes and scores (Table 5.20), fewer

data sets provide results compared with when RI is applied to all credit attributes (Table 5.13). When smaller numbers of attributes are used, the over-fitting problem occurs less frequently (Tables 5.19 and 5.20) for performing credit data sets than when all attributes are used (Table 5.13).

5.2.3.4 Summary of results – Bank B

Table 5.21 shows the data sets with best performance accuracy with the application of DT and RI. When all credit data attributes from Bank B are used, the application of RI provides the best average prediction performance with the entropy-based discretized data set BA12* (72.73%). When RI is applied to the Forward Feature Selection technique attributes and combining it with core loan attributes and scores the best average prediction accuracy is achieved with the manually discretized data set BM4 (80.15%). When DT is applied to such data, BA10, which consists of equal-width binning discretized data, provides the best average prediction of 66.90%. The non-performing loans in the training data sets are replicated five times for BA12* and BM4, and once for BA10.

The rules for performing and non-performing loans that are extracted by the application of DT are aligned with real-world interpretation of performing and non-performing loan characteristics/attributes. Some rules for performing loans contradict real-world interpretation of performing loan characteristics/attributes when RI is applied to the data sets. The quality of extracted rules for non-performing loans is lower than that for performing loans. The skewed proportion of the performing and non-performing loans in the training and testing data sets may have affected these results.

The performance accuracy with the application of DT slightly improves when the Forward Feature Selection Technique is used, compared with when all FI and QI attributes and scores are used. The extracted rules for performing loans are not all aligned with real-world interpretation of performing loan characteristics/attributes, and the rule for non-performing loans contradicts the real-world interpretation of non-performing loan characteristics/attributes. They are therefore excluded from the discussion. The performance accuracy of RI is slightly better when all FI-QI attributes and scores are used, compared with when the Forward Feature Selection Technique is used. The over-fitting problem with regard to non-performing loans is significantly more evident with the application of RI, but does not show significant differences between the use of all FI and QI attributes and the use of the Forward Feature Selection Technique.

Table 5.21 Best performance accuracy for Bank B credit data sets

Data Mining Techniques		Data Sets	Classification Accuracy with Training Credit Data Sets		Prediction Accuracy with Testing Credit Data Sets		Average of Prediction Accuracy
			Performing Loans	Non-Performing Loans	Performing Loans	Non-Performing Loans	
DT							
All credit attributes		BA10	87.85%	47.37%	97.44%	36.36%	66.90%
Core loan, FI and QI attributes		BR6	90.20%	50.00%	92.86%	25.00%	58.93%
Core loan and Scores		BA10	90.48%	57.14%	97.44%	36.36%	66.90%
RI							
All credit attributes		BA12*	87.00%	75.71%	95.45%	50.00%	72.73%
Core loan, FI and QI attributes		BR7	71.76%	41.18%	93.94%	17.65%	55.80%
Core loan and scores		BR8	95.40%	74.70%	97.37%	33.33%	65.35%
Forward Feature Selection and DT	Selected attributes	BM2	85.96%	50.00%	93.33%	40.00%	66.67%
	Combined attributes	BA10	90.29%	52.17%	97.44%	36.36%	66.90%
Forward Feature Selection and RI	Selected attributes	BA11*	71.94%	55.56%	93.33%	11.43%	52.38%
	Combined attributes	**BM4**	**62.03%**	**50.00%**	**93.62%**	**66.67%**	**80.15%**

5.2.4 Bank C

The content of the individual credit data sets of Bank C can be summarized as follows: ten numerical attributes of the financial position of customers, 34 attributes (a mixture of numerical and textual) of core loan information, and 117 attributes with scores (numerical). Bank C is a privately-owned bank, and the data sets they provided are of traders (43.58%) and service providers (31.44%) who are borrowers of the bank. The loan size of the major portion (55%) of the customers sampled for this research was between IDR 300 million and IDR 500 million. Bank C applied scoring to assess loan applications. The credit data sets provided for this research comprise the numerical and textual scoring conducted by Bank C and do not contain any descriptive unstructured text. Thus, the tests performed on Bank C credit data sets are limited to the application of DT and RI. The Forward Feature Selection Technique is not conducted to the data sets for the reasons discussed in Chapter 4, Section 4.5.4.

The imbalanced (skewed) proportion of performing and non-performing loans in the training and testing data sets affects the performance accuracy of the data mining techniques

irrespective of the levels of replication applied to the training data sets. Consequently, the application of DT and RI to the data sets primarily provides an indication of the comparative significance of such data sets for constructing prediction methods, and the contribution of the data mining techniques to the credit risk scoring system of Bank C. As such, the extracted rules selected and presented in this section are based on their usefulness in real-world scenarios as opposed to the number of predicted cases covered.

5.2.4.1 Decision tree

Table 5.22 shows the performance accuracy for all data sets achieved in prompt (0 second) processing time. The non-performing loans in the CM2 training data set are replicated three times and represent 16 of 110 total loans in the training data set. The non-performing loans in the CM3 training dataset are replicated seven times and represent 32 of 126 total loans in the training data set. The non-performing loans in the CM4 training dataset are replicated eleven times and represent 48 of 170 total loans in the training data set. The number of non-performing loans is three, out of a total of 42 loans in all testing data sets. The over-fitting problem related to performing and non-performing loans occurs in all the data sets that provide results (CM2, CM3 and CM4).

Data sets CM2 and CM3 produce identical average performances of 72.50%; however CM2, with thrice-replicated non-performing loans, shows better classification accuracy and is therefore regarded as the data set that provides the best results. Figure 5.9 displays the tree generated by Data Set CM2.

The model built by training data set CM2 has a 100% testing data set coverage rate. There are six extracted rules for performing and non-performing loans. An exemplary rule from CM2 that provides the highest prediction accuracy for performing loans is "debtlegal $\leq$ 2.500 AND liquidity > 1.375". This rule covers 28 credit data sets with 93.33% prediction accuracy. The credit parameters comprising the rule are excellent (score 1) or good (score 2) proof of the applicant's business legality, and reasonable ability to meet short-term financial obligations.

There are two rules for non-performing loans. These rules have distinct coverage rates in the training data set, resulting in low coverage rates in the testing data set due to the skewed proportion and limited number of credit data sets. The highest testing data set coverage is 33.33% for any non-performing loan rule. An exemplary rule that provides the highest prediction accuracy for non-performing loans is "deblegal > 2.500". This rule has a testing data set coverage rate of 33.33%, with 50% prediction accuracy. The single credit parameter for the prediction is substandard proof of the applicant's business legality.

Table 5.22 Performance accuracy of DT - Bank C

Data Sets	# of Leaf Nodes	Size of Tree	Classification Accuracy with Training Credit Data Sets		Coverage Rate	Prediction Accuracy with Testing Credit Data Sets		Coverage Rate
			Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
CM1	1	1	95.88%	0.00%	100.00%	92.86%	0.00%	100.00%
CM2	**6**	**10**	**100.00%**	**80.00%**	**100.00%**	**95.00%**	**50.00%**	**100.00%**
CM3	6	10	100.00%	82.05%	100.00%	95.00%	50.00%	100.00%
CM4	6	10	100.00%	82.76%	100.00%	92.86%	0.00%	100.00%

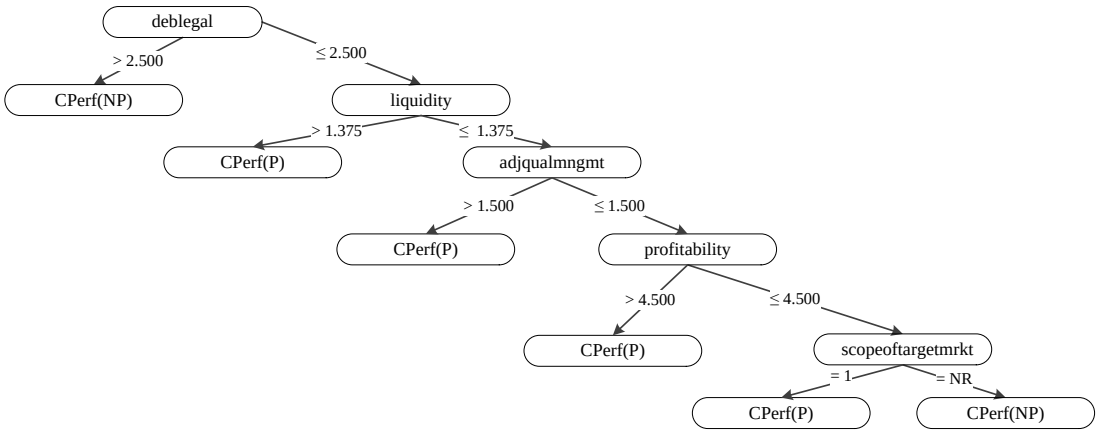


Figure 5.9 Decision tree from Data Set CM2 – Bank C

Key:

deblegal = debtor's legality

liquidity = liquidity

adjqualmngmt = adjustment to management

profitability = profitability

scopeoftargetmrkt = scope of target market

CPerf(P) = Credit Performance (Performing Loan)

CPerf(NP) = Credit Performance (Non-Performing Loan)

5.2.4.2 Rule induction

Table 5.23 shows the performance of RI applied to the data sets in one second of processing time. As with the results from the application of DT, the performance pertaining to the non-replicated data set (CM1) is the worst. With the RI application all data sets provide classification results in the training data sets, although none of the non-performing loans are predicted in the testing data sets. This can be ascribed to the very limited number of non-performing loans in both the training and testing data sets, and the effect this has in narrowing the spectrum of non-performance causes in the training data sets and mismatching it in the testing data sets. In this regard, the average classification performance result is used to select the best-performing data set since the assumption applies that the extracted rules would have performed better if the testing data set had contained more non-performing loan

data. As when DT is applied, Data Set CM2 provides the best average prediction accuracy (55.56%).

Table 5.23 Performance accuracy of RI – Bank C

Data Sets	# of Rules	Classification Accuracy with Training Credit Data Sets		Coverage rate	Prediction Accuracy with Testing Credit Data Sets		Coverage rate
		Performing Loans	Non-Performing Loans		Performing Loans	Non-Performing Loans	
CM1	1	95.92%	0.00%	100.00%	92.86%	0.00%	100.00%
CM2	**3**	**94.62%**	**64.71%**	**100.00%**	**94.44%**	**16.67%**	**100.00%**
CM3	3	100.00%	78.05%	100.00%	91.89%	0.00%	100.00%
CM4	3	100.00%	78.69%	100.00%	91.89%	0.00%	100.00%

The model built by training data set CM2 has a 100% testing data set coverage rate. There are three extracted rules for performing and non-performing loans. An exemplary rule with the highest prediction accuracy for performing loans is "if liquidity > 1.375". This rule covers 29 credit data sets with 100% prediction accuracy. The single credit parameter for the prediction is reasonable ability to meet short-term financial obligations.

5.2.4.3 Summary of results – Bank C

Table 5.24 shows the data sets with the best performance accuracy with the application of DT and RI. Data Set CM2 provides the best performance with the application of DT and RI. Overall, DT provides better results than RI. The over-fitting issue on performing and non-performing loans is evident in all data sets when DT and RI are applied (Tables 5.22 and 5.23).

Table 5.24 Best Performance accuracy for Bank C credit data sets

Data Mining Techniques	Data Sets	Classification Accuracy with Training Credit Data Sets		Prediction Accuracy with Testing Credit Data Sets		Average of Prediction Accuracy
		Performing Loans	Non-Performing Loans	Performing Loans	Non-Performing Loans	
DT	**CM2**	**100.00%**	**80.00%**	**95.00%**	**50.00%**	**72.50%**
RI	**CM2**	**94.62%**	**64.71%**	**94.44%**	**16.67%**	**55.56%**

5.2.5 Comparison of data mining technique results

Table 5.25 shows the best results when DT and RI are applied for all sample banks. Five out of six data sets contain replicated performing and non-performing loan credit data sets to overcome the problem regarding the skewed proportion of non-performing and performing loan credit data sets provided by the banks for this research. There is no single discretization technique that seems to improve the classification and prediction performance of data sets since the four best data sets in Table 5.25 show differences. This implies there is no superior discretization technique for the credit data sets used in this research. The best performance is achieved when only attributes that may be relevant are selected for the application of DT and RI. When the Forward Feature Selection Technique is applied to the financial and qualitative information of Bank A, better data set performance is achieved. In the case of Bank B the use of core loan attributes and scores produces much better data sets than when such information is used in combination with FI and QI attributes. This provides a definitive indication that the current credit attributes used by the respective banks could be streamlined for better credit risk assessment. Furthermore, the best results from Bank B (BA10 and BM4) are achieved with reduced attributes.

The results of DT and RI applied to Bank B and Bank C credit data sets are substantially affected by the skewed proportion of performing and non-performing loans. However, the use of the DT and RI techniques shows the potential to improve credit decisions by the banks. Based on the sample credit data sets of the banks the data mining techniques applied in this research improve the accuracy of classifying performing loans that serve as the primary source of non-performing risk. Using the testing data sets of Bank B as an example, the misclassification of performing loans based on their current practice is 10% (5 non-performing loans classified as performing from a total of 50 loans) while with the application of DT the misclassification is reduced to 3%, where a total of 39 loans are predicted as performing and only one of these loans is actually non-performing. For non-performing loans the misclassification of Bank B is regarded as 100% since all loan applications are approved in terms of their criteria, while the DT reduces the non-performing loan misclassification to 64% (seven performing loans predicted as non-performing loans form a total of eleven, when only four are actually non-performing).

Table 5.25 Summary of the best performance accuracy for all sample banks

Data Mining Techniques	Data Sets	Classification Accuracy with Training Credit Data Sets		Prediction Accuracy with Testing Credit Data Sets		Description
		Performing Loans	Non-Performing Loans	Performing Loans	Non-Performing Loans	
DT:						
Bank A	**AA7***	**77.34%**	**50.00%**	**73.68%**	**100.00%**	Original number of credit data sets (no oversampling is applied) Entropy-based discretization Selected attributes derived from Forward Feature Selection are tested together with core loan attributes
Bank B	BA10	90.48%	57.14%	97.44%	36.36%	Oversampling is applied Equal-width binning discretization Selected attributes derived from the assigned score attributes combined with core loan attributes.
Bank C	CM2	100.00%	80.00%	95.00%	50.00%	Oversampling is applied
RI:						
Bank A	**AM3**	**90.43%**	**96.51%**	**92.31%**	**81.25%**	Oversampling is applied Manually discretized Selected attributes derived from Forward Feature Selection are tested together with core loan attributes
Bank B	BM4	62.03%	50.00%	93.62%	66.67%	Oversampling is applied Manually discretized Selected attributes derived from Forward Feature Selection are tested together with core loan attributes
Bank C	CM2	94.62%	64.71%	94.44%	16.67%	Oversampling is applied

5.2.6 The contribution of data mining techniques used with qualitative data analysis

The importance of qualitative information in credit risk assessment is indicated by the highest average performance accuracy of the selected data mining techniques, achieved with data sets AA7* and AM3 of Bank A where descriptive qualitative text and financial information are used exclusively and in equal quantity. To confirm the contribution of descriptive qualitative information used in the data mining techniques applied in this research, NN analysis is applied as well to compare the performance accuracy achieved with text-based qualitative information and without it. NN is a well-known technique for (quantitative) credit risk assessment that serves as a good method for constructing credit assessment models, as shown by previous studies (see Chapter 2). It is also a supreme machine learning technique for MSME credit risk assessment when the inputs are in numerical (ratio and binary) format (Angelini, di Tollo, and Roli 2008; Bensic, Sarlija, and Zekic-Susac 2005; Wu and Wang 2000).

The NN technique is applied to the (quantitative) core loan and financial attributes (total of 42 numerical attributes) of Bank A credit data sets AR4, AR5 and AR6. These data sets are the most balanced in terms of proportion between performing and non-performing credit data sets and contain core loan information, financial information and descriptive qualitative information. The difference between the data used in previous sections is that the descriptive qualitative information has been removed from the input data for NN.

In order to obtain optimum NN performance accuracy, different numbers of hidden layers are trialed. One hidden layer for AR4, three hidden layers for AR5 and two hidden layers for AR6 provide the best classification and prediction accuracy. Tables 5.26 and 5.27 show the performance accuracy of the data mining techniques (DT and RI) applied to core loan information, financial information and descriptive qualitative information, against that of NN applied to core loan and financial information only.

The average classification performance (Table 5.26) of NN is lower than that of DT for all three data sets, but its prediction accuracy (Table 5.27) is significantly lower for all three data sets. On the other hand, the average classification performance of NN for Data Set AR4 is less than that of RI, but more than that of RI for Data Set AR5 and almost equal to RI for Data Set AR6. The prediction performance NN is lower for all three data sets, notwithstanding its classification performance compared with RI. This confirms the positive contribution of descriptive qualitative information in credit risk assessment. Considering the previously mentioned classification and prediction performance of DT, RI and NN, only in the case of Data Set AR4 is the training data set coverage rate of NN substantially worse than that of DT and RI.

Table 5.26 Summary of classification performance accuracy – Bank A

Data Sets	DT		Coverage Rate	RI		Coverage Rate	NN		Coverage Rate
	PLs*	NPLs+		PLs*	NPLs+		PLs*	NPLs+	
AR4	81.08%	53.33%	100.00%	72.55%	56.25%	98.51%	69.77%	41.67%	73.92%
AR5	95.35%	87.50%	100.00%	86.52%	75.56%	97.01%	93.18%	86.96%	100.00%
AR6	100.00%	92.00%	100.00%	97.50%	90.00%	96.11%	95.29%	92.63%	100.00%

* Performing Loans (PLs)

+ Non-Performing Loans (NPLs)

The testing data set coverage rate of NN in Data Sets AR5 and AR6 is better than that of DT and RI, but lower in Data Set AR4. This can be ascribed to the smaller number of attributes used with NN, but does not support the expectation that a higher coverage rate might represent better prediction performance. Overall, the classification and prediction performance accuracy of DT, RI and NN, confirm the contribution of qualitative information in credit risk assessment.

Table 5.27 Summary of prediction performance accuracy – Bank A

Data Sets	DT		Coverage Rate	RI		Coverage Rate	NN		Coverage Rate
	PLs*	NPLs+		PLs*	NPLs+		PLs*	NPLs+	
AR4	57.89%	70.00%	100.00%	56.25%	61.54%	96.55%	50.00%	47.62%	91.99%
AR5	73.33%	78.57%	89.66%	54.55%	71.43%	86.21%	52.17%	66.67%	100.00%
AR6	61.90%	87.50%	75.86%	68.42%	90.00%	79.31%	47.62%	50.00%	98.10%

* Performing Loans (PLs)

+ Non-Performing Loans (NPLs)

5.3 Summary

In this chapter, the application of selected data mining techniques (DT and RI) to training and testing credit data sets of the three sample banks have been conducted. Both DT and RI perform the best with the credit data sets provided by Bank A, where Forward Feature Selection is applied to financial information and descriptive qualitative text together with core loan information for credit risk assessment. When DT is applied, the best performance is achieved with the non-replicated original number of credit data sets where

numerical attributes values are discretized using the entropy-based discretization technique. When RI is applied, the best performance is achieved with replicated credit data sets where numerical attributes values are manually discretized. The use of selected attributes increases the performance accuracy for performing and non-performing loans.

The training data set coverage rates of NN are generally comparable to those of DT and RI, while the testing data set coverage rates are generally better. The application of DT and RI to Bank A credit data sets show higher performance accuracy compared to NN where qualitative information is excluded. The contribution of descriptive qualitative information to the performance of data mining technique results is confirmed by the application of NN analysis in addition to the selected data mining techniques used to compare the performance accuracy achieved with descriptive qualitative information and that achieved without it.

In the next chapter, conclusions and future research directions are discussed.

CHAPTER 6 – SUMMARY AND CONCLUSIONS

6.1 Introduction

This chapter first summarizes key points, focusing on the objectives of the research, its theoretical underpinnings and the research methodology applied. This is followed by a concise overview of the findings from which conclusions are drawn and directions for future research in the area of credit risk assessment for Indonesian MSMEs are provided.

6.2 Summary

The Indonesian banking industry has gone through three distinct stages of change in the past five decades. In the first stage (1966–1981), the lending market share was dominated by seven state-owned banks with financing prioritized for government projects and big-sized enterprises. Privately-owned banks focused on regional and international lending. During this period large businesses mostly obtained their financing from international banks. The situation continued to the second stage (1982–1997) with an upsurge in the lending market share of new, privately-owned domestic banks and foreign/joint venture banks, although the largest market share was still in the hands of state-owned banks. During this era, bad corporate governance occurred in the overall banking industry. In an effort to improve bank supervision a number of regulations were introduced, including a scoring system to assess the health of banks. One of the bonus assessment points was a minimum threshold for SME financing. The third stage (1998–2011) was marked by the repercussions of the AFC on the Indonesian currency. The plummeting exchange rate affected the capacity of large businesses to repay their financial obligations to banks. Many either closed down or merged, depending on the severity of their liquidity problems. In order to rectify the situation, BI enforced strict prudential banking, and the government increasingly engaged banks in providing more financing to MSMEs, as small businesses were seen to play an important role in the economy and showed better ability to survive the economic downturn in comparison with large businesses. Since then, MSME financing has increased.

Although the contribution of MSMEs to the Indonesian economy has been acknowledged for many years (even more in the aftermath of the 1996 AFC and the 2006 GFC), their financing needs were traditionally served by domestic formal and non-formal microfinance institutions. Only BRIs and BPRs consistently targeted and continue to target their lending business exclusively to MSMEs; thus, the majority of banks were at this stage not properly prepared to apply credit risk assessment for MSMEs. The mandated 5Cs

lending principles have since then been applied differently from the way used for large businesses. In addition to the collection of data via application forms and the available financial information of applicants, bank staff perform interviews and conduct physical observations of the business activities of applicants. These actions are taken to overcome the lack of reliable financial information resulting from the unsystematic bookkeeping of many MSMEs. The discretionary nature of this assessment methodology and the loan quality risk emanating from it are the dominant reasons for this research. The research is aimed at constructing an objective and more accurate credit risk assessment method for MSMEs than is applied by banks at present.

The same credit data that banks collect and apply in their credit risk assessment is used in this research. Previous studies regarding credit risk assessment methods for MSMEs were applied to reliable and complete credit data obtained from financial institutions and/or credit bureaus. Statistical (DA and logit) and machine learning (NN, SVM, RS and RSF) techniques were applied to increase the accuracy of credit-granting decisions. Good models were produced; but these models are confined to structured types of credit data and overlook the potential contribution of unstructured types of credit data (qualitative information).

Constructing a credit risk assessment model for Indonesian MSMEs brings a particular challenge since the credit data available to and used by banks in many cases is not exactly similar and in some cases is incomplete. The use of extensive qualitative data in combination with commonly used quantitative data for credit assessment purposes in this research addresses this gap. It also offers the first study of credit risk assessment for MSMEs carried out in Indonesia.

The three sample banks (Bank A, B and C) in this research apply the 5Cs credit assessment principles to their respective market segments, but not in similar ways. The principles are intertwined with their own forms of relationship lending. Bank A employs discretionary lending, while the main tools for Bank B and C are scoring systems. Although the literature shows that quantitative data is primarily used for scoring systems, Banks B and C assign scores to both quantitative and qualitative data. All banks collate quantitative and qualitative data, but each differs in detail and extent. In this study, both types of credit data are presented in tandem, and selected data mining techniques are applied to them while preserving their original context.

In the KDD process, data mining consists of the systematic and automated extraction of hidden and novel patterns from large databases. Prior to the application of data mining techniques, the data preparation step is required to ensure that the obtained data are ready for the mining tasks. Pre-processing methods such as data cleaning, data integration, data transformation and data reduction are performed to attain completeness, consistency, and organization of the data. Based on the type of data on hand (relational, sequential, semi-

structured and unstructured) and the purpose (descriptive and predictive) of the mining task, appropriate data mining techniques are selected. The patterns extracted from the application of these techniques express relationships among the attributes in the data sets. So that users may understand the patterns, they are transformed into predictive rules or decision trees. The post-processing step of KDD concerns the interpretation of these patterns with regard to their application and quality.

The steps in this KDD process are performed on the credit data sets obtained from the three sample banks. The components of the credit data sets of the banks are not the same, since each bank differs in their credit risk assessment methodologies, and focuses on different specific data. Bank A applies a discretionary lending method with credit data comprising financial, qualitative and core loan data. Bank B applies a scoring system combined with a discretionary lending method, so the credit data includes financial, qualitative and core loan data as well as scores. Bank C applies only a scoring system based on credit data containing financial data, core loan data and scores. Although the original credit data of the banks are similar in terms of the broad categories of financial, qualitative and core loan data, there are variations in how and to what extent the different components are used in their credit risk assessments. In this research all the collected structured financial and core loan data sets of banks A and B are manually and automatically discretized, while their descriptive unstructured qualitative data is manually categorized. The semi-structured credit data sets of Bank A and B are then presented in an XML document. Bank C credit data sets are not discretized and categorized, as the attributes values (scores and description of scores) are pre-categorized by the bank and kept in the score format.

The use of frequent sub-tree mining on credit data in this study is limited by an issue of combinatorial complexity. This is resolved by converting the credit data from a tree-structured format into a flat format, using the extracted general structure (DSM). DT, RI and Forward Feature Selection are then applied to the data sets. Another aspect affecting the study is the imbalance in the proportion of performing and non-performing loans in the credit data sets received from the banks: very few non-performing loan data sets were provided, compared to performing loan data sets. The original number of performing and non-performing loans is used in the study, as well as different levels of oversampling to reduce the impact of the imbalance. The analysis focuses on finding the best average prediction accuracy for the target classes (performing and non-performing loans).

The highest average DT and RI prediction accuracy is achieved with data sets from Bank A. The DT Data Set (AA7*) consists of the original number of credit data sets (no oversampling is applied) with discretized numerical attribute values using the entropy-based technique. The credit data for this data set constitutes selected attributes from the application of Forward Feature Selection that are combined with core loan attributes. The RI Data Set

AM3 contains replicated credit data sets (original number of performing loans and three times replicated non-performing loans) with manually discretized numerical attribute values. The credit data for this data set constitutes selected attributes from the application of Forward Feature Selection that are combined with core loan attributes. The worst results with DT (CM2) and RI (CM2) are from the data sets of Bank C, where the scoring applied by the bank is exclusively used for credit risk assessment. Despite the number of replications applied to such training data sets, the skewed proportion of the performing and non-performing loans in the training and testing data sets affect the results where RI is used more than when DT is used.

The application of Forward Feature Selection to Bank A and B data sets indicates that a reduced number of credit attributes improves the performing and non-performing loan prediction accuracy. The positive effect when fewer attributes are used for credit risk assessment is more apparent when DT is used than when RI is used. The role of qualitative information for credit risk assessment of MSMEs is substantiated by the higher prediction accuracy in Bank A and Bank B data sets where raw or interpreted qualitative information is included in the analysis, than when it is excluded from the assessment as shown by Bank C data sets and the application of NN to Bank A quantitative core loan and financial attributes.

6.3 Conclusions

Based on these findings, the application of data mining techniques improves the credit risk assessment methods currently applied by the three sampled Indonesian banks. DT and RI provide automated and objective credit-granting decisions where the accuracy of rejection of non-performing loans is increased. Another task that is also achieved by data mining is selection of credit attributes that are the most important in the classification of loans in terms of performance. The implication of applying the techniques is, however, that future performing loans may also be rejected; but the assumption is that the cost of foregoing a performing loan will be much less (interest income) than accepting a non-performing loan (capital loss, interest on the capital and litigation costs). This aspect is beyond the scope of this thesis and could be addressed in separate research.

The existing skewed proportion of performing and non-performing loans that banks provided for this study make it difficult to find optimum credit risk assessment models, but data mining allows the continuous inclusion of new credit data sets as the performance of more existing loans become evident over time, so skewness will reduce. With such a “live” system, banks will be equipped with continually updated knowledge from which to make more reliable decisions on new loan applications lodged by MSMEs.

The proposed credit risk assessment methodology for MSMEs consists of steps that are coherent with the research objectives of this study. The first research objective is to develop

a template to capture quantitative and qualitative types of credit data in a domain-specific way, and organize (contextualize) the available information effectively. This is achieved by the development of a XML document. The second research objective is attained by the use of DSM and creating a flat data format to overcome limitations from the application of traditional frequent tree mining. Based on the converted format, appropriate data mining techniques are selected and applied to the data sets. The third research objective is to develop a conceptual framework for the construction of comprehensive credit parameters for MSMEs from quantitative and qualitative information. The use of all credit data supplied by the sample banks (because of the structuring of all such data for data mining technique application) achieves the objective. Research objective four entails the development of an approach for constructing credit parameters for MSMEs. In this regard, Forward Feature Selection is used to select the parameters deemed prominent to assess MSME's credit worthiness. The predictive power of such attributes is tested within its own cohort and with core loan credit attributes. Afterwards, a comparison with the predictive power of all credit attributes is performed. This approach is validated using the testing credit data sets; thus, the fourth research objective is attained. The final research objective, to develop a Decision Support Methodology for MSME credit risk assessment based on knowledge patterns derived from the application of data mining, is also attained as the analysis shows that the inclusion of qualitative information is an essential factor in credit risk assessments of MSMEs, as opposed to the exclusive application of a scoring system.

6.4 Directions for future research

Throughout this research, limitations relating to the proportion of the provided credit data sets have been addressed. In this section, future research directions within the field of credit risk assessment methodology for Indonesian MSMEs are discussed. These directions are related to current practices of credit risk assessment for MSMEs, based on the use of extensive credit attributes. This research shows that accuracy increases when discriminating credit risk parameters for MSMEs are used. The potential of the proposed credit risk assessment methodology encourages banks towards extensive use of their credit data sets including both rejected loan applications and knowledge discovery from different credit data set groupings. Furthermore, exploration of techniques to manage the descriptive unstructured credit information effectively as well as to overcome the imbalance class problem is required.

6.4.1 The improvement of discriminating credit risk attributes

For future research, a development of credit risk parameters with higher discriminating power and intended for different purposes is warranted. To improve the application of data mining techniques in credit risk assessment, a more comprehensive collection of credit data

sets in terms of quantity and quality, and the application of grouping to the credit data sets, are warranted. The discussions in Chapter 5 clearly identify the over-fitting problem that affects the quality of the extracted rules, particularly in the cases of Banks B and C. This can be overcome by the inclusion of more credit data sets in the analysis by large Banks like B and C that have nationwide operations. The challenge is to develop a credit data base that includes historical and present credit information from different branches to enable comprehensive analysis.

Aside from a more balanced proportion of performing and non-performing credit data sets, the inclusion of rejected MSME loan applications may also be advantageous. The credit parameters that this type of rejection is based on can provide insight about attributes that are regarded prominent in the initial credit-granting decision by the bank's assessors. These findings can be compared between different banks. The relevance of these attributes can also be compared based on their different use by banks. In addition, details about the duration and amount of delayed payments may be valuable in identifying more specific credit risk parameters. This research uses delayed payment as a single indicator of non-performing loans. To enhance the results from this study, delayed payments could be categorized into more specific regulatory groupings, like special mentioned loans, substandard loans, doubtful loans and bad debts set by BI, thus presenting credit risk parameters relevant for different types of delayed payment prediction. This could be monitored over time to determine whether relationships exist between attributes of customers and movements between categories of delayed payments. The challenge in using a more detailed target class lies in maintaining a relatively balanced proportion of each target class when it is combined with performing credit data sets: this can be overcome by keeping the number of credit data sets equal for each target class, following the class with least available credit data sets.

The systematic grouping of credit data sets may improve the credit assessment method. Especially for banks with nationwide coverage such as Bank B, insight about credit risk parameters based on industry-specific or region-specific business risk exposures, or the combination of these, is valuable. With a more balanced proportion of performing and non-performing credit data sets, and more comprehensive credit data, the prospect of refining the extracted rules is increased; this may lead to more reliable credit-granting decisions. In order to increase the quality of the extracted rules, exploration should be directed toward constructing an effective way to integrate the domain working environment that includes constraints as discussed in Cao et al. (2010).

6.4.2 The integration of the system with natural language processing (NLP) tools

The considerable amount of descriptive unstructured text in the collated credit data sets warrants a tool that automatically converts the related text into the XML document for Bahasa. There are several text-to-XML annotators, such as Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally 2004), GLOSS (Kaye 2006), and PCFG parser (Klein and Manning 2003) that use statistical and rule-based annotators for text. The application of a domain ontology to such tools such as (1) an ontology-based (Alani et al. 2003; Embley et al. 1998) and (2) an ontology-driven (Handschuh, Staab, and Maedche 2001; Vargas-Vera et al. 2002) configuration will reduce the ambiguity of the terms. Unfortunately these tools are currently limited to English syntax, and the phrase structure trees that are generated and presented in an XML document are of limited use without an established bilingual corpus. In this regard, development of a language-independent representation-based dictionary (Dorr 1997) may be useful for continuing a live system on credit databases for large-scale banks like Banks B and C, notwithstanding the considerable amount of resources required to maintain the database. Taking into account the subjectivity involved in the descriptive unstructured text, a text-based modelling tool that can capture the sentiment within the text such as that provided by Latent Semantic Analysis (LSA) is a preferable option. LSA was introduced by Dumais et al. (1988), and works on the collection of sentences (i.e. text) as opposed to words. Rather than annotating the text, LSA uses a statistical approach to extract meaning by way of vector similarities among documents (Dumais et al. 1988). In the same manner, the use of cross-language Latent Semantic Indexing (CL-LSI) (Dumais et al. 1997) is potentially useful to overcome the limited English syntax. Although the prospect of ontology and LSA approaches is realistic, the challenge remains to find a method or tool to enable efficient concept extraction from descriptive unstructured credit data in Bahasa.

6.4.3 The challenge of big data

The increasing challenge in data-driven company management is the analysis of growing volume, velocity and variety of a database that is known as “big data” (McAfee and Brynjolfsson 2012). A dynamic credit risk assessment system as proposed in this thesis entails a similar challenge, from the exploitation of a rich compilation of data consisting of the real-time feed of new loan applications and the ongoing performance changes of existing loans, and that need to be integrated to allow real-time strategic decision-making (i.e. credit-granting decision). Cohen et al. (2009) propose the Magnetic, Agile, Deep (MAD) approach; from this, Herodotou et al. (2011) have developed Starfish, a tool for big data analysis based on MAD. In the credit risk domain, future research related to big data should be aimed at

increasing the quality of evidence-based decision-making and its scalability. In order to achieve this, the evaluation at the post-processing stage should include active involvement by domain experts, which iterates the relevance of D^3Mining (Cao et al. 2010). As data analytics become more efficient, the interpretation of its results and incorporation of such knowledge into business innovation and practice becomes a necessity.

REFERENCES

Abdou, H., J. Pointon and A. El-Masry. 2008. “Neural Nets Versus Conventional Techniques in Credit Scoring in Egyptian Banking.” *Expert Systems with Applications* 35 (3): 1275–1292.

Abe, K., S. Kawasoe, T. Asai, H. Arimura, and S. Arikawa. 2002. “Optimized Substructure Discovery for Semi-structured Data.” In *Principles of Data Mining and Knowledge Discovery-PKDD'02, Lecture Notes in Artificial Intelligence vol. 2431*, eds. T. Elomaa, H. Manilla and H. Toivonen, 1–14. Berlin: Springer-Verlag.

Abiteboul, S. 1997. “Querying Semi-Structured Data.” In *Database Theory—ICDT'97, Lecture Notes in Computer Science vol. 1186*, eds. F. N. Afrati and P. Kolaitis, 1–18. Berlin: Springer-Verlag.

Adrick, P. 2010. “First Greece, then Ireland: Europe’s Debt Problem has Gone from Bad to Worse.” *The Telegraph*, 27 November.

Agrawal, R., D. Gunopulos, and F. Leymann. 1998. “Mining Process Models from Workflow Logs’. In *Advances in Database Technology—EDBT'98, Lecture Notes in Computer Science vol. 1377*, eds. F. Saltor, G. Alonso, H.-J. Schek and I. Ramos, 467–483. Berlin: Springer-Verlag.

Agrawal, R., T. Imielinski, and A. Swami. 1993. “Mining Association Rules between Sets of Items in Large Databases.” In *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data - SIGMOD '93, Washington DC, USA, 26–28 May 1993*, edited by Peter Buneman and Sushil Jajodia, 207–216, New York, NY: ACM Press.

Agrawal, R., and R. Srikant. 1994. “Fast Algorithms for Mining Association Rules.” In *VLDB’94: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, 12–15 September 1994*, edited by Jorge B. Bocca, Matthias Jarke and Carlo Zaniolo, 487–499, San Francisco, CA: Morgan Kaufmann.

———. 1995. “Mining Sequential Patterns”. In *Proceedings of the Eleventh International Conference on Data Engineering (ICDE '95)*, Taipei, Taiwan, 6-10 March 1995, edited by Philip S. Yu and Arbee L. P. Chen, 3–14, Taipei, Taiwan: IEEE Computer Society Press.

Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis and N. R. Shadbolt. 2003. “Automatic Ontology-based Knowledge Extraction from Web Documents.” *IEEE Intelligent Systems* 18 (1): 14–21.

Alhabshi, S. M., A. A. A. Khalid and B. Bardai. 2009. “The Development of Corporate Credit Information Database and Credit Guarantee System.” http//www.asean.org/documents/ASEAN+3RG/0809/FR/13d.pdf.

Altman, E. and G. Sabato. 2007. “Modelling Credit Risk for SMEs: Evidence from the U.S. Market.” *Abacus* 43 (3): 332–357.

Altova. n.d. Icon information systems GmbH. *XMLSpy*. http://www.xmlspy.com.

Angelini, E., G. di Tollo and A. Roli. 2008. “A Neural Network Approach for Credit Risk Evaluation.” *The Quarterly Review of Economics and Finance* 48 (4): 733–755.

Angelini, P., R. Di Salvo and G. Ferri. 1998. “Availability and Cost of Credit for Small Businesses: Customer Relationships and Credit Cooperatives.” *Journal of Banking & Finance* 22 (6–8): 925–954.

Asai, T., K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, and S. Arikawa. 2002. “Efficient Substructure Discovery from Large Semi-structured Data.” In *Proceedings of the Second SIAM International Conference on Data Mining (SIAM 2002), Arlington, USA, 11–13 April 2002*, edited by Robert Grossman, 158–174, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Asch, L. 1995. “How the RMA/Fair, Isaac Credit-scoring Model was Built.” *The Journal of Commercial Lending* 77 (10): 10–15.

Asch, L. 2000. “Credit Scoring: a Tool for More Efficient SME Lending.” *SME Issues* 1 (2): 1–4.

Atkinson, A. B. 1970. “On the Measurement of Inequality”. *Journal of Economic Theory* 2 (3): 244–263.

Audretsch, D. B. and Z. J. Acs. 1994. ”New-firm Startups, Technology, and Macroeconomic Fluctuations.” *Small Business Economics* 6 (6): 439–449.

Auld, G. W., A. Diker, M. A. Bock, C. J. Boushey, C. M. Bruhn, M. Cluskey, M. Edlefsen, D. L. Goldberg, S. L. Misner, B. H. Olson, M. Reicks, C. Wang, and S. Zaghloul. “Development of a Decision Tree to Determine Appropriateness of NVivo in Analyzing Qualitative Data Sets.” *Journal of Nutrition Education and Behavior* 39 (1): 37–47.

Avner, S. 1995. “Discovery of Comprehensible Symbolic Rules in a Neural Network.” In *Proceedings of the First International Symposium on Intelligence in Neural and Biological Systems (INBS '95), Herndon, USA, 29–31 May 1995*, 64–71, Los Amitos, CA: IEEE Computer Society Press.

Baas, T. and M. Schrooten. 2006. “Relationship Banking and SMEs: a Theoretical Analysis.” *Small Business Economics* 27 (2-3): 127–137.

Baesens, B., T. V. Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen. 2003. “Benchmarking State-of-the-art Classification Algorithms for Credit Scoring.” *The Journal of the Operational Research Society* 54 (6): 627–635.

Baesens, B., R. Setiono, C. Mues and J. Vanthienen. 2003. “Using Neural Network Rule Extraction and Decision Tables for Credit-risk Evaluation. “*Management Science* 49 (3): 312–329.

Bahrammirzaee, A. 2010. “A Comparative Survey of Artificial Intelligence Applications in Finance: Artificial Neural Networks, Expert System and Hybrid Intelligent Systems.” *Neural Computing and Applications* 19 (8): 1165–1195.

Bank for International Settlements. 2011. Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems. Basel, Switzerland.

Bank Indonesia. 2001. Peraturan Bank Indonesia Nomor 3/2/PBI/2001 “Tentang Pemberian Kredit Usaha Kecil” [Bank Indonesia Regulation Number 3/2/PBI/2001 on Small Business Lending]. Jakarta, Indonesia.

———. 2005. “Peraturan Bank Indonesia Nomor 7/2/PBI/2005 Tentang Penilaian Kualitas Aktiva Bank Umum” [Bank of Indonesia Regulation Number 7/2/PBI/2005 on Quality Assessment of Commercial Banks Assets] , Jakarta, Indonesia.

———. 2006. “Peraturan Bank Indonesia Nomor 8/13/PBI/2006 Tentang Perubahan Atas Peraturan Bank Indonesia Nomor 7/3/PBI/2005 Tentang Batas Maksimum Pemberian Kredit Bank Umum” [Bank Indonesia Regulation Number 8/13/PBI/2006 on Amendment on Bank of Indonesia Regulation Number 7/3/PBI/2005 on Legal Lending Limit for Commercial Banks]. Jakarta, Indonesia.

———. 2009. “Peraturan Bank Indonesia Nomor 11/13/PBI/2009 Tentang Batas Maksimum Pemberian Kredit Bank Perkreditan Rakyat” [Bank Indonesia Regulation Number 11/13/PBI/2009 on Legal Lending Limit for Rural Banks]. Jakarta, Indonesia.

———. 2010a. “Peraturan Bank Indonesia Nomor 12/19/PBI/2010 Tentang Giro Wajib Minimum Bank Umum Pada Bank Indonesia Dalam Rupiah Dan Valuta Asing” [Bank of Indonesia Regulation Number 12/19/PBI/2010 on Statutory Reserves in Rupiah and Foreign Currency for Commercial Banks]. Jakarta, Indonesia.

———. 2010b. “Surat Edaran Bank Indonesia Nomor 12/36/DPNP Perihal Perubahan Izin Usaha Bank Umum Menjadi Izin Usaha Bank Perkreditan Rakyat Secara *Mandatory* Dalam Rangka Konsolidasi” [Bank of Indonesia Circular Number 12/36/DPNP on

Mandatory Change of Business License from Commercial Banks to Bank Perkreditan Rakyat for Consolidation Purposes]. Jakarta, Indonesia.

———. 2011. "Peraturan Bank Indonesia Nomor 13/26/PBI/2011 Tentang Perubahan Atas Peraturan Bank Indonesia Nomor 8/19/PBI/2006 Tentang Kualitas Aktiva Produktif Dan Pembentukan Penghapusan Aktiva Produktif Bank Perkreditan Rakyat" [Bank Indonesia Regulation Number 13/26/PBI/2011 on Amendment on Bank of Indonesia Regulation Number 8/19/PBI/2006 on Earning Assets Quality and Allowance for Earning Assets Losses for Rural Banks]. Jakarta, Indonesia.

———. 2012. "Peraturan Bank Indonesia Nomor 14/8/PBI/2012 Tentang Kepemilikan Saham Bank Umum" [Bank Indonesia Regulation Number 14/8/PBI/2012 on Capital Requirements for Commercial Banks]. Jakarta, Indonesia.

Barbará, D., Y. Li, J. Couto, J-L. Lin, and S. Jajodia. 2003. "Bootstrapping a Data Mining Intrusion Detection System." In *Proceedings of the 2003 ACM Symposium on Applied Computing, Melbourne, USA, 9–12 March 2003,* edited by Gary B. Lamont, Hisham Haddad, George A. Papadopoulos and Brajendra Panda, 421–425, New York, NY: ACM Press.

Barnett, V., and T. Lewis. 1984. *Outliers in Statistical Data*. 2nd ed. Chichester: John Wiley & Sons.

Barth, J., D. Lin and K. Yost. 2011. "Small and Medium Enterprise Financing in Transition Economies." *Atlantic Economic Journal* 39 (1): 19–38.

Bastide, Y., N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. 2000. "Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets." In *Computational Logic—CL 2000, Lecture Notes in Artificial Intelligence vol. 1861*, eds. J. Lloyd, V. Dahl, S. Fraser, U. Furbach, M. Kerber, K.-K. Lau, C. Palamidessi, L. M. Pereira, Y. Sagiv and P. J. Stuckey, 972–986. Berlin: Springer-Verlag.

Bay, S. D., and M. J. Pazzani. 1999. "Detecting Change in Categorical Data: Mining Contrast Sets." In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, 15–18 August 1999,* edited by Usama M. Fayyad, Surajit Chaudhuri and David Madigan, 302–306, New York, NY: ACM Press.

Bayardo, R. J., R. Agrawal, and D. Gunopulos. 2000. "Constraint-based Rule Mining in Large, Dense Databases." *Data Mining and Knowledge Discovery* 4 (2): 217–240.

Beaver, G. 2004. “Management, Strategy and Policy in the UK Small Business Sector: a Critical Review.” *Journal of Small Business and Enterprise Development* 11 (1): 34–49.

Beck, T. 2008. “Banking Services for Everyone?” Barriers to Bank Access and Use around the World.” *The World Bank Economic Review* 22 (3): 397.

Beck, T. and A. Demirguc-Kunt. 2006. “Small and Medium-size Enterprises: Access to Finance as a Growth Constraint.” *Journal of Banking & Finance* 30 (11): 2931–2943.

Beeri, C., and T. Milo. 1999. “Schemas for Integration and Translation of Structured and Semi-structured Data.” In *Database Theory—ICDT’99, Lecture Notes in Computer Science vol. 1540*, eds. C. Beeri and P. Buneman, 296–313. Berlin: Springer-Verlag.

Behr, P. 2011. “How Do Lending Relationships Affect Access to Credit and Loan Conditions in Microlending?” *Journal of Banking & Finance* 35 (8): 2169.

Bellotti, T. and J. Crook. 2009. “Support Vector Machines for Credit Scoring and Discovery of Significant Features.” *Expert Systems with Applications* 36 (2–2): 3302–3308.

Bensic, M., N. Sarlija and M. Zekic-Susac. 2005. “Modelling Small-business Credit Scoring by Using Logistic Regression, Neural Networks and Decision Trees.” *Intelligent Systems in Accounting, Finance and Management* 13: 133–150.

Berger, A. N., L. F. Klapper and G. F. Udell. 2001. “The Ability of Banks to Lend to Informationally Opaque Small Businesses.” *Journal of Banking & Finance* 25 (12): 2127–2167.

Berger, A. N. and G. F. Udell. 1990. “Collateral, Loan Quality and Bank Risk.” *Journal of Monetary Economics* 25 (1): 21–42.

———. 1995. “Relationship Lending and Lines of Credit in Small Firm Finance.” *The Journal of Business* 68 (3): 351–381.

———. 1998. “The Economics of Small Business Finance: the Roles of Private Equity and Debt Markets in the Financial Growth Cycle.” *Journal of Banking & Finance* 22 (6–8): 613–673.

———. 2006. ‘A More Complete Conceptual Framework for SME Finance.” *Journal of Banking & Finance* 30 (11): 2945–2966.

Berger, W. H., and F. L. Parker. 1970. “Diversity of Planktonic Foraminifera in Deep-sea Sediments.” *Science* 168 (3937): 1345–1347.

Berry, A., E. Rodriguez and H. Sandee. 2001. “Small and Medium Enterprise Dynamics in Indonesia.” *Bulletin of Indonesian Economic Studies* 37 (3): 363–384.

———. 2002. “Firm and Group Dynamics in the Small and Medium Enterprise Sector in Indonesia.” *Small Business Economics* 18 (1-3): 141–161.

Bester, H. 1985. Screening vs. Rationing in Credit Markets with Imperfect Information. *The American Economic Review* 75 (4): 850–855.

———. 1987. “The Role of Collateral in Credit Markets with Imperfect Information.” *European Economic Review* 31 (4): 887–899.

Bhasin, B. B. 2010. “Globalization of Entrepreneurship: Policy Considerations for SME Development in Indonesia.” *The International Business & Economics Research Journal* 9 (4): 95–103.

Bierman, H., Jr. and W. H. Hausman. 1970. “The Credit Granting Decision.” *Management Science* 16 (8): B519–B532.

Blöchlinger, A. and M. Leippold. 2006. “Economic Benefit of Powerful Credit Scoring.” *Journal of Banking & Finance* 30 (3): 851–873.

Bloodgood, L., N. Christ, D. Cook, J. D. l. Cruz, M. Ferrantino, D. Fravel, W. Greene et al. 2010. “Small and Medium-sized Enterprises: U.S. and EU Export Actitivities, and Barriers and Opportunities Experienced by U.S. Firms.” Washington, DC: U.S. International Trade Commission.

Boot, A. W. A. 2000. “Relationship Banking: What do We Know?” *Journal of Financial Intermediation* 9 (1): 7–25.

Boser, B. E., I. M. Guyon, and V. N. Vapnik. 1992. “A Training Algorithm for Optimal Margin Classifiers.” In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, USA, 27–29 August 1992*, edited by David Haussler, 144–152, New York, NY: ACM Press.

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone. 1984. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Brooks-Cole.

Brin, S., R. Motwani, and C. Silverstein. 1997. “Beyond Market Baskets: Generalizing Association Rules to Correlations.” In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Tucson, AZ, 13–15 May 1997*, edited by Joan Peckham, 265–276, New York, NY: ACM Press.

Brin, S., R. Motwani, J. D. Ullman, and S. Tsur. 1997. “Dynamic Itemset Counting and Implication Rules for Market Basket Data.” In *Proceedings of the ACM SIGMOD*

International Conference on Management of Data, Tucson, AZ, 13–15 May 1997, edited by Joan Peckham, 255–264, New York, NY: ACM Press.

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 2000. "Graph Structure in the Web." *Computer Networks* 33 (1–6): 309–320.

Buneman, P. 1997. "Semistructured Data." In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Tucson, AZ, 12–14 May 1997,* edited by Alberto O. Mendelzon and Meral Özsoyoglu, 117–121, New York, NY: ACM Press.

Buneman, P., S. Davidson, M. Fernandez, and D. Suciu. 1997. "Adding Structure to Unstructured Data." In *Database Theory—ICDT'97, Lecture Notes in Computer Science vol. 1186,* eds. F. N. Afrati and P. G. Kolaitis, 336–350. Berlin: Springer-Verlag.

Burges, C. J. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery* 2 (2): 121–167.

Cai, D., X. He, Z. Li, W-Y. Ma, and J-R. Wen. 2004. "Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information." In *Proceedings of the 12th Annual ACM International Conference on Multimedia, New York, USA, 10–16 October 2004,* edited by Henning Schulzrinne, Nevenka Dimitrova, Martina Angela Sasse, Sue B. and Rainer Lienhart, 952–959, New York, NY: ACM Press.

Calvanese, D., G. De Giacomo, and M. Lenzerini. 1999. "Modeling and Querying Semi-structured Data." *Networking and Information Systems Journal* 2: 253–273.

Cao, L., P. S. Yu, C. Zhang and Y. Zhao. 2010. *Domain Driven Data Mining.* Boston, MA: Springer.

Cao, L., and C. Zhang. 2006. "Domain-driven Actionable Knowledge Discovery in the Real World." In *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science vol. 3918,* eds. W.K. Ng, M. Kitsuregawa, J. Li and K.Chang, 821–830. Berlin: Springer-Verlag.

Cao, L., Y. Zhao, and C. Zhang. 2008. "Mining Impact: Targeted Activity Patterns in Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 20 (8): 1053–1066.

Carvalho, D., and A. Freitas. 2000. "Principles of Data Mining and Knowledge Discovery: a Genetic Algorithm-based Solution for the Problem of Small Disjuncts." In *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Artificial Intelligence vol.*

1910, eds. D. A. Zighed, J. Komorowski and J. Zytkow, 217–233. Berlin: Springer-Verlag.

Catlett, J. 1991. "Overpruning Large Decision Trees." In *Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, NSW, 24–30 August 2001,* edited by John Mylopoulos and Raymond Reiter, 764–769, San Francisco, CA: Morgan Kaufmann.

Chakrabarti, D., and C. Faloutsos. 2006. "Graph Mining: Laws, Generators, and Algorithms." *ACM Computing Surveys* 38 (1): 1–69.

Chakraborty, A. and C. X. Hu. 2006. "Lending Relationships in Line-of-credit and Nonline-of-credit Loans: Evidence from Collateral Use in Small Business." *Journal of Financial Intermediation* 15 (1): 86–107.

Chalos, P. 1985. "The Superior Performance of Loan Review Committee." *Journal of Commercial Bank Lending* 68: 60–66.

Chan, R., Q. Yang, and Y-D. Shen. 2003. "Mining High Utility Itemsets." In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003), Melbourne, FL, 19–22 December 2003,* edited by Xindong Wu, Alex Tuzhilin and Jude Shavlik, 19–26, Washington, DC: IEEE Computer Society Press.

Chandler, G. G. and J. Y. Coffman. 1979. "A Comparative Analysis of Empirical vs. Judgmental Credit Evaluation." *Financial Review* 14 (4): 23–23.

Chandler, G. N. and E. Jansen. 1992. "The Founder's Self-assessed Competence and Venture Performance." *Journal of Business Venturing* 7 (3): 223–236.

Chau, M., D. Zeng, H. Chen, M. Huang, and D. Hendriawan. 2003. "Design and Evaluation of a Multi-agent Collaborative Web Mining System." *Decision Support Systems* 35 (1): 167–183.

Chawla, N. V. 2003. "C4. 5 and Imbalanced Data Sets: Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure." In *Proceedings of the International Conference on Machine Learning '03 Workshop on Learning from Imbalanced Datasets II, Washington, DC, 21–24 August 2003,* edited by N. V. Chawla, N. Japkowicz, F. Provost and P. Turney, 1–8, USA: AAAI Press.

Chen, F-L. and F-C. Li. 2010. "Combination of Feature Selection Approaches with SVM in Credit Scoring." *Expert Systems with Applications* 37 (7): 4902–4909.

Chen, M-C. and S-H. Huang. 2003. “Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques.” *Expert Systems with Applications* 24 (4): 433–441.

Chen, M.S., J. Han, and P. S. Yu. 1996. “Data Mining: an Overview from a Database Perspective.” *IEEE Transactions on Knowledge and Data Engineering* 8 (6): 866–883.

Chen, W-D. and J-M. Li. 2009. “A Model Based on Factor Analysis and Support Vector Machine for Credit Risk Identification in Small-and-medium Enterprises.” In *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics, Baoding, China, 12–15 July 2009,* 913–918, Washington, DC: IEEE Computer Society Press.

Cheng, H., X. Yan, J. Han, and C-W. Hsu. 2007. “Discriminative Frequent Pattern Analysis for Effective Classification.” In *Proceedings of the 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007*, edited by Rada Chirkova, Asuman Dogac, M. Tamer Özsu and Timos K. Sellis, 716–725, Washington, DC: IEEE Computer Society Press.

Cheng, H., X. Yan, J. Han, and P. S. Yu. 2008. “Direct Discriminative Pattern Mining for Effective Classification.” In *Proceedings of the 24th International Conference on Data Engineering, Cancún, México, 7–12 April 2008,* edited by Gustavo Alonso, José A. Blakeley and Arbee L. P. Chen, 169–178, Washington, DC: IEEE Computer Society Press.

Chi, Y., S. Nijssen, R. R. Muntz and J. N. Kok. 2005. “Frequent Subtree Mining: an Overview.” *Fundamenta Informaticae, Special Issue on Graph and Tree Mining* 66 (1–2): 161–198.

Chi, Y., Y. Yang, Y. Xia and R. R. Muntz. 2004. “CMTreeMiner: Mining Both Closed and Maximal Frequent Subtrees.” In *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence vol. 3056,* eds. Honghua Dai, Ramakrishnan Srikant and Chengqi Zhang, 63-73. Berlin: Springer-Verlag.

Church, K. W., and P. Hanks. 1990. “Word Association Norms, Mutual Information, and Lexicography.” *Computational Linguistics* 16 (1): 22–29.

Chye, K. H., T. W. Chin and G. C. Peng. 2004. “Credit Scoring Using Data Mining Techniques.” *Singapore Management Review* 26 (2): 25–47.

Clark, P., and R. Boswell. 1991. “Rule Induction with CN2: Some Recent Improvements.” In *Machine Learning-EWSL-91, Lecture Notes in Computer Science vol. 482*, ed. Y. Kodratoff, 151–163. Berlin: Springer Verlag.

Clark, P., and T. Niblett. 1989. “The CN2 Induction Algorithm.” *Machine Learning Journal* 3 (4): 261–283.

Codd, E. F. 1970. “A Relational Model of Data for Large Shared Data Banks.” *Communications of the ACM* 13 (6): 377–387.

———. 1979. “Extending the Database Relational Model to Capture More Meaning.” *ACM Transactions on Database Systems* 4 (4): 397–434.

Cohen, J., B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton. 2009. “MAD Skills: New Analysis Practices for Big Data.” *The VLDB Endowment* 2 (2): 1481–1492.

Cohen, W. W. 1995. “Fast Effective Rule Induction.” In *Proceedings of the Twelfth International Conference of Machine Learning, Tahoe City, USA, 9–12 July 1995,* edited by Armand Prieditis and Stuart J. Russel, 115–123, San Francisco, CA: Morgan Kaufmann.

Cole, D. C. and B. F. Slade. 1998a. *Building a Modern Financial System: The Indonesian Experience*. Melbourne, Australia: Cambridge University Press.

Cole, D. C. and B. F. Slade. 1998b. “Why has Indonesia’s Financial Crisis Been so Bad?” *Bulletin of Indonesian Economic Studies* 34 (2): 61–66.

Cole, R. A. 1998. “The Importance of Relationships to the Availability of Credit.” *Journal of Banking & Finance* 22 (6–8): 959–977.

Coleman, G. B., and H. C. Andrews. 1979. “Image Segmentation by Clustering.” *The IEEE* 67 (5): 773–785.

Cook, D. J., and L. B. Holder. 1994. “Substructure Discovery using Minimum Description Length and Background Knowledge.” *Journal of Artificial Intelligence Research* 1: 231–255.

Cooper, A. C., F. J. Gimeno-Gascon and C. Y. Woo. 1994. “Initial Human and Financial Capital as Predictors of New Venture Performance.” *Journal of Business Venturing* 9 (5): 371–395.

Copestake, J. 2007. “Mainstreaming Microfinance: Social Performance Management or Mission Drift?” *World Development* 35 (10): 1721–1738”

Crawford, S. L. 1989. “Extensions to the CART Algorithm.” *International Journal of Man-Machine Studies* 31 (2): 197–217.

Crone, S. F. and S. Finlay. 2012. “Instance Sampling in Credit Scoring: an Empirical Study of Sample Size and Balancing.” *International Journal of Forecasting* 28 (1): 224–238.

Cronje, T. 2013. *11040, 10954 & Bank 22 Bank Lending*. For Curtin University and Open Universities Australia. Australia: Mc Graw Hill.

Danos, P., D. L. Holt and J. Eugene A. Imhoff. 1989. “The Use of Accounting Information in Bank Lending Decisions.” *Accounting Organizations and Society* 14 (3): 235–246.

Dasarathy, B. V. 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA, USA*:* IEEE Computer Society Press.

Dash, M. and H. Liu. 1997. “Feature Selection for Classification.” *Intelligent Data Analysis* 1 (3): 131–156.

Date, C. J. 1977. *An Introduction to Database Systems*. Reading, MA: Addison-Wesley.

Daubie, M., P. Levecq and N. Meskens. 2002. “A Comparison of the Rough Sets and Recursive Partitioning Induction Approaches: an Application to Commercial Loans.” *International Transactions in Operational Research* 9 (5): 681–694.

de la Torre, A., M. S. Martínez Pería and S. L. Schmukler. 2010. “Bank Involvement with SMEs: Beyond Relationship Lending.” *Journal of Banking & Finance* 34 (9): 2280–2293.

de Maindreville, C., and E. Simon. 1988. “A Production Rule Based Approach to Deductive Databases.” In *Proceedings of the Fourth International Conference on Data Engineering, Los Angeles, CA, 1–5 February 1988,* 234–241, Washington, DC: IEEE Computer Society Press.

Dean, J. W., and D. E. Bowen. 1994. “Management Theory and Total Quality: Improving Research and Practice through Theory Development.” *The Academy of Management Review* 19 (3): 392–418.

Degryse, H. and P. V. Cayseele. 2000. “Relationship Lending within a Bank-based System: Evidence from European Small Business Data.” *Journal of Financial Intermediation* 9 (1): 90–109.

Delen, D., D. Cogdell, and N. Kasap. 2012. “A Comparative Analysis of Data Mining Methods in Predicting NCAA Bowl Outcomes.” *International Journal of Forecasting* 28 (2): 543–552.

Derelioğlu, G., F. Gürgen and N. Okay. 2009. “A Neural Approach for SME’s Credit Risk Analysis in Turkey.” In *Machine Learning and Data Mining in Pattern Recognition,*

Lecture Notes in Artificial Intelligence vol. 5632, ed. P. Perner, 749–759. Berlin: Springer-Verlag.

Desai, V. S., J. N. Crook and G. A. Overstreet. 1996. "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment." *European Journal of Operational Research* 95 (1): 24–37.

Dikkers, H. and L. Rothkrantz. 2005. "Support Vector Machines in Ordinal Classification: an Application to Corporate Credit Scoring." *Neural Network World* 15 (6): 491–507.

Dima, A. M. and S. Vasilache. 2009. "ANN Model for Corporate Credit Risk Assessment." In *Proceedings of the International Conference on Information and Financial Engineering, Singapore, 17–20 April 2009*, 94–98, Los Alamitos, CA: IEEE Computer Society Press.

Dinh, T. H. T. and S. Kleimeier. 2007. "A Credit Scoring Model for Vietnam's Retail Banking Market." *International Review of Financial Analysis* 16 (5): 471–495.

Dinu, V., H. Zhao, and P. L. Miller. 2007. "Integrating Domain Knowledge with Statistical and Data Mining Methods for High-density Genomic SNP Disease Association Analysis." *Journal of Biomedical Informatics* 40 (6): 750–760.

Dirickx, Y. M. I. and L. Wakeman. 1976. "An Extension of the Bierman-Hausman Model for Credit Granting." *Management Science* 22 (11): 1229–1237.

Dobbs, M. 2007. "Small Business Growth: Recent Evidence and New Directions." *International Journal of Entrepreneurial Behaviour & Research* 13 (5): 296–322.

Dong, G., K. K. Lai and J. Yen. 2012. "Credit Scorecard Based on Logistic Regression with Random Coefficients." *Procedia Computer Science* 1 (1): 2463–2468.

Dong, G., and J. Li. 1999. "Efficient Mining of Emerging Patterns: Discovering Trends and Differences." In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, 15–18 August 1999*, edited by Usama M. Fayyad, Surajit Chaudhuri and David Madigan, 43–52, New York, NY: ACM Press.

Dorr, B. J. 1997. "Large-scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation." *Machine Translation* 12 (4): 271–322.

Dougherty, J., R. Kohavi, and M. Sahami. 1995. "Supervised and Unsupervised Discretization of Continuous Features." In *Proceedings of the Twelfth International Conference of Machine Learning, Tahoe City, USA, 9–12 July 1995,* edited by

Armand Prieditis and Stuart J. Russel, 194–202, San Francisco, CA: Morgan Kaufmann.

Dowla, A. and D. Alamgir. 2003. “From Microcredit to Microfinance: Evolution of Savings Products by MFIs in Bangladesh.” *Journal of International Development* 15 (8): 969–988.

Dowling, J. M. and C-F. Yap. 2008. “Indonesian Economic Development: Mirage or Miracle?” *Journal of Asian Economics* 19 (5–6): 474–485

Drummond, C. and R. C. Holte. 2003. “C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling Beats Over-sampling.” In *Proceedings of the International Conference on Machine Learning ‘03 Workshop on Learning from Imbalanced Datasets II, Washington, DC, 21–24 August 2003,* edited by N. V. Chawla, N. Japkowicz, F. Provost and P. Turney, 1–8, USA: AAAI Press.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern Classification.* 2nd ed. NY: John Wiley & Sons.

Dumais, S. T., G. W. Furnas, T. K. Landauer, S. Deerwester and R. Harshman. 1988. “Using Latent Semantic Analysis to Improve Access to Textual Information.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Washington, DC, 15–19 May 1988,* edited by J. J. O’Hare, 281–285, New York, NY: ACM Press.

Dumais, S. T., T. A. Letsche, M. L. Littman and T. K. Landauer. 1997. “Automatic Cross-language Retrieval Using Latent Semantic Indexing.” In *Proceedings of the AAAI Spring Symposium Series on Cross-language Text and Speech Retrieval, Stanford University, CA, 24–26 March 1997,* edited by J-Y. Nie, M. Simard, P. Isabelle and R. Durand,15–21, Menlo Park, CA: AAAI Press.

Durand, D. 1941. “Risk Elements in Consumer Instalment Financing.” In *Studies in Consumer Instalment Financing vol. 8*. NY: The National Bureau of Economic Research.

Edminster, R. H. 1971. “Financial Ratios and Credit Scoring for Small Business Loans.” *Journal of Commercial Bank Lending* (September): 10–23.

Edmister, R. O. 1972. “An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction.” *The Journal of Financial and Quantitative Analysis* 7 (2): 1477–1493.

Edmister, R. O. 1988. “Combining Human Credit Analysis and Numerical Credit Scoring for Business Failure Prediction.” *Akron Business and Economic Review* 19 (3): 6–14.

Eisenbeis, R. A. 1977. “Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics.” *The Journal of Finance* 32 (3): 875–900.

———. 1978. “Problems in Applying Discriminant Analysis in Credit Scoring Models.” *Journal of Banking & Finance* 2 (3): 205–219.

———. 1996. “Recent Developments in the Application of Credit-scoring Technique to the Evaluation of Commercial Loans.” *IMA Journal of Mathematics Applied in Business and Industry* 7: 271–290.

Elsas, R. and J. P. Krahnen. 1998. “Is Relationship Lending Special?” Evidence from credit-file Data in Germany.” *Journal of Banking & Finance* 22 (10–11): 1283–1316.

Embley, D. W., D. M. Campbell, R. D. Smith and S. W. Liddle. 1998. “Ontology-based Extraction and Structuring of information from Data-rich Unstructured Documents.” In *Proceedings of the Seventh International Conference on Information and Knowledge Management, Bethesda, USA, 3–7 November 1998,* edited by Georges Gardarin, James C. French, Niki Pissinou, Kia Makki and Luc Bouganim, 52–59, New York, NY: ACM Press.

Emel, A. B., M. Oral, A. Reisman and R. Yolalan. 2003. “A Credit Scoring Approach for the Commercial Banking Sector.” *Socio-Economic Planning Sciences* 37 (2): 103–123.

Enoch, C., O. Frécaut and A. Kovanen. 2003. “Indonesia’s Banking Crisis: What Happened and What Did We Learn?” *Bulletin of Indonesian Economic Studies* 39 (1): 75–92.

“European SMEs under Pressure.” 2010. In *Annual Report on EU Small and Medium-Sized Enterprises 2009*, Brussels, Belgium: European Commission.

Everett, J. and J. Watson. 1998. “Small Business Failure and External Risk Factors.” *Small Business Economics* 11 (4): 371–390.

Fantazzini, D. and S. Figini. 2009. “Random Survival Forests Models for SME Credit Risk Measurement.” *Methodology and Computing in Applied Probability* 11 (1): 29–45.

Fayyad, U. 1996a. “The KDD Process for Extracting Useful Knowledge from Volumes of Data.” *Communications of the ACM* 39 (11): 27–34.

Fayyad, U1996b. “Data Mining and Knowledge Discovery: Making Sense out of Data.” *IEEE Expert* 11 (5): 20–25.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. “From Data Mining to Knowledge Discovery in Databases.” *American Association for Artificial Intelligence* 17 (3): 37–54.

Feldman, R., and I. Dagan. 1995. “Knowledge Discovery in Textual Databases.” In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, 20–21 August 1995*, edited by Usama M. Fayyad and Ramasamy Uthurusamy, 112–117, Menlo Park, CA: AAAI Press.

Ferrucci, D. and A. Lally. 2004. “UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Eesearch Environment.” *Natural Language Engineering* 10 (3-4): 327-348.

Finlay, S. 2008. “Towards Profitability: a Utility Approach to the Credit Scoring Problem.” *The Journal of the Operational Research Society* 59 (7): 921–931.

———. 2009. “Are We Modelling the Right Thing? The Impact of Incorrect Problem Specification in Credit Scoring.” *Expert Systems with Applications* 36 (5): 9065–9071.

———. 2010. “Credit Scoring for Profitability Objectives.” *European Journal of Operational Research* 202 (2): 528–537.

Frame, W. S., A. Srinivasan and L. Woosley. 2001. “The Effect of Credit Scoring on Small-business Lending.” *Journal of Money, Credit, and Banking* 33 (3): 813–825.

Freitas, A. A. 1999. “On Rule Interestingness Measures.” *Knowledge-based Systems* 12 (5): 309–315.

Gallant, S. I. 1993. *Neural Network Learning and Expert Systems*. Cambridge, MA: MIT Press.

Gehrke, J., V. Ganti, R. Ramakrishnan, and W-Y. Loh. 1999. “Boat-optimistic Decision Tree Construction.” *ACM SIGMOD Record* 28 (2): 169–180.

Geibel, P., and F. Wysotzki. 1996. “Learning Relational Concepts with Decision Trees.” In *Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 3–6 July 1996*, edited by Lorenza Saitta, 166–174, San Francisco, CA: Morgan Kaufmann.

Geng, L., and H. Hamilton. 2006. “Interestingness Measures for Data Mining: a Survey.” *ACM Computing Surveys* 38 (3): 9–32.

———. 2007. “Choosing the right lens: Finding what is Interesting in Data Mining.” *Studies in Computational Intelligence* 43: 3–24.

Gertler, M. and S. Gilchrist. 1994. “Monetary Policy, Business Cycles and the Behavior of Small Manufacturing Firms.” *The Quarterly Journal of Economics* CIX (2): 309–340.

Glymour, C., D. Madigan, D. Pregibon and P. Smyth. 1997. “Statistical Themes and Lessons for Data Mining.” *Data Mining and Knowledge Discovery* 1: 11–28.

Godau, M., W. Hiemann and S. Jansen. 2004. “A Financial Sector Development Program: Fact-finding Mission.” Jakarta: Report to KfW-Jakarta office.

Goldberg, D. E. 1989. “Genetic Algorithms in Search, Optimization, and Machine Learning.” Reading, MA: Addison-Wesley.

Goldman, R., and J. Widom. 1997. “Dataguides: Enabling Query Formulation and Optimization in Semistructured Databases.” In *Proceedings of the 23rd International Conference on Very Large Data Bases, Athens, Greece, 25–29 August 1997,* edited by Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos and Manfred A. Jeusfeld, 436–445, San Francisco, CA: Morgan Kaufmann.

Good, I. J. 2003. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. Cambridge, MA: MIT Press.

Gool, J. V., B. Baesens, P. Sercu and W. Verbeke. 2009. “An Analysis of the Applicability of Credit Scoring for Microfinance.” In *Proceedings of the Academic and Business Research Institute Conference, Orlando, USA, 24–26 September 2009,* [online at: http://www.aabri.com/OC09manuscripts/OC09042.pdf].

Gorton, G., and A. Metrick. 2010. “Regulating the Shadow Banking System.” *Brookings Papers on Economic Activity* Fall: 261–312.

Gouda, K., and M. J. Zaki. 2001. “Efficiently Mining Maximal Frequent Itemsets.” In *Proceedings of the IEEE International Conference on Data Mining, San Jose, USA, 29 November–2 December 2001,* edited by Nick Cercone, T. Y. Lin and Xindong Wu, 163–170, Los Amitos, CA: IEEE Computer Society Press.

Grahne, G., and J. Zhu. 2003. “Efficiently Using Prefix-trees in Mining Frequent Itemsets.” In *Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, Melbourne, USA, 19 December 2003,* edited by Bart Goethals and Mohammed Javeed Zaki, 1–10, [online at:http:www.ceur-ws.org/Vol-90/].

Gray, C. and C. Mabey. 2005. “Management Development Key Differences between Small and Large Businesses in Europe.” *International Small Business Journal* 23 (5): 467–485.

Grenville, S. 2004. “What Sort of Financial Sector Should Indonesia Have?” *Bulletin of Indonesian Economic Studies* 40 (3): 307–327.

Guha, S., R. Rastogi, and K. Shim. 2001. “Cure: an Efficient Clustering Algorithm for Large Databases.” *Information Systems* 26 (1): 35–58.

Guyon, I., S. Gunn, M. Nikravesh and L. A. Zadeh, eds. 2006. *Feature Extraction: Foundations and Applications: Studies in Fuzziness and Soft Computing*. Berlin: Springer-Verlag.

Haber, S. and A. Reichel. 2007. "The Cumulative Nature of the Entrepreneurial Process: the Contribution of Human Capital, Planning and Environment Resources to Small Venture Performance." *Journal of Business Venturing* 22 (1): 119–145.

Hadzic, F. 2012. "A Structure Preserving Flat Data Format Representation for Tree-structured Data." In *New Frontiers in Applied Data Mining, Lecture Notes in Artificial Intelligence vol. 7104*, eds. L. Cao, J. Z. Huang, J. Bailey, Y. S. Koh and J. Luo, 221–233. Berlin: Springer-Verlag.

Hadzic, F., H. Tan and T. Dillon. 2011. "Mining of Data with Complex Structures." *Studies in Computational Intelligence Series vol. 333*. Berlin: Springer-Verlag.

Hairuddin, H., N. L. M. Noor and A. M. A. Malik. 2012 "Why Do Microenterprises Refuse to Use Information Technology: a Case of Batik Microenterprises in Malaysia." *Procedia – Social and Behavioral Sciences* 57: 494–502.

Hamada, M. 2003. *Transformation of the Financial Sector in Indonesia*. IDE Research Paper No. 6. Chiba, Japan: Institute of Developing Economies.

Han, J., M. Kamber and J. Pei. 2012. *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA: Morgan Kaufmann.

Han, J., J. Pei, and Y. Yin. 2002. "Mining Frequent Patterns without Candidate Generation." In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, USA, 16–18 May 2002*, edited by Wei-Dong Chen, Jeffrey F. Naughton and Philip A. Bernstein, 1–12, New York, NY: ACM Press.

Hand, D. J. 1998. "Data Mining: Statistics and More?" *The American Statistician* 52 (2): 112.

Hand, D. J. and W. E. Henley. 1997. "Statistical Classification Methods in Consumer Credit Scoring: a Review." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 160 (3): 523–541.

Handschuh, S., S. Staab and A. Maedche. 2001. "CREAM: Creating Relational Metadata with a Component-based, Ontology-driven Annotation Framework." In *Proceedings of the 1st International Conference on Knowledge Capture, Victoria, Canada, 21–23 October 2001*, edited by Siegfried Handschuh, Rose Dieng and Steffen Staab, 76–83, New York, NY: ACM Press.

Hao, Y., Z. Chi, D. Yan and X. Yue. 2012. “An Improved Fuzzy Support Vector Machine for Credit Rating.” In *Network and Parallel Computing*, eds. K. Li, C. Jesshope, H. Jin and J-L. Gaudiot, 495–505. Berlin: Springer-Verlag.

Harter, T. R. 1974. “Potentials of Credit Scoring: Myth or Fact.” *Credit and Financial Management* 76: 27–28.

Hashimoto, K., I. Takigawa, M. Shiga, M. Kanehisa, and H. Mamitsuka. 2008. “Mining significant tree patterns in carbohydrate sugar chains.” *Bioinformatics* 24 (16): i167–i173.

Hassler, H. W., J. H. Myers and M. Seldin. 1963. “Payment History as a Predictor of Credit Risk.” *Journal of Applied Psychology* 47 (6): 383–385.

He, H., and E. A. Garcia. 2009. “Learning from Imbalanced Data.” *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–1284.

He, Z., X. Xu, and S. Deng. 2003. “Discovering Cluster-based Local Outliers.” *Pattern Recognition Letters* 24 (9–10): 1641–1650.

Hecht-Nielsen, R. 1988. “Neurocomputing: Picking the Human Brain.” *IEEE Spectrum* 25 (3): 36–41.

Heckerman, D. 1996. “Bayesian Networks for Knowledge Discovery.” *Advances in Knowledge Discovery and Data Mining* 11: 273–305.

Henley, W. E. and D. J. Hand. 1996. “A K-nearest-neighbour Classifier for Assessing Consumer Credit Risk.” *Journal of the Royal Statistical Society. Series D (The Statistician)* 45 (1): 77–95.

Hens, A. B. and M. K. Tiwari. 2012. “Computational Time Reduction for Credit Scoring: an Integrated Approach Based on Support Vector Machine and Stratified Sampling Method.” *Expert Systems with Applications* 39 (8): 6774–6781.

Herodotou, H., H. Lim, G. Luo, N. Borisov, L. Dong, F. B. Cetin and S. Babu. 2011. “Starfish: a Self-tuning System for Big Data Analytics.” In *Proceedings of the Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, USA, 9–12 January 2011*, 261–272, [online at: http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper36.pdf].

Hilderman, R., and H. Hamilton. 2002. “Applying Objective Interestingness Measures in Data Mining Systems.” *Principles of Data Mining and Knowledge Discovery*: 432–439.

———. 2001. "*Knowledge Discovery and Measures of Interest*." Boston, MA: Kluwer Academic Publishers.

Hoffmann, F., B. Baesens, J. Martens, F. Put and J. Vanthienen. 2002. "Comparing a Genetic Fuzzy and a Neurofuzzy Classifier for Credit Scoring." *International Journal of Intelligent Systems* 17 (11): 1067–1083.

Hoffmann, F., B. Baesens, C. Mues, T. Van Gestel and J. Vanthienen. 2007. "Inferring Descriptive and Approximate Fuzzy Rules for Credit Scoring Using Evolutionary Algorithms." *European Journal of Operational Research* 177 (1): 540–555.

Holland, J., ed. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Holmes, G., A. Donkin and I. H. Witten. 1994. "Weka: a Machine Learning Workbench. In *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, Qld, 29 November–2 December 1994,* 357–361, Piscataway, NJ: IEEE Computer Society Press.

Hong, J., I. Mozetic, and R. S. Michalski. 1986. *AQ15: Incremental Learning of Attribute-based Descriptions from Examples, the Method and User's Guide*. Urbana, IL: Department of Computer Science, University of Illinois.

Hormozi, A. M. 2004. "Data Mining: a Competitive Weapon for Banking and Retail Industries." *Information Systems Management* 21 (2): 62.

Hsu, C-C., and Y-C. Chen. 2007. "Mining of Mixed Data with Application to Catalog Marketing." *Expert Systems with Applications* 32 (1): 12–23.

Huang, C-L., M-C. Chen and C-J. Wang. 2007. "Credit Scoring with a Data Mining Approach Based on Support Vector Machines." *Expert Systems with Applications* 33 (4): 847–856.

Huang, J-J., G-H. Tzeng and C-S. Ong. 2006. "Two-stage Genetic Programming (2SGP) for the Credit Scoring Model." *Applied Mathematics and Computation* 174 (2): 1039–1053.

Huang, Z. 1998. "Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values." *Data Mining and Knowledge Discovery* 2 (3): 283–304.

Hulme, D. 2002. "Is Microdebt Good for Poor People? a Note on the Dark Side of Microfinance." *Small Enterprise Development* 11 (1): 26–28.

Ikasari, N. and F. Hadzic. 2012. "An Assessment on Loan Performance from Combined Quantitative and Qualitative Data in XML." In *Discovery Science, Lecture Notes in*

Artificial Intelligence vol. 7569, eds. J-G. Ganascia, P. Lenca and J-M. Petit, 268–283. Berlin: Springer-Verlag.

Ikasari, N., F. Hadzic and T. S. Dillon. 2011. "Incorporating Qualitative Information for Credit Risk Assessment through Frequent Subtree Mining for XML." In *XML Data Mining: Models, Method, and Applications,* ed. A. Tagarelli, 467–503. Hershey, PA: IGI Global.

Ince, H. and B. Aktan. 2009. "A Comparison of Data Mining Techniques for Credit Scoring in Banking: a Managerial Perspective." *Journal of Business Economics and Management* (3): 233.

Inderst, R. and H. M. Mueller. 2007. "A Lender-based Theory of Collateral." *Journal of Financial Economics* 84 (3): 826–859.

Inokuchi, A., T. Washio, and H. Motoda. 2003. "Complete Mining of Frequent Patterns from Graphs: Mining Graph Data." *Machine Learning* 50 (3): 321–354.

Isern, J., L. B. Prakash, A. Pillai, S. Hashemi, R. P. Christen, G. J. Ivatury and R. Rosenberg. 2007. *Sustainability of Self-help Groups in India: Two Analyses.* Occasional Paper No. 12, Washington, DC: Consultative Group to Assist the Poor.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone and M. S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2 (3): 841–860.

Jain, A. K., and P. J. Flynn. 1996. *Image Segmentation Using Clustering*. Piscataway, NJ: IEEE Computer Society Press.

Jain, AK., M. N. Murty, and P. J. Flynn. 1999. "Data Clustering: a Review." *ACM Computing Surveys* 31 (3): 264–323.

Jarvis, R. 2002. "The Use of Quantitative and Qualitative Criteria in the Measurement of Performance in Small Firms." *Journal of Small Business and Enterprise Development* 7 (2): 123–134.

Jensen, H. L. 1992. "Using Neural Networks for Credit Scoring." *Managerial Finance* 18 (6): 15–26.

Jie, C., J. Huidong, H. Hongxing, D. McAullay, C. M. O'Keefe, R. Sparks, and C. Kelman. 2012. "Mining Consequence Events in Temporal Health Data." *Intelligent Data Analysis* 14 (2): 245–261.

Jiménez, G., V. Salas and J. Saurina. 2006. "Determinants of Collateral." *Journal of Financial Economics* 81 (2): 255–281.

Jones, N. C., and P. A. Pevzner. 2004. *An Introduction to Bioinformatics Algorithms.* Boston, MA: MIT Press.

Joos, P., K. Vanhoof, H. Ooghe and N. Sierens. 1998. "Credit Classification: a Comparison of Logit Models and Decision Trees." In *Proceedings of the Workshop on Application of Machine Learning and Data Mining in Finance, 10th European Conference on Machine Learning, Chemnitz, Germany, 21–24 April 1998,* edited by Gholamreza Nakhaeizadeh, 59–72, London, UK: Springer-Verlag.

Joshi, A., and Z. Jiang. 2002. "Retriever: Improving Web Search Engine Results Using Clustering." In *Managing Business with Electronic Commerce: Issues and Trends,* ed. Aryya Gangopadhyay, 59–81. Hershey, PA: IGI Global.

Kah, J. M. L., D. L. Olds and M. M. O. Kah. 2005. "Microcredit, Social Capital, and Politics: the Case of a Small Rural Town—Gossas, Senegal." *Journal of Microfinance* 7 (1): 119–149.

Kamath, C. 2003. "Mining Science and Engineering Data." In *The Handbook of Data Mining*, ed. N. Ye, 549–572. Mahwah, NJ: Lawrence Erlbaum.

Kangasharju, A. 2002. "Growth of the Smallest: Determinants of Small Firm Growth during Strong Macroeconomic Fluctuations." *International Small Business Journal* 19 (1): 28–43.

Kantardzic, M. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms.* Piscataway, NJ: Wiley-IEEE Press.

Karunaratne, N. D. 1999. "The Asian Miracle and Crisis: Rival Theories, the IMF Bailout and Policy Lessons." *Intereconomics* January/February: 19–26.

Kaufman, L., and P. J. Rousseeuw. 1992. *Finding Groups in Data: An Introduction to Cluster Analysis.* NY: John Wiley & Sons.

Kaye, R. 2006. "The Gloss System for Transformations from Plain Text to XML. Refereed paper for the Proceedings of MathUI 2006 at http://www.activemath.org/`paul/MathUI06/.

Kenward, L. R. 1999. "Assesing Vulnerability to Financial Crisis: Evidence from Indonesia." *Bulletin of Indonesian Economic Studies* 35 (3): 71–95.

Khan, A. M. and V. Manopichetwattana. 1989. "Innovative and Noninnovative Small Firms: Types and Characteristics." *Management Science* 35 (5): 597–606.

Kim, H. S. and S. Y. Sohn. 2012. “Support Vector Machines for Default Prediction of SMEs Based on Technology Credit.” *European Journal of Operational Research* 201 (3): 838–846.

Kirkos, E., C. Spathis, and Y. Manolopoulos. 2007. “Data Mining Techniques for the Detection of Fraudulent Financial Statements.” *Expert Systems with Applications* 32 (4): 995–1003.

Klapper, L. F., V. Sarria-Allende and R. Zaidi. 2006. “A firm-level Analysis of Small and Medium Size Enterprise Financing in Poland.” In *World Bank Policy Research Working Paper 3984*. Washington, DC: World Bank.

Klein, D. and C. D. Manning. 2003. “Accurate Unlexicalized Parsing.” In *Proceedings of the 41st Meeting of the Association for Computational Linguistics, Sapporo, Japan, 7–12 July 2003*, edited by Erhard W. Hinrichs and Dan Roth, 423–430, Stroudsburg, PA: Association for Computational Linguistics.

Knorr, E. M., and R. T. Ng. 1998. “Algorithms for Mining Distance-based Outliers in Large Datasets.” In *Proceedings of the Twenty-fourth International Conference on Very Large Data Bases, New York, USA, 24–27 August 1998,* edited by Ashish Gupta, Oded Shmueli and Jennifer Widom, 392–403, San Francisco, CA: Morgan Kaufmann.

Knorr, E. M., R. T. Ng, and V. Tucakov. 2002. “Distance-based Outliers: Algorithms and Applications.” *The VLDB Journal* 8 (3): 237–253.

Kohavi, R., and F. Provost. 1998. “Glossary of Terms.” *Machine Learning* 30 (2–3): 271–274.

Kohavi, R., and M. Sahami. 1996. “Error-based and Entropy-based Discretization of Continuous Features.” In *Proceedings of the Second International Conference on Knowledge Discovery and Data* Mining, *Portland, USA, 2–4 August 1996,* edited by Evangelos Simoudis, Jiawei Han and Usama M. Fayyad, 114–119, Menlo Park, CA: AAAI Press.

Kotey, B. and G. G. Meredith. 1997. “Relationships among Owner/Manager Personal Values, Business Strategies, and Enterprise Performance.” *Journal of Small Business Management* 35: 37–64.

Koza, J. R. 1994. “Genetic Programming as a Means for Programming Computers by Natural Selection.” *Statistics and Computing* 4 (2): 87–112.

Lahsasna, A., R. N. Ainon and T. Y. Wah. 2012. “Credit Scoring Models using Soft Computing Methods: a Survey.” *The International Arab Journal of Information Technology* 7 (2): 115–123.

Lai, K., L. Yu, S. Wang and L. Zhou. 2006a. “Neural Network Metalearning for Credit Scoring.” In *Intelligent Computing,* eds. D-S. Huang, K. Li and G. Irwin, 403–408. Berlin: Springer-Verlag.

———. 2006b. “Credit Risk Analysis Using a Reliability-based Neural Network Ensemble Model.” In *Artificial Neural Networks, Lecture Notes in Computer Science vol. 4132,* eds. S. Kollias, A. Stafylopatis, W. Duch and E. Oja, 682–690. Berlin: Springer-Verlag.

Lai, K. K., L. Yu, L. Zhou and S. Wang, eds. 2006. “Credit Risk Evaluation with Least Square Support Vector Machine.” In *Rough Sets and Knowledge Technology, Lecture Notes in Computer Science vol. 4062,* ed. G. Wang, 490-495. Berlin: Springer-Verlag.

Lallich, S., O. Teytaud, and E. Prudhomme. 2007. “Association Rule Interestingness: Measure and Statistical Validation.” In *Quality Measures in Data Mining,* eds. F. Guillet, and H. J. Hamilton, 251–275. Berlin: Springer-Verlag.

Lavrač, N., P. Flach, and B. Zupan. 1999. “Rule Evaluation Measures: a Unifying View.” *Inductive Logic Programming* 99: 174–185.

Lee, T-S. and I. F. Chen. 2005. “A two-stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines.” *Expert Systems with Applications* 28 (4): 743–752.

Lee, T-S., C-C. Chiu, C-J. Lu and I. F. Chen. 2002. “Credit Scoring Using the Hybrid Neural Discriminant Technique.” *Expert Systems with Applications* 23 (3): 245–254.

Lee, Y-C. 2007. “Application of Support Vector Machines to Corporate Credit Rating Prediction.” *Expert Systems with Applications* 33 (1): 67–74.

Lefebvre, E. and L. A. Lefebvre. 1992. “Firm Innovativeness and CEO Characteristics in Small Manufacturing Firms.” *Journal of Engineering and Technology Management* 9 (3–4): 243–277.

Lehmann, E. and D. Neuberger. 2001. Do Lending Relationships Matter?” Evidence from Bank Survey Data in Germany.” *Journal of Economic Behavior & Organization* 45 (4): 339–359.

Lenca, P., B. Vaillant, P. Meyer, and S. Lallich. 2007. “Association Rule Interestingness Measures: Experimental and Theoretical Studies.” *Quality Measures in Data Mining*: 51–76.

Lengnick-Hall, C. A. 1996. “Customer Contributions to Quality: a Different View of the Customer-oriented Firm.” *The Academy of Management Review* 21 (3): 791–824.

Lent, B., R. Agrawal, and R. Srikant. 1997. “Discovering Trends in Text Databases.” In *Proceedings of the Third International Conference on Knowledge Discovery and Data* Mining, *Newport Beach, USA, 14–17 August 1997,* edited by David Heckerman, Heikki Mannila and Daryl Pregibon, 227–230, Menlo Park, CA: AAAI Press.

Leung, D. and L. Rispoli. 2011. *The Contribution of Small and Medium-sized Businesses to Gross Domestic Product: a Canada–United States Comparison.* Ottawa: Statistics Canada.

Li, J., H. Shen, and R. Topor. 2002. “Mining the Optimal Class Association Rule Set.” *Knowledge-Based Systems* 15 (7): 399–405.

Li, L., H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark. 2004. “Data Mining Techniques for Cancer Detection Using Serum Proteomic Profiling.” *Artificial Intelligence in Medicine* 32 (2): 71–83.

Li, R-Z., S-L. Pang and J-M. Xu. 2002. “Neural Network Credit-Risk Evaluation Model Based on Backpropagation Algorithm.” In *Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, China, 4–5 November 2002,* 1702–1706, Piscataway, NJ: IEEE Computer Society Press.

Li, S-T., W. Shiue and M-H. Huang. 2006. “The evaluation of Consumer Loans using Support Vector Machines.” *Expert Systems with Applications* 30 (4): 772–782.

Liddle, R. W. 1991. “The Relative Autonomy of the Third World Politician: Soeharto and Indonesian Economic Development in Comparative Perspective.” *International Studies Quarterly* 35 (4): 403–427.

Limsombunchai, V., C. Gan and M. Lee. 2005. “An Analysis of Credit Scoring for Agricultural Loans in Thailand.” *American Journal of Applied Sciences* 2 (8): 1198–1205.

Lindgren, C-J., T. J. T. Baliño, C. Enoch, A-M. Gulde, M. Quintyn and L. Teo. 2002. *Financial Sector Crisis and Restructuring: Lessons from Asia.* Occasional paper 188. Washington, DC: International Monetary Fund.

Ling, C. X., T. Chen, Q. Yang, and J. Cheng. 2002. “Mining Optimal Actions for Profitable CRM.” In *Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 9–12 December 2002,* edited by Vipin Kumar, Shusaku Tsumoto, Ning Zhong, Philip S. Yu and Xindong Wu, 767–770, Los Alamitos, CA: IEEE Computer Society Press.

Liu, B., W. Hsu, and S. Chen. 1997. “Using General Impressions to Analyze Discovered Classification Rules.” In *Proceedings of the Third International Conference on*

Knowledge Discovery and Data Mining, *Newport Beach, USA, 14–17 August 1997*, edited by David Heckerman, Heikki Mannila and Daryl Pregibon, 31–36, Menlo Park, CA: AAAI Press.

Liu, B., W. Hsu, and Y. Ma. 1998. "Integrating Classification and Association Rule Mining." In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27–31 August 1998,* edited by Rakesh Agrawal, Paul E. Stolorz and Gregory Piatetsky-Shapiro, 80–86, Menlo Park, CA: AAAI Press.

Liu, B., W. Hsu, L-F. Mun, and H-Y. Lee. 1999. "Finding Interesting Patterns Using User Expectations." *IEEE Transactions on Knowledge and Data Engineering* 11 (6): 817–832.

Liu, H-H., and C-S. Ong. 2008. "Variable Selection in Clustering for Marketing Segmentation Using Genetic Algorithms." *Expert Systems with Applications* 34 (1): 502–510.

Loh, W-Y., and N. Vanichsetakul. 1988. "Tree-structured Classification via Generalized Discriminant Analysis." *Journal of the American Statistical Association* 83 (403): 715–725.

Longenecker, J. G., C. W. Moore and J. W. Petty. 1997. "*Credit Scoring and the Small Business: A Review and the Need for Research."* San Francisco, CA: United States Association for Small Business and Entrepreneurship.

Lowe, D. G. 2001. "Local Feature View Clustering for 3D Object Recognition." In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, USA, 8–14 December 2001,* 682–688, Los Alamitos, CA: IEEE Computer Society Press.

Lu, S., H. Hu, and F. Li. 2001. "Mining Weighted Association Rules." *Intelligent Data Analysis* 5 (3): 211–226.

Luo, D., L. Cao, C. Luo, C. Zhang, and W. Wang. 2008. "Towards Business Interestingness in Actionable Knowledge Discovery." In *Applications of Data Mining in E-Business and Finance*, eds. C. Soares, Y. Peng, J. Meng, T. Washio and Z.-H. Zhou, 99-109. Netherlands: IOS Press

Malhotra, R. and D. K. Malhotra. 2003. "Evaluating Consumer Loans Using Neural Networks." *Omega* 31 (2): 83–96.

Man, T. W. Y., T. Lau and K. F. Chan. 2002. “The Competitiveness of Small and Medium Enterprises: a Conceptualization with Focus on Entrepreneurial Competencies.” *Journal of Business Venturing* 17 (2): 123–142.

Mannila, H. 1997. “Discovery of Frequent Episodes in Event Sequences.” *Data Mining and Knowledge Discovery* 1 (3): 259–289.

Martens, D., B. Baesens, T. V. Gestel and J. Vanthienen. 2007. “Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines.” *European Journal of Operational Research* 183 (3): 1466–1476.

Matheus, C. J., P. K. Chan, and G. Piatetsky-Shapiro. 1993. “Systems for Knowledge Discovery in Databases.” *IEEE Transactions on Knowledge and Data Engineering* 5 (6): 903–913.

McAfee, A. and E. Brynjolfsson. 2012. “Big Data: the Management Revolution.” *Harvard Business Review* 90 (10): 60-66.

Mehta, M., R. Agrawal, and J. Rissanen. 1996. “SLIQ: A Fast Scalable Classifier for Data Mining.” In *Advances in Database Technology—EDBT'96, Lecture Notes in Computer Science vol. 1057,* eds. G. Gardarin, M. Bouzeghoub and P. Apers, 18–32. Berlin: Springer-Verlag.

Mei, Q., and C. X. Zhai. 2005. “Discovering Evolutionary Theme Patterns from Text: an Exploration of Temporal Text Mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, USA, 21–24 August 2005*, edited by Robert Grossman, Roberto J. Bayardo and Kristin P. Bennett, 198–207, New York, NY: ACM Press.

Mendelzon, A. O., and P. T. Wood. 1995. “Finding Regular Simple Paths in Graph Databases.” *SIAM Journal on Computing* 24 (6): 1235–1258.

Menkhoff, L., D. Neuberger and C. Suwanaporn. 2006. “Collateral-based Lending in Emerging Markets: Evidence from Thailand.” *Journal of Banking & Finance* 30 (1): 1–21.

Michalski, R. S. 1986. “Understanding the Nature of Learning: Issues and Research Directions.” *Machine Learning: An Artificial Intelligence Approach* 2: 3–25.

Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz and T. Euler. 2006. “YALE: Rapid Prototyping for Complex Data Mining Tasks.” In *Proceedings of the welfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 20–23 August 2006,* edited by Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven and Dimitrios Gunopulos, 935–940, New York, NY: ACM Press.

Miller, D. and J-M. Toulouse. 1986. “Chief Executive Personality and Corporate Strategy and Structure in Small Firms.” *Management Science* 32 (11): 1389–1389.

Mitchell, T. M. 1997. *Machine Learning*. NY: McGraw-Hill.

Miyahara, T., T. Shoudai, T. Uchida, K. Takahashi and H. Ueda. 2001. “Discovery of Frequent Tree Structured Patterns in Semistructured Web Documents. In *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence vol. 2035,* eds. D. Cheung, G. J. Williams and Q. Li, 47–52. Berlin: Springer-Verlag.

Mramor, D. and A. Valentincic. 2003. “Forecasting the Liquidity of Very Small Private Companies.” *Journal of Business Venturing* 18 (6): 745–771.

Myers, J. H. and E. W. Forgy. 1963. “The Development of Numerical Credit Evaluation Systems.” *Journal of the American Statistical Association* 58 (303): 799–806.

Nahm, U. Y., and R. J. Mooney. 2001. “Mining Soft-matching Rules from Textual Data.” In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, USA, 4–10 August 2001,* edited by Berhard Nebel, 979–986, San Francisco, CA: Morgan Kaufmann.

Nasukawa, T. 2001. “Text Analysis and Knowledge Mining System.” *IBM Systems Journal* 40 (4): 967.

Nasution, A. 2012. “Indonesia Imposes New Bank Ownership Caps.” *East Asia Forum: Economics, Politics and Public Policy in East Asia and the Pacific*. http://www.eastasiaforum.org/2012/08/24/indonesia-imposes-new-bank-ownership-caps (Accessed 26 November 2012).

Naudé, W. 2009. *The Financial Crisis of 2008 and the Developing Countries*. UNU-WIDER Project on New Directions in Development Economics Discussion Paper No. 2009/01. Helsinki, Finland: World Institute for Development Economics Research.

“The New SME Definition: User Guide and Model Declaration.” 2004. Brussels, Belgium: European Commission.

Ngai, E. W. T., L. Xiu, and D. C. K. Chau. 2009. “Application of Data Mining Techniques in Customer Relationship Management: a Literature Review and Classification.” *Expert Systems with Applications* 36 (2– 2): 2592–2602.

Nichter, S. and L. Goldmark. 2009. “Small Firm Growth in Developing Countries.” *World Development* 37 (9): 1453–1464.

Ohsaki, M., S. Kitaguchi, K. Okamoto, H. Yokoi, and T. Yamaguchi. 2004. “Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis.” In *Knowledge*

Discovery in Databases: PKDD 2004, Lecture Notes in Computer Science vol. 3202, eds. J.-F. Boulicaut, F. Esposito, F. Giannotti and D. Pedreschi, 362–373. Berlin: Springer-Verlag.

Ong, C-S., J-J. Huang and G-H. Tzeng. 2005. "Building Credit Scoring Models using Genetic Programming." *Expert Systems with Applications* 29 (1): 41–47.

Ono, A. and I. Uesugi. 2005. The Role of Collateral and Personal Guarantees in Relationship Lending: Evidence from Japan's Small Business Loan Market. RIETI Discussion Paper Series 05-E-027.

Orgler, Y. E. 1972. "A Credit Scoring Model for Commercial Loans." *Journal of Money, Credit and Banking* 2 (4): 435–445.

Ortiz-Molina, H. and M. F. Penas. 2008. "Lending to Small Businesses: the Role of Loan Maturity in Addressing Information Problems." *Small Business Economics* 30 (4): 361–383.

Ou, C. and G. W. Haynes. 2006. "Acquisition of Additional Equity Capital by Small Firms: Findings from the National Survey of Small Business Finances." *Small Business Economics* 27 (2-3): 157–168.

Ou, Y., L. Cao, C. Luo, and C. Zhang. 2008. "Domain-driven Local Exceptional Pattern Mining for Detecting Stock Price Manipulation." In *PRICAI 2008: Trends in Artificial Intelligence, Lecture Notes in Computer Science vol. 5351*, eds. T.-B. Ho and Z.-H. Zhou, 849–858. Berlin: Springer-Verlag.

Our history. http://www.sba.gov/about-sba-services/our-history.

Padmanabhan, B., and A. Tuzhilin. 1998. "A Belief-driven Method for Discovering Unexpected Patterns." In *Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 27–31 August 1998,* Rakesh Agrawal, Paul E. Stolorz and Gregory Piatetsky-Shapiro, 94–100, Menlo Park, CA: AAAI Press.

———. 2000. "Small Is Beautiful: Discovering the Minimal Set of Unexpected Patterns." In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, USA, 20–23 August 2000,* edited by Raghu Ramakrishnan, Salvatore J. Stolfo, Roberto J. Bayardo and Ismail Parsa, 54–63, New York, NY: ACM Press.

Pang, S-L., Y-M. Wang and Y-H. Bai. 2002. "Credit Scoring Model Based on Neural Network." In *Proceedings of the First International Conference on Machine Learning*

and Cybernetics, Beijing, China, 4–5 November 2002, 1742–1746, Piscataway, NJ: IEEE Computer Society Press.

Park, J. S., M-S. Chen, and P. S. Yu. 1995. "An Effective Hash-based Algorithm for Mining Association Rules." In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, USA, 22–25 May 1995,* edited by Michael J. Carey and Donovan A. Scheider, 175–186, New York, NY: ACM Press.

Patten, R. H. and J. K. Rosengard. 1992. *Progress with Profits: The Development of Rural Banking in Indonesia*. Cambridge, MA: Harvard Institute for International Development.

Patten, R. H., J. K. Rosengard and J. R. D. E. Johnston. 2001. "Microfinance Success amidst Macroeconomic Failure: the Experience of Bank Rakyat Indonesia during the East Asian Crisis." *World Development* 29 (6): 1057–1069.

Pawlak, Z. 1995. "Rough Sets: Theoretical Aspects of Reasoning about Data." In *Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science, Nashville, USA, 28 February–2 March 1995,* edited by C. Jinshong Hwang and Betty W. Hwang, 262–264, New York, NY: ACM Press.

Pei, J., J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang. 2001. "H-mine: Hyper-structure Mining of Frequent Patterns in Large Databases." In *Proceedings of the 2001 International Conference on Data Mining, San Jose, USA, 29 November–2 December 2001,* edited by Nick Cercone, Tsau Young Lin and Xindong Wu, 441–448, Los Alamitos, CA: IEEE Computer Society Press.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M-C. Hsu. 2004. "Mining Sequential Patterns by Pattern-growth: the Prefixspan Approach." *IEEE Transactions on Knowledge and Data Engineering* 16 (11): 1424–1440.

Pemerintah Republik Indonesia. 1998. "Undang-Undang Republik Indonesia Nomor 10 Tentang Perbankan" [Banking Law Number 10]. Jakarta, Indonesia: Government of Indonesia.

Pemerintah Republik Indonesia. 2004. "Undang Undang Republik Indonesia Nomor 24 Tahun 2004 Tentang Lembaga Penjamin Simpanan" [Law Number 24 Year 2004 on Deposit Guarantee Institution]. Jakarta, Indonesia: Government of Indonesia.

Peng, Y., G. Kou, Y. Shi and Z. Chen. 2008. "A Multi-criteria Convex Quadratic Programming Model for Credit Data Analysis." *Decision Support Systems* 44 (4): 1016–1030.

Petersen, M. A. and R. G. Rajan. 1994. “The Benefits of Lending Relationships: Evidence from Small Business Data.” *The Journal of Finance* 49 (1): 3–37.

Piatetsky-Shapiro, G. 1991a. “Discovery, Analysis, and Presentation of Strong Rules.” In *Knowledge Discovery in Databases*, ed. G. Piatetsky-Shapiro and W. Frawley, 229–238. Cambridge, MA: MIT Press.

———. 1991b. Knowledge Discovery in Real Databases: a Report on the IJCAI-89 Workshop.” *The AI Magazine* 11 (5): 68–70.

Piramuthu, S. 1999. “Financial Credit-risk Evaluation with Neural and Neurofuzzy Systems.” *European Journal of Operational Research* 112 (2): 310–321.

Piton, T., J. Blanchard, H. Briand, and F. Guillet. 2009. “Domain Driven Data Mining to Improve Promotional Campaign ROI and Select Marketing Channels.” In *Proceedings of the 18th ACM Conference on Information and Knowledge Managemen, Hong Kong, China, 2–6 November 2009,* edited by David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu and Jimmy J. Lin, 1057–1066, New York, NY: ACM Press.

Plaut, S. E. 1985. “The Theory of Collateral.” *Journal of Banking & Finance* 9 (3): 401–419”

“Principles for the Management of Credit Risk.” 2000. Basel, Switzerland: Basel Committee on Banking Supervision.

Quinlan, R. 1979. “Discovering Rules by Induction from Large Collections of Examples.” In *Expert Systems in the Micro Electronic Age*, ed. D. Michie, 168–201. Edinburgh: Edinburgh University Press.

———. 1983. “Learning Efficient Classification Procedures and Their Application to Chess End-games.” *Machine Learning: An Artificial Intelligence Approach* 1: 463–482.

———. 1986. “Induction of Decision Trees.” *Machine Learning* 1 (1): 81–106.

———. 1993. “*C4.5: Programs for Machine Learning*.” San Mateo, CA: Morgan Kaufmann.

Reichert, A. K., C-C. Cho and G. M. Wagner. 1983. “An Examination of the Conceptual Issues Involved in Developing Credit-scoring Models.” *Journal of Business & Economic Statistics* 1 (2): 101–114.

“Rencana Strategis Kementerian Koperasi Dan Usaha Kecil Dan Menengah Tahun 2010 – 2014” [Strategic Plan of Ministry of Cooperatives and Small Medium Enterprises Year 2010–2014]. 2010. Jakarta, Indonesia: Kementerian Koperasi dan Usaha Kecil dan Menengah.

Richbell, S. M., H. D. Watts and P. Wardle. 2006. “Owner-managers and Business Planning in the Small Firm.” *International Small Business Journal* 24 (5): 496–514.

Rosenberg, E. and A. Gleit. 1994. “Quantitative Methods in Credit Management: a Survey.” *Operations Research* 42 (4): 589–613.

Rosenkrantz, W. A. 1997. Introduction to Probability and Statistics for Scientists and Engineers. NY: McGraw-Hill.

Rudjito. 2003. “Financing Challenges of SMEs from the Policy Perspective.” Paper presented at *the 2nd Annual Conference of PECC Finance Forum: Issues and Challenges for Regional Financial Cooperation in the Asia-Pacific* at Hua Hin, Thailand: Finance Forum Pacific Economic Cooperation Council (PECC).

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. “Learning Internal Representation by Error Propagation.” In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition volume 1: Foundations*, eds. D. E. Rumelhart and J. L. McClelland, 1: 318–362. Cambridge, MA: MIT Press.

Sahar, S. 1999. “Interestingness via What is Not Interesting.” In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, 15–18 August 1999*, edited by Usama M. Fayyad, Surajit Chaudhuri and David Madigan, 332–336, New York, NY: ACM Press.

Salas, H. A., J. Azé, S. Bringay, F. Flouvat, N. Selmaoui-Folcher, F. Cernesson, and M. Teisseire. 2012. “Finding Relevant Sequences with the Least Temporal Contradiction Measure: Application to Hydrological Data.” Paper presented at *the AGILE’2012 International Conference on Geographic Information Science*, Avignon, France, 24–27 April 2012.

Savasere, A., E. Omiecinski, and S. Navathe. 1995. “An Efficient Algorithm for Mining Association Rules in large Databases.” In *VLDB’94: Proceedings of the 22nd International Conference on Very Large Data Bases, Zurich, Swizerland, 11–15 September 1995*, edited by Umeshwar Dayal, Peter M. D. Gray and Shojiro Nishio, 432–444, San Francisco, CA: Morgan Kaufmann.

Schaffernicht, E., C. Möller, K. Debes and H-M. Gross. 2009. “Forward Feature Selection Using Residual Mutual Information.” In *Proceedings of the 17th European Symposium on Artificial Neural Networks – Advances in Computational Intelligence and Learning, Bruges, Belgium, 22–24 April 2009*, 583–588, [online at https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2009-43.pdf].

Schebesch, K. and R. Stecking. 2005. “Support Vector Machines for Credit Scoring: Extension to Non Standard Cases.” In *Innovations in Classification, Data Science, and Information Systems*, eds. D. Baier and K-D. Wernecke, 498–505. Berlin: Springer-Verlag.

Schiffer, M. and B. Weder. 2001. *Firm Size and the Business Environment: Worldwide Survey Results*. Discussion Paper number 43. Washington, DC: The World Bank and International Finance Corporation.

Schreiner, M. 2002. *Scoring: The Next Breakthrough in Microcredit?”* St. Louis, MI: Consultative Group to Assist the Poorest.

Schutz, R. R. 1951. “On the Measurement of Income Inequality.” *The American Economic Review* 41: 107–122.

Sebag, M., and M. Schoenauer. 1988. “Generation of Rules with Certainty and Confidence Factors from Incomplete and Incoherent Learning Bases.” In *Proceedings of the 2nd European Knowledge Acquisition Workshop (EKAW'88), Bonn, Germany, 19–23 June 1988,* edited by John H. Boose, Brian R. Gaines and Marc Linster, 28.1–28.20, Germany: Gesellschaft fur Mathematik und Datenverarbeitung

Shanmugasundaram, J. 2001. “Efficiently Publishing Relational Data as XML Documents.” *The VLDB Journal* 10 (2): 133–154.

Shannon, C. E. 1997. “The Mathematical Theory of Communication.” 1963. *MD Computing: Computers in Medical Practice* 14 (4): 306.

Shannon, C. E., and W. Weaver. 1949. *The Mathematical Theory of Communication* vol. 19. Urbana, IL: University of Illinois Press.

Shapiro, B. A., and K. Zhang. 1992. “Comparing Multiple RNA Secondary Structures Using Tree Comparisons.” *Computer Applications in the Biosciences* 6 (4): 309–318.

Shaw, M. J. 2001. “Knowledge Management and Data Mining for Marketing.” *Decision Support Systems* 31 (1): 127.

Shawkat Ali, A. B. M. and S. A. Wasimi. 2007. *Data Mining: Methods and Techniques*. Melbourne, VIC: Thomson Learning Australia.

Shen, Y-D., Z. Zhang, and Q. Yang. 2002. “Objective-oriented Utility-based Association Mining.” In *Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 9–12 December 2002,* edited by Vipin Kumar, Shusaku Tsumoto, Ning Zhong, Philip S. Yu and Xindong Wu, 426–433, Los Alamitos, CA: IEEE Computer Society Press.

Sherwood, J. 2008. “Iceland Seen Turning to IMF.” *The Wall Street Journal Asia Edition*, 17 October. http://online.wsj.com/article/SB122425739356744673.html?”mod=sphere_ts&mod=sphere_wd.

Silberschatz, A. 1996. “What Makes Patterns Interesting in Knowledge Discovery Systems.” *IEEE Transactions on Knowledge and Data Engineering* 8 (6): 970–982.

Silberschatz, A., and A. Tuzhilin. 1995. “On Subjective Measures of Interestingness in Knowledge Discovery.” *Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, 20–21 August 1995*, edited by Usama M. Fayyad and Ramasamy Uthurusamy, 275–281, Menlo Park, CA: AAAI Press.

Simpson, E. H. 1949. “Measurement of Diversity.” *Nature* 163: 688.

“Skim Kredit Program Yang Dikeluarkan Pemerintah” [Government Credit Schemes]. www.bi.go.id (Accessed 28 September 2012).

Skowron, A., and C. Rauszer, eds. 1992. “The Discernibility Matrices and Functions in Information Systems.” In *Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory* vol. 11, ed. R. Slowinski, 331–362. Netherlands: Kluwer Academic Publishers.

“SME Market Access and Internationalization: Medium-term KPIs for the SMEWG Strategic Plan. 2010.” APEC Policy Support Unit. Singapore: Asia–Pacific Economic Cooperation.

Solomonoff, R. J. 1957. “An Inductive Inference Machine.” *IRE Convention Record, Section on Information Theory*, 2: 56-62.

Srinivasan, V. and Y. H. Kim. 1987a. The Bierman-Hausman Credit Granting Model: a Note.” *Management Science* 33 (10): 1361–1362.

Srinivasan, V. and Y. H. Kim. 1987b. Credit Granting: a Comparative Analysis of Classification Procedures.” *The Journal of Finance* 42 (3): 665–681.

Steijvers, T., W. Voordeckers and K. Vanhoof. 2012. “Collateral, Relationship Lending and Family Firms.” *Small Business Economics* 34 (3): 243–259.

Stewart, W. H., W. E. Watson, J. C. Carland and J. W. Carland. 1999. “A Proclivity for Entrepreneurship: a Comparison of Entrepreneurs, Small Business Owners, and Corporate Managers.” *Journal of Business Venturing* 14 (2): 189–214.

Stiglitz, J. E. and A. Weiss. 1981. “Credit Rationing in Markets with Imperfect Information.” *The American Economic Review* 71 (3): 393–410.

Šušteršič, M., D. Mramor and J. Zupan. 2009. “Consumer Credit Scoring Models with Limited Data.” *Expert Systems with Applications* 36 (3, Part 1): 4736–4744.

Tabachnick, B. G. and L. S. Fidell. 2007. *Using Multivariate Statistics.* 5th ed. Boston, MA: Pearson.

Tam, K. Y. and M. Y. Kiang. 1992. “Managerial Applications of Neural Networks: the Case of Bank Failure Predictions.” *Management Science* 38 (7): 926–947.

Tambunan, T. 2008. “SME Development, Economic Growth, and Government Intervention in a Developing Country: the Indonesian Story.” *Journal of International Entrepreneurship* 6: 147–167.

Tan, H., T. S. Dillon, F. Hadzic, L. Feng and E. Chang. 2006. “IMB3-Miner: Mining Induced/Embedded Subtrees by Constraining the Level of Embedding.” In *Advances in Knowledge Discovery and Data Mining,* eds. W. K. Ng, M. Kitsuregawa and J. Li, 450–461. Berlin: Springer-Verlag.

Tan, P-N., V. Kumar, and J. Srivastava. 2002. “Selecting the Right Interestingness Measure for Association Patterns.” In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 23–26 July 2002,* edited by Osmar R. Zaïane, Randy Goebel, David Hand, Daniel Keim and Raymond Ng, 32–41, New York, NY: ACM Press.

Taniguchi, K., H. Sakamoto, H. Arimura, S. Shimozono, and S. Arikawa. 2001. “Mining Semi-structured Data by Path Expressions.” In *Discovery Science, Lecture Notes in Artificial Intelligence vol. 2226,* eds. K. P. Jantke and A. Shinohara, 378–388. Berlin: Springer-Verlag.

Theil, H. 1967. *Economics and Information Theory*. Amsterdam: North-Holland Publishing Company

Thomas, L. C. 2002. “A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers.” *International Journal of Forecasting* 16 (2): 149–172.

Thomas, L. C., D. B. Edelman and J. N. Crook. 2002. *Credit Scoring and Its Applications.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Thong, J. Y. L. and C. S. Yap. 1995. “CEO Characteristics, Organizational Characteristics and Information Technology Adoption in Small Businesses.” *Omega* 23 (4): 429–442.

Tsai, C-F. and J-W. Wu. 2008. “Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring.” *Expert Systems with Applications* 34 (4): 2639–2649.

Tsai, M-C., S-P. L. Lin, C-C. Cheng and Y-P. Lin. 2009. “The Consumer Loan Default Predicting Model: an Application of DEA-DA and Neural Network.” *Expert Systems with Applications* 36 (9): 11682–11690.

Tufféry, S. 2011. *Data Mining and Statistics for Decision Making.* 1st ed. West Sussex: John Wiley.

Van Gestel, I. T., B. Baesens, I. J. Garcia and P. Van Dijcke. 2003. “A Support Vector Machine Approach to Credit Scoring.” *Bank en Financiewezen*, 2: 73-82.

Vapnik, V. 1999. The Naure of Statistical Learning Theory. NY: Springer.

Vapnik, V. N., and A. Y. Chervonenkis. 1971. “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities.” *Theory of Probability & Its Applications* 16 (2): 264–280.

Vargas-Vera, M., E. Motta, J. Domingue, M. Lanzoni, A. Stutt and F. Ciravegna. 2002. “MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup.” In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Lecture Notes in Artificial Intelligence vol. 2473,* eds. Gómez-Pérez, A. and V. R. Benjamin, 379–391. Berlin: Springer-Verlag.

Veloso, A., W. Meira, and M. J. Zaki. 2006. “Lazy Associative Classification.” In *Proceedings of the Sixth International Conference on Data Mining, Hong Kong, China, 18–22 December 2006,* edited by Shusaku Tsumoto, Christopher W. Clifton, Ning Zhong, Xindong Wu, Jiming Liu, Benjamin W. Walsh and Yiu-Ming Cheung, 645–654, Piscataway, NJ: IEEE Computer Society Press.

Viganò, L. 1993. “A Credit Scoring Model for Development Banks: an African Case Study.” *Savings and Development* 17 (4): 441–482.

Voordeckers, W. and T. Steijvers. 2006. “Business Collateral and Personal Commitments in SME Lending.” *Journal of Banking & Finance* 30 (11): 3067–3086.

Waite, D. 1973. “The Economic Significance of Small Firms.” *The Journal of Industrial Economics* 21 (2): 154–166.

Wang, J., J. Han, and J. Pei. 2003. “CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets.” In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, 24–27*

August 2003, edited by Lisa Getoor, Ted Senator, Pedro Domingos and Christos Faloutsos, 236–245, New York, NY: ACM Press.

Wang, K., S. Zhou, and J. Han. 2002. "Profit Mining: from Patterns to Actions." In *Advances in Database Technology—EDBT 2002, Lecture Notes in Computer Science vol. 2287*, eds. C. S. Jensen, K. G. Jeffrey, J. Pokorny, S. Šaltenis, E. Bertino, K. Bőhm and M. Jarke, 31–44. Berlin: Springer-Verlag.

Wang, Q., J. Yu, and K. F. Wong. 2002. "Approximate Graph Schema Extraction for Semi-structured Data." In *Advances in Database Technology—EDBT 2000, Lecture Notes in Computer Science vol. 1777*, eds. C. Zaniolo, P. C. Lockemann, M. H. Scholl and T. Grust, 302–316. Berlin: Springer-Verlag.

Wang, Y., S. Wang and K. K. Lai. 2005. "A New Fuzzy Support Vector Machine to Evaluate Credit Risk." *IEEE Transactions on Fuzzy Systems* 13 (6): 820–831.

Weaver, P. M. and C. D. Kingsley. 2001. *Banking and Lending Practice*. 4th ed. Sydney: Lawbook.

Webb, G. I., and D. Brain. 2002. "Generality is Predictive of Prediction Accuracy." In *The 2002 Pacific Rim Knowledge Acquisition Workshop (PKAW 2002)*, eds. T. Yamaguchi, A. Hoffmann, H. Motoda and P. Compton. Tokyo, Japan. *Tokyo, Japan*, edited by T. Yamaguchi, A. Hoffmann, H. Motoda and P. Compton, 117–130. Berlin: Springer-Verlag.

Wei, L., J. Li and Z. Chen. 2007. "Credit Risk Evaluation Using Support Vector Machine with Mixture of Kernel." In *Computational Science – ICCS 2007*, eds. Y. Shi, G. van Albada, J. Dongarra and P. Sloot, 431–438. Berlin: Springer-Verlag.

Weiss, S. M., and C. A. Kulikowski. 1992. *Computer Sytems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. San Mateo, CA: Morgan Kaufmann.

West, D. 2002. "Neural Network Credit Scoring Models." *Computers & Operations Research* 27 (11–12): 1131–1152.

Wette, H. C. 1983. "Collateral in Credit Rationing in Markets with Imperfect Information: Note." *The American Economic Review* 73 (3): 442–445.

Wiginton, J. C. 1982. "A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior." *The Journal of Financial and Quantitative Analysis* 15 (3): 757–770.

Wiklund, J. and D. Shepherd. 2003. “Knowledge-based Resources, Entrepreneurial Orientation, and the Performance of Small and Medium-sized Businesses.” *Strategic Management Journal* 24 (13): 1307–1314.

Wilson, A. M., L. Thabane, and A. Holbrook. 2004. “Application of Data Mining Techniques in Pharmacovigilance.” *British Journal of Clinical Pharmacology* 57 (2): 127–134.

Wu, C. and X-M. Wang. 2002. “A Neural Network Approach for Analyzing Small Business Lending Decisions.” *Journal Review of Quantitative Finance and Accounting* 15 (3): 259–276.

Wymenga, P., V. Spanikova, A. Barker, J. Konings and E. Canton. 2012. *EU SMEs in 2012: At the Crossroads. Annual Report on Small and Medium-sized Enterprises in the EU, 2011/12*. Rotterdam: Ecorys Macro & Sector Policies.

Xiao, D. 2011. “International Experience for SME Credit Risks Identification: Based on Monitoring-cashflow Method.” In *Proceedings of the 2011 International Conference on Management and Service Science, Wuhan, China, 12–14 August 2011*, 1–4, Piscataway, NJ: IEEE Computer Society Press.

Yan, X., H. Cheng, J. Han, and P. S. Yu. 2008. “Mining Significant Graph Patterns by Leap search.” In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, Canada, 10–12 June 2008* edited by Jason Tsong-Li Wang, 433–444, New York, NY: ACM Press.

Yan, X., J. Han, and R. Afshar. 2003. “Clospan: Mining Closed Sequential Patterns in Large Datasets.” In *The Third SIAM International Conference on Data Mining*, eds. D. Barbará and C. Kamath, 166–177. USA: Society for Industrial and Applied Mathematics.

Yang, Q., J. Yin, C. X. Ling, and T. Chen. 2003. “Postprocessing Decision Trees to Extract Actionable Knowledge.” In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), Melbourne, FL, 19–22 December 2003,* edited by Xindong Wu, Alex Tuzhilin and Jude Shavlik, 685–688, Washington, DC: IEEE Computer Society Press.

Yao, H., H. J. Hamilton, and C. J. Butz. 2004. “A Foundational Approach to Mining Itemset Utilities from Databases.” In *The Fourth Siam International Conference on Data Mining*, eds. M. W. Berry, U. Dayal, C. Kamath and D. Skillicorn, 482–486. USA: Society for Industrial and Applied Mathematics.

Yao, P. 2009. “Hybrid Classifier Using Neighborhood Rough Set and SVM for Credit Scoring.” In *Proceedings of the 2009 International Conference on Business Intelligence and Financial Engineering, Beijing, China, 24–26 July 2009,* edited by Shouyang Wang, Lean Yu, Fenghua Wen, Shaoyi He, Yong Fang and K. K. Lai, 138–142, Piscataway, NJ: IEEE Computer Society Press.

Yao, P. and Y. Lu. 2011. “Neighborhood Rough Set and SVM Based Hybrid Credit Scoring Classifier.” *Expert Systems with Applications* 38 (9): 11300–11304.

Yin, X., and J. Han. 2003. “CPAR: Classification Based on Predictive Association Rules.” In *Proceedings of the Third SIAM International Conference on Data Mining (SIAM 2002), San Francisco, USA, 22–24 April 2003,* edited by Daniel Barbará and Chandrika Kamath, 331–335, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Yobas, M. B., J. N. Crook and P. Ross. 2002. “Credit Scoring Using Neural and Evolutionary Techniques.” *IMA Journal of Mathematics Applied in Business and Industry* 11 (2): 111–125.

Yoshida, K., H. Motoda, and N. Indurkhya. 1994. “Graph-based Induction as a Unified Learning Framework.” *Applied Intelligence* 4 (3): 297–316.

You, J-I. 1995. “Small Firms in Economic Theory.” *Cambridge Journal of Economics* 19: 441–462.

Yu, L., K. Lai, S. Wang and L. Zhou. 2007. “A Least Squares Fuzzy SVM Approach to Credit Risk Assessment.” In *Fuzzy Information and Engineering*, ed. B-Y. Cao, 865–874. Berlin: Springer-Verlag.

Yu, L., S. Wang and K. K. Lai. 2008. “Credit Risk Assessment with a Multistage Neural Network Ensemble Learning Approach.” *Expert Systems with Applications* 34 (2): 1434–1444.

Yu, L., S. Wang, K. K. Lai and L. Zhou. 2008. Bio-inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines. Berlin: Springer-Verlag.

Yu, L., S. Wang, F. Wen, K. K. Lai and S. He. 2008. “Designing a Hybrid Intelligent Mining System for Credit Risk Evaluation.” *Journal of Systems Science and Complexity* 21 (4): 527–539.

Zadeh, L. A. 1965. “Fuzzy Sets.” *Information and Control* 8: 338–353.

———. 1983. “Commonsense Knowledge Representation Based on Fuzzy Logic.” *Computer* 16: 61–65.

Zaki, M. 2001. “Spade: an Efficient Algorithm for Mining Frequent Sequences.” *Machine Learning* 42 (1): 31–60.

———. 2003. “XRules: an Effective Structural Classifier for XML Data.” In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, 24–27 August 2003,* edited by Lisa Getoor, Ted Senator, Pedro Domingos and Christos Faloutsos, 316–325, New York, NY: ACM Press.

———. 2005. “Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications.” *IEEE Transactions on Knowledge and Data Engineering* 17 (8): 1021–1035.

Zamir, O., and O. Etzioni. 1999. “Grouper: a Dynamic Clustering Interface to Web Search Results.” *Computer Networks* 31 (11–16): 1361–1374.

Zhang, C., and S. Zhang. 2001. “Collecting Quality Data for Database Mining.” *AI 2001: Advances in Artificial Intelligence* 131–142.

Zhang, L. and X. Hui. 2009. “Application of Support Vector Machines Method in Credit Scoring.” In *The Sixth International Symposium on Neural Networks,* eds. H. Wang, Y. Shen, T. Huang and Z. Zeng, 283–290. Berlin: Springer-Verlag.

Zhang, T., ed. 2002. “Association Rules.” In *Knowledge Discovery and Data Mining: Current Issues and New Applications, Lecture Notes in Computer Science vol. 1805,* eds. T. Terano, H. Liu and A. L. P. Chen, 245–256. Berlin: Springer-Verlag.

Zhang, T., R. Ramakrishnan, and M. Livny. 1996. “BIRCH: an Efficient Data Clustering Method for Very Large Databases.” In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Canada, 4–6 June 1996,* edited by H. V. Jagadish and Inderpal Singh Mumick, 103–114, New York, NY: ACM Press.

Zhong, N., Y. Y. Y. Yao, and M. Ohishima. 2003. “Peculiarity Oriented Multidatabase Mining.” *IEEE Transactions on Knowledge and Data Engineering* 15 (4): 952–960.

Zhou, L., K. K. Lai and L. Yu. 2012. “Least Squares Support Vector Machines Ensemble Models for Credit Scoring.” *Expert Systems with Applications* 37 (1): 127–133.

Zhou, W., and G. Kapoor. 2011. “Detecting Evolutionary Financial Statement Fraud.” *Decision Support Systems* 50 (3): 570–575.

Zhou, X., D. Zhang and Y. Jiang. 2008. “A New Credit Scoring Method Based on Rough Sets and Decision Tree.” In *Advances in Knowledge Discovery and Data Mining,* eds. T. Washio, E. Suzuki, K. Ting and A. Inokuchi, 1081–1089. Berlin: Springer-Verlag.