Sudan University of Science & Technology

College of Graduate Studies

Credit Scoring Using Data Mining Classification: Application on Sudanese Banks

تصنيف الإئتمان بإستخدام تقنيات تنقيب البيانات: تطبيق على المصارف السودانية

Submitted for the Degree of Doctor of Philosophy in Computer Science

By

Eiman Mohammed El Hassan

B.Sc. in computer science, University of Khartoum

M.Sc. in computer science, University of Khartoum

Supervisor:Prof. Izzeldin Mohammed Osman

July 2014

Sudan University of Science & Technology

College of Graduate Studies

جامعة السودان للعلوم والتكنولوجيا

كلية الدراسات العليا

Approval Page

Name of Candidate: Eman Mahmoud Elhassan Abdelrahman

Thesis title: Credit Scoring Using Data Mining Classification: Application to Sudanese Banks

تصنيف المخاطر الائتمانية باستخدام تقنيات التنقيب في البيانات: تطبيق على المصارف السودانية

Approved by:

1. External Examiner

Name: Abubaker Elsiddig Mughram Alhassan

Signature: ..... Date: 16th Jan '17

2. Internal Examiner

Name: Mohamed Elhafiz Mustafa

Signature: ..... Date: 10/2/16

3. Supervisor

Name: Ibrahim Mahmoud Omer

Signature: ..... Date: 10/2/2016

(وَمَا تَوْفِيقِي إلا باللّهِ عَلَيْهِ تَوَكّلْتُ وَإِلَيْهِ أُنِيب)

صدَقَ الله العَظِيم

سورة هود الآية88

Abstract

The main aim of this thesis is to develop suitable and high performance Credit Scoring Models (CSMs) to assess credit risk of personal loans for the Sudanese commercial banks using data mining techniques.

Two Sudanese credit datasets were constructed. These datasets were provided by Agricultural Bank of Sudan and Al Salam Commercial Bank. In addition to these two datasets, a German credit dataset was also employed in this research as a benchmarking dataset.

Three data mining classification techniques were employed in this research: Artificial Neural Network (ANN), Support Vector Machine(SVM) and Decision Tree (DT). Genetic Algorithm (GA) is also applied as a feature selection technique. Two validation methods (split validation with two ratios (70:30 and 60:40) and 10-cross validation) were used to validate the proposed credit scoring models.

As a result of combining GA with the specified classification techniques, tables of attributes and their weights were produced. By using these tables new reduced sets of features were identified for each dataset (i.e. new reduced datasets were produced from the original datasets).

Experiments in this research were conducted in three stages. In stage 1, classification techniques were applied individually to each dataset .In stage 2, these techniques were combined with GA and in stage 3 these techniques were applied to the reduced datasets.

Nine proposed credit scoring models for each dataset were developed for each stage. These models were compared for each dataset in terms of

fiveevaluation measures: Accuracy, Precision (Defaulter), Precision (Non-defaulter), Type I and Type Π errors. As a result of these comparisons, the suggestions for the best models for each dataset were given.

The experiments carried out in this research show that:

- For all datasets, combining GA as a wrapper-feature selection technique with ANN, SVM and DT classification techniques is more beneficial than applying these techniques individually. Applying specified classification techniques to the reduced datasets does not bring a significant improvement to the major models in terms of the specified five measure indicators compared to the resulting models from applying these techniques to the original datasets.In addition, and as well-known fact the performance of each technique heavily depends on the nature of datasets.

مستخلص الأطروحة

تهدف هذه الأطروحة لتطوير نماذج تصنيف إئتمان عالية الكفاءة وملائمة لتقويم مخاطر الإئتمان للبنوك التجاريه السودانية بإستخدام تقنيات التنقيب عن البيانات .

تم إنشاء مجموعتين من البيانات الإئتمانية حيث اخذت هذه البيانات من البنك الزراعي السوداني ومن بنك السلام التجاري. ثم إستخدمت البيانات الأئتمانية الألمانية كمجموعة بيانات للمعايرة.

وقد استخدمت في هذا البحث ثلاثة تقنيات تنقيب بيانات للتصنيف وهي الشبكات العصبانيه الإصطناعية (Artificial Neural Networks) و آلة المتجه الداعم (Support Vector Machine) وشجرة القرار (Decision Tree) . ثم استخدمت الخوارزمية الجينية (Genetic Algorithm) لإختيار الخصائص . وتم تطبيق إثنين من طرق التحقيق وهي التحقيق المنقسم(split validation)بنسبتين (30:70 و40:60) والتحقيق المتبادل (10-cross-validation) لتقييم هذه النماذج المقترحة.

نتج من دمج الخوارزمية الجينية (GA) مع تقنيات التصنيف جداول تحوي الخصائص وأوزانها لكل مجموعة بيانات. وبإستخدام هذه الجداول تم استخلاص عدد مخفض من الخصائص لكل مجموعة بيانات، وبالتالي تم الحصول على مجموعات بيانات إئتمانية مصغرة من المجموعات الأصلية .

أجريت تجارب هذا البحث على ثلاث مراحل لكل مجموعة بيانات. في المرحلة الأولى تم تطبيق تقنيات التصنيف المحددة بصورة فردية علي إي مجموعة بيانات. في المرحلة الثانية تم دمج الخوارزمية الجينية مع تقنيات التصنيف المحددة. في المرحلة الثالثة تم تطبيق تقنيات التصنيف علي المجموعات الإئتمانية المصغرة. ثم تطوير تسعة نماذج مقترحة لتصنيف الإئتمان لكل مجموعة بيانات في كل مرحلة.

وقد تمت مقارنة كل من النماذج المقترحة لكل مجموعة بيانات بناءا على اساس خمسة إجراءات للتقييم وهي الصحة والدقة (للمتعثرين) والدقة(لغير المتعثرين) والأخطاء من النوع الأول والنوع الثاني . ونتيجة لهذه المقارنات تم إقتراح النماذج الأفضل لكل مجموعة بيانات.

وقد أوضحت النتائج المستخلصة من هذه التجارب التي أُجريت في هذا البحث الآتي :

- لكل مجموعات البيانات الإئتمانية إتضح أن دمج الخوارمية الجينية (GA) مع تقنيات التصنيف (الشبكات العصبانيه الإصطناعيه (ANN) و آلة المتجه الداعم (SVM) وشجرة القرار (DT)) أكثرفاعلية من تطبيق هذه التقنيات بشكل فردي. كما أن تطبيق تقنيات التصنيف على مجموعات البيانات الإئتمانية المصغرة لايحقق تحسنا كبيرا في نماذج تصنيف الإئتمان بالمقارنة مع النماذج الناتجة من تطبيق هذه التقنيات على مجموعات البيانات الأصلية.
- إن أداء تقنيات التصنيف يعتمد بشكل كبير على طبيعة مجموعة البيانات الإئتمانية.

Dedications

This dissertation is lovingly dedicated to my parents, husband and to my all wonderful family members.

Acknowledgements

This dissertation would not have been completed, obviously, without support from Allah: to Him I am, first and foremost most grateful.

I should also like to express my infinite gratitude to my supervisor Professor Izzeldin Mohammed Osman for his scholarly and humane guidance, his patience and continuous encouragement.

I am forever indebted to my parents for everything they have done for me.

The College of Computer Science and Information Technology deserve lot of thanks for what they have taught me over the years.

Special thanks also go to Dr. Mohammed El hafiz, Ph.D. Program Coordinator (Batch 1) and to Dr. Mahmoud Ali Ahmed, Dean of the Faculty of Mathematical Sciences (KhartoumUniversity) and to all his kind staff-members.

I also fully appreciate the help I have received from several of my students and officials at several banks in Khartoum with regard to collecting the necessary data; to all of them I remain deeply indebted. Also to my close friend Dr. NisreenBeshir Osman and to everyone who has, directly or indirectly, contributed to this research endeavor, I offer my sincere thanks.

List of Publications

- A paper entitled **"Data Mining Techniques in Credit Scoring: A Survey"**Submitted for publication.

- Eiman Kambal, Izzeldin Osman, Methag Taha, Noon Mohammed, Sara Mohammed, paper entitled "**Credit Scoring Using Data Mining Techniques with Particular Reference to Sudanese Banks"** presentedin proceedings of IEEE International Conference on Computer, Electrical and Electronics Engineering (ICCEEE 2013). Khartoum, Sudan.

- A paper under preparation entitled "Credit Scoring Using Data Mining Techniques: The Case of Islamic Banks".
- A paper under preparation entitled "Multistage Credit Scoring Models".

Table of Contents

List of Figures

List of Tables

Table **Page**

List of Abbreviations

CS	Credit Scoring
CSM	Credit Scoring Model
CSS	Credit Scoring System
DM	Data Mining
KDD	knowledge Discovery in Database
AI	Artificial Intelligence
LDA	Linear Discriminant Analysis
SNDA	Skew Normal Discriminant Analysis
STDA	Skew T Discriminant Analysis
SDA	Stepwise Discriminant Analysis
FD	Flexible Discriminant Analysis
MDA	Mixture Discriminant Analysis
LR	Logistic Regression
DT	Decision Tree
CHAID	chi Square Automatic Interaction
CART	Classification and Regression Tree
ANN	Artificial Neural Network
BP	Back Propagation
MLP	Multilayer Perceptron
SVM	Support Vector Machine
GA	Genetic Algorithm
NBC	Naive Bayes Classifier
RS	Rough Set
FSRT	Feature selection based on rough set and tabu search

DA	Discriminant Analysis
KNN	K-Nearest Neighbor
RBF	Radial Basis Function
GP	Genetic Programming
BFS	Best First Search
PSO	Particle Swarm Optimization
FSVM	Fuzzy Support Vector Machine
HCSM	Hybrid Credit Scoring Model
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC
PCC	Percentage Correctly Classified
SCD1	Sudanese Credit Dataset1
SCD2	Sudanese Credit Dataset2
RSCD1	Reduced Sudanese Credit Dataset1
RSCD2	Reduced Sudanese Credit Dataset2
Tpos	True Positive
Tneg	True Negative
Fpos	False Positive
SCSM	Sudanese Credit Scoring Model

CHAPTER ONE

1. Introduction

1.1 Overview

Banks and other lending organizations are ‘‘profit-seeking’’ organizations, that make money for their shareholders. They provide financial products and services to clients while managing a diversity of risks [38].

Credit risk is the most challenging risk to which financial institution are exposed. International Convergence of Capital Measurement and Capital Standards defined credit risk as the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms [3].

Credit risk appears when wrong decisions associated with the approval of the loan application are taken. The wrong credit risk assessment leads to increase in the number of defaulters and as a consequence could derive banks towards bankruptcy.

An evidence of the potential social and economic impact of credit risk decisions locally and globally is the U.S. Sub-prime mortgage crisis (2007-2008)[90]. The main causes of this crisis are that financial institutions offered mortgage loans to higher–risk borrowers without appropriate review and documentation. The consequence of this crisis is the incidence of the global financial crisis[90].

As a conclusion of the aforementioned, credit risk evaluation decisions are of the main success keys for financial institutions.

In Sudan, commercial banks use Islamic financing modes to provide loans to different industries such as manufacturing, agricultural, commercial, service enterprises, and others[48, 38]. Hence economic and, therefore, the social well-being of the Sudan is highly dependent on the behavior of commercial banking sector[48, 38].

The main objective of this research is to develop highly accurate credit scoring models (CSMs) by using data mining (DM) classification techniques to enhance the quality of loan credit decisions. The proposed models will be tested and validated using real life credit datasets of Sudanese commercial banks.

This chapter focuses on the motivation, problem statement, objectives, scope, contributions of the research, and organization of the thesis.

1.2 Motivation of the Research

Credit risk is the most challenging risk to which financial institution are exposed[38]. Unfortunately, credit-risk evaluation decisions are complex and unstructured problems, which cannot be readily solved formally[76]. In the past this problem was solved through judgmental and subjective decision which was issued by loan officers using their experience and analysis of data. Hence loan granting decisions are expected to be inaccurate, more subjective, and time consuming[53].

Recently the number of lender organizations has increased, and also demand for loans. Hence, in spite of difficulties and complexity, accurate and faster credit risk decision support systems are most demanded for the lending organization to survive and face drastic competition among other organizations.

Hence the productive direction is to exploit advances in information technology and the vast amount of collected data to automate the lending decision through Credit Scoring (CS) as a formal credit risk evaluation method which helps to lower the costs of credit and increase reliability and speed of lenders' credit decision making.

1.3 Problem Statement

In recent years, the number of banks in the Sudan has increased gradually, so banks are facing high competition. In spite of various modern diversified banking services, loan granting still constitutes the core of the income of commercial banks [81] even though it is more risky. The risk appears when wrong decisions associated with the approval of the borrower loan application are taken. The wrong credit risk assessment leads to increase in the number of bank defaulters and as a consequence bankruptcy of banks.

Sudanese banks are Islamic banks. Islamic banks invest the funds on the basis of profit and loss sharing paradigm. Therefore, they have to be more careful toward loan granting decisions than conventional banks. For these banks to avoid the loss a high performance credit risk evaluation systems are required.

Nowadays in Sudanese commercial banks credit risk assessment for borrowers is evaluated judgmentally and manually by loan officers. Analysis of borrowers' applications is done manually and depends on loan officers' characters and experience. Judgmental assessments rely on more subjective (not free from bias), inaccurate, inconsistent, and informal decisions. Furthermore the loan approval process is becoming time consuming.

1.4 Research Objectives

The main aim of this research is to enhance the process of loan granting in Sudanese commercial banks by developing and introducing CSM(s) to evaluate their personal loans. This is achieved by fulfilling the following objectives:

1. To identify the currently used credit risk evaluation systems used in the Sudanese banking**.**
2. To investigate loan variables used in the loan granting decision-making process in the Sudanese banks.
3. To build high quality credit dataset(s) for Sudanese banks.
4. To review the different DM techniques which are applied to CS problem and address their advantages, shortcomings and their potential influence in improving the loan granting process.
5. To identify the more appropriate DM classification techniques needed for developing the proposed CSMs for each dataset.
6. To compare these proposed CSMs and select the optimal one for each dataset.

1.5 Research Scope

The scope of this research is as follows:

- The proposed models will be applied to assess credit worthiness of loan applicants (individual borrowers) in Sudanese banks.
- The datasets for these proposed models were collected from two Sudanese banks containing historical information for personal loans.
- This study considers all type of loans equal irrespective of the loan conditions.

1.6 Contributions

- This thesis contributes to the research on the CSMs by constructing two new credit datasets which are different from the datasets used in previous researches. These datasets are derived from Sudanese banks (Islamic banks) which are deriving their rules from Islamic Sharia. All credit datasets in the literature are derived from conventional banks.
- Another contribution of this research is the design of CS readiness test that determines readiness of banks to apply CSM.
- This is the first study that, has investigated the use of CSMs in the Sudanese banking sector. These banks are currently using personal judgmental techniques to evaluate borrower credit risk.
- An important contribution of this thesis is the comprehensive literature review on the most used techniques in CS and specifies their pros and cons. In addition two approaches of hybridization of these techniques are also presented.

- Introducing three DM classification techniques Artificial Neural Networks (ANNs), Support Vector Machine (SVM) and Decision Tree (DT) in addition to feature selection technique genetic algorithm (GA) to develop CSMs for the data collected from two Sudanese commercial banks and German credit dataset (benchmarking dataset) and using two validating methods (split validation with two ratios (70:30 and 60:40) and 10-cross validation) to validate these models. In stage 1, classification techniques were applied individually to each dataset .In stage 2, these techniques were combined with a feature selection technique and in stage 3; these techniques were applied to reduced datasets (extracted by a feature selection technique). All resulting models in all stages for each dataset are compared in terms of five measures (Accuracy, Precision (Defaulter), Precision (Non-defaulter), Type I and Type Π errors). As a result of these comparisons, the suggestions for the best models for each dataset are given.
- The development process of the proposed models for Sudanese banks through three stages of experiments; together with the final results from these experiments and outcomes of the research are expected to provide a valuable dimension not only for the Sudanese baking sector but also for CSM developers in environments similar to that of Sudanese banks e.g. Islamic banks.

1.7 Organization of the Thesis

The thesis is organized as follows:

Chapter One: is an introductory chapter that discusses the motivation, problem statement, objectives, scope and contributions of the research.

Chapter Two is the background chapter which provides general concepts of banking systems, risk management in banks, credit risk and credit risk evaluation systems. It also discusses broadly the concept of CS which is the main issue of this research. DM concepts are also clarified in this chapter.

Chapter Three is the literature review chapter. It discusses the most widely used DM classification techniques in CS. These techniques are categorized into three approaches: Statistical, Artificial Intelligence (AI), and Hybrid. The pros and cons of each technique are also presented.

Chapter Four is the research methodology chapter. It describes how the research work was conducted. The chapter also discusses the different phases in this research work and the methodology followed during each phase.

Chapter Five describes the implementation phase of this research. Collection and construction of datasets stages are presented in this chapter. This chapter also presents the detailed description of the datasets and their proposed CSMs.

Chapter Six presents and discusses the results of all proposed CSMs of all stages for each dataset. Results of comparisons between all proposed CSMs for each dataset are also discussed.

Chapter Seven provides the summary and conclusions derived from this research along with recommendations for the future work.

A list of references presents the references used in the thesis.

Appendices include the types of Islamic financing modes and parts of the Sudanese credit datasets.

CHAPTER TWO

2. Background

2.1 Overview

This chapter provides a solid background for this research. It starts by general concepts of banking systems. This concept includes services provided by banks, challenges facing banks, and risk management in banks. Credit risk and credit risk evaluation systems are discussed. The concept of CS which is the main issue of this research is broadly and clearly illustrated in this chapter. Furthermore comparisons between Credit Scoring System (CSS) and other credit risk evaluation systems are also provided. As the result of these comparisons the benefits of CSS are fully identified. In this research the CS problem is modeled as a DM classification problem. Therefore, DM concepts are also clarified. In addition DM classification prediction tasks and their techniques are presented at the end of this chapter.

2.2 General Concepts of Banking System

The banking system is considered as one of the most important economic sectors in the country for its role in the economic development process. The main role of it is providing capital to various economic sectors through credit facilities and (long and short terms) loans[48].

The banking system in any society, mainly consists of Central Bank, the commercial banks, specialized banks, and in addition to social banks, savings funds, cooperative societies, and financial institutions that deal in

accordance with the laws of banking in the country[48]. All these types of banks share the following functions to some extent[57] :

1. Maturity transformation: banks accept deposits from savers; guarantee to return these on demand .These deposits are used to make loans for longer durations. Banks improve the productivity of the economy by this transformationof short term savings into long term investments.
2. Credit creation: For banks, each deposit they get is split into 2 parts. One part (smaller one) usually stays in the bank as reserves in case the depositor wants some of their money back at a short notice. The other part of deposit is lent on to an investor.
3. Credit allocation: Demand for credit is always higher than the part of the savings that the bank is allowed to invest. So banks have to take decision to provide loan. To take such decisions, many questions have to be answered: What is client going to do with the money? ; What are the risks of project the money will go to finance? What is the possible return? ; What is the likelihood that the client will be able to repay the money as agreed with the bank?

2.3 Challenges Facing Banks

One of the main challenges to banks is competition. Recently, the number of banks increased. So banks have to enhance the quality of their services and expand in new geographies according to business distribution and the customer valuable opinions[31].

Banks have to know their customers and classify them. They have to keep them by providing best services for them. The provision of services has to ensure the efficiency and profitability[31].

A fast changing market place forced banks to invest in technological innovation so as to survive and gain competitive advantage over others[31].

Managing risks is the major challenge in banking sector. Hence Banks cannot live without risk management systems that are capable of identifying, measuring, controlling business exposure to guarantee sustainable growth for bank[31].

All these challenges are not independent but each one is related to the others directly or indirectly.

2.4 Risk Management in Banks

"Risk is the potentiality that both the expected and unexpected events may have an adverse impact on the bank's capital or earnings"[87].

Banks are profit seeking organizations. They provide variety of products and services to customers while managing a diversity of risks [38]. So banks have to turn these risks into profits to make money for their shareholders[38, 101]. Risk taking is the basic requirement for future profitability.

There are three main categories of risks in banks[87] :

1. Market Risk: Defined as the possibility of loss to bank caused by the changes in the market variables.

2. Operational Risk: defined as the risk of loss arising from inadequate or failed internal processes, people and systems or from external events. In order to diminish this, internal control and internal audit systems are used.
3. Credit Risk: defined as the risk of loss due to the lack of commitment by the creditor to the repayment according to the agreed terms. As this is the focus of this research it willbe discussed in details later.

2.4.1 Credit Risk

Credit (or loan) is defined as “Transaction between two parties in which one (the creditor or lender) supplies money, goods, services, or securities in return for a promised future payment by the other (the debtor or borrower)”[1]. Credit may be extended by public or private institutions to financebusiness activities,agricultural operations, consumer expenditures, or government projects[1].

Most revenues of lending organizations in general and banks in particular are generated by lending (credits or loans) operations. Hence loans constitute a cornerstone of the banking industry. Lenders to provide this (or other) services in their decision-making processes, try to optimize their ‘‘risk-return’’ trade-off [38]. Credit risk is the most challenging risk to which financial institution are exposed. The ability to quantify credit risk for borrowers is central to the core aspects of the lending process.

According to International Convergence of Capital Measurement and Capital Standards [3] credit risk is most simply defined as “the potential that

a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms".

Wrong decisions associated with credit risk evaluation may lead to bankruptcy of lending organizations which leads to business failure and losses to many stakeholders (shareholders, managers, lenders (banks), suppliers, clients, the financial community, government, competitors, and regulatory bodies, among others). Therefore, loan granting decision requires extremely careful scrutiny. The ability of a financial institution to successfully overcome challenges actually depends on credit risk evaluation systems that are used to assess credit risk for borrowers.

2.4.2 Credit-Risk Evaluation Systems

Financial institutions in general and banks in particular use a variety of credit evaluation systems which are used to differentiate between loans that are more likely to be repaid from those that are less likely to be repaid[4], there are three types:

1. **Judgmental Systems**: Systems in which the credit decision is made manually by loan officers or other persons [4].

2. **Credit Rating**: Systems in which credit ratings are set by an Internal rating system in banks (produce ratings only for their own business and institutional loans) or a public rating agency(produce ratings for worldwide companies, financial instruments and Sovereigns) [74]. Credit rating agencies use their judgment and experience to determine a rating to a particular borrower[74].

3. **Credit Scoring Systems(CSSs):** Systems in which the credit decision is made mechanically on the basis of statistical (or others) models [4].

2.4.3 Similarities and Differences between Judgmental and Credit Scoring Systems

There are many differences and similarities between judgmental and CSSs as follows[4] :

- Both systems may employ the similar data in credit decision.
- Both systems assume that past experience can be used to predict future performance, but not with certainty.
- Judgmental systems generally rely on less standardized information (subjectively evaluated, depends on loan officers' experiences and common senses). So they may not produce consistent decisions between applicants who have similar data.
- CS draws on types of information that will be similar for all borrowers (consistency). So a given set of such information produces a similar credit score for all borrowers.
- Loan application may be rejected by judgmental system because of weaknesses in only one distinct criterion (such as maximum debt to income ration or minimum loan size), while in CSSs weaknesses in one criterion may be overcome by strength in one or more other criteria. Therefore, the judgmental and CS methods will not always produce the same results when they are applied to the same loan application.

2.4.4 Historical Background for Credit Scoring Method

- The history of CS begins with the study of David Durand in 1941 who examined car loan applications (National Bureau of Economic Research)[21].
- The year 1956 was the second important landmark in the history of CS. Fair, Isaac and Company was founded on the principles that data can improve business decisions if used intelligently.
- In 1963, a paper published in the Journal of American Statistical Association noted that “numerical rating systems are not in widespread use” (Myers and Forgy, 1963). The authors attempted to prove that “statistical credit scoring” represented an improvement over the judgmental evaluation of credit risk [21].

In United States:

- By the end of the 1970s, most of the U.S’s largest commercial banks, finance companies, and credit card issuers used CSSs [4].
- By the late 1980s much had changed. Lenders could purchase the generic credit history scores of individuals who were not their account holders [4].
- The use of CS then spread to additional loan products including home mortgage and small-business lending [4].
- In the 90s, score-cards were introduced to CS.

2.4.5 Credit Scoring Approach Definitions

CS is defined in [22] as a main analytical scientific method for credit risk assessment. This method uses quantitative measures of the performance and characteristics of past loans to predict the future performance of loans with

similar characteristics. Another definition by [75, 95] state “Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques assess, and therefore help to decide, who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders”. Moreover, CS is "a quantitative evaluation system employed by banks to assess the creditworthiness of an individual or firm that applies for a loan” [54].

Furthermore, Anderson in [12] suggested that to define credit scoring, the term should be broken down into two components**,** credit and scoring**.** Firstly, simply the word “credit” means “buy now, pay later”. It is derived from the Latin word “credo”, which means “I believe” or “I trust in”. Secondly, the word “scoring” refers to “the use of a numerical tool to rank order cases according to some real or perceived quality in order to discriminate between them, and ensure objective and consistent decisions”. Therefore, scores might be presented as “numbers, or grades” which may be presented as “letters” or “labels”.

2.4.6 Benefits of Credit Scoring

The following points summarize the benefits of the CS method:

- CS models are built on much larger samples than a loan analyst can remember [7].

- It reduces the time needed in the loan approval process and increases its consistency. The time saving means cost savings to the bank and benefits to the customer as well the lenders[80].
- It improves objectivity in the loan approval process. This objectivity helps lenders ensure that, the same criteria are applied to all borrowers regardless of race, gender, or other factors [80, 7]. This is an important factor particularly in countries where nepotism prevails.
- CS aids financial institutions in determination of appropriate interest rate and the amount of loan. Lower-risk consumers are charged a lower interest rate and vice versa [82].
- An additional vital advantage of CS is that the same data can be analyzed by different credit analysts or statisticians and give the same decisions [7].
- Making the securitization of loans more feasible [80].

2.4.7 Weaknesses of Credit Scoring

Although CS has significant benefits, it has also some shortcomings as follows:

- A CSM may be built using a biased sample of borrowers who have been granted loans. This may occur because applicants who are rejected will not be included in the data for constructing the model. Hence, the sample will be biased [80].
- Misclassification (accuracy) problem: Accuracy is a very important consideration in CS. An improvement in accuracy of even a

fraction of a percent translates into significant future savings. Inaccurate CSS leads to poorly performing loans. And hence using CSmethod becomes more harmful than beneficial [80, 7].

- CSMs are not standardized and differ from one organization to another; expensive to buy and to train credit analysts [7].
- CSMs actually depend on historical data. Unless it is frequently updated it will expire and become less accurate [7].

Despite the limitations highlighted above, there is no doubt that CS will continue to be a major tool in assessing credit risk in lending organizations[7].Using CS appropriately allow banks and other lending organization to gain important competitive advantage over other competitors[80]. In addition CS method is of particular importance in countries with prevailing nepotism and a lack of transparency in financial transactions.

2.5 Data Mining Concepts

The amount of data collected by businesses has grown rapidly in recent years. Availability of automated data collection tools and evolution of database technology are the major reasons for collecting this vast amount of data. We are truly in the age of big data, but there are real challenges in extracting useful knowledge from huge amount of data.

DM or knowledge discovery in database (KDD) is a process of extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from vast amount of data[47]. It combines techniques

from many fields, including databases, artificial intelligence, machine learning, pattern recognition, statistics and visualization[47]. See Figure 2.1

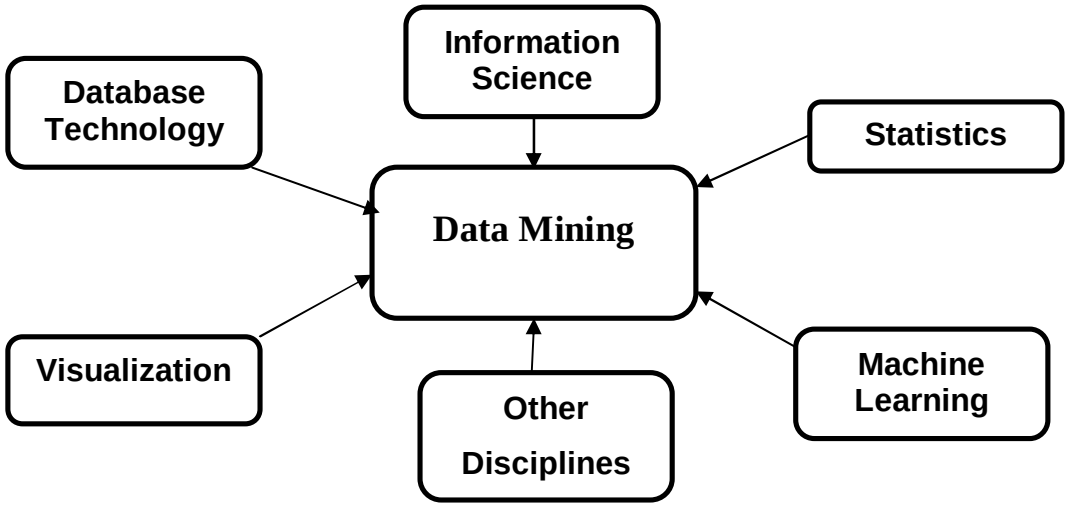

Figure 2.1 Data Mining as A Confluence of Multiple Disciplines[46]

According to [39] DM process can be described as: Iterative (the results of one step may mean that a previous step needs to be revisited) and Semi- automatic (because it involves by many decisions made by humans such as determine of the objective of the process , select the appropriate tools ,measuring patterns …..etc).

DM process consists of an iterative sequence of the following steps[39, 47] : (See Figure 2.2)

1. Defining the goal of the process: Developing an understanding of the application domain and identifying the goals of the DM process from the customer's view point.

2. Data selection: Selecting a target dataset (suitable to the goal determined in 1) by focusing only on subset of data variables or data sample.

3. Data preparation (cleaning, integration, and transformation): This may involve the removal of noise from the data, handling missing fields, integrating data from multiple sources, using transformation methods to reduce the search space, deriving new attributes, etc.

4. Choosing the DM task: This decision can depend upon the goal of the DM process, the type of data available (e.g. it may be ordered) and the available techniques.
5. Choosing the DM algorithm(s): The choice of a DM algorithm depends on the DM task chose in 4. A DM task may have more than one available algorithm.
6. Pattern evaluation: identification of the truly interesting patterns.

   An interestingness patterns have to be easily understood by humans, valid on test or future data with some degree of certainty, beneficial, and novel.

7. Interpreting mining results: The presentations of the DM results are important, as evaluation is difficult. Different visualization techniques may be used.
8. Consolidating discovered knowledge: Incorporating derived knowledge into the organization.

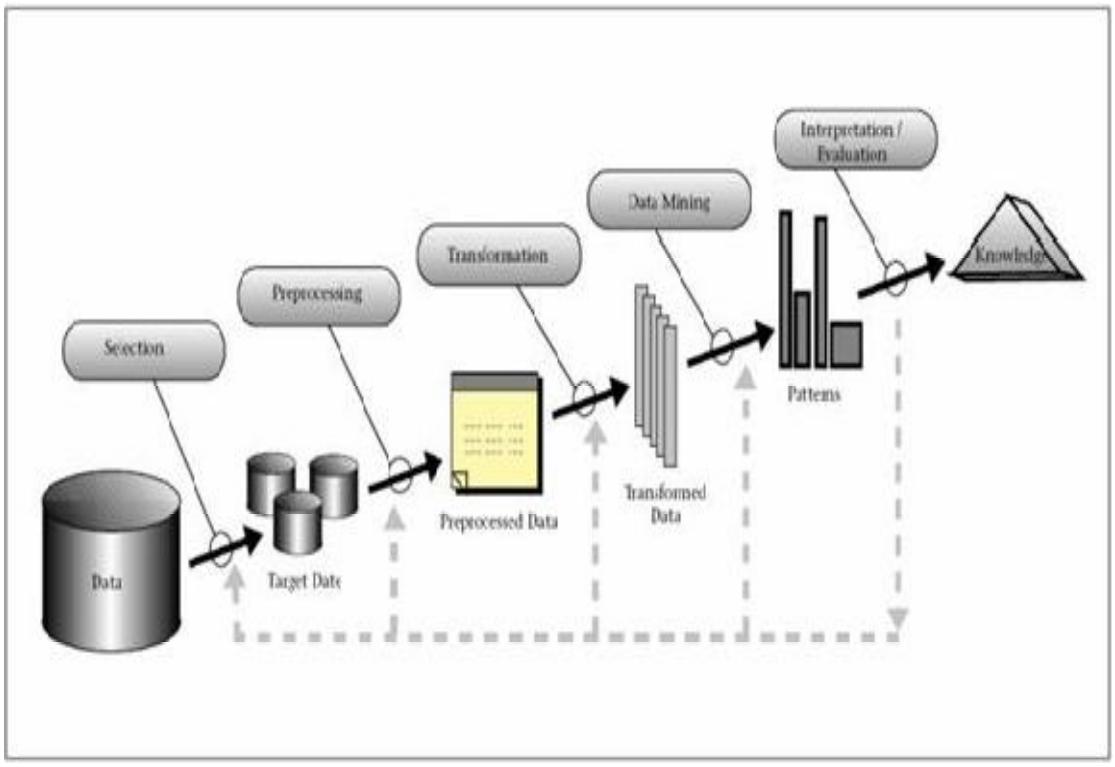

Figure 2.2 An Overview of the Steps of Data Mining Process [39]

2.5.1 Data Mining Functionalities

DM functionalities are used to specify the kind of patterns to be found in DM tasks. DM tasks can be classified into two categories:

1. Descriptive tasks: Tasks that are employed to find human-interpretable patterns that describe the data. Clustering, summarization and association rule discovery are examples of descriptive tasks[47].
2. Predictive tasks: Tasks that are used to perform inference (find patterns) on the current data for predicting the future behavior of some entities. Classification, regression, and deviation detection are examples of predictive tasks[47].

2.5.2 Modeling Credit Scoring as a Classification Problem

The main objective of CS method is to classify customers according to their different risk levels based on the available credit history information. Therefore, CS problems are basically in the scope of the more general and widely discussed classification problems [73]. In statistics this classification is known as a prediction, and in the field of machine learning it is often called supervised learning [73].

CS can be modeled as a DM classification problem for the following reasons:

1. DM (or knowledge discovery from data (KDD)) is one of the recent oncoming data analysis techniques, which consists of many steps. These steps startwith preprocessing of data and end by producing interpretable useful knowledge [47].

2. Classification is one of DM predictive tasks [47].

2.6 Data Mining Classification and prediction Techniques

Many classification and prediction methods have been proposed earlier by researchers in machine learning, pattern recognition, and statistics. The shortcoming of these methods is that, they cannot handle large datasets (memory resident algorithm)[47]. Recent DM research has been developed scalable classification and prediction techniques which are capable of handling large data[47].

2.6.1 Classification and Numeric Prediction

Classification and numeric prediction are two forms of DM tasks that can be used to predict future data trends. Classification is a task of predicting categorical (discrete, unordered) labels while numeric prediction predicting continuous (or values) for given inputs[47]. Prediction and classification also differ in the methods that are used to build their respective models.

2.6.2 Definition of Classification

Classification is a two-step process[47]:

1. Learning step: where a classification algorithm builds the classifier by analyzing or learning from a training set (tuples i.e. records or rows of tables) which are selected from the database under analysis) made up of selected tuples and their associated class labels. Each tuple X (in database) is assumed to belong to a predefined class as determined by database attribute called the class label attribute. (In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects).

   This first step of the classification process can also be viewed as the learning of a mapping or function, $y = f(\boldsymbol{X})$, that can predict the associated class label y of a given tuple $\boldsymbol{X}$.

2. Testing step: in this step a test set is used, made up of tuples and their associated class labels. These tuples are randomly selected from the general dataset. They are independent of the training tuples, meaning that they are not used to construct the classifier. Test set is used to

evaluate the accuracy of the classifier which has been built in the first step. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. If the accuracy of the classifier is considered acceptable (depending on the problem domain), the classifier can be used to classify previously unseen tuples (data) for which the class label is not known.

2.6.3 Data Mining Classification and Prediction Techniques

There are many DM classification and prediction techniques such as ANN, DT, SVM, Case-Based Reasoning(CBR), Rough Set (RS), Liner and Logistic Regression(LG),Discriminant Analysis(DA), Bayesian, *k*-Nearest-Neighbor(KNN), …etc. Each method has its own characteristics advantages and disadvantages[47].

2.7 Summary

As a conclusion of this chapter it is clear that banks cannot survive without reliable risk management systems. Among different types of risks, credit risk is the most challenging risk that faces banks. When compared with other systems CSS is currently the best credit risk evaluation system. However it has many shortcomings.

DM is a one of the recent oncoming data analysis techniques. Hence CS is molded in this research as a DM classification problem. A DM process comprises many steps starting with problem identification passing through the DM step and ending with potential useful knowledge extraction. There are many DM classification and prediction techniques that can be employed in DM step.

CHAPTER THREE

3. Literature Review

3.1 Overview

CS models have been applied by many researchers to improve the process of assessing credit worthiness by differentiating between prospective loans on the basis of the likelihood of repayment. Thus, CS is a very typical DM classification problem. A wide range of DM classification techniques have been used to develop CSMs [4-6]. This chapter discusses the most widely used DM classification techniques in CS. These techniques are categorized into three approaches: Statistical, Artificial intelligence (AI), and Hybrid. The pros and cons of each technique are also presented. See Figure 3.1.

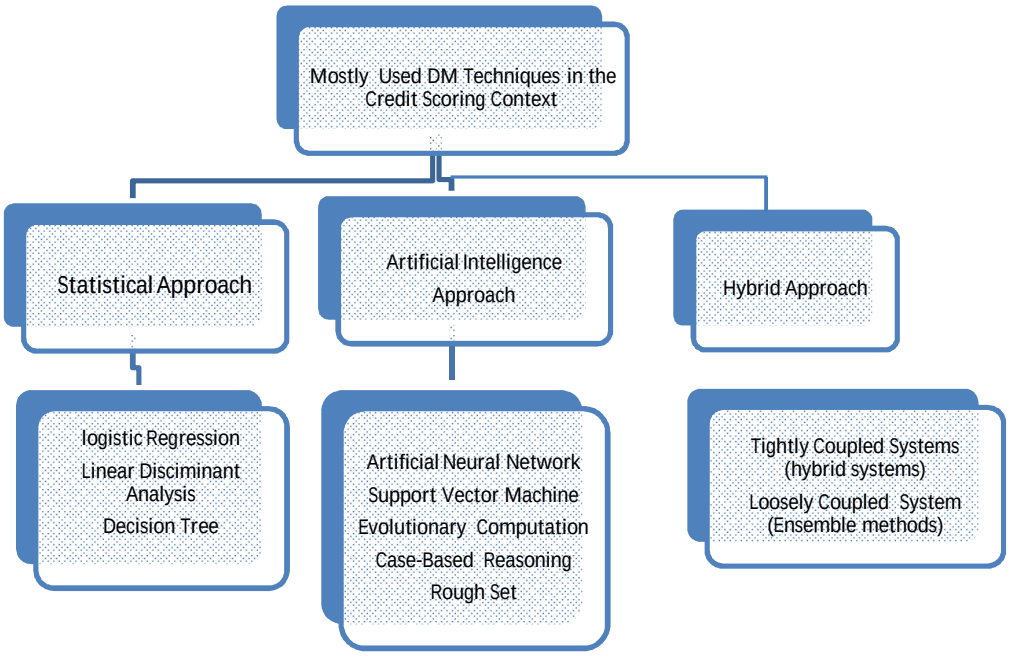

Figure 3.1: Data Mining Techniques in Credit Scoring

3.2 Statistical approach

Parametric and non-parametric statistical techniques have been successfully applied to build scoring models. This section surveys two parametric techniques and one non-parametric, namely: linear discriminant analysis (LDA), LR and DT

3.2.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a multivariate statistical technique that leads to the development of a linear discriminant function maximizing the difference between two populations' means per unit of dispersion about those means and that minimizes the likelihood of misclassification[34].

According to [71] LDA can be expressed as:

$$D= \beta_0+\beta_1 X_1+\beta_2 X_2 + \cdots + \beta_n X_n$$

Where D represents the discriminant score, β_0 is the intercept term, β_1 , … … … … … , β_n represent the coefficient associated with the corresponding independent variables X_1 , … … … , X_n .

LDA, a simple parametric statistical model, was one of the first CSMs. The first scoring model was developed by Durand [36], who examined car loan applications. In addition, West, and Baesens et al. [110, 16] proposed LDA in building a CSM. In these studies LDA performed well in many cases when it was compared with other techniques such as LR, KNN, DT, SVM and ANN.

LDA was combined with back-propagation NN (BP-NN) by Lee et al. [69] giving a hybrid model. The proposed hybrid approach came together with conventional neural networks and outperformed traditional DA and LR in terms of classification rate. Furthermore Chen et al. [24] developed CSMs using recent discriminant techniques such as Skew-normal discriminant analysis (SNDA), Skew-t discriminant analysis (STDA), Stepwise discriminant analysis (SDA), Sparse discriminant analysis (Sparse DA), Flexible discriminant analysis (FDA), and Mixture discriminant analysis (MDA). The results show that SNDA, STDA, and SDA outperformed other techniques in terms of total percentage of correctly classified cases (total PCC) and the bad rate among accepts (BRA).

Because of simplicity, LDA is still one of the most commonly used techniques in developing CSMs[7]. However, LDA sometimes suffers fromlack of accuracy due to the presumptions of linear relationship between response and independent variables, normal distribution of variables, and the equality of covariance matrices of the good and bad credit classes [110].

3.2.2 Logistic Regression

Logistic regression (LR) is a commonly used statistical modeling technique in which the probability of a dichotomous outcome is related to a set of potential predictor variables in the form [69]:

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0+\beta_1 X_1+\beta_2 X_2 + \cdots + \beta_n X_n$$

Where p= Probability of the outcome of interest,

$X_1, \ldots\ldots\ldots\ldots, X_n$Independent variables, β_0:Constant ,

β_1 , … … … … , β_n : Coefficients of independent variables,

In contrast to LDA, LR model does not require the assumptions of LDA. LR models have been widely used for developing CSMs[102, 110, 16, 62]. In these studies the LRCSMs achieved better in terms of accuracy when they are compared with other models such as LDA, ANN, KNN, and DTs. LR was also identified as a good alternative to ANN. Duki et al. [35] presented aCS decision support system based on a LR model. The obtained results from the simulated proposed system were used to determine the confidence interval for the mean probability of default, which is actually the basis for loan applicant assessment.

LR suffers from the weakness due to the model assumption, that independent variables must be linearly related to the logit of the dependent variable[67].

3.2.3 Decision Tree

Decision trees (DTs) are the popular non-parametric statistical models which are utilized for classification and prediction purposes[45].

DT learns from class-labeled training tuples. Each internal node (non-leaf node) in DT represents a test on an attribute; each branch denotes an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node[46].

Construction of DT is very easy and does not require any domain knowledge or parameter setting. Moreover DTs do not require the assumption about

probability distribution of response variable and are also applicable irrespective of the nature of response and explanatory variables [45].

In spite of the greater flexibility of DTs, they have the disadvantage of greater demand for computational resources. Furthermore DTs structure depends on the observed data, thus a small change in data alters the structure of the tree [45].

Many DT algorithms have been developed during the last few decades such as Iterative Dichotomiser (ID3), Chi Square Automatic Interaction(CHAID), Classification and Regression Trees (CART), C4.5, and C5.0 [45, 46]. These DT induction algorithms have been used for classification in several application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology [46].

The DT model is of white box nature; so it is simple to understand and explain. For this reason and the other aforementioned advantages, DTs have been applied for CS applications in a number of studies [70, 103, 68, 121, 125, 102, 13, 115].

Yu et al. and Wah et al. [102, 115] applied CHAID and C4.5 decision trees to CSand comparedthem with other models such as LR, SAS scorecard, ANN, and SVM. DTs did not achieve better results in terms of accuracy, but they had good explanatory capability. Contrary tothese results, the results achieved by Sultan et al. [13]showed that DT CSM yielded better result than ANN model in terms of accuracy .This contradiction may be because of the different datasets that were employed in these studies.

In 2010, Zurada et al. [125] applied six classification models LR, ANN, radial base function(RBF) ANN, SVM, KNN and DT to five datasets to develop six CSMs for each dataset. The assessment of these models revealed that DT achieved high accuracy in most developed CSMs.

C4.5 and C5.0 decision trees were compared with DA,BP-ANN,LR by Zhong-Yin and Li et al [121, 70], where DTs outperformed all other techniques in terms of classification rate. Actually in the first study C4.5 yielded 100% classification rate for testing and training sets. Furthermore Li et al. also compared C5.0 to CART and to CHAID DTs, C5.0 also outperformed these two types of DTs. CART was the worst one in that study. The experiment of Wah et al. [103]showed that ANN and LR were slightly better than CART. In 2006, Lee et al[68] demonstrated of effectiveness CS using CART and multivariate adaptive regression splines (MARS), the result of that study revealed that CART and MARS outperformed traditional DA, LR, ANN andSVM techniques in terms of accuracy.

It can be concluded from the aforementioned studies, where the DT model is applied to CS problem as a single classification technique that the accuracy of DTs was not stable and was easily affected by noise data and by the redundancy of the data attributes. Therefore, some researchers considered combining DT with the other data DM techniques so as to enhance its accuracy. For example, Chiua et al[27] developed a hybrid CSM by combining DT with RS. RS was used for reducing the number of features and, as a consequence, the reduction of computational resources needed for DT model. The hybrid model in that study achieved better results in terms of accuracy and transparency when it was compared with C4.5 alone in the

same study , BP, genetic programming(GP), and SVM+GA from another study. Zhang et al[119] developedanother hybrid CSM by hybridization of C4.5 with GA and k-means. GA was used to reduce the redundancy attributes of data and K-means was used to remove noise data. The result of this experiment showed that GA and K-means can effectively improve the classification accuracy of the DT CSM.

3.3 Artificial Intelligence Approach

A wide range of AI techniques have been used in the context of CS.These techniques provide a better alternative for conventional statistical techniques. Techniques, such as ANN, SVM, Evolutionary computational, CBR and RS are all widely used techniques in building CSMs.

3.3.1 Artificial Neural Network

Artificial neural network (ANN) is a mathematical technique that simulates the neurophysiology of human brain, which consists of a billion interconnected neurons working in parallel. An ANN consists of a number of very simple highly connected processors (neurons). These neurons are connected by weighted links passing signals from one neuron to another. Weights express the importance of each neuron input. ANN learns through repeated adjustment of these weights. ANN is made up of multiple layers (input, hidden, and output) [82].

ANN is one of the modern techniques that have been widely used in financial applications[18] . CS is the most famous of these applications in which ANN hasbeen used[8, 86, 77, 13, 58, 114, 121, 59, 83, 62, 16, 20].

In most studies, researchers compared ANN with traditionalstatistical methods such as DA, LR, Probit regression, Naive Bayes (NB), CART, and KNN [114, 8, 16, 20]. ANN achieved better performance than these techniques, so it is considered to be the proper alternative to these conventional techniques in CS[64, 7].

In terms of accuracy, computational complexity and processing time, Nwulu et al[83] compared ANN to SVM as a current technique. Australian dataset was utilized in that study. The Experimental results obtained indicated that although both techniques are highly efficient, ANNs obtained slightly better results and in relatively shorter times.

On the other hand, some researches compared CSSs which applied ANN with different training-to-validation ratios, different learning algorithms, different activation functions, and different number of hidden layers. Nine ANNs with different training-to-validation data ratios were developed by Khasman[59]. ANNwithtraining-to-validation ratio of (40%:60%) outperformed the others. That study acknowledged that the success of ANN was dependent upon the training –to-validation ratio. Two types of ANN were applied to the well-known Australian and German datasets by Marcano-Cedeño et al[77]; the first one was the ArtificialMetaplasticity implementation (AMP) on Multilayer Perceptron (MLP) (AMMLP) and the second one was theclassic MLP trained with BP algorithm. The AMMLP achieved impressive results compared to traditional MLP. While Pacelli[86] compared two feed-forward multi-layers neural networks: one was developed for that study and the second was built for research conducted in 2004. Two networks were developed to forecast the credit risk of a panel of Italian manufacturing companies. They used the same learning algorithms

(namely BP) but using different activation functions and a different number of hidden layers. The first network was made up of two hidden layers with sigmoid symmetric stepwise activation function while the second network made up of three hidden layers with logistic activation function. Actually this research addresses one of the main disadvantages of CSSs, that is, CSSs are actually built using historical data, so CSSs must beupdated to comply with any new changes and address deficiencies in the old system.

3.3.1.1 Limitations of Artificial Neural Networks

Despite the high classification rate of ANN in the development of CSMs, ANN is criticized for:

1. its poor performance in case of irrelevant and large number of attributes [84].
2. lack of a formal method to select optimum topology for network by setting suitable parameters[64].
3. long learning time (computational cost)[58].
4. black box nature (lack of transparency); there is no explanation why certain borrowers are classified as good and others classified as bad[64].

3.3.1.2 Efforts to Overcome Limitations

Feature selection is one of the major preprocessing steps in DM process which is solving the curse of dimensionality problem for data by removing irrelevant and redundant attributes. In context of CS several researchers

addressed this problem by applying one of the feature selection techniques to enhance ANN accuracy rate.

Šušterši and Stjepan[92, 93] attempted to enhance ANN accuracy by applying GA as the feature selection technique and compared it with other techniques such as principal component analysis (PCA), forward selection, information gain ,Gini index,….etc. These experiments concluded with the finding, that GA when applied to ANN was significantly better than other techniques. The only shortcoming of GA-ANN was that it took long time to run.

In [123] where the actual question of the study was " Does Feature Reduction Help Improve the Classification Accuracy Rates?", six models were built with different classification techniques (LR, ANN, RBFNN, SVM,……etc), and two scenarios were compared; in the first one the classification models were applied to a reduced dataset (German dataset was reduced by different feature selection techniques); in the second one the classification models were applied to the whole original independent variables. As a result of these two scenarios, it was concluded that feature reduction does not always enhance the accuracy of classification models.

Raghavendra et al[88] evaluated the effectiveness of feature selection for ANN by using other feature selection techniques such as best first search(BFS),info gain etc. For this study, three datasets were employed (German, Australian, and Japanese). As a result of that study the hypothesis (effectiveness of feature selection for neural networks) has been proved. In addition BFS, Wrapper Subset Eval, and Random Search were found to be more efficient than other used techniques.

Parameter setting is one of the factors affecting the performance of ANN because the determination of the best architecture of ANN is done informally by a trial and error method which is tedious and time consuming. Correa B. et al[30] proposed two evolutionary algorithms: GA and Binary Particle Swarm Optimization (BPS), to optimize the architecture of a (MLP network) in order to improve the predictive power of the CSM. These two models were compared with SAS minor default MLP, and LR. GA-ANN and BPS-ANN outperformed default MLP and LR. BPS outperformed GA in CPU time.

In order to speedup the learning phase Khashman[58] normalized input values of the Australian credit dataset; all numerical attributes values were normalized separately (to values between 0 and 1) to assure that normalized attributes are meaningful for the network after normalization.

Neural networks have also been criticized for their poor explanation capability, specifically when applied to CSSs because the reasoning of their decision is not available. Hence, enhancement of the transparency of neural networks acts as one of the success factors for ANN in developing CSSs. Hybridization and rule extraction from trained neural networks are the most used methods to enhance transparency for ANN[64]. Neuro-Fuzzy is a model which combines parallel computation and learning ability of ANNs with the human like knowledge representation and explanation of fuzzy systems[82]. This model was employed by Lahsasna et al. and Akkoç[66, 11] to develop interpretable CSMs. In addition to enhancing transparency, these models achieved better results in terms of accuracy when were compared with other traditional classification models.

In March 2003, Baesens et al[15] provided another treatment of the transparency problem by developing user-friendly CSSs using neural network rule extraction techniques. In that study three artificial neural network rule extraction techniques namely, Neurorule, Trepan, and Nefclass, were contrasted for credit-risk evaluation. The experiments were conducted on three real-life financial credit-risk datasets. It was shown that, in general, both Neurorule and Trepan yielded better classification accuracy when compared to the popular C4.5 algorithm and the LR classifier. In addition to rule extraction techniques, a decision table was also used in that study to visualize extracted rules in graphical format to facilitate easy explanation.

3.3.2 Support Vector Machine

Support vector machines (SVMs), which were introduced by Vapnik and his colleagues in 1995[46]have proved to be effective and promising techniques for DM [96].

SVMs have been applied for classification and prediction problems of both linear and nonlinear data. In case of linear inseparable data SVM algorithm uses a nonlinear mapping to transform the original training data into a higher dimension, and then searches for the linear optimal hyperplane for classification of the data using essential training tuples called support vectors[46].

As a promising recent competitor SVMs have been successfully applied to the CS problem. In [83, 120, 118, 19, 16] studies SVM was used to construct CSMs and compare them with other classifications techniques in terms of accuracy, misclassification, computational cost, and other evaluation

criteria, SVM classifier yielded good results in most of these studies in terms of specific criteria.

Well known classificationalgorithms such as LR, DA, KNN, DT, ANN, and the recent least square SVM (LS-SVMs) were applied to eight real life CS datasets by Baesens et al[16]. These techniques were evaluated using PCC, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) criteria. Radial basis functionLS-SVM achieved better results in term of PCC and AUC but not significant improvement over ANN, LR, and LDA. Furthermore in that study SVM was compared with ANN in terms of accuracy, computational complexity, and processing time. Experimental results showed that ANN outperformed SVM in terms of accuracy while SVM required shorter training time.

BP-ANN and GP were also compared with SVM by Zhang, D. et al[118] and achieved better results than SVM in terms of accuracy but the classification accuracy of SVM was stable(i.e. in every run for the same dataset ,SVM yielded the same results). On the other hand SVMyielded the same result as ANN when Zhang, L et al[120] applied to Australian and German credit datasets. In 2009 Bellotti et al [19] tested three SVM models (linear, polynomial, and Gaussian RBF) by applying them to a larger dataset and then compared their performance against LR, LDA, and KNN. SVM with linear or Gaussian RBF achieved highest AUC, slightly better than LR, but not significantly so.

3.3.2.1 SVM Parameters Optimization and Feature Selection

Parameters optimization and feature selection for SVM model are two major factors to enhance SVM classification performance. Several studies have introduced some techniques regarding these factors [116, 111, 51].

A combination of GA with SVMs byHuang et al[51] resulted in a hybrid model GA-SVMs, which could simultaneously perform feature selection and model parameters optimization. This study illustrated that GA-SVM model was very competitive with BP-ANN and GP in terms of classification accuracy. The only drawback of GA-SVM CSM was the long training time. It is noteworthy in this study is the improvement of the performance of the SVM model when compared with the results achieved from[118], especially since these two studies were applied to the same datasets (UCI datasets). Xu et al[111] also applied GA to optimize parameters setting of four SVM CSMs with different kernel functions. In addition to GA,PCA was also employed as a feature selection technique.

Yun et al [116] proposed a hybrid model to optimize SVM kernel function parameter and input feature subset simultaneously. This model combined SVM and PSO resulting in (PSO-SVM) which was simple and accurate when it was compared to GA-SVM, DT, SVM, LDA, etc. Accordingly, PSO-SVM was adopted as a promising approach for CS.

3.3.2.2 Support Vector Machine Main Drawbacks

Despite the successful results achieved by this technique in the context of credit scoring, it still suffers from two major drawbacks:

1. Sensitivity toward outliers and noisy data:In standard SVM, each input point is assigned to one of two classes[96]. However in CS applications, the credit data may contain outliers and noisy data[94]. These outliers may not be exactly assigned to one of these classes.

Fuzzy SVM model (FSVM) was proposed to deal with this problem.By using FSVM, each instance in the training dataset is assigned with a membership degree, if one instance is detected as an outlier, it is assigned with a low membership degree, so contribution to total error term decreases. Unlike the equal treatment in standard SVMs[96].

In many studies [109, 94, 113]FSVM model with different flavors was applied to different datasets to construct CSMs.

Tang et al. [94] applied FSVM model to two datasets. The model outperformed standard SVM and ANN in terms of accuracy and Type I and Type Ⅱ error rates. Wang et al[109] proposed new FSVM model. This new model used the same idea of the original FSVM, in addition to that, each input instance in the training dataset can be treated as both positive and negative class but with different membership degrees. The reason for this new model is that, actually in credit risk analysis, we cannot say that one borrower is absolutely good or bad. The results of this study showed that, the new FSVM achieves better performance than traditional methods if it uses RBF kernel and membership generated by LR. One of the major drawbacks of this model was the computational complexity. Furthermore, Yao [113] attempted to enhance accuracy of FSVM by using CART and MARS as feature selection techniques and using GA as a parameter optimization technique. As a result, two hybrids FSVM-based CSMs, CART-SVM and

MARS-SVM were developed and compared with each other and with CART,MARS,and FSVM. The result of Yao's study showed that the hybrid SVM not only had a best classification but also had a lower TypeΠ error rate.

2. SVM Black–box nature: "The Support Vector Machine (SVM) is a state-of-the-art classification technique that generally provides accurate models, as it is able to capture non-linearities in the data. However, this strength is also its main weakness, as the generated non-linear models are typically regarded as incomprehensible black-box models" [79].

Martens et al. [79] addressed this problem in the application of CS and solved it by applying SVM rule extraction techniques and suitable ANN rule extraction techniques. Two rule extraction techniques namely Trepan tree and RIPPER were employed in that study. German credit dataset was utilized to evaluate these techniques. These two rule extraction techniques were compared in terms of their rules outputs expressiveness. In addition, in this study RIPPER rule set was transformed into a decision table to enhance the explanation capability of RIPPER technique.

In conclusion, SVM black-box problem still needs further research regarding many issues, such as the need for intuitive rule sets, handling high dimensional data, and ranking for rules outputs expressiveness[79].

3.3.3 Evolutionary Computational Techniques

Evolutionary computing is a set of problem-solving techniques based on the Darwinian principles of natural selection and evolution[82].All evolutionary computation techniques were inspired by biological processes of inheritance,

mutation, natural selection, and the genetic crossover that occurs when parents mate to produce offspring. It makes use of the concept of survival of the fittest by progressively accepting better solutions to the problem [5]. The most popular evolutionary computational techniques are GA and GP [82].

Differences between GP and GAs refer to the particular representation of the solution: GP produces computer programs or programming language expressions as the solution, whereas GAs give a string of numbers that represent the solution. [5]

Using GA and GP in solving a problem, does not require specification of all the details of a problem in advance, because solutions are evaluated by a fitness function representing the problem to be solved. Hence GA and GP approaches have been used successfully in CS applications[5].

GA and GP in CS are either used as classification techniques or combined with other classification techniques for some tasks such as feature selection, parameter optimization, or to resolve other problems.

Desai [32] applied GA to the CSM. The GA outperformed LDA, LR, and ANN in terms of classification accuracy. Similarly, Finlay [40] utilized GA to generate CSMs. Experimental results showed that GAs can perform as well as, if not marginally better than, liner regression and LR.

In addition, GA can be combined with other classification techniques to enhance their accuracy. For examples Bahnsen et al. [17] combined GA with ANN as a parameter setting technique. In order to enhance ANN classification accuracy, Oreski et al. [85] investigated the performance of seven feature selection techniques over a dataset from a Croatian bank. The

experimental results concluded that GA-NN model was significantly better in feature selection for classification compared to some other techniques used for selecting features.

GA was also combined with SVM to perform feature selection task and model parameters optimization simultaneously by Huang et al. [51]. Furthermore, Hoffmann and Lahsasna et al. [65, 49] combined GA with fuzzy logic to extract optimal fuzzy rules in the former and to setting parameters in the latter study. Vukovic[100] also combined GA with CBR model that uses preference theory functions. In this model GA was employed for setting the parameters of each preference function, and to set attribute weights.

In developing CSMs models, GP was utilized and achieved better in terms of accuracy when it was compared with other methods such as ANN, LR , probit analysis ,etc. [6, 91, 84]. Huang et al. [52] presented a two-stage GP (2SGP) to address CS problems by integrating the function-based and the induction-based methods. First, the IF–THEN rules were derived using GP. Next, the reduced data was fed into GP again to form the discriminant function for providing the capability of forecasting. Zhang, D. et al. [117] developed another hybrid CSM (HCSM) to deal with the CS problem by synthesizing the advantages of GP and SVMs. Two credit datasets in UCI database were selected for the experiment. HCSM obtained better classification accuracy when it was compared with SVM, GP, DT, LR, and ANN.

However, the success of evolutionary computation techniques in CSSs have often been criticized because they are black-box techniques whose

resulting decisions are not easily interpretable for the financial and business analysts[5].

3.3.4 Case–Based Reasoning

Case-based reasoning (CBR) as a learning method was presented by Roger Shank, the professor of Yale University, in 1982[108]. It is a classifier which uses a database of problem solutions to solve new problems. CBR stores the tuples or "cases" for problem with their own classes as a symbolic description. When given a new case to classify, a case-based reasoner will first check if an identical training case exists. If one is found, then the associated class to that case is returned. If no identical case is found, then the case-based reasoner will search for training cases that are similar to those of the new case [46]. Hence, the learning approach of a CBR is just like the reasoning process of human beings [108].

CBR has been applied successfully to many applications, for examples bankruptcy prediction [28], medical[23] and manufacturing [98]. In addition to these applications CBR has also been applied to CS problems.

Zurada[124] applied CBR to a balanced credit dataset (ie the number of defaulters is equal to non-defaulters). In his study CBR employed KNN to classify cases (similarity measurement) and classification accuracy of CBR CSM with different k values was computed. The results achieved from this study were consistent with the results of the earlier studies. Vukovic et al. [100] proposed the CBR model that used preference theory functions for similarity measurements between cases. A GA was used for setting the parameters of each preference function, as to set attribute weights. Three

different benchmark datasets were employed to evaluate the model. The experimental results showed that the proposed approach can, in some cases, outperform the traditional KNN (based on the Euclidean distance measure) classifier. Dong[33] conducted a comparative analysis of similarity measures. The main objective of this study was to investigate the effect of similarity metrics on the performance of the CBR CS proposed system. Six distances were used as similarity metrics for case retrieval. The experiment results showed that the system's performance was almost not sensitive to the choice of similarity metrics. Furthermore García et al. [43] investigated the effect of filtering algorithms when they are applied to case-based classifiers in the context of credit risk assessment. The experiment tested twenty different algorithms in eight credit databases. The results of the study showed that, the use of filters led to significant improvement in performance and saving in storage resources when compared to the nearest neighbor prediction model with no filtering. In order to combine the advantages of CBR and ANN, Chuang et al. and Wang et al. [29, 108] developed a hybrid ANN-CBR CSMs. These hybrid models outperformed many other techniques such as ANN, SVM, MARS, and CBR in terms of accuracy. In addition decreasing TypeIand TypeΠerrors was yielded in [29].

In spite of the success and feasibility of CBR in CS problems, there are many challenges in using CBR; these include finding a good similarity metric, the selection of salient features for indexing training cases, and the development of efficient indexing techniques [46].

3.3.5 Rough Set

Rough sets(RSs), originally proposed by Pawlak in1982, are a mathematical classification tool used to deal with vagueness or uncertainty in data[84]. Rough set theory is based on the establishment of equivalence classes within the given training data. All of the data tuples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data. Given real world data, it is understood that some classes cannot be distinguished in terms of the available attributes. Rough sets can be used to approximately or "roughly" define such classes[46].

The main advantage of RSsis that, it does not need any pre-assumptions or preliminary information about the data.

Rough sets in CS are seldom used as a stand-alone solution; they are usually combined with other classification methods. Feature selection based on rough set and tabu search (FSRT) was proposed by Wang et al. [107]. FSRT was combined with classification methods such as SVM to develop a CSM. Results from this experiment have revealed that FSRT was promising and less expensive in computational cost.

The classical RS model can just deal with nominal data butCS datasets are always mixed (nominal and numerical data). So, to elevate this shortcoming, a RS was combined with fuzzy set by Yao [112]. Zhou et al. [122] also combined RS with GA-SVM to develop CSM.

Despite successfully dealing with vagueness or uncertainty in data, in some cases RS is criticized for forecasting ability, when a new object does not match any extracted rule[84].

3.4 Hybrid Approach in Credit Scoring Models

Both statistical techniques and Artificial Intelligence (AI) techniques have been explored for credit scoring, but there are no reliable conclusions on which ones are better. The reason behind this is that, the performance of CS problem depends on the details of the problem, the data structure, the characteristics used, the extent to which it is possible to identify the classes by using those characteristics, and the objective of the classification[105].

Recently, there has been a growing interest that existing applications of single AI techniques can be further improved by two approaches of hybridization [37], these are:

1. Tightly coupled systems (hybrid systems): Simple methods are combined in an inseparable unit. A combined method can overcome the limitation of method and gain advantage from another one.
2. Loosely coupled (Ensemble methods): Using multiple learners to solve the same problem where each learner can be identified as a separate unit.

3.4.1 Hybrid Systems in Credit Scoring

Actually the goal of hybrid CSS is to overcome weaknesses of the specific simple method by gaining strengths from the other participant methods in the system.

In a hybrid CSM, each simple method has such a specific role such as selection of features, classification, optimization of parameter setting, detection of outliers and noisy data, enhancement of transparency, etc. In this chapter many hybrid CSS are discussed in previous sections such as Neuro-Fuzzy, GA-ANN, PBS-ANN, SVM-GA, DT-RS, FSVM-CART-MARS…etc. All of these hybrid systems were suggested by many researchers to overcome limits of simple techniques to enhance the performance of CSMs.

According to Tasi et al. [97] hybrid systems are categorized into four types:

1. classification + classification,
2. clustering + classification,
3. classification + clustering,
4. clustering + clustering.

The first two categories of hybrid systems have been used widely in CS in which at least two simple methods are combined. Some examples of these hybrid systems are summarized in Table 3.1. The "category" in Table1 follows the classification of [97].

Table 3.1: Summary of Studies in Hybrid Credit Scoring Models

Title	Category	Methods and their Roles	Reference
Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment.	1	GA: Feature selection. ANN: Classifier.	[85]
An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for Credit Scoring Analysis: The case of Turkish credit card data.	1	ANN: Learner, Fuzzy logic: Enhance explanation capability of model	[11]

Constructing a reassigning credit scoring model.	1	MARS: Feature selection, ANN: Classifier, and CBS: Reassigning the rejected good credit applicants to reduce type 1 error of model.	[29]
A hybrid approach to integrate GA into dual scoring model in enhancing the performance of CSM.	1	GA: Feature selection. LR: Classifier (predictor)	[26]
Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier.	1	GA: Optimized fuzzy rules. Fuzzy logic: Basic classifier and enhancing transparency of the model.	[9]
A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines	1	GP: Classifier SVM: Classifier	[117]
A New Method for Estimating Bank Credit Risk.	1	DT: Classifier , RS: Feature reduction.	[27]
Credit Risk Assessment Using Rough Set Theory and GA-based SVM.	1	GA: Optimization parameter setting,RS: Attributes reduction SVM: Classifier	[122]
Hybrid Fuzzy SVM Model Using CART and MARS for Credit Scoring.	1	FSVM: Classifier CART and MARS : Select input features GA: Optimize model parameters.	[113]
Hybrid Mining Approach in the Design of Credit Scoring Models.	2	SOM: Determine the number of clusters and the starting points of each cluster. k-means: Generate clusters of samples belonging to new classes and eliminate the unrepresentative samples from each class . ANN: Classifier.	[51]
Credit risk Evaluation by Hybrid Data Mining Technique	2	k-means: Clustering techniques to re label inconsistent samples. SVM: Classifier.	[25]
A New Hybrid Method for Credit Scoring Based on Clustering and SVM (ClsSVM).	2	Subtractive clustering method: Divide dataset into clusters. SVM: classifier.	[60]
Cluster-based dynamic scoring model.	2	K-means: clustering technique. ANN: classifier.	[71]

3.4.2 Ensemble Systems in Credit Scoring

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. Learners of an ensemble are usually called base learners [105].

Multiple classifier systems can be classified into one of three architectural types:

1. Static parallel (SP),
2. Multi-stage (MS) ,
3. Dynamic classifier selection (DCS).

For these architectures, a large number of ensemble learning algorithms has been developed such as bagging, boosting, stacking, random subspace, decorate, rotation forest and so on. These methods can be applied to any base classifiers [105, 78]. The outputs from each base classifier are combined to deliver a final classification decision. A large number of combination functions are available. These include: voting; rank based, and probabilistic methods (Bayesian methods) [99].

Many studies have shown that such ensemble methods performed better than single AI techniques for CS [105, 99, 44]. See Table 3.2.

Wang et al.[107] used LR, DT, ANN, and SVM as a base learners. In their study a comparative assessment of the performance for three ensemble methods (bagging, boosting, and stacking) was conducted. The result of their experiment revealed that bagging achieved better than boosting across all datasets. Stacking and bagging DT yielded the best result in terms of average accuracy, Type I error, and Type Π error. Finlay[41] evaluated the performance of several multiple classifier systems in terms of their ability to

classify borrowers correctly. Although some multiple classifiers achieved better than the single best classifier, but many did not. In addition the new boosting algorithm (Error Trimed Boosting) was exploited and outperformed bagging and Ad a Boost by a significant margin.

To reduce theinfluence of the noisy data and theredundant attributes on the accuracy of DT Wang et al.[106] integrated two ensemble strategies: bagging and random subspace and proposed two ensemble classifiers (RS-Bagging DT and Bagging-RS DT). Two real world credit datasets were selected to demonstrate the effectiveness and feasibility of proposed methods. Experimental results revealed that the single DT got the lowest average accuracy among five single classifiers LR, LDA, MLP and Radial Basis Function Network (RBFN). Moreover, RS-Bagging DT and Bagging-RS DT got better results compared to the five single classifiers and four popular ensemble classifiers such as Bagging DT, Random Subspace DT, Random Forest and Rotation Forest. In attempting to answer the question: what base classifiers should be employed in each ensemble in order to achieve the highest performance?, Marqués et al. [78] studied the behavior of several well-known prediction models when used to construct classifier ensembles. In this study seven classification methods and five ensemble approaches have been applied to six credit datasets. The experimental results showed that, C4.5 decision tree performed the best in terms of both accuracy and TypeI error. The result of C4.5 was closely followed by MLP, LR, and SVM. The KNN and the NB models were significantly the worst in all ensembles. Wang et al.[42] proposed a new hybrid ensemble approach, called RSB-SVM, which was based on two popular ensemble strategies, i.e., bagging and random subspace and uses (SVM) as base learner. The dataset

from 239 companies' financial records which was collected by the Industrial and Commercial Bank of China, was selected to test the performance of the proposed method. RSB-SVM was compared with other seven common used methods in enterprise credit risk assessment, e.g., LRA, DT,ANN, SVM, bagging SVM, random subspace SVM (RS SVM) and Boosting SVM. The proposed RSB-SVM yielded the best performance among these methods.

In spite of superiority of ensemble CSMs in terms of accuracy when they are compared with single classifier models, they suffer from many drawbacks[63] as follows:

1. increased storage ,(multiple classifiers are employed).
2. increased computation to classify a new object (more than one classifier are processed) .
3. lack of transparency (multiple classifiers contribute in the last classification decision).

This lack of transparency may lead to limit the usage of ensemble learning methods in CS because the transparency is one of the success factors for CSMs. Rule extraction can be employed to enhance interpretability of ensemble learners [104].

3.5 Summary

This chapter presents a literature review of the mostly used DM techniques in solving the CS problem. These techniques are categorized into three approaches, statistical, artificial intelligence and hybrid approaches. The pros and cons of statistical and artificial intelligence techniques are identified and listed in Table 3.3. Two approaches of hybridization namely,

tightly coupled systems (hybrid systems) and loosely coupled (Ensemble methods) are also discussed. Studies of these systems are summarized in Tables3.1 and 3.2.

It is possible to draw the following conclusions from the literature review:

- CS can be modeled as DM classification problem.
- Many traditional statistical and modern computational intelligence classification techniques have been presented in the literature to tackle this problem.
- Traditional statistical such as LDA and LR methods are available to manage data and identify patterns but these methods work most effectively under some presumptions. These presumptions are not often realized in credit data. Therefore these techniques are inappropriate to develop CSM for the real data.
- In contrast to traditional techniques, modern intelligence techniques do not need presumptions in data.
- DTs have been applied successfully for CS applications because of their white box nature and they do not need for presumptions in data. The major drawback of DTs is that, their accuracy is affected by noise data and by the redundancy of attributes of data.
- ANN and SVM have been applied to a wide range of data types in the finance area and, in general, have had good predictive results. However, they have been criticized for their poor explanation capability.
- Recently, two approaches (tightly or loosely coupled) of hybridization have been adopted to improve the performance of single AI techniques.
- In a hybrid CSM, each simple method has such a specific role as selectionof features, classification, enhancement of transparency...etc.

- GA as a feature selection technique has been successfully used with classification techniques such as ANN, SVM, …..etc. (enhancing accuracy for CSMs).
- In these hybrid models, beside the accuracy, the transparency and the knowledge extracted from CSM will be an important feature because it will help understanding the lending process, the relations between customer features and their creditworthiness and improving the loan granting decision.
- There are a lot of DM methodologies that have been utilized to manage the problems of CS. However, each method has its advantages and limitations, and there has not been a comprehensive approach to reveal the most utilized DM technique that addresses the CS issue.
- Therefore, till now there is no best technique for CS problems for all situations .The capability of CSMs depends on the details of the problems, the data structure, the characteristics used, the extent to which it is possible to identify the classes by using those characteristics, and the objective of the classification.
- Lastly, as a result of shortage of published credit datasets, most papers have applied their experiments on the German and Australian credit datasets. Accurate assessments of different techniques need intensive experiments. These tests have to be conducted on a large number of credit datasets.

Table 3.2 Summary of Studies in Ensemble Credit Scoring Models

Title	Base Learners	Ensemble Methods	Reference
A comparative assessment of ensemble learning for credit scoring.	LR, DT and ANN	Bagging, Boosting and Stacking.	[105]
Multiple classifier architectures and their application to credit risk assessment.	LR, LDA, CART, ANN and KNN	Bagging, AdaBoost and Error Trimmed boosting.	[41]
Two credit scoring models based on dual strategy ensemble trees.	DT	Bagging and Random subspace.	[106]
Exploring the behavior of base classifiers in credit scoring ensembles.	KNN, NBC, LR, MLP and RBF, SVM with a linear kernel, and C4.5	bagging, AdaBoost, Random subspace, decorate and rotation forest.	[78]
A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine.	SVM	bagging and random subspace.	[42]

Table 3.3 Summary of Pros and Cons of Data Mining Techniques in Credit Scoring

Technique	Pros	Cons	Studies
LDA	Simplicity	Lack of accuracy due to presumptions of: 1. Linear relationship between response and independent variables. 2. Normal distribution of variables.	[69, 36, 110, 24]
LR	Simplicity	Lack of accuracy due to presumptions of: 1. linear relationship between independent variables and the logit response variable	[110, 35, 102, 62, 16]
DT	1. No need for presumptions in data 2. Applicable whatever the nature of response and explanatory variables. 3. White box nature.	1. The accuracy is not stable affected by noise data and the redundancy of attributes of data. 2. Greater demand for computational resources. 3. Structure of DT depend on the observed data, small change alter the structure of tree.	[13, 125, 70, 103, 68, 121, 102, 115] .
ANN	1. Non linear classification model 2. No need for presumptions in data. 3. High accuracy 4. High Robustness.	1. Poor performance in case of irrelevant and large number of attributes. 2. Black box nature.	[83, 59, 86, 77, 13, 58, 8, 121, 114, 62, 16, 20].

SVM	• Nonlinear classification model. • No need for presumptions in data. • High accuracy. • Overcome curse of dimensionality problem	1. Sensitivity toward outliers and noisy data. 2. Black box nature.	[83, 120, 118, 19, 109, 94, 16]
GA &GP	• No need for presumptions in data. • High accuracy	• High computational cost • Black box nature.	[40, 91, 52, 32, 84, 6]
CBR	• No need for presumptions in data. • Learning approach is like reasoning process of human beings. • White box nature	Difficulties in: • Finding good similarity metric and development of efficient indexing techniques specially in case of vast amount of data	[100, 33, 43, 124]
RS	• Deal with vague and uncertainty. • No need for: presumptions in data or preliminary information about data. • White box nature.	• Forecasting ability is weak when new object does not match any extracted rules	[107, 112, 122]

CHAPTER FOUR

4. Research Methodology

4.1 Overview

This chapter presents the different phases of this research work and discusses the methodology during the development of the proposed CSMs to achieve the objective of this research. In this study a DM framework is proposed to solve the CS problem. The problem domain is Sudanese commercial banks. Hence the preliminary stage of this research started with studying the banking sector in the Sudan to identify the currently used credit risk evaluation systems and to determine the readiness often Sudanese banks to apply credit scoring. This was followed by an intensive literature survey to identify the mostly used DM classification techniques in credit scoring.

At present a Sudanese credit dataset does not exist. For this reason one of the contributions of this research is the preparation two initial Sudanese credit datasets. As a result of the literature survey, four DM techniques have been chosen to construct the proposed CSMs. German credit dataset is also employed in this research. GA is combined with selected DM techniques as a feature selection technique. Different feature sets are selected by GA for each technique. These sets are intersected with each other (for each dataset) to produce datasets with new (reduced) set of features. Selected DM classification techniques are applied to these reduced datasets. One package is identified to simulate these models. The last stage in this research is an evaluation phase in which five measures are identified to evaluate the proposed CSMs models. Figure 4.1 illustrates the main phases of this

research.

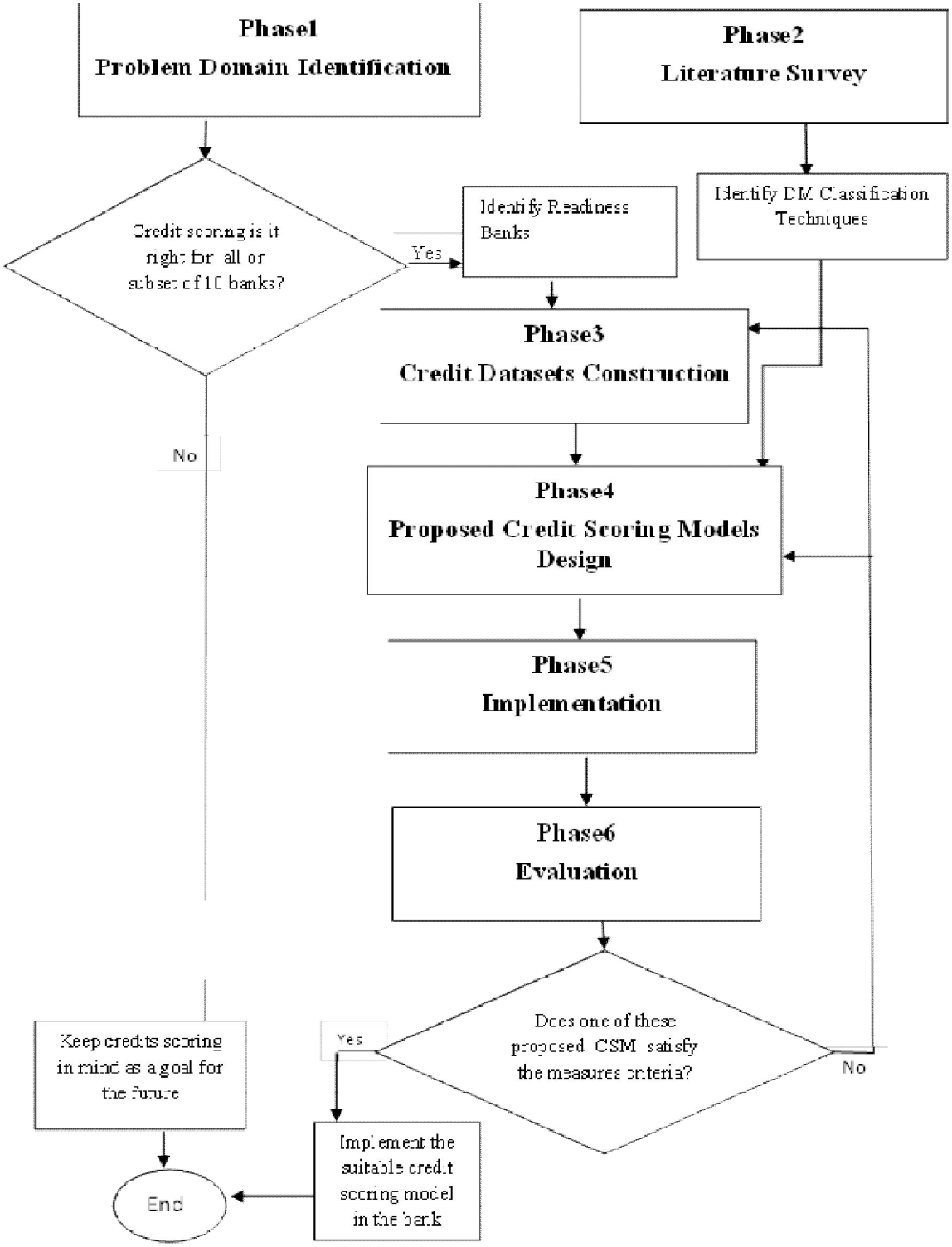

Figure 4.1 Methodology of the Research Work

4.2 Phase 1: Problem Domain Identification

This phase contains two stages. The first one is the studying of Sudanese banking sector which is the domain of this research. The second stage presents the early stage of this research which includes surveying and interviewing loan officers in many banks. The results of this survey and interviews act as a one of the leading motivation for this research.

4.2.1Sudan's Banking Sector

The emergence of Sudanese banks was at the beginning of the twentieth century in 1903[48]. The structure of these banks has evolved through different periods of time [48]. Currently, the structure of banking system consists of the Sudanese Central Bank in addition to 34 commercial banks. The banking sector in Sudan forms the backbone of Sudan's financial system and is the primary source of financing for the domestic economy. Figure 4.2 presents the structure of Sudanese banking system.

4.2.1.1 Islamization of the Sudanese Banks and Islamic Financial Modes

The decision of islamization of Sudanese banks was issued in October 1984[48]. Consequently the Central bank of Sudan issued a decree in December 1984 that ordering financial transactions to be in line with the new regulations according to ribba-free (interest-free) Islamic financing modes[48]. The Islamic financial systems derive their rules from Islamic Sharia[55].One of the main differences between Islamic banks and conventional banks is that the Islamic banks invest the funds on the basis of profit and loss sharing paradigm whereas the conventional banks resort to the

rate of interest [55]. There are many different Islamic financing modes. For more details of these modes see Appendix A.

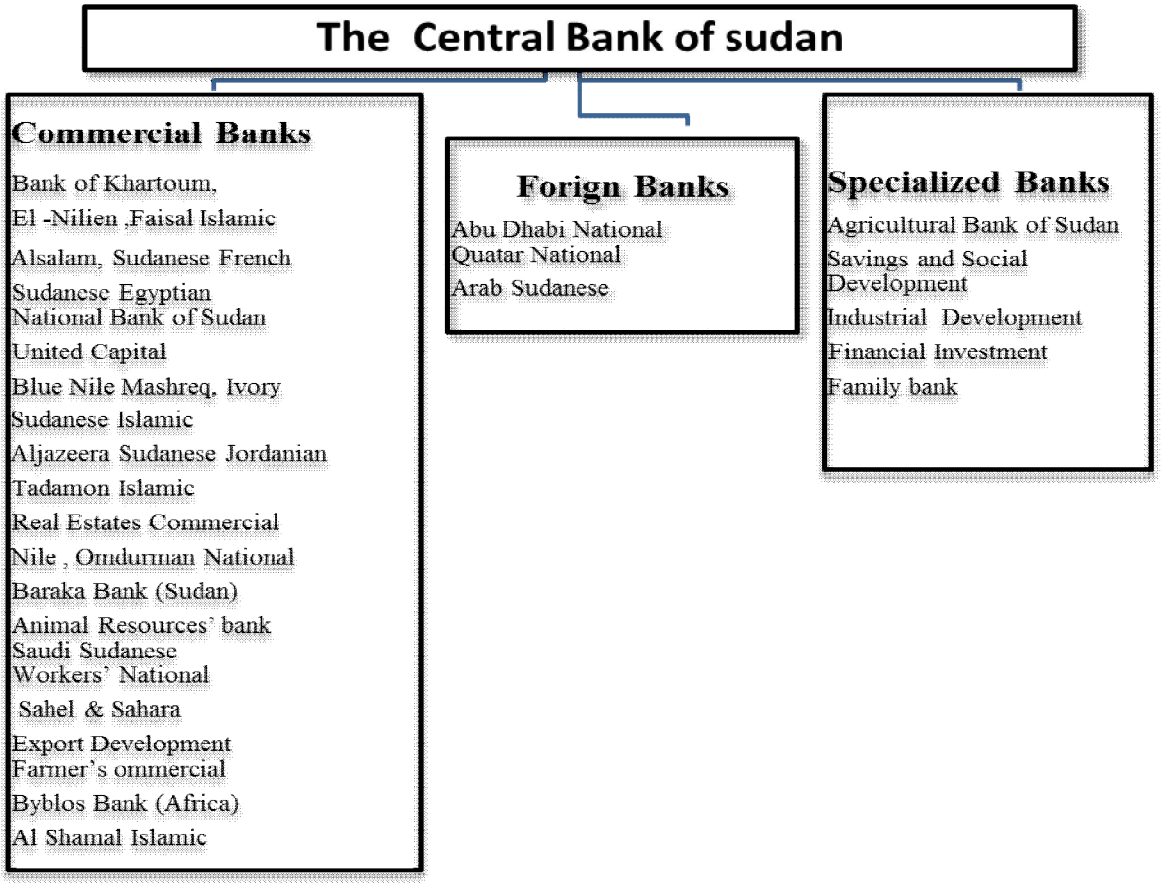

Figure 4.2 The current structure of the Sudanese Banking System [48]

4.2.2 Surveys and Interviews

In order to understand and recognize problems concerning the loan granting process and identifying the currently used credit risk evaluation systems in Sudanese banks. A fact finding surveys and non- structured interviews (in different department such as Finance, Investment etc.) were conducted in ten banks. In addition structured personal interviews were designed and conducted with loan officers in 10 banks. See Table 4.1.

The main objectives of these interviews are as follows:

- Identifying the currently used credit risk evaluation systems in Sudanese banks.
- Gathering opinions of loan officers regarding these systems.
- Examining the readiness of banks for credit scoring.
- Identifying goals of (DM process) the proposed Sudanese CSMs (SCSMs) from the loan officer's view point. The objectives and the uses of scoring models (goals) drive the entire DM process. They are the basis on which the DM project is established and decide the criteria by which the final model will be judged.

Table 4.1 structured Interview Questions

1.	Do you know what CS is, and does your bank use CS to evaluate the client's loan application?
2.	What are the key characteristics included in the currently used credit evaluation systems, describe the main characteristics of your bank's credit policy?(steps taken , methods)
3.	What are the advantages and disadvantages of the currently used credit evaluation system?
4.	Why is your bank not using credit scoring?
5.	How much time is spent on taking a loan granting decisions?
6.	How many persons contribute to this decision?
7.	Is there a need to automate credit risk evaluation decisions?
8.	What about data concerning closed loan process, is it stored? manually or electronically?
9.	Do you use historical data regarding closed loan processes in decision making in the future?

These are the following deliverables from this phase:

- Shortcomings of currently used systems.
- The selection of two banks in the basis of their readiness for CS.

- Obtaining real credit datasets is a problematic and time consuming task due to data sensitivity and the privacy preservation of data. Hence in this research only two banks were selected to apply CS. In the next chapter many challenges that facing building Sudanese credit datasets will be illustrated.

4.3 Phase Two: Literature survey

In this phase more than 100 scientific papers were read and analyzed to give a solid background of CS and to identify DM classification techniques that are used to develop CSMs. By the end of this phase the mostly used DM techniques have been identified and classified. Pros and cons for each technique in CS problem are also determined. As the result of this intensive analysis a good planning for a survey paper with the entitled “Credit Scoring Using Data Mining Techniques: A survey” was ready. By the end of this phase appropriate techniques that have been employed for proposed CSMs were chosen.

4.4 Phase Three: Credit Datasets Construction

In Sudan credit agencies and credit bureaus do not exist at present. Financial organizations have not built credit datasets from the performing and non-performing loans. Hence the main objective of this phase is to construct Sudanese credit datasets. Three stages were conducted in this phase to achieve the objective of this phase.

4.4.1 Creation of Datasets

The main objective of this stage is to prepare two Sudanese credit datasets for the selected two banks in phase one.

The following are the activities conducted to prepare two credit datasets:

- Identifying data concerning closed past loans from two banks under the direction of loan officers.
- Merging data, if it is found into separate (files) departments.
- Interviewing loan officers to identify relevant features that affect the loan granting process in context of Sudanese banks.
- Selecting the attributes determined by loan officers.
    - Classifying each closed loan process (defaulted or not) according to the financing mode and the procedures of the banks.

The outputs of this stage are two initial Sudanese credit datasets.

4.4.2 Preprocessing of Datasets

Quality of mining results depends on quality of data. Therefore, the data preprocessing step is one of the main steps in DM process that is applied to enhance the quality of data[46].

According to [46] preprocessing techniques consist of:

- **Data cleaning:**which can be applied to fill missing values, remove noise and outliers, and correct inconsistencies in the data.

- **Data integration:** which merges data from multiple sources into a one data store.
- **Data transformations**: In data transformation, the data are transformed into forms appropriate for mining. Data transformation consistsof normalization where the attribute data are scaled so as to fall within a small specified range. Other data transformations are aggregation or summarization, and generalization.
- **Data reduction:**which can reduce the data size by aggregating, eliminating redundant features, or clustering for instance. A reduced dataset that is much smaller in volume but have to produce the same (or almost the same) analytical results of original dataset.

Data preprocessing techniques have to be applied before mining to improve the overall quality of the patterns mined and/or the time required for the actual mining[46].

In this research Data preprocessing stage consists of the following activities:

- Manipulation of missing values.
- Normalization of the numerical attributes.
- Removing the outliers.
- Transformation of the categorical attributes to numerical (to gain numerical version of a dataset).
- Labeling of instances.
- Creation of new attribute

4.5 Phase Four: Design of the Proposed Credit Scoring Models

In this phase many proposed CSMs (single and hybrid) for each dataset were designed by using different DM classification techniques. Three stages were accomplished in this phase.

4.5.1Building Credit Scoring Models Using Single Techniques

DM classification techniques are identified to build CS models. Three commonly discussed DM classification techniques were chosen in this stage namely: ANN, SVM, and DT were applied to the two Sudanese credit datasets. German credit dataset was also employed in this stage as a benchmarking dataset. Holdout and 10-fold cross-validation methods were used to randomly split the data. Holdout method divides the given data into two independent sets, a training set and a test set. The training set is used to derive the model. The accuracy is estimated using the test set[47]. For this research two ratios for training:testing were chosen. These are 60:40 and 70:30. For K-Cross-validation method the initial data are randomly partitioned into k mutually exclusive subsets or "folds," each of approximately equal size. Training and testing is performed k times[46]. For this research K was set to 10.

The output of this stage was nine CSMs for each dataset (Two Sudanese credit datasets and German dataset).

4.5.2Building Credit Scoring Models Using Hybrid Techniques

In this stage GA is combined with techniques used in section 4.5.1 as a feature selection technique. The main objective of this stage is to enhance performance of single techniques in 4.5.1.These hybrid techniques were

applied to two Sudanese credit datasets and German credit dataset. Holdout and 10-fold cross-validation were also used.

The output of this stage is nine hybrid CSMs (GAANN, GADT, and GASVM with 60:40 and 70:30 and 10-cross validation) for each dataset. Furthermore a list of attributes and their weights for each model is output.

4.5.3 Datasets Reduction

For each model (GA ANN, GADT, GASVM) the mostaccurate one is chose for each dataset.For example, if the accuracies for GANN are 78, 80, 70 for 70:30,60:40, and cross validation respectively. Then the GAANN with 60:40 is chosen and their reduced subsets of features are identified for this model. This step was repeated for other models (GADT and GASVM). Therefore three sets of features (S1, S2, and S3) are identified for each dataset. These sets are intersected to produce new reduced dataset (R1). The feature is selected in RI if it appears in at least two sets.

Therefore, new set of features are identified for each dataset. The output of this stage is three reduced datasets.

4.5.4 Building Credit Scoring Models Using Reduced Datasets

Apply ANN, SVM and DT tothe datasets derived from 4.5.3.

4.6 Phase 5: Implementation

The main objective of this phase is to identify simulation tools to develop the proposed CSMs of phase4. RapidMiner package was used in

experiments for this study as simulation modeling technique. An addition an Excel sheets are also used to store the datasets.

4.7 Phase 6: Evaluation

The main objective of this phase is to identify evaluation criteria for the proposed CS models and validate them to choose the suitable model for each dataset. In order to achieve this objective the measures criteria have to be identified and the results have to be validated.

4.7.1 Identification of Measures Criteria

In this research CS is modeled as a classification problem. There are many criteria that are used to evaluate the classification model such as accuracy, robustness, interpretability …..etc. [46].

The following criteria that were used to evaluate the proposed models[56]:

- **Accuracy:** Evaluates how accurately a classifier will classify future data, that is, data on which the algorithm has not been trained.
- **Precision** (of the class defaulter): The percentage of tuples classified as "defaulter" that are actually defaulter tuples.
- **Precision** (of the class non-defaulter): The percentage of tuples classified as "non-defaulter" that are actually non-defaulter tuples.
- **Type Ⅰ error:** The rate of classifying customers as defaulters" when they are non-defaulters.
- **Type Ⅱ error:** The rate of classifying customers as "non-defaulters" when they are defaulters.

4.7.2Validation Results

The main objective of this stage is to identify the most suitable DM classification technique forSudanese banks CSMs. All CSMs for the two Sudanese and German credit datasets are compared using measures criteria identified in 4.7.1. In addition results are discussed and the suitable DM classification techniques for each dataset are identified.

4.8 Summary

This chapter presented the research phases, how each phase was conducted, and how these phases are related. The activities, objectives, and outputs for all stages in different phases were also illustrated. This methodology is compliant with the objectives of this research as stated in chapter one. A general overview of this research methodology is summarized in Table 4.2.

Table 4.2 A Quick Review of the Research Methodology

Stage	Activities	Objective(s)	Outputs
Phase1 Problem Domain Identification			
1. Sudan's Banking Sector	- Studying of Sudanese banking sector	- To illustrate the financing modes that are employed in Sudanese banks.	- The main difference between Islamic financing modes and other traditional modes.
2. Surveying and making Interviews	- Design structured interviews to loan officers in banks - Making personal interviews (structured and not structured) with loan officers in 10 banks. - Examining the readiness of banks for CS	- To identify the currently used credit risk evaluation systems. - To gather opinions of loan officers in currently used credit risk evaluation systems and identifying goals of (DM process) the proposed l (SCSMs) from their view point. - To identify banks those are ready to employ CS.	- Shortcomings of currently used credit evaluation systems - Two selected ready banks
Phase 2 Literature Survey			
Literature Survey	- Reading and analysis scientific papers - Prepare the planning of survey paper.	- To build Back ground of CS - To identify the mostly used DM classification techniques in CS problem.	- Benefits and shortcomings of CS approach - The mostly used DM classification techniques. - The pros and cons of these techniques
Phase Three: Credit Datasets Construction			
1. Creation of Dataset	- Identifying data concerning closed loans from two banks. - Merging data - Interviewing loan officers to identify relevant features that affect the loan granting process. - Selecting the attributes determined by loan officers. - Classified the each loan status (defaulted or not) according to the financing mode.	- To Prepare the two initial Sudanese credit datasets	- Two initial Sudanese credit datasets

2. Preprocessing of Datasets	- Manipulation of missing values. - Normalization of the numerical attributes. - Removing the outliers. - Transformation of the categorical attributes to numerical.	- To Prepare high quality datasets	- Two Sudanese credit datasets. - Two versions for each dataset(numerical and categorical)
Phase Four: Design of the proposed CSMs			
1. Building CSMs using single techniques	- Applying DT, SVM, ANN to Sudanese credit datasets and to German credit data. - Applying cross validation and handout techniques.	- To identify the suitable technique for the Sudanese credit datasets. - To compare the results of Sudanese and German credit datasets	- Nine CSMs models for each dataset.
2. Building CSMs using hybrid techniques	- Applying GA as a feature technique to proposed and German models - Comparing the results of single models with hybrid models.	- To enhance the performance of proposed single CSMs.	- Nine hybrid CSMs (GAANN, GADT, and GASVM) for each dataset. - A list of attributes and their weights for each model.
3. Datasets Reduction	- Identifying hybrid CSMs with highest accuracy. - Determine the features selected by GA for these models. - Making intersection reduces datasets	- To determine the effect of feature selected by GA.	- One reduced dataset for each dataset.
4. Building CSMs using reduced datasets	- Applying DT, SVM and ANN to the reduction datasets.	-To construct CSM using reduced datasets	- Nine CSMs for each dataset.
Phase 5: Implementation			
1. Implementation	- Identification of simulation tools for the proposed CSMs.	- To simulate the proposed CSMs.	- Implemented CSMs using RapidMiner, and Excel sheet
Phase 6: Evaluation			
1. Identification of measures criteria	- Identify the suitable evaluation criteria	- To validate CSMs results	- Set of classification measures criteria: accuracy, Precision and Type I and Type II errors.
2. Validate the results	- Compare the models results with evaluation criteria	- To choose the suitable DM classification techniques for Sudanese banks CSMs	- Most appropriate DM classification techniques for Sudanese banks CSMs.

CHAPTER FIVE

5. Data Collection, Datasets and Models Construction

5.1 Overview

This chapter describes the implementation phase of this research. This phase consists of stage for collecting data from different banks. Surveys and interviews with, managers, loan officers, and others in ten banks were conducted. As a result of these surveys, interviews, and CS readiness test two banks were identified. Two Sudanese credit datasets were constructed. These two datasets and the German dataset were employed to build and evaluate the proposed CSMs.

5.2 Data Collection

5.2.1 Surveys and Interviews' Outcomes

As mentioned in the previous chapter, this research was started by conducting fact finding surveys and interviewing loan officers and others in ten different Sudanese banks.

This survey was started by visiting ten banks to represent our idea. Many challenges that faced these surveys were as follows:

- Entering banks as a researcher. There is no formal procedure to do that and it depends on personal relations. This explains why we could visit ten banks only.

- Welcoming and understanding our subject and idea. We make a little presentation to clarify objectives of our project, CS method, information that are needed to do our project ….etc. Based on the success of this step we were directed to the specific departments that are appropriate for this research. Sometimes we found that this was not the appropriate department and we were directed to another one and so on till find the appropriate ones. In each department we were repeating the same presentation.
- Accepting the CS as a credit risk evaluation system. While it is very common for bankers unfamiliar with scoring to initially react to it with considerable skepticism, equally often we have found experienced bankers quickly come to appreciate the great potential scoring holds for improving loan granting process.

These challenges limit the number of banks to be surveyed to only ten banks.

5.2.2 Structured Interviews' Findings

In order to understand the bank's credit policies and procedures, steps taken to grant loan, and the role played by each participant in the process, structured interviews were conducted with ten loan officers in different banks. Table 5.1 represents the summery of findings for these interviews.

Table 5.1 Summary of the Key Findings from the Structured Interviews for Ten Sudanese banks

Question	Responses	Percentages%
1. Do you know what CS is, and does your bank use CS to evaluate the client's loan application?	Yes (for the 1st part)	50
	No (for the 1st part)	50
	Yes (for the 2nd part)	0
	No (for the 2nd part)	100
2. What are the key characteristics included in the currently used credit evaluation systems, describe the main characteristics of your bank's credit policy?(steps taken , methods)	Steps taken : • Studying the feasibility study for the project. • Making financial analysis for the loan application. • Studying the appropriateness of guarantees introduced by applicant. • Field visit.	100
	Credit risk evaluation system Using judgmental	40
	Using credit rating	30
	Using hybrid system (judgmental + rating)	30
3. What are the advantages and disadvantages of the currently used credit evaluation system?	**Advantages :** Reduced the likelihood of defaulting.	80
	The granting decisions are taken after negotiation with many persons in different positions	10
	The granting decisions are compatible with general standard.	10
	Disadvantages: Long time taken to take decision	40
	Depends on loan officers experiences.	10
	Difficulties in determine rating and making calculations of scores.	10
	The difficulty of making loan granting decision because of lack of information.	10
	No disadvantages	30
4. Why is your bank not using credit scoring?	It is unknown to the bank.	20
	International CSS is not suitable to Sudanese bank.	20
	Keeping CS as a goal for the future.	50
	Lack in information	10
5. How much time is spent on taking a loan granting decision?	3-7 days	40
	3 days	40
	3-30 days	10
	7-14 days	10

6. How many persons contribute in (to) this decision?	3-7 persons	50
	9-10 persons	20
	1-3 persons	20
	13persons	10
7. Is there a need to automate credit risk evaluation systems?	Yes	80
	No	10
	Yes, but in addition to personal evaluated	10
8. What about data concerning closed loan process, is it stored? Manually or electronically?	Manually	40
	Electronically	50
	Part is electronically and another part is sored manually	10
9. Do you use historical data regarding closed loan processes in decision making in the future?	Yes, in loan granting decisions and other analysis.	80
	No, it is not employed in granting decisions but for the other purposes.	20

The outcomes of these surveys and interviews can be summarized as follows:

1- All ten banks employ either judgmental or credit rating systems to evaluate the credit risk.

2. The loan granting process consists of the following major steps:

A.1 The loan applicant submits an application for a fund to the Bank together with all the needed information, guarantees and feasibility study for his project. Application forms contain demographic information (name, age, number of wives, etc.) and financial information (Finance duration, Finance form, Loan type, etc.).

B.1 The Bank determines whether this request is in line (consistent) with the internal bank and government financial policies.

C.1 The finance officer prepares a financial analysis for the application. Information about the client's business activities is also collected and used for this analysis.

D.1 After that, the risk management department determines the level of risk of the project (borrower credit worthiness). Based on the result of credit risk assessment, the director of local finance issues an approval within his authorities or gives a recommendation to the general manager. The decision in this step is taken judgmentally by loan officers in some banks or using the credit rating adopted by the bank. The final decision mainly depends on the output of this step.

E.1 In case of major loans, the Board of Director of the bank issues the final decision (approval or rejection).

F.1 The finance officer informs applicant by the decision .

G.1 In case of approving the loan request, the bank sends the customer's data to the Central Bank of Sudan, which gives the customer a unique code called "Credit Code". This code is unique and is used to identify borrowers.

H.1 After the bank has received the Code, a bank's employee sits with the customer to define the profitability of this project and the amount of benefit from it to determine the required payment and profit margin based on the Loan application information.

I.1 Then IT Department opens a file for the process and records all information about it.

J.1 Lastly, the follow-up process starts by the Compliance Officer to follow up the repayment process.

5.2.3 Loan Granting Process Shortcomings in Sudanese Banks

These are the general problems for the loan granting process:

- Loan granting decision is taken judgmentally based on the loan officers experiences. (Disadvantages of judgmental credit risk evaluation systems were stated before in chapter 2).

- Time consuming: time is spent on taking a loan granting decision is about (3-4) days. Some kind of loans needs quick approving decisions.

- Large number of participants in the loan decisions.

- The difficulty of making use of the historical data (old cases) in the decision of new similar loans. Historical data is too huge to be analyzed by human.

5.2.4 Readiness Factors for Credit Scoring

To select Sudanese banks for applying CS, a CS readiness test was applied to these ten banks. This test includes the following factors:

- Acceptance: which is determined by the degree welcoming, understanding, and the support of employing CSSs instead of the judgmental systems.
- Historical data: From the definition of CSMs, these models rely absolutely on historical data. Historical data in banks may or may not exist. Some banks delete all data concerning the previously made loans (closed loans). In new banks there are not enough closed loans. In addition, the data must include a fairly large number of each type of outcome (defaulters & non-defaulter).

- Data consistency: CS draws on types of similar attributes for all borrowers (consistency). Loan data in some banks is not consistent for all borrowers.
- Type of storage: historical data in banks may stored manually, electronically, or hybrid (part is stored electronically and other is manual).
- Data Providence: some banks refused to provide the data because of privacy preservation reasons.

Based on the aforementioned factors and the findings of the structured and non-structured interviews, the CS readiness test was applied to these banks. Tables5.2 and 5.3 illustrate the result of this test. To preserve the privacy of these banks, in this table we named banks (bank1, bank2,….., bank10) instead of the actual names of the banks. Actually only two banks (Bank5 and bank9) were the two banks that passed this test.

Table 5.2 Results of Credit Scoring Readiness Test

Bank#	**Acceptance**	**Historical Data(EE,D,EN)**	**Consistency**	**Storing Data (M,E,H)**	**Data Providence**
Bank1	x	EE	x	M	x
Bank2	✓	D	✓	M	✓
Bank3	✓	EE	✓	M	✓
Bank4	✓	EN	✓	E	✓
Bank5	✓	EE	✓	E	✓
Bank6	✓	EN	x	M	✓
Bank7	✓	EN	✓	E	✓
Bank8	✓	EE	✓	H	✓
Bank9	✓	EE	✓	E	✓
Bank10	x	EN	✓	E	✓

Table 5.3 Key to Abbreviations of Table 5.2

Acceptance	✓ : The CS is accepted by the bank x: CS is not accepted	**Historical Data**	**EE**: Existing and Enough. **EN**: Existing and Not enough, **D**: Deleted.	**Consistency**	**✓:**Data is similar for all borrowers **x**:Data is not similar for all borrowers	**Storing data**	M:Manually E:Electronically H: Hybrid	**Data Providence**	✓: Data is provided x: Data is not provided

5.3 Datasets Construction and Description

5.3.1 Datasets Construction

Two credit datasets for two selected Sudanese banks (that passed the CS readiness test) were built for this research.

5.3.1.1 Sudanese Credit Dataset1

Sudanese Credit Dataset1 (SCD1) was provided by Agricultural Bank of Sudan. This bank is one of the largest Sudanese specialized banks. Loan data of this bank was stored in two Excel sheets under the management of statistics and information department in the bank. The first sheet contains demographic information of borrowers such as birth date, material status, number of spouses…etc. The second sheet contains financial information such as loan type, finance size, financing form … etc. This sheet also contains the follow-up repayment information such as repaid installments, the remaining installments… etc. The bank provided us with loan data for the past three years. The data contained 2500 records for past closed loans.

5.3.1.2 Sudanese Credit Dataset2

Sudanese Credit Dataset2 (SCD2) was provided by Al Salam Commercial Bank. This bank is one of the recent commercial banks in the Sudan. Like the first bank loan data is stored into Excel sheets under the management of the IT department in the bank. The first sheet also contains demographic information of borrowers. The second sheet contains financial information, and follow- up information of the repayment for the loan process. The data contains past loan processes for retail financing (i.e. the financing mode for these loans is Murabaha). The data contain 3299 records for past closed and current loans.

5.4Datasets Construction

For constructing these two datasets, similar steps were taken as follows:

5.4.1 Identification of Data

The first step to construct these two datasets started with interviewing loan officers in the two selected banks to explain the meaning of attributes and to identify the attributes that existed when the borrower introduced his loan application i.e. before the loan process started. Therefore, many attributes such as the attributes related to follow up of the process were removed. Furthermore under directions of loan officers the loan variables which were used in the loan granting decision-making process were identified. In addition, attributes that can identify the clients such as names, telephone numbers, residential addresses, credit codes,….etc. were removed. Furthermore in SCD2 all tuples for the current loans were also removed.

5.4.2 Data Integration

The Data saved electronically in the two files (Excel sheets). Credit code attribute was used to join the two sheets. Thus the data was integrated into one Excel sheet. After that Credit codes were removed to preserve privacy of clients.

5.4.3Missing Values Manipulation

Six methods were suggested by[46] to fill the missing values:

1. Ignore the record.
2. Fill the missing value manually.
3. Use a global constant.
4. Replace the missing value with the mean.

5. Replace the missing value with the mean of all samples of that category.

6. Use the most likely value through the help of regression.

For method 2 actually we do not know the actual data to be filled. Methods 3 to 6 bias the data.Therefore, in this research we adopted the first method by eliminating the tuples that contain the missing value in the two datasets. The advantage of this method is that, the model would be based on actual data and not guessed data.

In addition, all attributes in which the percentage of missing values was more than 40% were removed from datasets. After this step the number of tuples in SCD1 and SCD2 datasets was reduced to 1310 and 960 tuples respectively.

5.4.4 Numerical Attributes Normalization

Normalization process scaled values of continuous attributes to fall within a small, specified range[46].

Normalization is particularly useful for some classification algorithms such as neural networks. Normalizing the input values for classification techniques will help speed up the learning phase[46].

Three methods to normalize the data were discussed by [46]namely, Min-max normalization; Zero-mean normalization method where the normalization is based on the mean and standard deviation of the attribute; and Normalization by decimal scaling.

Min-Max normalization method was adopted in this research. Two attributes in SCD1 (Monthly Salary Value and Monthly Expenditures Value) and three attributes in SCD2 (Approved Amount, Profit Margin, Periodical Instalment Amount) were normalized to values between 0 and 1.

Min-max normalization performs a linear transformation on the original data. Suppose that minA and maxA are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v, of A to v0 in the range [newminA, newmaxA] by computing:

$$v0 = \frac{v - minA}{\text{maxA} - \text{minA}}(newmaxA\text{-}newminA)+new\ minA)$$

Example:

A= attribute approved amount in SCD2

minA=3900,new minA=0

maxA=685400,new max A=1

Table 5.4 presents the values of attribute (a part of A values) before and after normalization.

Table 5.4 Result of Normalization for Approved Amount Attribute

Approved Amount (before normalization)	**Approved Amount (after normalization)**
3900	0
120000	0.016948
46600	0.006233
89232	0.012456
199000	0.02848
149845	0.021304
20000	0.00235
32000	0.004102
152165	0.021643
20000	0.00235
10000	0.00089
6854400	1

5.4.5 Outliers Removing

An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data [89]. To detect outliers in two datasets, detect outlier distance operator in RapidMiner was employed. This operator depends on the outlier detection approach recommended by Ramaswamy,S; et al. [89] . In their paper, a formulation for distance-based outliers is proposed that is based on the distance of a point from its k-th nearest neighbor. Each point is ranked on the basis of its distance to its k-th nearest neighbor and the top n points in this ranking are declared to be outliers. The values of k and n can be specified by the number of neighbors and number of outliers parameters respectively. This search is based on simple and intuitive distance-based definitions for outliers by Knorr and Ng [61] which in simple words is: “A point p in a dataset is an outlier with respect two parameters k and d if no more than k points in the dataset are at a distance of d or less from p”. K (number of neighbors) and n(number of outliers) are chosen to be 10 in this research. Euclidean distance function was chose as a distance function (used for calculating the distance between two tuples). Table 5.5 presents part of SCD2 after using Rapid-Miner detecting outlier operator. This operator adds a new Boolean attribute named 'outlier' to the given dataset. Outlier value False means that the tuple is not outlier. In Table 5.4 Tuple 73 was identified as an outlier tuple. 10 outliers were identified for each Sudanese Credit Datasets.

Table 5.5 Result of Rapid Miner Detecting Outlier Operator

Row NO	Status	**Outlier**	Gender	Age	Material Status	#Children	#Spouses	occupation phone	ID Type
71	Non-Defaulter	FALSE	0	33	1	0	0	13	1
72	Non-Defaulter	FALSE	0	39	2	0	1	13	1
73	Non-Defaulter	**TRUE**	0	72	2	0	1	12	1
74	Non-Defaulter	FALSE	0	29	1	0	0	13	1

5.4.6 Transformation

The two original credit sets have both categorical and numerical attributes. Many classification techniques such as ANN and SVM cannot accept categorical attributes. Therefore, all categorical attributes in the two datasets were transformed to numerical attributes. This led to availability of two versions for each dataset (original and numeric version). Table 5.6 presents the transformation process of “Occupation” attributes in SCD1 to numerical attribute. All categorical attributes in the two datasets were transformed in this way.

Table 5.6 Transformation of the Occupation Attribute

Occupation values in SCD1 original version	Occupation values in SCD1 numerical version
Farmer	1
Free business	2
Worker	3
Teacher	4
Employee	5
Policeman	6
Merchant	7
Pensioner	8
Lawyer	9
Other	10

5.4.7 Instance Labeling

Data collected for the closed loans in two selected banks were not labeled (this data is not collected specially for CSM).DM classification techniques cannot be applied to unlabeled tuples (instances). Therefore, tuples in two datasets were labeled manually in this research.

For labeling data two dates were compared in datasets, actual due date for the loan process (date on which an installment loan must be paid in full) and the actual date on which the creditor repaid the loan. According to the result of this comparison (the difference between these two dates) and the grace period for finance mode of the loan process, the status of the tuple was identified. Grace period is a provision in most loan contracts which allows payment to be received after a certain period of time after the actual due date. During this period no late fees will be charged, and the late payment

will not result in default or cancellation of the loan. Grace periods for finance modes are specified by the Central Bank of Sudan.

For example if the period between two dates is more than one month and the finance mode is murabahah (Grace period of murabahah is only one month) then the status of this loan is defaulter, otherwise is not defaulter. For other modes the grace period is three months.

5.4.8Age Attribute Creation

Based on the two dates, Birth date and start-date of loan, a new attribute "Age" has been created. According to loan officers viewpoint, age is one of the most effective attribute in loan granting decisions.

By the end of these processes two Sudanese Credit Datasets (two versions for each dataset) were constructed and ready to be employed in CSMs.

5.5 Datasets Description

This section presents the detailed description for the three credit datasets namely SCD1, SCD2 and the German credit dataset. Appendix B presents parts of these Sudanese datasets.

5.5.1 Description of the Sudanese Credit Dataset 1

Title: Sudanese Credit dataset1 (SCD1).
Source Information: Agricultural Bank of Sudan.
Classes of dataset: Instances of this dataset was classified into two classes, Defaulter and Non-defaulter.

Number of Instances (tuples): 1300 instances. 720 are classified as non-defaulters and 580 as defaulters.

Versions of Dataset: Two versions of this dataset were provided. The first version (SCD1version1) contains numerical and categorical attributes (original dataset). The second version (SCD1 version2) of this dataset contains numerical attributes (all categorical attributes in SCD1version1 were transformed to numeric ones).

Attributes of SCD1: Number of Attributes in SCDS1 is 17 attributes in addition to class attribute (Status).

Types Attributes in SCD1version1: (5 numerical, 12 categorical)

Attribute description for SCD1:

For all categorical attributes the values is transformed to integer values in version 2.

Attribute 1: (categorical)

Have phone

1: Have.
2: Not have.

Attribute 2: (categorical)

ID Type

1: Personal Card.

2: Military card.

3: Permanent Residence.

4: Driving license.

5: Passport.

6: Juridical card.

Attribute 3: (categorical)

Gender

0: Male
1: Female

Attribute 4: (numerical)
Age

Attribute 5: (categorical)
Occupation

1: Farmer.
2: Free Business.
3: Worker.
4: Teacher.
5: Employee.
6: Policeman.
7: Merchant.
8: Pensioner.
9: Lawyer.
10: Other.

Attribute 6: (categorical)
Material Status

1:Single
2:Married
3:Divorced

Attribute 7: (numerical)
Number of Dependents

Attribute 8: (numerical)
Number of Spouses

Attribute 9: (numerical)
Monthly Salary Value

Attribute 10: (numerical)
Monthly Expenditures Value

Attribute 11: (categorical)
Finance size

1: small.
2: micro.
3: normal.

Attribute 12: (categorical)
Finance duration

1: long.
2: short.
3: medium.

Attribute 13: (numerical)
Payment Method

1 (payment at the end of duration).
30 (payment every 1 month).
90 (payment in quarter of year).
360(payment every 1 year).

Attribute 14: (categorical)
Finance Form

1: Murabaha.
2: Salam.

Attribute 15: (categorical)
Loan Type

1: Auto.
2: Irrigate.
3: Traditional.
4: Local Trade.
5: Transport.
6: Professional.
7: Industry**.**

Attribute 16: (categorical)
Insurance Description

1: Future checks.
2: Guarantee letter.
3: Mortgage mechanic.
4: Direct storage.
5: Mortgage car.
6: Mortgage real state.
7: other.
8: Adoption confidence.
9: Personal guarantee.
10: No guarantee.
11: Mortgages trunk.

Attribute 17: (categorical)
0perational Type

0: Non Installment.
1: Installment.

Attribute 18: (categorical)
Status

0: Defaulter.
1: Non-defaulter.

5.5.2 Description of the Sudanese Credit Dataset 2

Title: Sudanese Credit Dataset2 (SCD2).

Source Information: Al Salam Commercial Bank.

Classes of dataset: Instances of this dataset was classified into two classes, Defaulter and Non-defaulter.

Number of Instances: 950 instances. 745 are classified as non-defaulters and 205 as defaulters.

Versions of Dataset:

Two versions of this dataset are provided. The first version (SCD2version1) contains numerical and categorical attributes (original dataset). The second version (SCD2 version2) of this dataset contains numerical attributes (all categorical attributes in SCD1version1 were transformed to numeric ones).

Attributes of SCD2:

Number of Attributes in SCDS2: 17 attributes in addition to class attribute Status).

Types Attributes in SCD2version1: (6 numerical, 12 categorical)

Attribute description for SCD2:

For all categorical attributes the values is transformed to integer values in version 2.

Attribute 1: (categorical)

Gender:

1: Male
2: Female

Attribute 2: (numerical)
Age

Attribute 3: (categorical)
Marital status :

1:Single

2:Married

3:Divorced

Attribute 4: (numerical)
Number of children

Attribute 5: (numerical)
Number of Spouses

Attribute 6: (categorical)
Occupation

1: Employee

2: Engineer

3: Pharmacist

4: Worker

5: Policeman

6: Physician

7: Merchant

8: Lawyer

9: Free business

10: Farmer

11: Jurist

12: Pensioner

13: Other

Attribute 7: (categorical)
Phone:

1: Holder

2: Unholder

Attribute 8: (categorical)
ID Type

1: Personal Card
2: Driving License
3: Passport
4: Military Card

Attribute 9: (numerical)
Approved Amount

Attribute 10: (numerical)
Profit Margin

Attribute 11: (numerical)
Periodical Instalment Amount

Attribute 12: (categorical)
Finance Duration

1: Short term
2: Medium term
3: Long term

Attribute 13: (categorical)
Periodicity of Payment

1: Monthly Installments 30 Days

2: Six MonthInstallments180Days

Attribute 14: (categorical)
Purpose of credit

1: Commercial
2: Construction
3: Service
4: Others

Attribute 15: (categorical)
Sector:

1: Industrial_ Building
2: Services_Others
3: Services_Transportation

Attribute 16: (categorical)
Guarantee type

1: real state
2: car
3: Deferred cheaques
4: Deposit
5: Mechanics mortgage

Attribute 17: (categorical)
Finance size

1: Micro
2: Normal

Attribute 18: (categorical)
Status

0: Defaulter.
1: Non-defaulter.

5.5.3 Description of the German credit dataset[14]

Title: German Credit data

Source Information

Professor Dr. Hans Hofmann

Institutf"urStatistik und "Okonometrie

Universit"at Hamburg

Classes of dataset: Instances of this dataset was classified into two classes, Good(1) and Bad(2).

Number of Instances: 1000, 700 instances are classified as good borrower and 300 instances as bad borrower.

Versions of Dataset:

Two versions of this dataset datasets are provided. The first version contains numerical and categorical attributes (original dataset). The second version (numeric version) of this dataset contains numerical attributes.in this version Several attributes that are ordered categorical (such as attribute 17)

have been coded as integer.in the numeric version several indicator variables added. There the number of attributes in numeric version is greater the number of attributes in original version.

Number of Attributes (original version): 20 (7 numerical, 13 categorical) in addition to the label (Status) attribute.

Number of Attributes (numeric version): 24 (24 numerical) in addition to the label attribute.

Attribute description for German credit dataset (original version):

Attribute 1: (qualitative)
Status of existing checking account
A11 : ... < 0 DM
A12 : 0 <= ... < 200 DM
A13 : ... >= 200 DM /
salary assignments for at least 1 year
A14 : no checking account

Attribute 2: (numerical)
Duration in month

Attribute 3: (qualitative)
Credit history
A30 : no credits taken/
all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now
A33 : delay in paying off in the past
A34 : critical account/
other credits existing (not at this bank)

Attribute 4: (qualitative)
Purpose
A40 : car (new)
A41 : car (used)
A42 : furniture/equipment
A43 : radio/television

A44 : domestic appliances
A45 : repairs
A46 : education
A47 : (vacation - does not exist?)
A48 : retraining
A49 : business
A410 : others

Attribute 5: (numerical)
Credit amount

Attribute6: (qualitative)
Savings account/bonds
A61 : ... < 100 DM
A62 : 100 <= ... < 500 DM
A63 : 500 <= ... < 1000 DM
A64 : .. >= 1000 DM
A65 : unknown/ no savings account

Attribute 7: (qualitative)
Present employment since
A71 : unemployed
A72 : ... < 1 year
A73 : 1 <= ... < 4 years
A74 : 4 <= ... < 7 years
A75 : .. >= 7 years

Attribute 8: (numerical)
Installment rate in percentage of disposable income

Attribute 9: (qualitative)
Personal status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Attribute 10: (qualitative)
Other debtors / guarantors
A101 : none
A102 : co-applicant
A103 : guarantor

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)
Property

A121 : real estate
A122 : if not A121 : building society savings agreement/
life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no property

Attribute 13: (numerical)
Age in years

Attribute 14: (qualitative)
Other installment plans
A141 : bank
A142 : stores
A143 : none

Attribute 15: (qualitative)
Housing
A151 : rent
A152 : own
A153 : for free

Attribute 16: (numerical)
Number of existing credits at this bank

Attribute 17: (qualitative)
Job
A171 : unemployed/ unskilled - non-resident
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/
highly qualified employee/ officer

Attribute 18: (numerical)
Number of people being liable to provide maintenance for

Attribute 19: (qualitative)
Telephone
A191 : none
A192 : yes, registered under the customers name

Attribute 20: (qualitative)
foreign worker
A201 : yes
A202 : no

Attribute 21:(qualitative)
Status
1: Good
2: Bad

5.6 Credit Scoring Models Construction

In this section experiments for constructing the proposed CSMs are presented. Experiments for this research were conducted into three stages (stage1, stage2, and stage3) for each dataset. The experiments described in the following sections were performed on a PC with a 1.80GHz intel® core™ i7-2677 CPU and 6.0 GB RAM using Windows 7 64bit operating system.

5.6.1 Software Package

The main software package used in our experiments is RapidMiner 5.3.007. It is open-source Java-based DM software. It is free software. It can be downloaded and installed from RapidMiner home page http://rapid-i.com[2].

5.6.2 Datasets

Experiments for all stages were applied to two Sudanese credit datasets (SCD1, SCD2) and German credit dataset.

5.6.3 Validation Methods

Holdout(split) and K-fold cross-validation methods which are called split validation operator and X-validation respectively in RapidMiner were employed in experiments. **In RapidMiner** split validation operator was used to perform a simple validation i.e. it randomly splits up the dataset into a training set and a testing set and evaluates the model. In these experiments two different split ratios of Training: Testing (70:30 and 60:40) were applied. In RapidMiner X-Validation operator performs a cross-validation

in order to estimate the statistical performance of a learning operator. This operator partitions the input dataset into *k* subsets of equal size. From the*k* subsets, a single subset is retained as the testing dataset (i.e. input of the testing), and the remaining $k - 1$ subsets are used as training dataset . The cross-validation process is then repeated *k* times, with each of the *k* subsets used exactly once as the testing data. The *k* results from the *k* iterations can then be averaged (or otherwise combined) to produce a single estimation. The value *k* was chosen to be 10 in these experiments [2].

5.6.4 Sampling type

RapidMiner provides several types of sampling for building the training and testing subsets. The available options of sampling types are as follows[2]:

- **Linear sampling**: Linear sampling simply divides the dataset into partitions (training and testing according to specified split ratio) without changing the order of the instances (tuples) i.e. subsets with consecutive instances are created.
- **Shuffled sampling**: Shuffled sampling builds random subsets of a dataset. Instances are chosen randomly for making subsets.
- **Stratified sampling**: Stratified sampling builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset. For example, in the case of a binominal classification, stratified sampling builds random subsets such that each subset contains roughly the same proportions of the two values of class labels.

  In all experiments these types were tested. The sample type which achieved the best accuracy for the given model was chosen.

5.6.5 Data Mining Classification Techniques

Three DM classification techniques were applied in the experiments of this research namely, ANN, SVM, and DT. In addition to these techniques GA was as also employed in stage2 of these experiments as a feature selection technique.

5.6.5.1 Artificial Neural Network Parameters

By applying Neural Net operator in RapidMiner the model learns by means of a feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron). In all experiments in this research networks with one and two hidden layers, learning rate (This parameter determines how much we change the weights at each step) of 0.2 and 0.3 were tested (see Table 5.7). All other default parameters in RapidMiner were used [2].

Twelve different ANN models for each dataset and validation method were tested. The model with best accuracy was picked.

Table 5.7 ANN Models' Options

ANN model	Sampling type	# hidden layers	Learning rate
Model1	Stratified	One	0.3
Model2	Stratified	One	0.2
Model3	Shuffled	One	0.3
Model4	Shuffled	One	0.2
Model5	linear	One	0.3
Model6	Linear	One	0.2
Model7	Stratified	Two	0.3
Model8	Stratified	Two	0.2
Model9	Shuffled	Two	0.3
Model10	Shuffled	Two	0.2
Model11	linear	Two	0.3
Model12	Linear	Two	0.2

5.6.5.2 Support Vector Machine Parameters

SVM operator in RapidMiner uses the Java implementation of the support vector machine *mySVM*. This learning method can be used for both regression and classification and provides a fast algorithm and good results for many learning tasks.

In all SVM experiments in this research four Kernel types (Dot, Radial, Polynomial, Anova) were tested with different validation and sampling types. Therefore, twelve SVM models were tested for each dataset and validation method. The model with best accuracy was picked. See Table 5.8.All other default parameters in RapidMiner [2] were used.

Table 5.8 SVM Models' Options

SVM model	Sampling type	Kernel type
Model1	Stratified	Dot
Model2	Shuffled	Dot
Model3	Linear	Dot
Model4	Stratified	Radial
Model5	Shuffled	Radial
Model6	Linear	Radial
Model7	Stratified	Polynomial
Model8	Shuffled	Polynomial
Model9	Linear	Polynomial
Model10	Stratified	Anova
Model11	Shuffled	Anova
Model12	Linear	Anova

5.6.5.3 Decision Tree Parameters

A decision tree is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of a target attribute (often called *class* or *label*) based on several input attributes of the dataset.

In all experiments all split criteria (Information gain, Gain ratio, Gini index, and Accuracy) were tested with different validations and sampling types. For the other parameters default values in RapidMiner wereused [2]. Therefore, twelve DT models were tested for each dataset and validation method. The model with best accuracy was picked. See Table 5.9.

Table 5.9 DT Models' Options

DT MODELS	Sampling type	Kernel type
Model1	Stratified	Information gain
Model2	Shuffled	Information gain
Model3	Linear	Information gain
Model4	Stratified	Gain ratio
Model5	Shuffled	Gain ratio
Model6	Linear	Gain ratio
Model7	Stratified	Gini index
Model8	Shuffled	Gini index
Model9	Linear	Gini index
Model10	Stratified	Accuracy
Model11	Shuffled	Accuracy
Model12	Linear	Accuracy

5.6.6Feature Selection Techniques

Feature selection is one of the important and frequently used techniques in data preprocessing for DM [74]. Feature selection is a process of identifying and selecting a useful subset from original features. It reduces the number of features by removing redundant and irrelevant attributes [72].

In stage 2 experiments of this research, a wrapper based approach that integrates GA and the aforementioned classification techniques was adopted in order to enhance the performances for these classification techniques.

In RapidMiner Optimize Selection (Evolutionary) is an operator which selects the most relevant attributes of the given dataset by using GA. All default parameters for GA in RapidMiner were used [2].

5.6.7 Experiments Stages

There are three stages for each dataset. In stage1, ANN, SVM, and DT were applied to develop CSMS. Hybrid scoring models were constructed in stage 2 by combining classification techniques in stage1 with GA. In stage3 all techniques of stage1 were applied to intersected reduced datasets.

5.6.7.1 Stage1 Experiments

This section presents all experiments for stage1 to all datasets. For each experiment two tables are presented. The first table illustrates the details (dataset, technique, sampling type,…. etc.) and accuracy for the model which achieved the highest accuracy among other options (for each technique there are many options that are tested as we mentioned above,

table). The second table presents the confusion matrix for this model. The template for the confusion matrix is shown in Table 5.10.

Table 5.10 Confusion Matrix Template

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	TNeg	FNeg	%
Pred. Defaulter	FPos	Tpos	%

Defaulter: Positive class (Pos), **Non-defaulter**: Negative class (Neg).

pos: The number of positive ("defaulter") samples.

Tpos: The number of true positives ("defaulter" customers that were correctly classified as such).

Fpos: The number of false positives ("non-defaulters" were incorrectly labeled as ("defaulters")

neg: The number of negative ("non-defaulter") samples,

Tneg: The number of true negatives ("non-defaulter " customers that were correctly classified as such) .

Fneg: The number of false negatives ("defaulter" customers that were incorrectly labeled as "non-defaulter").

Precision (*pos*) = *Tpos/ (Tpos+ Fpos);*

Precision (*neg*) = *Tneg/ (Tneg+ Fneg);*

Accuracy = (Tpos+Tneg)/ (pos + neg).

For the German dataset Good and Bad were used instead of Non-defaulter and Defaulter respectively.

A. Sudanese Credit Dataset1stag1 experiments

A.1 ANN Experiments

Three experiments were conducted using ANN with two different split validation ratios and 10-fold cross validation.

A.1.1 ANN Experiment1

In this experiment split validation ratio 70:30 was used. Table 5.11 illustrates properties of the model and the accuracy achieved .Table 5.12 presents the confusion matrix for the model.

Table 5.11 SCD1 ANN Experiment1

Dataset	**SCD1**
Model name	ANN
Validation	Split 70:30
# Hidden layer	1
# neurons	3
Sampling type	Stratified
Learning rate	0.2
Accuracy	66.67%

Table 5.12 SCD1 ANN Experimet1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	178	92	65.93%
Pred. Defaulter	38	82	68.33%

A.1.2 ANN Experiment 2

In this experiment split validation ratio 60:40 was used. Table 5.13 illustrates the properties of this model and the accuracy achieved. Table 5.14 presents the confusion matrix for the model.

Table 5.13 SCD1 ANN Experiment2

Dataset	SCD1
Model name	ANN
Validation	Split 60:40
# Hidden layer	1
# neurons	3
Learning rate	0.3
Sampling type	Stratified
Accuracy	64.62%

Table 5.14 SCD1 ANN Experimet2 Confusion Matrix

	True Non defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	245	141	63.47%
Pred. Defaulter	43	91	67.91%

A.1.3 ANN Experiment 3

In this experiment 10 fold cross validation was used. Table 5.15 illustrates properties of this model and the accuracy achieved. Table 5.16 presents the confusion matrix for the model.

Table 5.15 SCD1 ANN Experiment1

Dataset	SCD1
Model name	ANN
Validation	10- cross-validation
# Hidden layer	2
# neurons	3
Sampling type	shuffled
Learning rate	0.3
Accuracy	63.62% +/- 3.15% (mikro: 63.62%)

Table 5.16 SCD1 ANN Experimet3 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
pred. Non-defaulter	545	298	64.65%
pred. Defaulter	175	282	61.71%

A.2 SVM Experiments

Like ANN experiments, three experiments were conducted using SVM with two different split validation ratios and 10-fold cross validation.

A.2.1 SVM Experiment 1

Table 5.17 and 5.18 present the properties and the confusion matrix for the model respectively. SVM model with stratified sampling technique and kernel type anova yielded the highest accuracy.

Table 5.17 SCD1 SVM Experiment1

Dataset	SCD1
Model name	SVM
Validation	Split70:30
Sampling type	Stratified
Kernel type	Anova
Accuracy	66.41%

Table 5.18 SCD1 SVM Experiment1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	189	104	64.51%
Pred. Defaulter	27	70	72.16%

A.2.2 SVM Experiment 2

Table 5.19 and Table 5.20 illustrate the properties and confusion matrix for the model respectively. In this experiment anova kernel type with stratified sampling achieved the highest accuracy than other tested options.

Table 5.19: SCDI SVM Experiment 2

Dataset	SCD1
Model name	SVM
Validation	Split 60:40
Sampling type	Stratified
Kernel type	anova
Accuracy	66.73%

Table 5.20 SCD1 SVM Experiment2 Confusion Matrix

	True Non defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	242	127	65.58%
Pred. Defaulter	46	105	69.54%

A.2.3 SVM Experiment 3

Table 5.21 and Table 5.22 illustrate the properties and confusion matrix for the model respectively. Anova with stratified sampling produced the highest accuracy SVM model with than other options.

Table 5.21 SCD1 SVM Experiment 3

Dataset	SCD1
Model name	SVM
Validation	10-cross validation
Sampling type	Stratified
Kernel type	anova
Accuracy	65.00% +/-3.69% (Mikro: 65.00%)

Table 5.22 SCD1 SVM Experiment3 Confusion Matrix

	True Non defaulter	True Defaulter	Class precision
Pred. Non-defaulter	589	324	64.51%
Pred. Defaulter	131	256	66.15%

A.3.3 DT Experiments

Three experiments were conducted using DT with two different split validation ratios and 10-fold cross validation.

A.3.1 DT Experiment 1

In this experiment split criteria Gini index with linear sampling resulted in the highest accuracy DT model. Table 5.23 and Table 5.24 present the properties and confusion matrix for the model respectively.

Table 5.23 SCD1 DT Experiment1

Dataset	SCD1
Model	DT
Validation	Split 70:30
Criterion	Gini index
Sampling type	Linear
Accuracy	60.00%

Table 5.24 SCD1 DT Experiment1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	214	139	60.62%
Pred. defaulter	17	20	54.05%

A.3.2 DT Experiment 2

In this experiment DT model with Gini index and linear sampling achieved the highest accuracy. Table 5.25 and Table 5.26 for present the properties and confusion matrix for the model respectively.

Table 5.25 SCD1 DT Experiment 2

Dataset	SCD1
Model name	DT
Validation	Split 60:40
Criterion	Gini index
Sampling type	Linear
Accuracy	58.85%

Table 5.26 SCD1 DT Experiment2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	271	180	60.09%
Pred. Defaulter	34	35	50.72%

A.3.3 DT Experiment 3

In this experiment Gini index with linear sampling produced the highest accuracy DT model. Table 5.27 and 5.28 present the properties and confusion matrix for the model respectively.

Table 5.27 SCD1 DT Experiment 3

Dataset	SCD1
Model name	DT
Validation	10-cross validation
Criterion	Gini index
Sampling type	linear
Accuracy	58.08% +/- 5.00% (mikro: 58.08%)

Table 5.28 SCD1 DT Experiment 3 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	496	321	60.71%
Pred. Defaulter	224	259	53.62%

B. Stage1 Experiments for Sudanese Credit Dataset2

B.1 ANN Experiments

Three experiments were conducted using ANN with two different split validation ratios and 10-fold cross validation.

B.1.1 ANN Experiment1

In this experiment split validation ratio 70:30 was used. Table 5.29 illustrates details of the model which achieved highest accuracy. Table 5.30 presents the confusion matrix for the model.

Table 5.29 SCD2 DT Experiment 1

Dataset	SCD2
Model name	ANN
Validation	Split 70:30
Sampling type	Stratified
# Hidden layer	2
# neurons	3
Learning rate	.3
Accuracy	79.65%

Table 5.30 SCD2 DT Experiment 1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
pred. Non-Defaulter	216	51	80.90%
pred. Defaulter	7	11	61.11%

B.1.2 ANN Experiment2

In this experiment split validation ratio 60:40 was used. Table 5.31 illustrates properties of the model which achieved the highest accuracy. Table 5.32 presents the confusion matrix for this model.

Table 5.31 SCD2 ANN Experiment2

Dataset	SCD2
Model name	ANN
Validation	Split 60:40
Sampling type	Stratified
# Hidden layer	2
# neurons	3
Learning rate	.2
Accuracy	78.68%

Table 5.32 SCD2 DT Experiment 2 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	283	67	80.86%
Pred. Defaulter	14	16	53.33%

B.1.3 ANN Experiment3

In this experiment 10 fold cross validation was used. Table 5.33 illustrates properties of the model and the accuracy achieved. Table 5.34 presents the confusion matrix for the model.

Table 5.33 SCD2 ANN Experiment3

Dataset	SCD2
Model name	ANN
Validation	10 Cross validation
Sampling type	Stratified
# Hidden layer	2
# neurons	3
Learning rate	.3
Accuracy	78.00%

Table 5.34SCD2 ANN Experiment3 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	709	175	80.20%
Pred. Defaulter	34	32	48.48%

B.2 SVM Experiments

Three experiments were conducted using SVM with two different split validation ratios and 10-fold cross validation.

B.2.1 SVM Experiment 1

In this experiment two kernel types Polynomial and Dot using stratified sampling resulted in the highest accuracy SVM models. Table 5.35 and Table 5.36 illustrate the properties and confusion matrix of the model respectively .

Table 5.35 SCD2 SVM Experiment1

Dataset	SCD2
Model name	SVM
Validation	Split70:30
Sampling type	Stratified
Kernel type	Polynomial and dot
Accuracy	79.65%

Table 5.36 SCD2 SVM Experiment1 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	222	57	79.57%
Pred. Defaulter	1	5	83.33%

B.2.2 SVM Experiment 2

In this experiment two kernel types Polynomial and Dot using stratified sampling resulted in a highest accuracy SVM model. Table 5.37 and Table 5.38 illustrate the properties and confusion matrix of the model respectively.

Table5.37 SCD2 SVM Experiment2

Dataset	SCD2
Model name	SVM
Validation	Split60:40
Sampling type	Stratified
Kernel type	Polynomial and dot
Accuracy	79.74%

Table 5.38 SCD2 SVM Experiment2 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	294	74	79.89%
Pred. Defaulter	3	9	75.00%

A.2.3 SVM Experiment 3

In this experiment kernel type Dot and shuffled sampling resulted in the highest accuracy SVM model. Table 5.39 and Table 5.40 present the properties and confusion matrix for the model respectively.

Table 5.39 SCD2 SVM Experiment3

Dataset	SCD2
Model name	SVM
Validation	10-cross-validation
Sampling type	Shuffled
Kernel type	Dot
Accuracy	79.26% +/- 4.16% (mikro: 79.26%)

Table 5.40 SCD2 SVM Experiment3 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	730	184	79.87%
Pred. Defaulter	13	23	63.89%

B.3 DT Experiments

Three experiments were conducted using DT with two different split validation ratios and 10-fold cross validation.

B.3.1 DT Experiment 1

In this experiment split criteria accuracy with shuffled sampling resulted in highest accuracy DT model. Table 5.41 and Table 5.42 present the properties and confusion matrix for the model respectively.

Table 5.41 SCD2 DT Experiment 1

Dataset	SCD2
Model name	DT
Validation	Split 70:30
Criterion	Accuracy
Sampling type	Shuffled
Accuracy	81.40%

Table 5.42 SCD2 DT Experiment1 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	225	49	82.12%
Pred. Defaulter	4	7	63.64%

B.3.3 DT Experiment 2

In this experiment split criteria accuracy with shuffled sampling also resulted in highest accuracy DT model. Table 5.43 and Table 5.44 present the properties and confusion matrix for the model respectively.

Table 5.43 SCD2 DT Experiment 2

Dataset	SCD2
Model name	DT
Validation	Split 60:40
Criterion	accuracy
Sampling type	Shuffled
Accuracy	80%

Table 5.44 SCD2 DT Experiment2 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	293	67	81.39%
Pred. Defaulter	9	11	55.00%

B.3.3 DT Experiment 3

In this experiment split criteria Gain ratio with shuffled sampling resulted in highest accuracy DT model. Table 5.45 and Table 5.46 present the properties and confusion matrix for the model respectively.

Table 5.45 SCD2 DT Experiment3

Dataset	SCD2
Model name	DT
Validation	10-cross validation
Criterion	Gain ratio
Sampling type:	Shuffled
Accuracy	78.42% +/- 3.47% (mikro: 78.42%)

Table 5.46 SCD2 DT Experiment3 Confusion Matrix

	True Non-Defaulter	True Defaulter	Class Precision
Pred. Non-Defaulter	745	205	78.42%
Pred. Defaulter	0	0	0.00%

C. Stage1 Experiments for German Dataset

The same stage 1 experiments for the two Sudanese credit datasets are also applied to the German dataset.

C.1 ANN Experiments

Three experiments were conducted using ANN with two different split validation ratios and 10-fold cross validation.

C.1.1 ANN Experiment1

ANN model with two hidden layers and three neurons for each layers, stratified sampling, and learning rate .3 achieved the highest accuracy. Table 5.47 and Table 5.48 present the properties and confusion matrix for model respectively.

Table 5.47 German ANN Experiment 1

Dataset	German
Model name	ANN
Validation	Split 70:30
# Hidden layer	2
# neurons	3
Learning rate	0.3
Sampling type	Stratified
Accuracy	75.67%

Table 5.48 German ANN Experiment 1Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	177	40	81.57%
Pred. Bad	33	50	60.24%

C.1.2 ANN Experiment2

In this experiment ANN model with two hidden layers and three neurons for each layers, stratified sampling, and learning rate .3 also achieved the highest accuracy. Table 5.49 and Table 5.50 present the details and confusion matrix respectively.

Table 5.49 German ANN Experiment 2

Dataset	German
Model name	ANN
Validation	Split 60:40
# Hidden layer	2
# neurons	3
Learning rate	0.3
Sampling type	Stratified
Accuracy	74.00%

Table 5.50 German ANN Experiment 2 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	241	65	78.76%
Pred. Bad	39	55	58.51%

C.1.3 ANN Experiment3

In this experiment ANN model with two hidden layers and three neurons for each layers, stratified sampling, and learning rate .2 achieved the highest accuracy. Table 5.51 and Table 5.52 present the properties and confusion matrix for the model respectively.

Table 5.51 German ANN Experiment 3

Dataset	German
Model name	ANN
Validation	10-cross-validation
# Hidden layer	2
# neurons	3
Sampling type	Stratified
Learning rate	0.2
Accuracy	74.60% +/- 4.52% (mikro: 74.60%)

Table 5.52 German ANN Experiment 3 Confusion Matrix

	true Good	true Bad	class precision
pred. Good	583	137	80.97%
pred. Bad	117	163	58.21%

C.2 SVM Experiments

Three experiments were conducted using two different split validation ratios and 10-fold cross validation.

C.2.1 SVM Experiment1

SVM polynomial kernel and stratified sampling achieved the highest accuracy. Table 5.53 and Table 5.54 present the properties and confusion matrix for the model respectively.

Table 5.53 German SVM Experiment 1

Dataset	German
Model name	SVM
Validation	Split 70:30
Sampling type	Stratified
Kernel type	Polynomial
Accuracy	75.33%

Table 5.54 German SVM Experiment 1Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	202	67	75.10%
Pred. Bad	8	23	74.79%

C.2.2 SVM Experiment 2

SVM polynomial kernel and stratified sampling achieved the highest accuracy. Table 5.55 and Table 5.56 present the properties and confusion matrix for the model respectively.

Table 5.55 German ANN Experiment 2

Dataset	German
Model name	SVM
Validation	Split60:40
Sampling type	Stratified
Kernel type	Polynomial
Accuracy	74.75%

Table 5.56 German SVM Experiment 2 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	264	85	75.64%
Pred. Bad	16	35	68.63%

C.2.3 SVM Experiment3: polynomial kernel and shuffled sampling achieved the highest accuracy SVM model. Table 5.57 and Table 5.58 present the properties and confusion matrix for the model respectively.

Table 5.57 German SVM Experiment 3

Dataset	German
Model name	SVM
Validation	10-cross validation
Sampling type	Shuffled
Kernel type	Polynomial
Accuracy	75.70% +/- 4.50% (mikro: 75.70%)

Table 5.58 German SVM Experiment 3 Confusion Matrix

	True Good	True Bad	Class Precision
pred. Good	665	208	76.17%
pred. Bad	35	92	72.44%

C.3 DT Experiments

Three experiments were conducted using DT with two different split validation ratios and 10-fold cross validation.

C.3.1 DT Experiment1

Gini index split criteria and stratified sampling achieved the highest accuracy DT model. Table 5.59 and Table 5.60 present the properties and confusion matrix for the model respectively.

Table 5.59 German DT Experiment 1

Dataset	German
Model name	DT
Validation	Split 70:30
Criterion	Gini index
Sampling type	stratified
Accuracy	73.00%

Table 5.60 German DT Experiment 1 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	189	60	75.90%
Pred. Bad	21	30	58.82%

C.3.2 DT Experiment2

DT with Gini index split criteria and shuffled sampling achieved the highest accuracy DT model. Table 5.61 and Table 5.62 present the properties and confusion matrix for the model respectively.

Table 5.61 German DT Experiment2

Dataset	German
Model name	DT
Validation	Split 60:40
Criterion	Gini index
Sampling type	Shuffled
Accuracy	68.50%

Table 5.62 German DT Experiment 2 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	233	93	71.47%
Pred. Bad	33	41	55.41%

A.3.3 DT Experiment3

DT with Gini index split criteria and stratified sampling achieved the highest accuracy. Table 5.63 and Table 5.64 present the details and confusion matrix respectively.

Table 5.63 German DT Experiment 3

Dataset	German
Model name	DT
Validation	10-cross-validation
Criterion	Gini index
Sampling type	Stratified
Accuracy	70.70% +/- 2.61% (mikro: 70.70%)

Table 5.64 German DT Experiment3 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	598	191	75.79%
Pred. Bad	102	109	51.66%

5.6.7.2 Stage 2 Experiments

Hybrid CSMs were developed in this stage by combining classification techniques in stage1 with GA . These hybrid models are abbreviated as GAANN, GASVM, and GADT.

This section presents all experiments for this stage to all datasets. An additional table is presented for each experiment. It contains weights of the attributes which were selected by GA.

A. Sudanese Credit Dataset1 Stag2 Experiments

A.1 GAANN Experiments

Three experiments were conducted using GA and ANN with two different split validation ratios and 10-fold cross validation.

A.1.1 GAANN Experiment1

GAANN with stratified sampling, one hidden layer, and learning rate 0.2 achieved the highest accuracy. Tables 5.65, 5.66, present the properties and confusion matrix for the model respectively. Table 5.67 presents the weights for each attribute. Ten attributes have weight one.

Table 5.65 SCD1 GAANN Experiment1

Dataset	SCD1
Model name	GAANN
Validation	Split70:30
# Hidden layer	1
# neurons	3
Learning rate	0.2
Sampling type	Stratified
Accuracy	71.03%

Table 5.66 SCD1 GAANN Experiment1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	185	82	69.29%
Pred. Defaulter	31	92	74.80%

Table 5.67 SCD1 GAANN Experiment1 Attributes Weights

Attribute Name	**Weight**
Have phone	1.0
IDType	0.0
Gender	0.0
Age	0.0
Occupation	1.0
Material Status	1.0
Number of Dependents	0.0
Number of Spouses	1.0
Monthly Salary Value	0.0
Monthly Expenditures Value	1.0
Finance Size	1.0
Finance Duration	1.0
Payment Method	1.0
Finance Form	0.0
Loan Type	0.0
Insurance Description	1.0
0perational Type	1.0

A.1.2 GAANN Experiment2

GAANN model with stratified sampling, one hidden layer and learning rate 0.2 achieved the highest accuracy. Tables 5.68, 5.69 present the properties and confusion matrix respectively. Table 5.70 presents weights for each attribute. Six attributes have weight one.

Table 5.68SCD1 GAANN Experiment2

Dataset	SCD1
Model name	GAANN
Validation	Split 60:40
# Hidden layer	1
# neurons	3
Sampling type	Stratified
Learning rate	0.2
Accuracy	69.62%

Table 5.69 SCD1 GAANN Experiment2

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	239	109	68.68%
Pred. Defaulter	49	123	71.51%

Table 5.70 SCD1 GAANN Experiment2 Attributes Weights

Attribute name	Weight
Have phone	0.0
IDType	0.0
Gender	1.0
Age	1.0
Occupation	1.0
Material Status	0.0
Number of Dependents	0.0
Number of Spouses	0.0
Monthly Salary Value	0.0
Monthly Expenditures Value	1.0
Finance Size	0.0
Finance Duration	0.0
Payment Method	1.0
Finance Form	0.0
Loan Type	0.0
Insurance Description	0.0
0perational Type	1.0

A.1.3 GAANN Experiment3

GAANN model with stratified sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.71 and 5.72 present the properties and confusion matrix for the model respectively. Table 5.73 presents the weights for each attribute. Seven attributes have weight one.

Table 5.71 SCD1 GAANN Experiment3

Dataset	SCD1
Model name	GAANN
Validation	10-cross-validation
# Hidden layer	2
# neurons	3
Learning rate	0.2
Sampling type	Stratified
Accuracy	67.31% +/- 2.90% (mikro: 67.31%)

Table 5.72 SCD1 GAANN Experiment 3 Confusion Matrix

	True Non defaulter	True Defaulter	Class Precision
Pred. Non defaulter	584	289	66.90%
Pred. Defaulter	136	291	68.15%

Table 5.73 SCD1 GAANN Experiment3 Attributes Weights

Attribute name	**Weight**
Have phone	0.0
IDType	1.0
Gender	0.0
Age	0.0
Occupation	0.0
Material Status	1.0
Number of Dependents	1.0
Number of Spouses	0.0
Monthly Salary Value	0.0
Monthly Expenditures Value	0.0
Finance Size	1.0
Finance Duration	0.0
Payment Method	1.0
Finance Form	0.0
Loan Type	1.0
Insurance Description	0.0
0perational Type	1.0

A.2 GASVM Experiments

Three experiments to develop GSVM models with two different validation ratios and 10 cross validation.

A.2.1 GASVM Experiment1

GASVM model using shuffled sampling and Radial kernel achieved the highest accuracy. Tables 5.74 and 5.75 present the properties and confusion matrix respectively. Table 5.76 presents the weights for each attribute. Five attributes have weight one.

Table 5.74 SCD1 GASVM Experiment1

Dataset	SCD1
Model name	GASVM
Validation	Split 70:30
Sampling type	Stratified
Kernel type	Radial
Accuracy	69.49%

Table 5.75 SCD1 GASVM Experiment1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	185	88	67.77%
Pred. Defaulter	31	86	73.50%

Table 5.76 SCD1 GASVM Experiment1 Attributes Weights

Attribute name	Weight
Have phone	0.0
IDType	0.0
Gender	0.0
Age	1.0
Occupation	0.0
Material Status	0.0
Number of Dependents	0.0
Number of Spouses	0.0
Monthly Salary Value	0.0
Monthly Expenditures Value	0.0
Finance Size	1.0
Finance Duration	0.0
Payment Method	1.0
Finance Form	0.0
Loan Type	1.0
Insurance Description	0.0
0perational Type	1.0

A.2.2 GASVM Experiment2

GASVM model using shuffled sampling and Anova kernel achieved the highest accuracy. Tables 5.77 and 5.78 present the properties and confusion matrix respectively. Table 5.79 present the weights for each attribute. Nine attributes have weight one.

Table 5.77 SCD1 GASVM Experiment2

Dataset	SCD1
Model name	GASVM
Validation	Split 60:40
Sampling type	Shuffled
Kernel type	anova
Accuracy	69.81%

Table 5.78 SCD1 GASVM Experiment2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
pred. Non-defaulter	251	109	69.72%
pred. Defaulter	48	112	70.00%

Table 5.79 SCD1 GASVM Experiment2 Attributes Weights

Attribute name	**weight**
Have phone	0.0
IDType	0.0
Gender	0.0
Age	1.0
Occupation	1.0
Material Status	1.0
Number of Dependents	0.0
Number of Spouses	0.0
Monthly Salary Value	0.0
Monthly Expenditures Value	0.0
Finance Size	1.0
Finance Duration	0.0
Payment Method	1.0
Finance Form	1.0
Loan Type	1.0
Insurance Description	1.0
0perational Type	1.0

A.2.3 GASVM Experiment3

GASVM model using shuffled sampling and Anova kernel achieved the highest accuracy. Table 5.80, 5.81 present the properties and confusion matrix respectively. Table 5.82 presents weights for each attribute. Eleven attributes have weight one.

Table 5.80 SCD1 GASVM Experiment 3

Dataset	SCDS1
Model name	GASVM
Validation	10-cross-validation
Sampling type	Shuffled
Kernel type	Anova
Accuracy	67.46% +/- 4.85% (mikro: 67.46%)

Table 5.81 SCD1 GASVM Experiment2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	595	298	66.63%
Pred. Defaulter	125	282	69.29%

Table 5.82 SCD1 GASVM Experiment3 attributes weights

Attribute name	**Weight**
Have phone	0.0
IDType	0.0
Gender	0.0
Age	1.0
Occupation	1.0
Material Status	1.0
Number of Dependents	1.0
Number of Spouses	1.0
Monthly Salary Value	0.0
Monthly Expenditures Value	0.0
Finance Size	1.0
Finance Duration	1.0
Payment Method	1.0
Finance Form	0.0
Loan Type	1.0
Insurance Description	1.0
0perational Type	1.0

GADT Experiments

Three experiments to develop GADT models with two different validation ratios and 10 cross validation.

A.3.1 GADT Experiment1

GADT model using stratified sampling and Gini index split criteria achieved the highest accuracy. Tables 5.83 and 5.84 present the properties and confusion matrix respectively. Table 5.85 presents weights for each attribute. Eight attributes have weight one.

Table 5.83 SCD1 GADT Experiment1

Dataset	SCD1
Model name	GADT
Validation	Split 70:30
Criterion	Gini index
Sampling type	Stratified
Accuracy	71.03%

Table 5.84 SCD1 GADT Experiment 1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	151	48	75.88%
Pred. Defaulter	65	126	65.97%

Table 5.85 SCD1 GADT Experiment1 attributes weights

Attribute name	Weight
Phone	1.0
IDType	0.0
Gender	1.0
Age	0.0
Occupation	1.0
Marital Status	0.0
Number Of Children	0.0
Number Of Spouses	1.0
Monthly Salary Value	0.0
Monthly Expenditures Value	0.0
Finance Size	0.0
Finance Duration	1.0
Payment Method	1.0
Finance Form	0.0
Loan Type	1.0
Insurance Description	0.0
0perational Type	1.0

A.3.2 GADT Experiment2

GADT model using stratified sampling and Gini index split criteria achieved the highest accuracy. Tables 5.86 and 5.87 present the properties and confusion matrix for the model respectively. Table 5.88 presents weights for each attribute. Six attributes have weight one.

Table 5.86 SCD1 GADT Experiment 2

Dataset	SCD1
Model name	GADT
Validation	Split 60:40
Criterion	Gini index
Sampling type	Stratified
Accuracy	68.65%

Table 5.87 SCD1 GADT Experiment 2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	233	108	68.33%
Pred. Defaulter	55	124	69.27%

Table 5.88 SCD1 GADT Experiment2 Attributes Weights

Attribute name	**Weight**
Phone	1.0
IDType	1.0
Gender	0.0
age	0.0
Occupation	1.0
MaritalStatus	0.0
Number ofChildren	0.0
Number ofSpouses	0.0
MonthlySalaryValue	0.0
MonthlyExpendituresValue	0.0
Finance Size	1.0
Finance Duration	0.0
Payment Method	1.0
Finance Form	0.0
Loan Type	0.0
Insurance Description	0.0
0perational Type	1.0

A.3.3 GADT Experiment3

GADT model using shuffled sampling and accuracy split criteria achieved the highest accuracy. Tables 5.89 and 5.90 present the properties and confusion matrix respectively. Table 5.91 presents the weights for each attribute. Seven attributes have weight one.

Table 5.89 SCD1 GADT Experiment 3

Dataset	SCD1
Model name	GADT
Validation	10-cross validation
Criterion	Accuracy
Sampling type	Shuffled
Accuracy	67.08% +/- 5.26% (mikro: 67.08%)

Table 5.90 SCD1 GADT Experiment 3 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-defaulter	581	289	66.78%
Pred. Defaulter	139	291	67.67%

Table 5.91 SCD1 GADT Experiment3 attributes weights

Attribute name	Weight
Phone	0.0
ID Type	1.0
Gender	0.0
Age	0.0
Occupation	0.0
Marital Status	0.0
Number Of Children	0.0
Number of Spouses	0.0
Monthly Salary Value	0.0
Monthly Expenditures Value	0.0
Finance Size	1.0
Finance Duration	1.0
Payment Method	1.0
Finance Form	1.0
Loan Type	0.0
Insurance Description	1.0
0perational Type	1.0

B. Sudanese Credit Dataset2stag2 experiments

B.1 GAANN Experiments

Three experiments were conducted using GA and ANN with two different split validation ratios and 10-fold cross validation.

B.1.1 GAANN Experiment1

GAANN model using shuffled sampling, one hidden layer and learning rate 0.3 achieved the highest accuracy. Tables 5.92 and 5.93 present the properties and confusion matrix for the model respectively. Table 5.94 presents weights for each attribute. Nine attributes have weight one.

Table 5.92 SCD2 GAANN Experiment 1

Dataset	SCD2
Model name	GAANN
Validation	Split70:30
# Hidden layer	1
# neurons	3
Learning rate	0.3
Sampling type	Shuffled
Accuracy	84.21%

Table 5.93 SCD2 GAANN Experiment 1 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
pred. Non-defaulter	235	42	84.84%
pred. Defaulter	3	5	62.50

Table 5.94 SCD2 GAANN Experiment1 Attributes Weights

Attribute name	Weight
Gender	1.0
Age	1.0
Marital Status	1.0
#Children	0.0
#Spouses	1.0
Occupation	0.0
Phone	0.0
Id Type	0.0
Approved Credit Amount	0.0
Profit Margin	0.0
Periodical instalment Amount	0.0
Finance Duration	1.0
Periodicity of Payments	1.0
Purpose of Credit	1.0
Sector	1.0
Guarantee Type	0.0
Finance Size	1.0

B.1.2 GAANN Experiment2

GAANN using shuffled sampling, one hidden layer and learning rate 0.2 achieved the highest accuracy. Tables 5.95 and 5.96, present the properties and confusion matrix respectively. Table 5.97 presents weights for each attribute. Eleven attributes have weight one.

Table 5.95 SCD2 GAANN Experiment 2

Dataset	SCD2
Model name	GAANN
Validation	Split60:40
# Hidden layer	1
# neurons	3
Learning rate	0.2
Sampling type	Shuffled
Accuracy	82.89%

Table 5.96 SCD2 GAANN Experiment 2 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-Defaulter	309	58	84.20%
Pred. Defaulter	7	6	46.15%

Table 5.97 SCD2 GAANN Experiment2 Attributes Weights

Attribute name	**Weight**
Gender	0.0
Age	0.0
Marital Status	0.0
#Children	1.0
#Spouses	1.0
Occupation	0.0
Phone	1.0
IdType	1.0
Approved Credit Amount	1.0
Profit Margin	1.0
Periodical instalment Amount	1.0
Finance Duration	1.0
Periodicity of Payments	0.0
Purpose of Credit	1.0
Sector	1.0
Guarantee Type	1.0
Finance Size	0.0

B.1.3 GAANN Experiment3

GAANN using stratified sampling, one hidden layer and learning rate 0.3 achieved the highest accuracy. Tables 5.98and 5.99 present the properties

and confusion matrix for the model respectively.Tables 5.100presents theweights for each attribute. Nine attributes have weight one.

Table 5.98 SCD2 GAANN Experiment 3

Dataset	SCD2
Model name	ANN
Validation	10 cross-validation
Sampling type	Stratified
# Hidden layer	1
# neurons	3
Learning rate	.3
Accuracy	80.21% +/- 1.55%(micro:80.21%)

Table 5.99 SCD2 GAANN Experiment 3 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-Defaulter	732	177	80.53%
Pred. Defaulter	11	30	73.17%

Table 5.100 SCD2 GAANN Experiment3 Attributes Weights

Attribute name	**Weight**
Gender	0.0
Age	1.0
Marital Status	1.0
#Children	0.0
#Spouses	0.0
Occupation	0.0
Phone	1.0
IdType	0.0
Approved Credit Amount	0.0
Profit Margin	1.0
Periodical Instalment Amount	1.0
Finance Duration	1.0
Periodicity of Payments	0.0
Purpose of Credit	1.0
Sector	1.0
Guarantee Type	0.0
Finance Size	1.0

B.2 GASVM Experiments

Three experiments to develop GSVM models with two different validation ratios and 10 cross validation.

B.2.1 GASVM Experiment1

GASVM with shuffled sampling and Anova or Dot kernel achieved the highest accuracy. Tables 5.101,5.102 and 5.103 present the properties, confusion matrix for the model and weights for attributes respectively .Ten attributes have weight one.

Table 5.101 SCD2 GASVM Experiment 1

Dataset	SCD2
Model name	GASVM
Validation	Split70:30
Sampling type	Shuffled
Kernel type	A nova or dot
Accuracy	84.91%

Table 5.102 SCD2 GASVM Experiment 1 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
pred. Non-Defaulter	235	39	85.77%
pred. Defaulter	4	7	63.64%

Table 5.103 SCD2 GASVM Experiment1 Attributes Weights

Attribute name	Weight
Gender	1.0
Age	0.0
MaritalStatus	1.0
#Children	0.0
#Spouses	1.0
Occupation	0.0
Phone	0.0
IdType	1.0
Approved Credit Amount	1.0
Profit Margin	1.0
Periodical Instalment Amount	1.0
Finance Duration	1.0
Periodicity of Payments	0.0
Purpose of Credit	0.0
Sector	1.0
Guarantee Type	1.0
Finance Size	0.0

B.2.2 GASVM Experiment2

GASVM model using shuffled sampling and Radial kernel achieved the highest accuracy. Tables 5.104 and 5.105 present the properties and confusion matrix respectively. Table 5.106 presents the weights for each attribute. Five attributes have weight one.

Table 5.104 SCD2 GASVM Experiment 2

Dataset	SCD2
Model name	GASVM
Validation	Split60:40
Sampling type	Shuffled
Kernel type	Radial
Accuracy	84.21%

Table 5.105 SCD2 GASVM Experiment 2 Confusion Matrix

	true Non-defaulter	True defaulter	Class precision
pred. Non-defaulter	314	59	84.18%
pred. Defaulter	1	6	85.71%

Table 5.106 SCD2 GASVM Experiment2 Attributes Weights

Attribute name	**Weight**
Gender	0.0
Age	1.0
Marital Status	1.0
#Children	0.0
#Spouses	1.0
Occupation	0.0
Phone	0.0
IdType	0.0
Approved Credit Amount	0.0
Profit Margin	0.0
Periodical instalment Amount	0.0
Finance Duration	0.0
Periodicity of Payments	1.0
Purpose of Credit	0.0
Sector	1.0
Guarantee Type	0.0
Finance Size	0.0

B.2.3 GASVM Experiment3

GASVM using shuffled sampling and Radial kernel achieved the highest accuracy. Table 5.107 and 5.108 present the properties and confusion matrix respectively. Table 5.109 presents the weights for each attribute. Eight attributes have weight one.

Table 5.107 SCD2 GASVM Experiment 3

Dataset	SCD2
Model name	GASVM
Validation	10-cross-validation
Sampling type	Shuffled
Kernel type	Radial
Accuracy	80.21% +/- 4.18% (mikro: 80.21%)

Table 5.108 SCD2 GASVM Experiment 3 Confusion Matrix

	true Non-defaulter	True defaulter	Class precision
pred. Non-defaulter	739	184	80.07%
pred. Defaulter	4	23	85.19%

Table 5.109 SCD2 GASVM Experiment3 Attributes Weights

Attribute name	**Weight**
Gender	1.0
Age	0.0
Marital Status	0.0
#Children	0.0
#Spouses	1.0
Occupation	0.0
Phone	1.0
Id Type	0.0
Approved Credit Amount	0.0
Profit Margin	0.0
Periodical Instalment Amount	1.0
Finance Duration	0.0
Periodicity of Payments	1.0
Purpose of Credit	1.0
Sector	1.0
Guarantee Type	0.0
Finance Size	1.0

B.3 GADT Experiments

Three experiments to develop GADT models with two different validation ratios and 10 cross validation.

B.3.1 GADT Experiment1

GADT using shuffled sampling and split criteria Gini index achieved the highest accuracy. Tables 5.110 and 5.111 present the properties and confusion matrix for the model respectively. Table 5.112 presents the weights for each attribute. Six attributes have weight one.

Table 5.110 SCD2 GADT Experiment 1

Dataset	SCD2
Model name	GADT
Validation	Split 70:30
Criterion	Gini index
Sampling type	Shuffled
Accuracy	85.26%

Table 5.111 SCD2 GADT Experiment1 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-defaulter	231	32	87.83%
Pred. Defaulter	10	12	54.55%

Table 5.112 SCD2 GADT Experiment1 Attributes Weights

Attribute Name	**Weight**
Gender	0.0
Age	0.0
Marital Status	0.0
#Children	0.0
#Spouses	0.0
Occupation	1.0
Phone	0.0
IdType	0.0
Approved Amount	0.0
Profit Margin	0.0
Periodical Instalment Amount	1.0
Finance Duration	0.0
Periodicity of Payment	1.0
Purpose for Credit	1.0
Sector	1.0
Guarantee Type	0.0
Finance Size	1.0

B.3.2 GADT Experiment2

GADT with shuffled sampling and Accuracy split criteria achieved the highest accuracy. Tables 5.113 and 5.114 present the properties and confusion matrix for the model respectively. Table 5.115 presents the weights for each attribute. Seven attributes have weight one.

Table 5.113 SCD2 GADT Experiment 2

Dataset	SCD2
Model name	GADT
Validation	Split 60:40
Criterion	Accuracy
Sampling type	Shuffled
Accuracy	83.16%

Table 5.114 SCD2 GADT Experiment2 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
pred. Non-defaulter	307	55	84.81%
pred. Defaulter	9	9	50.00%

Table 5.115 SCD2 GADT Experiment2 Attributes Weights

Attribute name	**Weight**
Gender	1.0
Age	1.0
Marital Status	1.0
#Children	0.0
#Spouses	1.0
Occupation	0.0
Phone	0.0
Id Type	0.0
Approved Amount	0.0
Profit Margin	0.0
Periodical Instalment Amount	0.0
Finance Duration	1.0
Periodicity of Payment	0.0
Purpose for Credit	1.0
Sector	1.0
Guarantee Type	0.0
Finance Size	0.0

B.3.3 GADT Experiment3

GADT using stratified sampling and split criteria accuracy achieved the highest accuracy. Table 5.116 and 5.117 present the properties and confusion matrix respectively. Table 5.118 presents the weights for each attribute. Nine attributes have weight one

Table 5.116 SCD2 GADT Experiment 3

Dataset	SCD2
Model name	GADT
Validation	10-cross-validation
Criterion	Accuracy
Sampling type	Stratified
Accuracy	81.05% +/- 2.35% (mikro: 81.05%)

Table 5.117 SCD2 GADT Experiment3 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	732	167	81.42%
Pred. Defaulter	13	38	74.51%

Table 5.118 SCD2 GADT Experiment3 Attributes Weights

Attribute name	**Weight**
Gender	0.0
Age	0.0
Marital Status	0.0
#Children	0.0
#Spouses	0.0
Occupation	1.0
Phone	1.0
IdType	1.0
Approved Amount	0.0
Profit Margin	0.0
Periodical Instalment Amount	1.0
Finance Duration	1.0
Periodicity of Payment	1.0
Purpose for Credit	1.0
Sector	1.0
Guarantee Type	0.0
Finance Size	1.0

C. German Credit Dataset Stag2 Experiments

C.1 German Credit Dataset GAANN Experiments

Three experiments to develop GANN models with two different validation ratios and 10 cross validation.

C.1.1 German Credit Dataset GAANN Experiment1

GAANN model using shuffled sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.119 and 5.120 present the properties and confusion matrix respectively. Table 5.121 presents the weight for each attribute. Fifteen attributes have weight one.

Table 5.119 German GAANN Experiment1

Dataset	German
Model name	GAANN
Validation	Split70:30
# Hidden layer	2
# neurons	3
Learning rate	0.2
Sampling	Shuffled
Accuracy	80.00%

Table 5.120 German GAANN Experiment1 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	195	30	86.67%
Pred. Bad	30	45	60.00%

Table 5.121 German GAANN Experiment1 Attributes Weights

Attribute name	Weight
A1	1.0
A2	1.0
A3	1.0
A4	0.0
A5	1.0
A6	1.0
A7	1.0
A8	1.0
A9	0.0
A10	1.0
A11	0.0
A12	0.0
A13	0.0
A14	0.0
A15	0.0
A16	1.0
A17	1.0
A18	1.0
A19	1.0
A20	0.0
A21	0.0
A22	1.0
A23	1.0
A24	0.0

C.1.2 German Credit Dataset GAANN Experiment2

GAANN model using with shuffled sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.122 and 5.123 present the properties and confusion matrix respectively. Table 5.124 presents the weight for each attribute. Thirteen attributes have weight one.

Table 5.122 German GAANN Experiment 2

Dataset	German
Model name	GAANN
Validation	split 60:40
# Hidden layer	2
# neurons	3
Learning rate	0.2
Sampling type	Shuffled
Accuracy	79.50%

Table 5.123 German GAANN Experiment2 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	265	54	83.07%
Pred. Bad	28	53	65.43%

Table 5.124 German GAANN Experiment2 Attributes Weights

Attribute name	**Weight**
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	1.0
A7	1.0
A8	1.0
A9	0.0
A10	0.0
A11	1.0
A12	1.0
A13	1.0
A14	0.0
A15	0.0
A16	1.0
A17	0.0
A18	0.0
A19	1.0
A20	0.0
A21	0.0
A22	0.0
A23	0.0
A24	0.0

C.1.3 German Credit Dataset GAANN Experiment3

GAANN with stratified sampling, one hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.125 and 5.126 present the properties and confusion matrix for the model. Table 5.127 presents the weight for each attribute. Thirteen attributes have weight one.

Table 5.125 German GAANN Experiment 3

Dataset	German
Model name	GAANN
Validation	10-cross-validation
# Hidden layer	1
# neurons	3
Learning rate	0.2
Sampling type	Stratified
Accuracy	77.70% +/- 3.90% (mikro: 77.70%)

Table 5.126 German GAANN Experiment3 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	623	146	81.01%
Pred. Bad	77	154	66.67%

Table 5.127 German GAANN Experiment3 Attributes Weights

Attribute name	Weight
A1	1.0
A2	1.0
A3	1.0
A4	0.0
A5	1.0
A6	1.0
A7	1.0
A8	0.0
A9	1.0
A10	1.0
A11	1.0
A12	0.0
A13	0.0
A14	1.0
A15	0.0
A16	1.0
A17	0.0
A18	0.0
A19	0.0
A20	0.0
A21	0.0
A22	1.0
A23	1.0
A24	0.0

C.2 German Credit Dataset GASVM Experiments

Three experiments to develop GASVM models with two different validation ratios and 10 cross validation.

C.2.1 German Credit Dataset GASVM Experiment1

GASVM model using shuffled sampling and Anova kernel achieved the highest accuracy. Tables 5.128 and 5.129 present the properties and confusion matrix respectively. Table 5.130 presnts the weight for each attribute. Thirteen attributes have weight one.

Table 5.128German GASVM Experiment 1

Dataset	German
Model name	GASVM
Validation	Split 70:30
Sampling type	Shuffled
Kernel type	Anova
Accuracy	81.33%

Table 5.129 German GASVM Experiment1 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	199	37	84.32%
Pred. Bad	19	45	70.31%

Table 5.130German GASVM Experiment1 Attribute Weights

Attribute name	Weights
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	1.0
A7	1.0
A8	0.0
A9	0.0
A10	0.0
A11	0.0
A12	0.0
A13	0.0
A14	0.0
A15	0.0
A16	1.0
A17	1.0
A18	1.0
A19	0.0
A20	0.0
A21	0.0
A22	1.0
A23	1.0
A24	1.0

C.2.2 German Credit Dataset GASVM Experiment2

GASVM using shuffled sampling and Polynomial kernel achieved the highest accuracy. Tables 5.131 and 5.132 present the properties and confusion matrix respectively. Table 5.133 presents the weight for each attribute. Fourteen attributes have weight one.

Table 5.131German GASVM Experiment 2

Dataset	German
Model name	GASVM
Validation	split 60:40
Sampling type	shuffled
Kernel type	Polynomial
Accuracy	80.25%

Table 5.132 German GASVM Experiment2 Confusion Matrix

	True Good	True Bad	Class precision
pred. Good	273	66	80.53%
pred. Bad	13	48	78.69%

Table 5.133 German GASVM Experiment2 Attribute Weights

Attribute name	**Weight**
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	1.0
A7	1.0
A8	1.0
A9	0.0
A10	0.0
A11	0.0
A12	0.0
A13	0.0
A14	0.0
A15	0.0
A16	1.0
A17	0.0
A18	1.0
A19	0.0
A20	1.0
A21	0.0
A22	1.0
A23	1.0
A24	1.0

C.2.3 German Credit Dataset GASVM Experiment3

GASVM with shuffled sampling and Polynomial kernel achieved the highest accuracy. Tables 5.134 and 5.135 present the properties and confusion matrix for the model respectively . Table 5.136 presents the weight for each attribute. Sixteen attributes have weight one.

Table 5.134German GASVM Experiment 3

Dataset	German
Model name	GASVM
Validation	10-cross-validation
Sampling type	Stratified
Kernel type	Polynomial
Accuracy	77.60% +/- 2.58% (mikro: 77.60%)

Table 5.135 German GASVM Experiment3 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	678	202	77.05%
Pred. Bad	22	98	81.67%

Table 5.136German GASVM Experiment3 Attributes Weights

Attribute name	Attribute weight
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	0.0
A7	1.0
A8	1.0
A9	0.0
A10	1.0
A11	1.0
A12	1.0
A13	0.0
A14	0.0
A15	1.0
A16	1.0
A17	0.0
A18	1.0
A19	0.0
A20	1.0
A21	1.0
A22	0.0
A23	1.0
A24	0.0

GADT Experiments

For the German credit dataset, the Attributes in numeric version and categorical version are not equivalent. Therefore,three experiments were conducted using the categorical dataset and addition three experiments were conducted using the numerical version. All these experiment used two split ratios and 10 cross validation.

C.3.1 GADT Experiment1

GADT using shuffled sampling and Gini index split criteria achieved the highest accuracy. Tables 5.137 and 5.138 present the properties and confusion matrix respectively. Table 5.139 presents the weight for each attribute. Eight attributes have weight one.

Table 5.137 German GADT Experiment 1

Dataset	German (categorical version)
Model name	GADT
Validation	Split 70:30
Criterion	Gini index
Sampling type	Shuffled
Accuracy	79%

Table 5.138 German GADT Experiment1 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	197	40	83.12%
Pred. Bad	23	40	63.49%

Table 5.139 German GADT Experiment1 Attributes Weights

Attribute name	Weight
Status of existing checking account	1.0
Duration in month	0.0
Credit history	1.0
Purpose	0.0
Credit amount	1.0
Savings account/bonds	1.0
Present employment since	1.0
Installment rate in percentage of disposable income	0.0
Personal status and sex	0.0
Other debtors / guarantors	0.0
Present residence since	0.0
Property	1.0
Age in years	1.0
Other installment plans	1.0
Housing	0.0
Number of existing credits at this bank	1.0
Job	1.0
Number of people being liable to provide maintenance for	1.0
Telephone	0.0
Foreign worker	1.0

C.3.2 GADT Experiment2

GADT using shuffled sampling and Gini index split criteria achieved the highest accuracy. Tables 5.140 and 5.141 present the properties and confusion matrix respectively. Table 5.142 presents the weight for each attribute. Thirteenth attributes have weight one.

Table 5.140 German GADT Experiment2

Dataset	German (categorical version)
Model name	GADT
Validation	Split 60:40
Criterion	Gini index
Sampling type	Shuffled
Accuracy	75.25%

Table 5.141 German GADT Experiment2 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	245	65	79.03%
Pred. Bad	34	56	62.22%

Table 5.142 German GADT Experiment2 Attributes Weights

Attribute name	Weights
Status of existing checking account	1.0
Duration in month	1.0
Credit history	1.0
Purpose	0.0
Credit amount	0.0
Savings account/bonds	1.0
Present employment since	1.0
Installment rate in percentage of disposable income	0.0
Personal status and sex	0.0
Other debtors / guarantors	1.0
Present residence since	0.0
Property	1.0
Age in years	1.0
Other installment plans	1.0
Housing	0.0
Number of existing credits at this bank	1.0
Job	1.0
Number of people being liable to provide maintenance	1.0
Telephone	1.0
Foreign worker	1.0

C.3.3 GADT Experiment3

GADT using shuffled sampling and Accuracy split criteria achieved the highest accuracy. Tables 5.143 and 5.144 present the properties and confusion matrix for the model respectively. Table 5.145 presents the weight for each attribute. Six attributes have weight one.

Table 5.143 German GADT Experiment3

Dataset	German (categorical version)
Model name	GADT
Validation	10-cross- validation
Criterion	accuracy
Sampling type	Shuffled
Accuracy	75.00% +/- 5.08% (mikro: 75.00%)

Table 5.144 German GADT Experiment3 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	609	159	79.30%
Pred. Bad	91	141	60.78%

Table 5.145 German GADT Experiment3 Attributes Weights

Attribute name	**Weight**
Status of existing checking account	1.0
Duration in month	0.0
Credit history	0.0
Purpose	0.0
Credit amount	0.0
Savings account/bonds	0.0
Present employment since	0.0
Installment rate in percentage of disposable income	0.0
Personal status and sex	0.0
Other debtors / guarantors	0.0
Present residence since	0.0
Property	1.0
Age in years	1.0
Other installment plans	1.0
Housing	0.0
Number of existing credits at this bank	0.0
Job	1.0
Number of people being liable to provide maintenance for	0.0
Telephone	0.0
Foreign worker	1.0

C.3.4 GADT Experiment4

GADT with shuffled sampling and Accuracy split criteria achieved the highest accuracy. Tables 5.146 and 5.147 present the properties and confusion matrix for the model respectively .Table 5.148 presents the weight for each attribute. Eleven attributes have weight one.

Table 5.146 German GADT Experiment4

Dataset	German (numeric)
Model name	GADT
Validation	Split 70:30
Criterion	Accuracy
Sampling type	Shuffled
Accuracy	77.33%

Table 5.147 German GADT Experiment4 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	189	42	81.82%
Pred. Bad	26	43	62.32%

Table 5.148 German GADT Experiment4 Attributes Weights

Attribute name	**Weight**
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	0.0
A6	1.0
A7	0.0
A8	1.0
A9	0.0
A10	1.0
A11	0.0
A12	0.0
A13	0.0
A14	0.0
A15	0.0
A16	0.0
A17	0.0
A18	1.0
A19	0.0
A20	0.0
A21	0.0
A22	1.0
A23	1.0
A24	1.0

C.3.5 GADT Experiment5

GADT using shuffled sampling and Accuracy split criteria achieved the highest accuracy. Tables 5.149 and 5.150, present the properties and confusion matrix for the model respectively .Table 5.151 the weights for each attribute respectively. Twelve attributes have weight one.

Table 5.149 German GADT Experiment5

Dataset	German (numeric)
Model name	GADT
Validation	Split 60:40
Criterion	accuracy
Sampling type	Shuffled
Accuracy	76.75%

Table 5.150 German GADT Experiment5 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	242	55	81.48%
Pred. Bad	38	65	63.11%

Table 5.151 German GADT Experiment5 Attributes Weights

Attribute name	**Weight**
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	1.0
A7	0.0
A8	1.0
A9	0.0
A10	1.0
A11	0.0
A12	0.0
A13	0.0
A14	0.0
A15	0.0
A16	0.0
A17	0.0
A18	1.0
A19	0.0
A20	0.0
A21	0.0
A22	1.0
A23	1.0
A24	1.0

C.3.6 GADT Experiment6

GADT using shuffled sampling and Accuracy split criteria achieved the highest accuracy. Tables 5.152 and 5.153, present the properties and confusion matrix for the model respectively .Table 5.154 the weights for each attribute respectively. Thirteen attributes have weight one.

Table 5.152 German GADT Experiment 6

Dataset	German (numeric)
Model name	GADT
Validation	10-cross
Criterion	accuracy
Sampling type	Shuffled
Accuracy	75.30% +/- 2.83% (mikro: 75.30%)

Table 5.153 German GADT Experiment6 Confusion Matrix

	True Good	True Bad	Class Precision
Pred. Good	628	175	78.21%
Pred. Bad	72	125	63.45%

Table 5.154 German GADT Experiment6 Attributes Weights

Attribute name	**Weight**
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	1.0
A7	1.0
A8	1.0
A9	0.0
A10	1.0
A11	0.0
A12	1.0
A13	0.0
A14	0.0
A15	1.0
A16	0.0
A17	0.0
A18	0.0
A19	0.0
A20	0.0
A21	0.0
A22	0.0
A23	1.0
A24	1.0

5.6.7.3 Stage3 Experiments

From Stage2 experiments (GAANN , GASVM, and GADT models with two different split validation ratios and 10-cross validation) three CSMs which achieved highest accuracies were identified for each dataset . The attributes which have weight one in at least two identified models were picked. Therefore, new reduced sets of features have been identified for each dataset. Table 5.155, Table 5.156, and Table 5.157 present the models for stage2 and their accuracies for SCD1, SCD2, and German dataset respectively. The best models in these tables are underlined.

Table 5.155 SCD1GA Models for Stage 2

Technique	GAANN			GASVM			GADT		
Validation	70:30	60:40	10- cross	70:30	60:40	10- cross	70:30	60:40	10-cross
Accuracy %	71. 03	69.62	67.31	69.49	69.81	67.46	71.03	68.65	67.08

Table 5.156 SCD2 GA Models for Stage 2

Technique	GAANN			GASVM			GADT		
Validation	70:30	60:40	10-cross	70:30	60:40	10-cross	70:30	60:40	10-cross
Accuracy %	84.21	82.89	80.21	84.91	84.21	80.210	85.26	83.16	81.05

Table 5.157 GermanDataset GA Models for Stage2

Technique	GAANN			GASVM			GADT		
Validation	70:30	60:40	10- cross	70:30	60:40	10- cross	70:30	60:40	10- cross
Accuracy %	80.00	79.50	77.70	81.0.3	80.25	77.60	77.33	76.75	75.30

From the tables 5.155, 5.156, and 5.157: GAANN (70:30), GASVM (60:40), and GADT (70:30) yielded highest accuracies for SCD1. Similarly GAANN (70:30), GASVM (70:30), and GADT (70:30) for SCD1 and German dataset.

From tables of attribute and their weights for these models (from stage 2) new reduced datasets were produced. Tables 5.158, 5.159, and 5.160 illustrate these reduced sets for SCD1, SCD2, and German dataset respectively.

Table 5.158 Reduced SCD1

Attribute name	**Weight**
Have phone	1.0
Occupation	1.0
Material Status	1.0
Number of Spouses	1.0
Finance Size	1.0
Finance Duration	1.0
Payment Method	1.0
Loan Type	1.0
Insurance Description	1.0
0perational Type	1.0

Table 5.159 Reduced SCD2

Attribute name	**Weight**
Gender	1.0
MaritalStatus	1.0
#Spouses	1.0
Periodical instalment Amount	1.0
Finance Duration	1.0
Periodicity of Payments	1.0
Purpose of Credit	1.0
Sector	1.0
Finance Size	1.0

Table 5.160 Reduced German Dataset

Attribute name	Weight
A1	1.0
A2	1.0
A3	1.0
A4	1.0
A5	1.0
A6	1.0
A7	1.0
A8	1.0
A10	1.0
A16	1.0
A17	1.0
A18	1.0
A22	1.0
A23	1.0
A24	1.0

A. Reduced Sudanese credit dataset1 (RSCD1) stage 3 experiments

In these experiments three classification techniques were applied namely ANN, SVM, and DT to RSCD1.

A.1 ANNExperiments

Three experiments were conducted using ANN with two different split validation ratios and 10-fold cross validation.

A.1.1 ANN Experiment1

ANN3 model with stratified sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.161 and 5.162 present the properties and confusion matrix for the model respectively.

Table 5.161 RSCD1 ANN3 Experiment 1

Dataset	RSCD1
Model name	ANN3
Validation	Split 70:30
# Hidden layer	2
# neurons	3
Learning rate	0.2
Sampling type	Stratified
Accuracy	65.13%

Table 5.162 RSCD1 ANN3 Experiment 1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	185	105	63.79%
Pred. Defaulter	31	69	69.00%

A.1.2 ANN Experiment2

ANN3 model with stratified sampling, two hidden layers and learning rate 0.3 achieved the highest accuracy. Table 5.163, 5.164 present the properties and confusion matrix for the model respectively.

Table 5.163 RSCD1 ANN3 Experiment 2

Dataset	RSCD1
Model name	ANN3
Validation	Split 60:40
# Hidden layer	1
# neurons	3
Learning rate	0.3
Sampling type	Stratified
Accuracy	65.19%

Table 5.164 RSCD1 ANN3 Experiment 2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	250	143	63.61%
Pred. Defaulter	38	89	70.08%

A.1.3 ANN Experiment 3

ANN model using shuffled sampling, two hidden layers and learning rate 0.3 achieved the highest accuracy. Table 5.165, 5.166 present the properties and confusion matrix for the model respectively.

Table 5.165 RSCD1 ANN3 Experiment 3

Dataset	RSCD1
Model name	ANN3
Validation	10-cross- validation
# Hidden layer	2
# neurons	3
Learning rate	0.3
Sampling type	Shuffled
Accuracy	63.77% +/- 4.36% (mikro: 63.77%)

Table 5.166 RSCD1 ANN3 Experiment 3 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	250	143	63.61%
Pred. Defaulter	38	89	70.08%

A.2 SVM Experiments

Three experiments were conducted using SVM with two different split validation ratios and 10-fold cross validation.

A.2.1 SVM Experiment1

In this experiment anova kernel and shuffled sampling achieved the highest accuracy. Table 5.167 and Table 5.168 present the properties and confusion matrix for the model respectively.

Table 5.167 RSCD1 SVM3 Experiment1

Dataset	RSCD1
Model name	SVM3
Validation	Split70:30
Sampling type	Shuffled
Kernel type	anova
Accuracy	64.10%

Table 5.168 RSCD1 SVM3 Experiment1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	173	98	63.84%
Pred. Defaulter	42	77	64.71%

A.2.2 SVM Experiment2

In this experiment anova kernel and stratified sampling achieved the highest accuracy. Table 5.169 and Table 5.170 present the properties and confusion matrix for the model respectively.

Table 5.169 RSCD1 SVM3 Experiment2

Dataset	RSCD1
Model name	SVM3
Validation	Split 60:40
Sampling type	Stratified
Kernel type	anova
Accuracy	65.00%

Table 5.170 RSCD1 SVM3 Experiment2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class Precision
Pred. Non-defaulter	270	164	62.21%
Pred. Defaulter	18	68	79.07%

A.2.3 SVM Experiment3

In this experiment anova kernel and stratified sampling achieved the highest accuracy. Table 5.171 and Table 5.172 present the properties and confusion matrix for the model respectively.

Table 5.171 RSCD1 SVM3 Experiment3

Dataset	RSCD1
Model name	SVM3
Validation	10-cross-validation
Sampling type	Stratified
Kernel type	Anova
Accuracy	66.62% +/- 3.02% (mikro: 66.62%)

Table 5.172 RSCD1 SVM3 Experiment3 Confusion Matrix

	true Non-defaulter	true Defaulter	class precision
pred. Non-defaulter	593	307	65.89%
pred. Defaulter	127	273	68.25%

A.3 DT Experiments

Three experiments were conducted using DT with two different split validation ratios and 10-fold cross validation.

A.3.1 DT Experiment 1

Gini index split criteria and stratified sampling achieved the highest accuracy DT3 model. Table 5.173 and Table 5.174 present the properties and confusion matrix for the model respectively.

Table 5.173 RSCD1 DT3 Experiment 1

Dataset	RSCD1
Model name	DT3
Validation	Split 70:30
Criterion	Gini index
Sampling type	stratified
Accuracy	61.79%

Table 5.174 RSCD1 DT3 Experiment 1 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	157	90	63.56%
Pred. Defaulter	59	84	58.74%

A.3 DT Experiment 2

Gini index split criteria and shuffled sampling achieved the highest accuracy DT3 model. Table 5.175 and Table 5.176 present the properties and confusion matrix for the model respectively.

Table 5.175 RSCD1 DT3 Experiment 2

Dataset	RSCD1
Model name	DT3
Validation	Split 60:40
Criterion	Gini index
Sampling type	Shuffled
Accuracy	63.08%

Table 5.176 RSCD1 DT3 Experiment 2 Confusion Matrix

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	210	113	65.02%
Pred. Defaulter	79	118	59.90%

A.1.2 DT Experiment 2

Gini index split criteria and stratified sampling achieved the highest accuracy DT3 model. Table 5.177 and Table 5.178 present the properties and confusion matrix for the model respectively.

Table 5.177 RSCD1 DT3 Experiment 3

Dataset	RSCD1
Model	DT3
Validation	10-cross-validation
Criterion	Gini index
Sampling type:	Stratified
Accuracy	63.23% +/- 3.07% (mikro: 63.23%)

Table 5.178 RSCD1 DT3 Experiment 3

	True Non-defaulter	True Defaulter	Class precision
Pred. Non-defaulter	511	269	65.51%
Pred. Defaulter	209	311	59.81%

B. Reduced Sudanese Credit dataset 2 (RSCD2) Stage 3 Experiments

In these experiments three classification techniques were applied namely ANN, SVM, and DT to RSCD2.

B.1 ANN Experiments

Three experiments were conducted using ANN with two different split validation ratios and 10-fold cross validation.

B.1.1 ANN Experiments1

ANN3 model with stratified sampling, one hidden layers and learning rate 0.3 achieved the highest accuracy. Tables 5.179 and 5.180 present the properties and confusion matrix for the model respectively.

Table 5.179 RSCD2 ANN3 Experiment 1

Dataset	RSCD2
Model name	ANN3
Validation	Split 70:30
# Hidden layer	1
# neurons	3
Learning rate	0.3
Sampling type	Stratified
Accuracy	78.95%

Table 5.180RSCD2 ANN3 Experiment 1 Confusion Matrix

	True Non-Defaulter	True defaulter	Class precision
Pred. Non-Defaulter	218	55	79.85%
Pred. Defaulter	5	7	58.33%

B.1.2 ANN Experiments2

ANN3 model with stratified sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.181 and 5.182 present the properties and confusion matrix for the model respectively

Table 5.181RSCD2 ANN3 Experiment 2

Dataset	RSCD2
Model name	ANN3
Validation	Split 60:40
# Hidden layer	2
# neurons	3
Learning rate	.2
Sampling type	Stratified
Accuracy	78.68%

Table 5.182RSCD2 ANN3 Experiment 2 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-Defaulter	289	73	79.83%
Pred. Defaulter	8	10	55.56%

B.1.3 ANN Experiments3

ANN3 model with stratified sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.183 and 5.184 present the properties and confusion matrix for the model respectively.

Table 5.183 RSCD2 SVM3 Experiment1

Dataset	RSCD2
Model name	ANN3
Validation	10-cross validation
# Hidden layer	1
# neurons	3
Learning rate	.2
Sampling type	shuffled
Accuracy	78.74% +/- 3.76% (mikro: 78.74%)

Table 5.184RSCD2 ANN3 Experiment 3 Confusion Matrix

	True Non-defaulter	True defaulter	Class Precision
Pred. Non-Defaulter	730	189	79.43%
Pred. Defaulter	13	18	58.06%

B.2 SVM Experiments

Three experiments were conducted using SVM with two different split validation ratios and 10-fold cross validation.

B.2.1 SVM Experiment1

In this experiment SVM3 with anova kernel and stratified sampling achieved the highest accuracy. Table 5.185 and Table 5.186 present the properties and confusion matrix for the model respectively.

Table 5.185 RSCD2 SVM3 Experiment1

Dataset	RSCD2
Model name	SVM3
Validation	Split70:30
Sampling type	Stratified
Kernel type	anova
Accuracy	80.00%

Table 5.186 RSCD2 SVM3 Experiment 2 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-Defaulter	223	57	79.64%
Pred. Defaulter	0	5	100.00%

B.2.2 SVM Experiment 2

In this experiment SVM3 with Polynomial, Dot kernels, and stratified sampling achieved the highest accuracy. Table 5.187 and Table 5.188 present the properties and confusion matrix for the model respectively

Table 5.187 RSCD2 SVM3 Experiment2

Dataset	RSCD2
Model name	SVM3
Validation	Split 60:40
Sampling type	Stratified
Kernel type	Polynomial or Dot
Accuracy	79.74%

Table 5.188 RSCD2 SVM 3 Experiment 2 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-Defaulter	294	74	79.89%
Pred. Defaulter	3	9	75.00%

B.2.3 SVM Experiment3

In this experiment SVM3 with Dot kernel and shuffled sampling achieved the highest accuracy. Table 5.189 and Table 5.190 present the properties and confusion matrix for the model respectively.

Table 5.189 RSCD2 SVM3 Experiment3

Dataset	RSCD2
Model name	SVM3
Validation	10-cross-validation
Sampling type	Shuffled
Kernel type	Dot
Accuracy	79.37% +/- 4.35% (mikro: 79.37%)

Table 5.190 RSCD2 SVM3 Experiment3 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
Pred. Non-Defaulter	731	184	79.89%
Pred. Defaulter	12	23	65.71%

B.3 DT Experiments

Three experiments were conducted using DT with two different split validation ratios and 10-fold cross validation.

B.3.1 DT Experiment 1

Gini index split criteria and shuffled sampling achieved the highest accuracy DT3 model. Table 5.191 and Table 5.192 present the properties and confusion matrix for the model respectively.

Table 5.191 RSCD2 DT3 Experiment 1

Dataset	RSCD2
Model name	DT3
Validation	Split 70:30
Criterion	Gini index
Sampling type	Shuffled
Accuracy	82.46%

Table 5.192 RSCD2 DT3 Experiment 1 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
pred. Non-Defaulter	224	45	83.27%
pred. Defaulter	5	11	68.75%

B.3.2 DT Experiment 2: Accuracy split criteria and shuffled sampling achieved the highest accuracy DT3 model. Table 5.193 and Table 5.194 present the properties and confusion matrix for the model respectively.

Table 5.193 RSCD2 DT3 Experiment2

Dataset	RSCD2
Model name	DT3
Validation	Split 60:40
Criterion	accuracy
Sampling type:	Shuffled
Accuracy	81.58%

Table 5.194 RSCD2 DT3 Experiment2 Confusion Matrix

	True Non-Defaulter	True defaulter	Class precision
Pred. Non-Defaulter	298	66	81.87%
Pred. Defaulter	4	12	75.00%

B.3.3 DT Experiment 3

Accuracy split criteria and shuffled sampling achieved the highest accuracy DT3 model. Table 5.195 and Table 5.196 present the properties and confusion matrix for the model respectively.

Table 5.195 RSCD2 DT3 Experiment 3

Dataset	RSCD2
Model name	DT3
Validation	10-cross-validation
Criterion	accuracy
Sampling type:	Shuffled
Accuracy	78.42% +/- 4.37% (mikro: 78.42%)

Table 5.196 RSCD2 DT3 Experiment3 Confusion Matrix

	True Non-defaulter	True defaulter	Class precision
pred. Non-Defaulter	730	190	79.35%
pred. Defaulter	15	15	50.00%

C. Reduced German Credit Dataset Stage 3 Experiments

In these experiments three classification techniques were applied namely ANN, SVM, and DT Reduced German dataset.

C.1 ANN Experiments

Three experiments were conducted using ANN with two different split validation ratios and 10-fold cross validation.

C.1.1 ANN Experiment1

ANN3modelwith stratified sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.197 and 5.198 present the properties and confusion matrix for the model respectively.

Table 5.197 Reduced German Dataset ANN3 Experiment 1

Dataset	Reduced German dataset
Model name	ANN3
Validation	Split 70:30
# Hidden layer	2
# neurons	3
Learning rate	0.2
Sampling type	Stratified
Accuracy	76.00%

Table 5.198Reduced German Dataset ANN3 Experiment 1 Confusion Matrix

	True Good	True Bad	Class precision
pred. Good	184	46	80.00%
pred. Bad	26	44	62.86%

C.1.2 ANN Experiment2

ANN3modelwith shuffled sampling, two hidden layers and learning rate 0.2 achieved the highest accuracy. Tables 5.199and 5.200 present the properties and confusion matrix for the model respectively.

Table 5.199 Reduced German Dataset ANN3 Experiment 2

Dataset	Reduced German dataset
Model name	ANN3
Validation	Split 60:40
# Hidden layer	2
# neurons	3
Learning rate	0.2
Sampling type	Shuffled
Accuracy	75.00%

Table 5.200Reduced German Dataset ANN3 Experiment 1 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	231	65	78.04%
Pred. Bad	35	69	66.35%

C.1.3 ANN Experiment3

ANN3model with stratified sampling, one hidden layers and learning rate 0.3 achieved the highest accuracy. Tables 5.201 and 5.202 present the properties and confusion matrix for the model respectively

Table 5.201 Reduced German dataset ANN3 Experiment 3

Dataset	Reduced German dataset
Model name	ANN3
Validation	10-cross-validation
# Hidden layer	1
# neurons	3
Learning rate	0.3
Sampling	Stratified
Accuracy	74.70% +/- 2.28% (mikro: 74.70%)

Table 5.202Reduced German Dataset ANN3 Experiment 3 Confusion Matrix

	True Good	True Bad	Class precision
pred. Good	602	155	79.52%
pred. Bad	98	145	59.67%

C.2 SVM Experiments

Three experiments were conducted using SVM with two different split validation ratios and 10-fold cross validation.

C.2.1 SVM Experiment1

In this experiment SVM3 with Dot kernel and stratified sampling achieved the highest accuracy. Table 5.203 and Table 5.204 present the properties and confusion matrix for the model respectively.

Table 5.203 Reduced German Dataset SVM3 Experiment1

Dataset	Reduced German dataset
Model name	SVM3
Validation	Split70:30
Sampling type	Stratified
Kernel type	Dot
Accuracy	75.33%

Table 5.204Reduced German Dataset SVM3 Experiment1 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	192	56	77.42%
Pred. Bad	18	34	65.38%

C.2.2 SVM Experiment2

In this experiment SV3 model with polynomial kernel and stratified sampling achieved the highest accuracy. Table 5.205 and Table 5.206 present the properties and confusion matrix for the model respectively.

Table 5.205 Reduced German Dataset SVM3 Experiment2

Dataset	Reduced German dataset
Model name	SVM3
Validation	Split60:40
Sampling type	Stratified
Kernel type	Polynomial
Accuracy	75%

Table 5.206 Reduced German Dataset SVM3 Experiment2 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	259	79	76.63%
Pred. Bad	21	41	66.13%

C.2.3 SVM Experiment3

In this experiment SVM3 model with Dot kernel and stratified sampling achieved the highest accuracy. Table 5.207 and Table 5.208 present the properties and confusion matrix for the model respectively.

Table 5.207 Reduced German Dataset SVM3 Experiment 3

Dataset	Reduced German dataset
Model name	SVM3
Validation	10-cross-validation
Sampling type	Stratified
Kernel type	Dot
Accuracy	75.70% +/- 3.00% (mikro: 75.70%)

Table 5.208 Reduced German Dataset SVM3 Experiment3 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	641	184	77.70%
Pred. Bad	59	116	66.29%

C.3 DT Experiments

Three experiments were conducted using DT with two different split validation ratios and 10-fold cross validation.

C.3.1 DT Experiment 1

Accuracy split criteria and linear sampling achieved the highest accuracy DT3 model. Table 5.209 and Table 5.210 present the properties and confusion matrix for the model respectively.

Table 5.209 Reduced German Dataset DT3 Experiment1

Dataset	Reduced German dataset
Model name	DT3
Validation	Split 70:30
Criterion	Accuracy
Sampling type	Linear
Accuracy	74.67%

Table 5.210 Reduced German Dataset DT3 Experiment 1Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	187	56	76.95%
Pred. Bad	20	37	64.91%

C.3.2 DT Experiment 2

Accuracy split criteria and linear sampling achieved the highest accuracy DT3 model. Table 5.211 and Table 5.212 present the properties and confusion matrix for the model respectively.

Table 5.211 Reduced German Dataset DT3 Experiment2

Dataset	Reduced German dataset
Model name	DT3
Validation	Split 60:40
Criterion	Accuracy
Sampling type	Linear
Accuracy	75.00%

Table 5.212Reduced German Dataset DT3 Experiment2 Confusion Matrix

	True Good	True Bad	Class precision
pred. Good	242	65	78.83%
pred. Bad	35	58	62.37%

C.3.3 DT Experiment 3

Accuracy split criteria and shuffled sampling achieved the highest accuracy DT3 model. Table 5.213 and Table 5.214 present the properties and confusion matrix for the model respectively.

Table 5.213 Reduced German Dataset DT3 Experiment3

Dataset	Reduced German dataset
Model name	DT3
Validation	10-cross-validation
Criterion	Accuracy
Sampling type:	Shuffled
Accuracy	73.70% +/- 4.38% (mikro: 73.70%)

Table 5.214Reduced German Dataset DT3 Experiment3 Confusion Matrix

	True Good	True Bad	Class precision
Pred. Good	615	178	77.55%
Pred. Bad	85	122	58.94%

5.7 Summary

In this chapter two credit datasets for the two identified Sudanese banks were constructed. These datasets and the German credit dataset were employed to evaluate the proposed CSMs. To develop these proposed CSMs, three stages of different experiments for each dataset were conducted. In all experiments two types of validation were used, namely split validation (with two different ratios of 70:30 and 60:40) and K-fold cross validation (K was chosen to be 10). Three sampling techniques for validation were also tested in these experiments, namely linear, shuffled and stratified.

ANN, SVM and DT as DM classification techniques were employed in stage 1 experiments.GA as a feature selection technique was employed in stage 2. As a result of using GA in stage2, tables of attributes and their weights were produced. By using these tables new reduced sets of features were identified for each dataset. These new reduced datasets were employed in stage 3.

All experiments in these different stages were repeated for each dataset. In each stage of these experiments nine CSMs were developed for each dataset. Hence for all stages 27 CSMs were developed for each dataset. All these CSMs will be evaluated and compared in the next chapter.

CHAPTER SIX

6. Results and Discussion

6.1 Overview

This chapter presents the final results of the proposed CSMs. It starts by presenting the evaluation measures that are identified to evaluate the proposed CSMs in this research. General characteristics for each dataset of this research are also illustrated.

The results of all experiments for each stage and for each dataset were compared in terms of specified evaluation measures and discussed.

General concluding remarks about the experiments and their techniques in different stages for each dataset are also presented. As the result of these comparisons best technique(s) is (are) identified for each dataset. Finally, the conclusions of the all experimental results for this research are illustrated.

6.2 Evaluation measures

The classification performance of the proposed CSMs were identified using Confusion matrix .The confusion matrixis a useful tool for analyzing how well the classifier can recognize tuples (instances) of different classes

Accuracy was used as a dominant evaluation measure for the models in this research for the following reasons:

- It can be used to estimate the classifier future prediction accuracy.

- It can be used for choosing the best classifier from a given set of classifiers (selecting the “best” classification model from two or more models).

Beside accuracy, other evaluation criteria which are adopted from the established standard measures in the fields of CS were used to assess the performance of the proposed models. These measures include precision (defaulter), precision (non-defaulter), Type I error (the rate of misclassifying a good credit customer or non-defaulter as a bad customer or defaulter) and Type Π error (the rate of misclassifying a bad credit customer or defaulter as a good customer or non-defaulter).

Type I error rate= *Fpos/(pos + neg) ;*

Type II error rate=*Fneg / (pos + neg);*

6.3 General Characteristics of the Datasets

In this research three credit datasets were employed to evaluate the proposed CSMs. Table 6.1 presents the main characteristics of these datasets. Number of attributes of SCD1 and SCD2 are equal while the number attribute of German is greater than the Sudanese datasets.

While SCD1 is a balanced dataset to some extent (percentage of defaulter is approximately equal to percentage of non-defaulter), SCD2 and German are imbalanced datasets.

It is clear that these datasets have different structures and different attributes (predictors).

According to literature survey that, in spite of the success of many statistical and AI techniques in developing CSMs, there are no reliable conclusions on which ones are better [105].Each technique has its drawbacks and advantages. Therefore, capability of CS problems depends on the data structure, the characteristics used, the extent to which it is possible to identify the classes by using those characteristics, and the objective of the classification[105].

One of the main objectives of this research is to search for an optimal DM classification technique for each dataset and to identify the optimal one for all if it is found.

In order to achieve this objective, all proposed models (in the previous chapter) in the different stages of this research were compared using the aforementioned measures.

Table 6.1 The General Characteristics of the Datasets

Dataset	# Attributes	# Instances	# Instances for each class	Non-defaulter (%)	Defaulter (%)
SCD1	17	1300	720:Non-defaulters, 580: Defaulters	55.38	44.62
SCD2	17	950	745:Non-defaulters, 205:Defaulters	21.58	78.42
German	20 (original), 24 (numerical version)	1000	700:Good, 300:Bad	70	30

6.4 Comparisons and Discussion of Results for Proposed CSMs

Different comparisons between the proposed models were conducted and discussed as follows:

1. The resulting models of the stage 1 experiments for each dataset were compared and discussed.
2. The resulting models of the stage 2 experiments for each dataset were compared and discussed.
3. The resulting models of stage 3 experiments for each dataset were discussed and compared.
4. The resulting models of the stage 1, stage 2 and stage 3 experiments for each dataset were compared and discussed.

6.4.1Comparisons and Discussion of Stage 1 Resulting Models

6.4.1.1Comparisons and Discussion of the SCD1 Stage 1 Resulting Models

Table 6.2 presents the final results for the SCD1 stage1 experiments. These results reveal that:

- SVM(60:40) model outperforms all others in terms of accuracy and precision (Non-defaulter).
- SVM(70:30) model outperforms all others in terms of precision (Defaulter).
- However, ANN (10-cross) and DT (70:30) models achieve lower Type Π error and Type I rates respectively; they yield lower accuracy than SVM models.

- Therefore, SVM (70:30) and SVM (60:40) are the optimal models because they are slightly different in accuracy, precision of defaulter and non-defaulter.
- Type Π error is most costly than Type I error. Therefore, in our opinion the latter one is the best because it achieves lower type Π error than the former.

Table 6.2 Results of Scoring Models for SCD1 Stage 1 Experiments

Techniques	VALIDATION			Accuracy%	Precision (%)		Type I error (%)	Type II error (%)
					Non-defaulter	Defaulter		
ANN	**Split validation**	**Training : Testing**	**70: 30**	66.67	65.93	68.33	9.74	23.59
			60: 40	64.62	63.47	67.91	8.27	27.12
	X-validation	**10 –cross**		63.62	64.65	61.71	13.46	**22.92**
SVM	**Split validation**	**Training : Testing**	**70:30**	66.41	64.51	**72.16**	6.92	26.67
			60:40	**66.73**	**65.58**	69.54	8.85	24.42
	X-validation	**10 –cross**		65	64.51	66.15	10.08	24.92
DT	**Split validation**	**Training : Testing**	**70:30**	60	60.62	54.05	**4.36**	35.64
			60:40	58.85	60.09	50.72	6.54	34.62
	X-validation	**10-cross**		58.08	60.71	53.62	17.23	24.69

6.4.1.2 Comparisons and Discussion of the SCD2 Stage 1 Resulting Models

Table 6.3 presents the final results for the SCD2 stage1 experiments. These results reveal that:

- DT (70:30) model outperforms all others in terms of accuracy and precision (Non-defaulter). It also yields the lowest Type Π error rate.
- SVM (70:30) model outperforms all others in terms of precision (Defaulter).
- DT (10-cross) model yields the lowest Type I error rate but it achieves the rate of zero for precision (Defaulter).
- DT (70:30) and SVM (70:30) are the optimal models for this experiment. They are slightly different in accuracy, precision (Non-defaulter), Type I and Type Π error rates but they have variations in their percentages for precisions of Defaulter. The former one achieves 63.63% for the percentages (Defaulter) while the latter achieves 83.33%.
- Therefore, in this case the choice of the best model depends on the requirements of the bank.

Table 6.3 Results of Scoring Models for SCD2 Stage 1 Experiments

Techniques	VALIDATION			Accuracy%	Precision (%)		Type I error (%)	Type II error (%)
					Non-defaulter	Defaulter		
ANN	**Split validation**	**Training : Testing**	**70: 30**	79.65	80.9	61.11	2.46	17.89
			60: 40	78.68	80.86	53.33	3.68	17.63
	X-validation	**10 –cross**		78	80.29	48.48	3.58	18.31
SVM	**Split validation**	**Training : Testing**	**70:30**	79.65	79.57	**83.33**	0.35	20
			60:40	79.74	79.89	75	0.79	19.47
	X-validation	**10 –cross**		79.26	79.87	63.89	1.37	19.37
DT	**Split validation**	**Training : Testing**	**70:30**	**81.4**	**82.11**	63.64	1.4	**17.19**
			60:40	80	81.39	55	2.37	17.63
	X-validation	**10 –cross**		78.42	78.42	0	**0**	21.58

6.4.1.3Comparisons and Discussion of the German Stage 1 Resulting Models

Table 6.4 presents the final results for the German dataset stage1 experiments. These results reveal that:

- SVM (10-cross) model outperforms all others in terms of accuracy.
- ANN (70:30) model outperforms all others in terms precision (GOOD) and yields the lowest Type Π error rate.
- SVM (70:30) model outperforms all others in terms precision (BAD) and yields the lowest Type I error rate.

- In these experiments it is so difficult to select the best model. The optimal models are ANN (70:30), SVM (70:30) and SVM (10-cross). The weak point of the first model is that it yields lowest precision (BAD) than other two models. For the second one it yields the highest rate of Type Π error among others. For the latter model, it is slightly different from the second model in all measures and yields Type Π error rate slightly less than the second model and much more than the first model.

Table 6.4 Results of Scoring Models for the German Dataset Stage 1 Experiments

Techniques	**VALIDATION**			**Accuracy%**	**Precision (%)**		**Type I error (%)**	**Type II error (%)**
					GOOD	BAD		
ANN	**Split validation**	**Training : Testing**	**70: 30**	75.67	**81.57**	60.24	11	**13.33**
			60: 40	74	78.76	58.51	9.57	16.25
	X-validation	**10 –cross**		74.6	80.97	58.21	11.7	13.7
SVM	**Split validation**	**Training : Testing**	**70:30**	75	75.09	**74.19**	**2.67**	22.33
			60:40	74.75	75.64	68.63	4	21.25
	X-validation	**10 -cross**		**75.7**	76.17	72.44	3.5	20.8
DT	**Split validation**	**Training : Testing**	**70:30**	73	75.9	58.82	7	20
			60:40	68.5	71.47	55.41	8.25	23.25
	X-validation	**10 -cross**		70.7	75.79	51.66	10.2	19.1

6.4.2 Comparisons and Discussion of the Stage 2 Resulting Models

6.4.2.1Comparisons and Discussion of the SCD1 Stage 2 Resulting Models

Table 6.5 presents the final results for the SCD1 stage2 experiments. These results reveal that:

- GAANN(70:30) and GADT(70:30) models outperform all others in terms of accuracy.
- In addition GAANN(70:30) outperforms others in terms of precision(Defaulter) and yields the lowest Type I error rate.
- GADT (70:30) also outperforms all others in terms of precision (Non-defaulter) and yields the lowest Type Π error rates.
- GAANN(70:30) and GADT(70:30) models can be chosen to be the optimal models. Each model of these models outperforms the other in terms of some of the evaluation measures.

Table 6.5 Resulting Scoring Models for the SCD1 Stage 2 Experiments

Techniques	VALIDATION			Accuracy%	Precision (%)		Type I error (%)	Type II error (%)
					Non-defaulter	Defaulter		
GAANN	**Split validation**	**Training : Testing**	**70: 30**	**71.03**	69.29	**74.8**	**7.95**	21.03
			60: 40	69.62	68.68	71.51	9.42	20.96
	X-validation	**10 –cross**		67.31	66.9	68.15	10.46	22.23
GASVM	**Split validation**	**Training : Testing**	**70:30**	69.49	67.77	73.5	7.95	22.56
			60:40	69.81	69.72	70	9.23	20.69
	X-validation	**10 –cross**		67.46	66.63	69.29	9.62	22.92
GADT	**Split validation**	**Training : Testing**	**70:30**	**71.03**	**75.88**	65.97	16.67	**12.31**
			60:40	68.65	68.33	69.27	10.58	20.77
	X-validation	**10 –cross**		67.08	66.78	67.67	10.69	22.23

6.4.2.2Comparisons and Discussion of the SCD2 Stage 2 Resulting Models

Table 6.6 presents the final results for the SCD2 stage2 experiments. These results reveal that:

- GADT(70:30) model outperforms all others in terms of accuracy and precision (Non-defaulter) and yields the lowest Type Π error rate.
- GASVM (60:40) model outperforms all others in terms of precision (Defaulter) and yields the lowest Type I error rate.
- GADT (70:30) and GASVM (60:40) models can be chosen to be the optimal models. These models are not significantly different in all measures except for the precision (Defaulter). For this measure the latter one is substantially outperform the former.
- Therefore, in our opinion GASVM (60:40) can be chosen as the best model.

Table 6.6 Resulting Scoring Models for the SCD2 Stage 2 Experiments

Techniques	**VALIDATION**			**Accuracy%**	**Precision (%)**		**Type I error (%)**	**Type II error (%)**
					Non-defaulter	Defaulter		
GAANN	**Split validation**	**Training : Testing**	**70: 30**	84.21	84.84	62.5	1.05	14.74
			60: 40	82.89	84.2	46.15	1.84	15.26
	X-validation	**10 –cross**		80.21	80.53	73.17	1.16	18.63
GASVM	**Split validation**	**Training : Testing**	**70:30**	84.91	85.77	63.64	1.4	13.68
			60:40	84.21	84.18	**85.71**	**0.26**	15.53
	X-validation	**10 –cross**		80.21	80.07	85.19	0.42	19.37
GADT	**Split validation**	**Training : Testing**	**70:30**	**85.26**	**87.83**	54.55	3.51	**11.23**
			60:40	83.16	84.81	50	2.37	14.47
	X-validation	**10 –cross**		81.05	81.42	74.51	1.37	17.58

6.4.2.3Comparisons and Discussion of the German Dataset Stage 2 Resulting Models

Table 6.7 presents the final results for the German dataset stage 2 experiments. These results reveal that:

- GASVM (70:30) model outperforms all others in terms of accuracy.
- GASVM (10-cross) model outperforms all others in terms of precision (BAD) and achieves the lowest Type I error rate.
- GAANN (70:30) model outperforms all others in terms of precision (GOOD) and achieves the lowest Type Π error rate.
- The optimal models are GAANN (70:30), GASVM (70:30) and SVM (10-cross) models. The weak point of the first model is that it yields lowest precision (BAD) than other two models. For the latter one it yields the highest rate of Type Π rate among others. Therefore, the second model can be chosen the best model out of these models.

Table 6.7 Results of Scoring Models for the German Dataset Stage 2 Experiments

Techniques	VALIDATION			Accuracy%	Precision (%)		Type I error (%)	Type II error (%)
					GOOD	BAD		
GAANN	**Split validation**	**Training : Testing**	**70: 30**	80	**86.67**	60	10	**10**
			60: 40	79.5	83.07	65.43	7	13.5
	X-validation	**10 –cross**		77.7	81.01	66.67	7.7	14.6
GASVM	**Split validation**	**Training : Testing**	**70:30**	**81.33**	84.32	70.31	6.33	12.33
			60:40	80.25	80.53	78.69	3.25	16.5
	X-validation	**10 –cross**		77.6	77.05	**81.67**	**2.2**	20.2
GADT	**Split validation**	**Training : Testing**	**70:30**	79	83.12	63.49	7.67	13.33
			60:40	75.25	79.03	62.22	8.5	16.25
	X-validation	**10 –cross**		75	79.3	60.78	9.1	15.9

6.4.3Results of Comparisons for Stage 3 Models

6.4.3.1Results of Comparisons for SCD1 Stage 3 Models

Table 6.8 presents the final results for the SCD1 stage3 experiments. These results reveal that:

- ANN3 (60:40) model outperforms all others in terms of accuracy.
- SVM3 (60:40) model outperforms all others in terms of precision (Defaulter) and yields lowest Type I error rate.
- DT3 (10-cross) model outperforms all others in terms of precision (Non-defaulter) and yields the lowest Type Π error rate.
- In these experiments it is so difficult to select the best model. The candidates to be an optimal models are ANN3 (60:40), SVM (60:40) and DT (10-cross). The weak point of the first model is that it yields higher Type Π error rate. For the second one it yields highest rate of Type Π rate among others. For the latter mode, it yields higher Type I error rate among the first and the second model.

Table 6.8 Results of Scoring Models for the SCD1 Stage 3 Experiments

Techniques	VALIDATION			Accuracy %	Precision (%)		Type I error (%)	Type II error (%)
					Non-defaulter	Defaulter		
ANN3	Split validation	Training : Testing	70: 30	65.13	63.79	69	7.95	26.92
			60: 40	**65.19**	63.61	70.08	7.31	27.5
	X-validation	10 –cross		63.77	63.94	63.39	11.46	24.77
SVM3	Split validation	Training : Testing	70: 30	64.1	63.84	64.71	10.77	25.13
			60: 40	65	62.21	**79.07**	**3.46**	31.54
	X-validation	10 -cross		65.17	63.71	68.25	10.19	24.64
DT3	Split validation	Training : Testing	70: 30	61.79	63.56	58.74	15.13	23.08
			60: 40	63.08	65.02	59.9	15.19	21.73
	X-validation	10 -cross		63.23	**65.51**	59.81	16.08	**20.69**

6.4.3.2Results of Comparisons for SCD2 Stage 3 Models

Table 6.9 presents the final results for the SCD2 stage3 experiments. These results reveal that:

- DT3 (70:30) model outperforms all others in terms of accuracy and precision (Non-defaulter). It also yields the lowest Type Π error rate.
- SVM3(70:30) model outperforms all others in terms of precision (Defaulter) and yields lowest Type I error rate.
- These two models have been candidate to be the optimal models among other models in this stage.

Table 6.9 Results of Scoring Models for the SCD2 Stage 3 Experiments

Techniques	VALIDATION			Accuracy %	Precision (%)		Type I error (%)	Type II error (%)
					Non-defaulter	Defaulter		
ANN3	**Split validation**	**Training : Testing**	**70: 30**	78.95	79.85	58.33	1.75	19.3
			60: 40	78.68	79.83	55.56	2.11	19.21
	X-validation	**10 –cross**		78.74	79.43	58.06	1.37	19.9
SVM3	**Split validation**	**Training : Testing**	**70: 30**	80	79.64	**100**	**0**	20
			60: 40	79.74	79.89	75	0.79	19.47
	X-validation	**10 –cross**		79.37	79.89	65.71	1.26	19.37
DT3	**Split validation**	**Training : Testing**	**70: 30**	**82.46**	**83.27**	68.75	1.75	**15.79**
			60: 40	81.58	81.87	75	1.05	17.37
	X-validation	**10 –cross**		78.42	79.35	50	1.58	20

6.4.3.3Results of Comparisons for the German Dataset Stage 3 Models

Table 6.10 presents the final results for the German credit dataset stage3 experiments. These results reveal that:

- ANN3 (70:30) model outperforms all others in terms of accuracy and precision (GOOD). It also achieves the lowest Type II error rate.
- ANN3 (60:40) model outperforms all others in terms of precision (BAD).
- SVM3 (60:40) achieves the lowest Type I error rate.
- Among all these models of this stage, ANN3 (70:30) can be chosen to be the optimal one.

Table 6.10 Results of Scoring Models for the German Dataset Stage 3 Experiments

Techniques	VALIDATION			Accuracy %	Precision (%)		Type I error (%)	Type II error (%)
					GOOD	BAD		
ANN3	**Split validation**	**Training : Testing**	**70 : 30**	**76**	**80**	62.86	8.67	**15.33**
			60 : 40	75	78.04	**66.35**	8.75	16.25
	X-validation	**10 –cross**		74.7	79.52	59.67	9.8	15.5
SVM3	**Split validation**	**Training : Testing**	**70: 30**	75.33	77.42	65.38	6	18.67
			60: 40	75	76.63	66.13	**5.25**	19.75
	X-validation	**10 –cross**		75.7	77.7	66.29	5.9	18.4
DT3	**Split validation**	**Training : Testing**	**70: 30**	74.67	76.95	64.91	6.67	18.67
			60: 40	75	78.83	62.37	8.75	16.25
	X-validation	**10 –cross**		73.7	77.55	58.94	8.5	17.8

6.4.4Comparisons between Stages 1, 2 and 3 Experiments Resulting Models and Discussion

As a result of using GA in stage 2, tables of attributes and their weights were produced in the previous chapter. By using these tables new reduced sets of features were identified for each dataset. These new reduced datasets were employed in stage 3 experiments.

The reasons behind these comparisons is to answer the question: Is it efficient to develop the proposed CSMs by applying the identified single classification techniques to the original datasets, or to combine these techniques with GA as a feature selection technique or applying these single techniques to the reduced datasets (more details about these reduced datasets were presented in the previous chapter)?

Each dataset is considered separately in the following sections.

6.4.4.1Comparisons between Stages 1, 2 and 3 Experiments Resulting Models and Discussion for SCD1

The first part of this section conducts the comparisons of stage 2 models to stage 1 models as follows:

- In terms of accuracy, precision (Non-defaulter and Defaulter) and Type Π error, all models in stage 2 outperformed the corresponding models of stage 1. See figures 6.1, 6.2, 6.3 and 6.5. Among all models, DT models in stage 2 got the biggest improvements in terms of these performance measures.

- The highest accuracy for this dataset (71.03%) was achieved by the two models namely, GAANN (70:30) and GADT (70:30).
- The highest precision (Non-defaulter) for this dataset (75.88%) was achieved by the GADT (70:30) model.
- GAANN (70:30) ranked second. It yielded 74.80% for precision (Defaulter).
- GADT (70:30) outperformed all other models of all stages in terms of Type Π error. It achieved the lowest Type Π error rate which was 12.31%.
- Four models in stage 2 GAANN (70:30), GAANN (10-cross), GASVM (10-cross) and GADT (10-cross) got lower Type I error rates than the corresponding models in stage1. For the other five models of stage 2 Type I rates were higher (with different degrees) than those of the corresponding models of stage 1. Among all these models GADT (70:30) model of stage 2 got the maximum increase rate for this performance measure compared to the corresponding model of stage 1. (Type I error rates for DT (70:30) and GADT (70:30) are 4.36% and 16.67% respectively). The reason behind this increase in Type I rate is that, from the previous chapter (Tables 5.22 and 5.83) the Fneg of the DT model increased from 17 to 65 for GADT model (i.e. more than threetimes) which is compatible with the increase in the Type I error rate.

The second part of this section conducts the comparisons of stage 2 models to stage 3 models as follows:

- It is very clear from figures 6.1 and 6.2 that, the resulting models from stage 2 experiments outperform all other corresponding models of stage 3 in terms of accuracy, precision(Non-defaulter).
- Except for GASVM (60:40), all other models in stage 2 outperform other corresponding models of stage 3 in terms of precision (Defaulter). While SVM3 (60:40) achieved 79.07% for this measure indicator, GASVM (60:40) achieved 70%. See figure 6. 3.
- The highest precision (Defaulter) for this dataset was achieved by the SVM3 (60:40) model.
- In terms of Type I error the resulting models from stage 2 achieved irregular results (see figure 6.4) as follows:
  - Five models GAANN (10-cross), GASVM (70:30), GASVM (10-cross), GADT (60:40) and GADT (10-cross)) yielded the best results out of the corresponding models of stage 3.
  - One model (GAANN (70:30) yielded the same result as the corresponding model of stage 3.
  - Three models (GAANN (60:40), GASVM (60:40) and, GADT (70:30) and yielded the worst results out of the corresponding models of stage 3.
- Except for one model GADT (10-cross), all other models of stage 2 outperformed all corresponding models in stage 3 in terms of Type Π error measure. See figure 6.5.

The third part of this section conducts the comparisons of stage 3 models to stage 1 models as follows:

– Three models of stage 3 achieved slightly worse accuracy than the corresponding models of stage 1. For the other models two models (DT3 (60:40) and DT3 (10-cross)) achieved substantial improvement in accuracy over the corresponding models of stage 1. See figure 6.1.

– Five models of stage 3 achieved slightly worse precision (Non-defaulter) than the corresponding models of stage 1. For the other models only two models (DT3(60:40) and DT(10-cross)) achieved substantial improvement in precision(Non-defaulter) over the corresponding models in stage 1. See figure 6.2.

– Eight models of stage 3 obtained higher precision (Defaulter) than the corresponding models of stage 1. SVM3 (60:40), DT3 (60:40) and DT3 (10-cross) achieved substantial improvement in precision (Defaulter) over the corresponding models of stage 1. See figure 6.3.

– Five models in stage 3 obtained lower Type I error rates than the corresponding models of stage 1. While SVM3 (60:40) achieved substantial improvement in terms of this measure indicator, DT3 (70:30) obtained the maximum increase in Type I error rate out of the corresponding model in stage 1. See figure 6.4.

– Five models in stage 3 obtained lower Type Π error rates than the corresponding models of stage 1. While SVM3 (60:40) obtained the maximum increase in Type Π error rate out of the corresponding model of stage 1, DT3 (70:30) and DT3 (60:40) achieved substantial improvement in terms of this measure indicator. See figure 6.5.

The concluding remarks for these comparisons are as follows:

1. From the first and second parts of these comparisons, it is clear that applying GA as a feature selection technique to single techniques in stage 1 leads to:

- Superiority of the all models of stage 2 in terms of accuracy and precision (Non-defaulter) to other resulting models from stage 1 and stage 3.
- Superiority of the eight models of stage 2 in terms of precision (Defaulter) and Type Π error to other resulting models from stage 1 and stage 3.
- Superiority of the three models of stage 2 in terms of Type I error to other resulting models from stage 1 and stage 3. One models of stage 2 achieved the same Type I error rate as the corresponding model of stage 3and at the same time lower than the corresponding model of stage 1.
- Applying GA has brought a substantial improvement for DT models (DT (70:30), DT (60:40) and DT (10-cross)) in terms of four measure indicators. These measures are accuracy, precision (Non-defaulter), precision (Defaulter) and Type Π error.
- These results reveal that combining GA with the ANN, SVM and DT is more beneficial than applying these techniques individually for this dataset.

2. The concluding remarks from the third part of these comparisons are as follows :

- Superiority of the six models of stage 3 in terms of accuracy to other resulting (corresponding) models from stage 1. (These six models

were outperformed by the corresponding models of stage 2 in terms of accuracy).

- Superiority of the four models of stage 3 in terms of precision (Non-defaulter) to other resulting (corresponding) models from stage 1. (These four models of stage 3 were outperformed by the corresponding models of stage 2 in terms of precision (Non-defaulter))
- Superiority of the eight models of stage 3 in terms of precision (Defaulter) to other resulting (corresponding) models from stage 1. (Only one model (SVM3 (60:40)) among these eight models got higher precision (Defaulter) than the corresponding model in stage 2).
- Superiority of the five models of stage 3 in terms of Type I error to other resulting (corresponding) models from stage 1. (Two models among these five models outperformed by the corresponding models of stage 2).
- Superiority of the five models of stage 3 in terms of Type Π error to other resulting (corresponding) models from stage 1. (four models among these five models outperformed by the corresponding models of stage 2).
- In terms of accuracy, precision (Non-defaulter) and precision (Defaulter) DT3 (60:40) and DT3 (10-cross) models got substantial improvement than the corresponding models of stage 1. In addition DT3 (70:30) and DT3 (60:40) got substantial improvement than the corresponding models in stage 1 in terms of Type Π error. SVM3 (60:40) got substantial improvement than the corresponding models in stage 1 in terms of precision (Defaulter) and Type I error.

– The improvements in some of the models of stage 3 were only slight improvements in most cases. See figures 6.1-6.5.

– It is notably that SVM3 (60:40)model is the only one model in stage3 which can compete with the models of stage 2.

3. Final concluding remarks for all comparisons for all experiments of all stages :

– These experiments indicate that combining GA to the individual techniques and applying these hybrid models to the original dataset is better than applying individual techniques to the reduced dataset.

– GAANN (70:30) and GADT (70:30) outperformed all other models of all stages it terms of accuracy.

– GADT (70:30) also outperformed all other models of all stages it terms of precision (Non-defaulter).

– SVM3 (60:40) outperformed all other models of all stages it terms of precision (Defaulter) and Type I error. The weakness of this model is that, it obtained the higher Type II error rate than GAANN (70:30) and GADT (70:30).

– Under the theme that, there is no free lunch (each model has its weaknesses and strong points) we can choose GAANN (70:30), GADT (70:30) and SVM3 (60:40) as the optimal models for the SCD1.

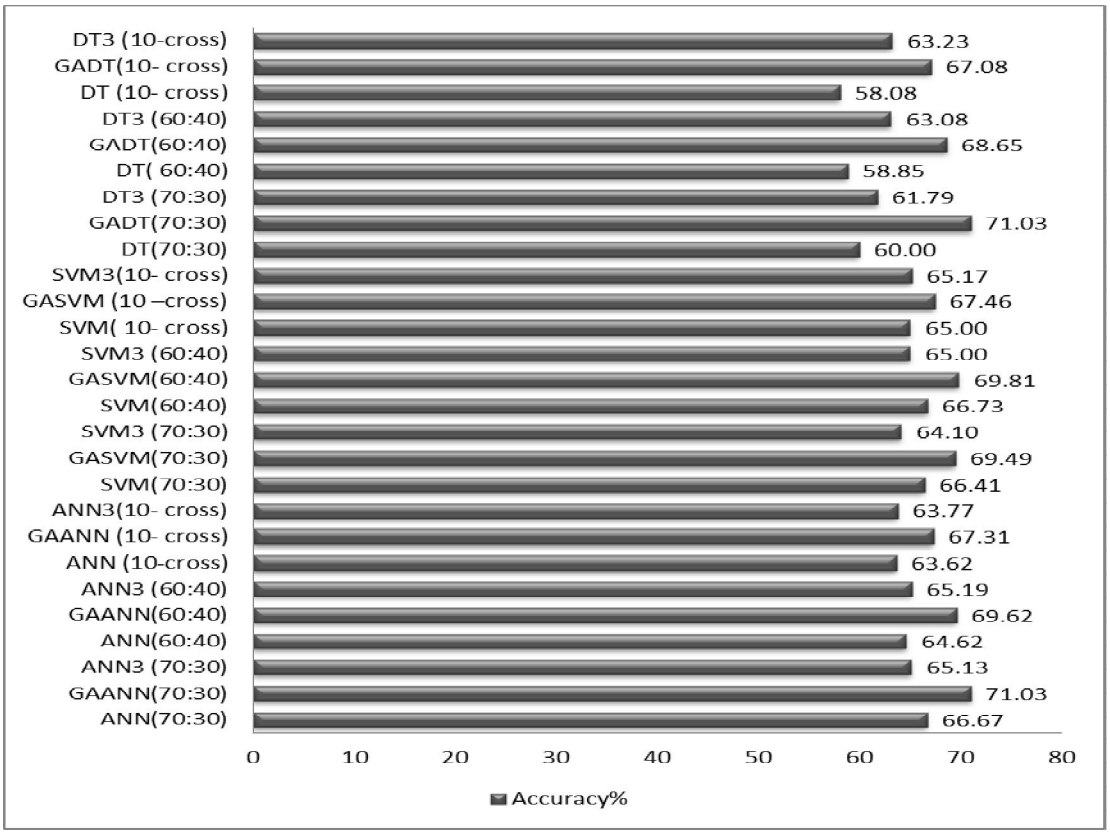

Figure 6.1 Results of Accuracy for the SCD1 Models of Stages 1, 2 &3

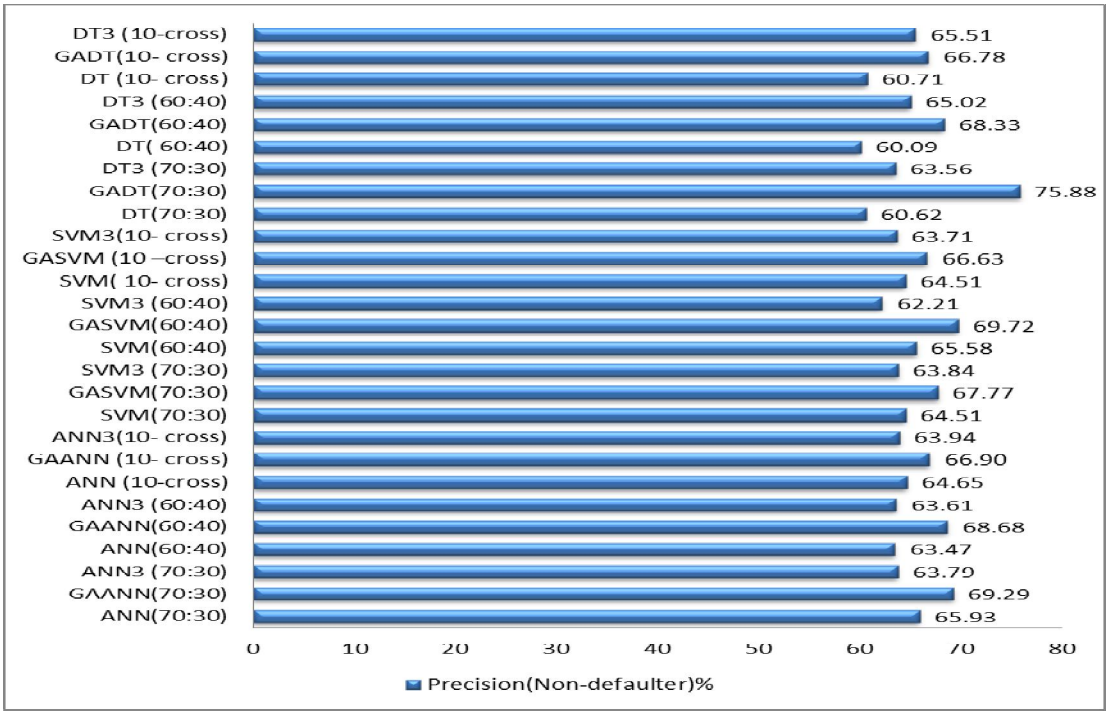

Figure 6.2 Results of Precision (Non-defaulter) for the SCD1 Models of Stages 1, 2 &3

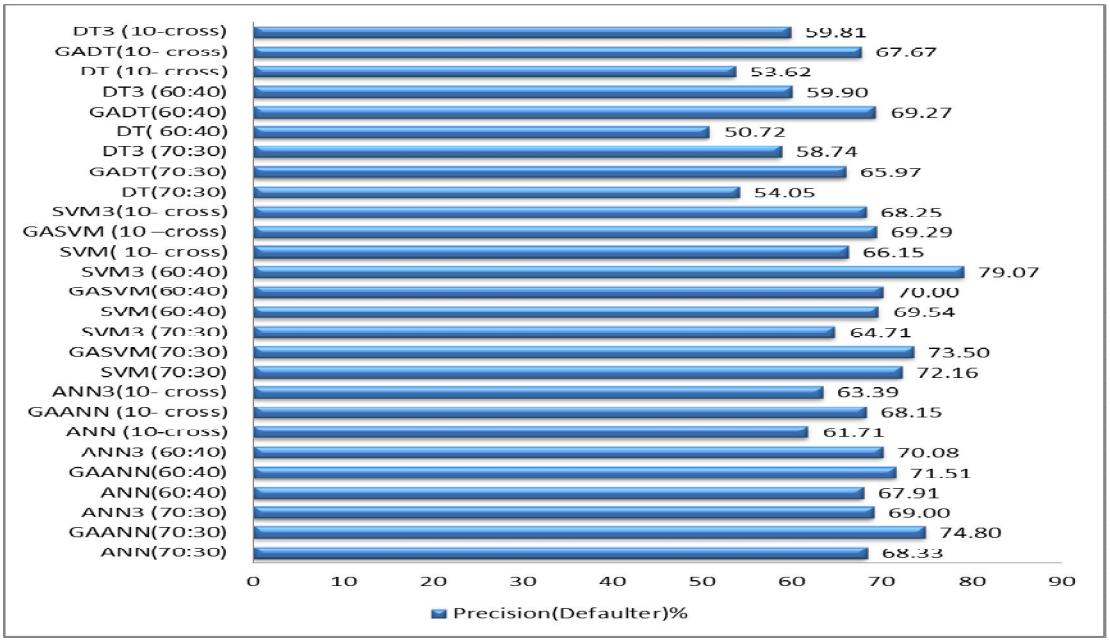

Figure 6.3 Results of Precision (Defaulter) for the SCD1 models of stages 1, 2 &3

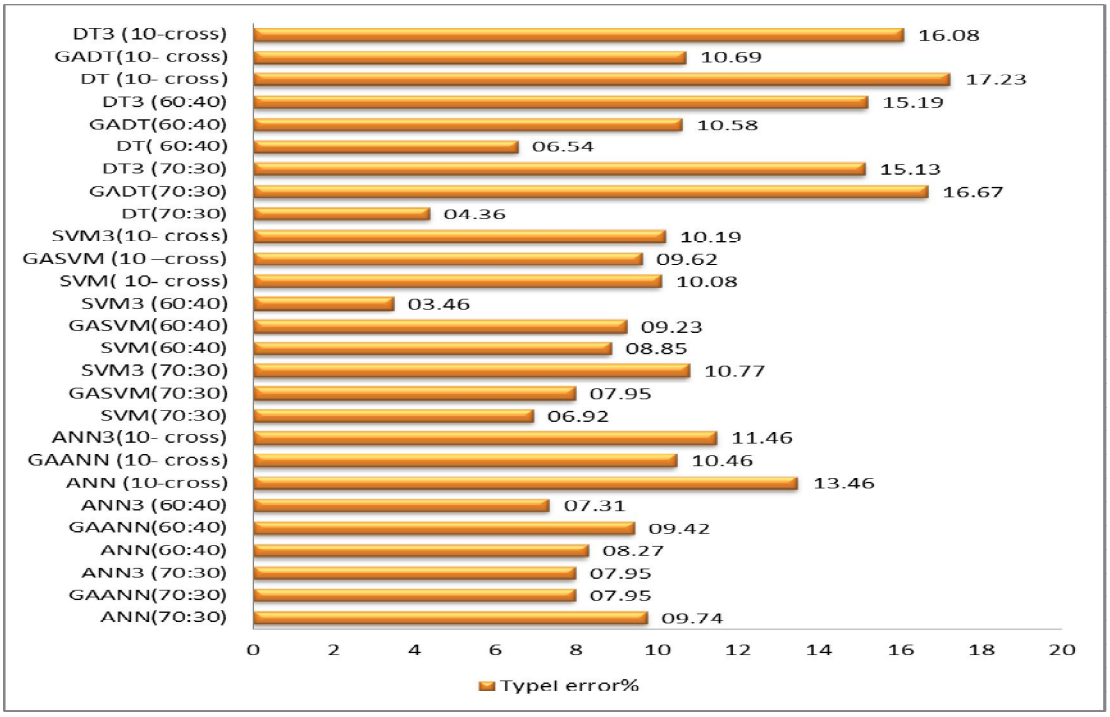

Figure 6.4 Results of Type I Error for the SCD1 models of stages 1, 2 &3

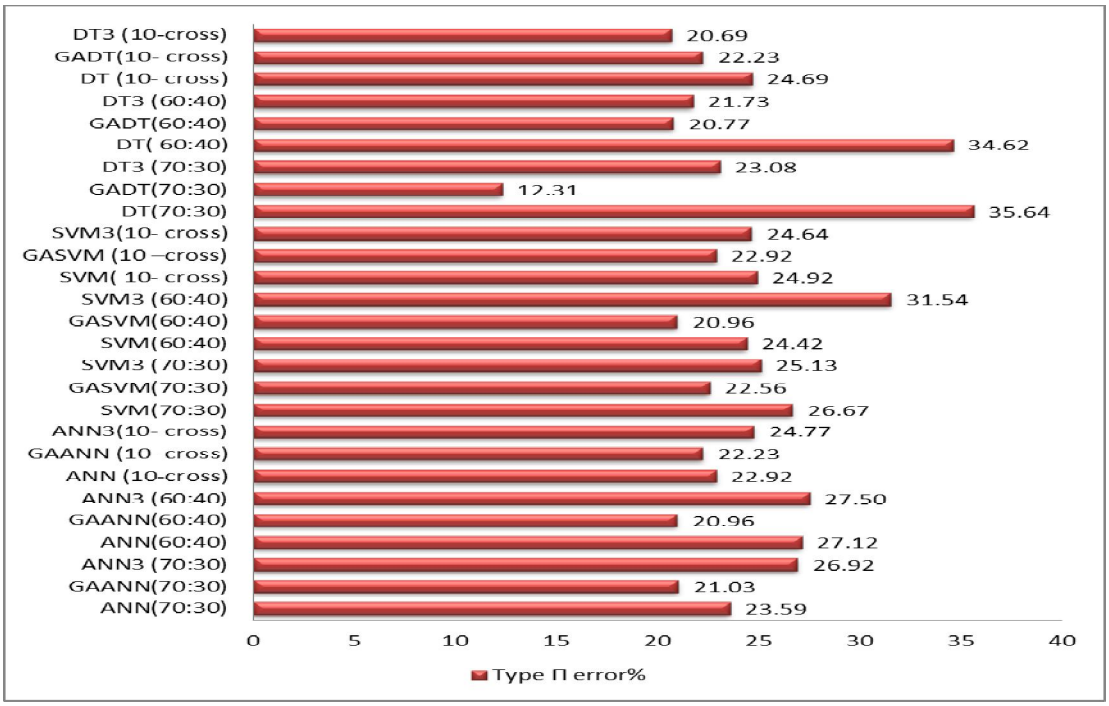

Figure 6.5 Results of Type II Error for the SCD1 models of stages 1, 2 &3

6.4.4.2Comparisons between Stages 1, 2 and 3 Experiments Resulting Models and Discussion for SCD2

The first part of this section conducts the comparisons of stage 2 models to the stage 1 models as follows:

- In terms of Accuracy and precision (Non-defaulter), all models in the stage 2 outperformed all corresponding models of stage 1. See figures 6.6 & 6.7.
- Except for four models GAANN (60:40), GASVM(70:30), GADT (70:30) and GADT (60:40), all other models in stage 2 outperformed the corresponding models of stage 1 in terms of Precisions (Defaulter).DT (10-cross) model got a substantial improvement in

terms of this measure indicator (increasing in the precision (Defaulter) from 0% to 74.51% in the GADT (10-cross) model). See figure 6.8.

- Except for GASVM (70:30), GADT (10-cross), GADT (70:30) and GADT (60:40) models, all other models in the stage 2 got lower Type I error rates than the corresponding models in stage 1. For the first two models Type I error slightly increased in stage 2 corresponding models while for the last one the Type I error rate remains as it is in stage 2 corresponding model. See figure 6.9.
- DT (10-cross) obtained the lowest Type I error rate (0.0%) among all other models in three stages.
- Except for two models GAANN (10-cross) and GASVM (10-cross), all other models in the stage 2 got lower Type Π error rates than the corresponding models in stage 1. For GAANN (10-cross) model in stage2 Type Π error increased (slightly) over stage 1 corresponding model and for GASVM (10-cross) Type Π error rate remained as it is in the stage 1 corresponding model. See figure 6.10.

The second part of this section conducts the comparisons of stage 2 models to the stage 3 as follows:

- Figures 6.6 and 6.7 reveal that that, the resulting models from stage 2 experiments outperform all other corresponding models in stage 3 in terms of accuracy and precision(Non-defaulter).
- The highest accuracy for the SCD2 (85.26%) and the highest precision (Non-defaulter) (87.83%) were achieved by GADT (70:30) model.
- While five resulting models from stage 2 experiments outperformed the corresponding models of stage 3 in terms precision (Defaulter), the other

four models achieved the worst results in the corresponding models of stage 3. See figure 6.8.

- GADT(10-cross) achieved substantial improvement in precision (Defaulter) over the corresponding models in stage 1. The precision (Defaulter) for DT(10-cross) and GADT(10-cross) is 0% and 74.51 % respectively. See figure 6.8.
    - GAANN (70:30), GAANN (60:40), GAANN (10-cross), GASVM (60:40), GASVM (10-cross) and GADT (10-cross) outperformed the corresponding models in stage 3 in terms Type I error. Three models of stage 2 ((GASVM (70:30), GADT (70:30) and GADT (60:40)) achieved the worst results in terms of this measure indicator out of the corresponding models in stage 3. See figure 6.9.
    - SVM3 (70:30) obtained the lowest Type I error rate (0.0%) among all other models in three stages.
    - Seven resulting models from stage 2 experiments outperformed the corresponding models in stage1 and stage 3 in terms Type Π error. GAANN (10-cross) was slightly worse than ANN3 (10-cross) in terms of this measure indicator. GASVM (10-cross) and SVM3 (10-cross) obtained the same Type Π error rates. See figure 6.10.
    - GADT (70:30) got the lowest Type Π error rate among all model of all stages.

In the third part of these comparisons when the stage 3 resulting models were compared with stage 1 models , the following was observed :

– Five models in stage 3 achieved slightly better accuracy than the corresponding models in stage 1. Only one model ANN3 (70:30) achieved slightly worse accuracy over the corresponding models in

stage 1. The accuracies for the other three models in the stage 3 and the corresponding models in the stage 1 were equal. See figure 6.6.

– Four models in stage 3 achieved slightly worse precision (Non-defaulter) than the corresponding models in stage 1. Four models in stage 3 achieved a slight improvement over corresponding models in stage 1. One model achieved the same results as the corresponding model of stage 1. See figure 6.7.

– Seven models in stage 3 obtained higher precision (Defaulter) than the corresponding models of stage 1. While ANN3 (70:30) model achieved lower precision (Defaulter) than the corresponding models of stage 1, SVM3 (60:40) model achieved the same precision (Defaulter) rate as the corresponding models in stage 1. See figure 6.8.

– Seven models in stage 3 obtained lower Type I error rates than the corresponding models in stage 1. While DT3 (70:30) was slightly worse than DT (70:30) in terms of this measure indicator, SVM3 (60:40) achieved the same result as SVM (60:40).

– SVM3 (70:30) and DT (10-cross) models achieved the lowest TypeI error rates (0.0%) among all models. See figure 6.9.

– Three models of stage 3 obtained higher Type Π error rates than the corresponding models of stage 1. While DT3 (70:30), DT3 (60:40) and DT3 (10-cross) were slightly better than the corresponding models of stage 1 in terms of Type Π error rate, the other three models achieved the same Type Π error rates as the corresponding models of stage 1. See figure 6.10.

The concluding remarks for these comparisons are as follows:

1. From the first and second parts of these comparisons, it is clear that applying GA as a feature selection technique to single techniques in stage 1 leads to:
   - Superiority of the all models of stage 2 in terms of accuracy and precision (Non-defaulter) to other resulting models from stage 1 and stage 3.
   - Superiority of the five models of stage 2 in terms of precision (Defaulter) and Type I error to other resulting models from stage 1 and stage 3.
   - Superiority of the seven models of stage 2 in terms of Type Π error to other resulting models from stage 1 and stage 3.
   - These results reveal that combining GA with the ANN, SVM and DT is more beneficial than applying these techniques individually for this dataset.
2. The concluding remarks from the third part of these comparisons are as follows :

– Superiority of the five models of stage 3 in terms of accuracy to other resulting (corresponding) models from stage 1 (slight improvement). (These five models were outperformed by the corresponding models of stage 2 in terms of accuracy).

– Superiority of the four models of stage 3 in terms of precision (Non-defaulter) to other resulting (corresponding) models from stage 1. Two models of stage 3 got the same result as the corresponding models of stage 1 in terms of this measure indicator. (These six

models of stage 3 were outperformed by the corresponding models of stage 2 in terms of precision (Non-defaulter)).

- Superiority of the seven models of stage 3 in terms of precision (Defaulter) to other resulting (corresponding) models from stage 1. (Four models ANN3 (60:40), SVM3 (70:30), DT3 (70:30) and DT3 (60:40) among these seven models got higher precision (Defaulter) than the corresponding models in stage 2).
- Superiority of the six models of stage 3 in terms of Type I error to other resulting (corresponding) models from stage 1. (Only one model (SVM3(70:30)) among these four models were better than the corresponding model of stage 2).
- Superiority of the three models of stage 3 in terms of Type Π error to other resulting (corresponding) models from stage 1. (These three models were outperformed by the corresponding models of stage 2).
- In terms of precision (Defaulter), SVM3 (70:30), DT3 (60:40) and DT3 (10-cross) models got substantial improvement than the corresponding models of stage 1.
- The improvements in some of the models of stage 3 were only slight improvements in most cases. See figures 6.7-6.10.
- It is notable that SVM3 (70:30)model is the only one model in stage3 which can compete with the models in stage2.

3. Final concluding remarks for all comparisons for all experiments of all stages :

- These experiments indicate that combining GA to the individual techniques and applying these hybrid models to the original dataset is better than applying individual techniques to the reduced dataset.

– GADT (70:30) outperformed all other models of all stages it terms of accuracy, precision (Non-defaulter) and Type Π error. GASVM (70:30) ranked at second in terms of these measure indicators.

– SVM3 (70:30) outperformed all other models of all stages in terms of precision (Defaulter) and Type I error. The weak point of this model is that, it obtained the higher Type Π error rate than GADT (70:30) and GASVM(70:30).

– Therefore, GADT (70:30) is strongly recommended to be the optimal model for SCD2. GASVM (70:30) and SVM3 (70:30) can be also chosen as the second optimal models for this dataset.

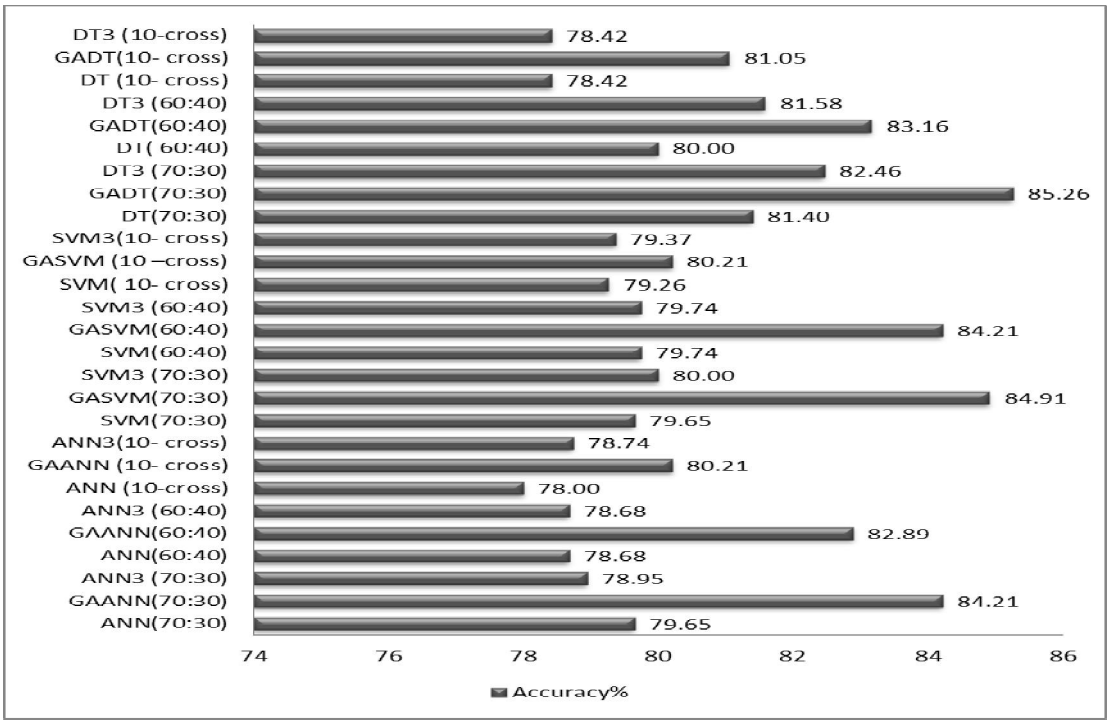

Figure 6.6 Results of Accuracy for the SCD2 models of stages 1, 2 &3

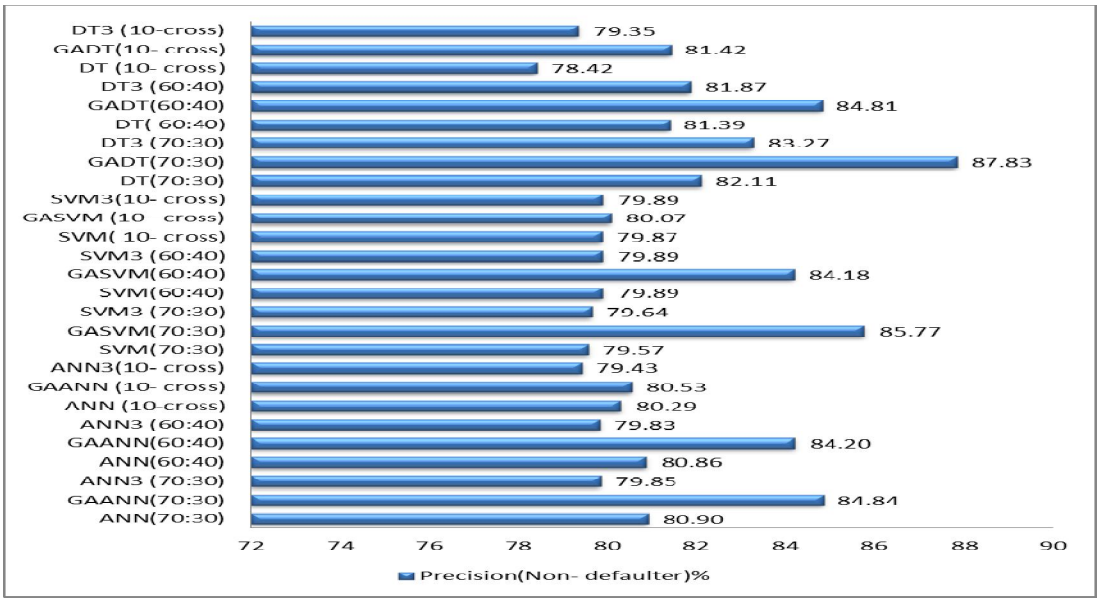

Figure 6.7 Results of Precision (Non-defaulter) for the SCD2Models of Stages 1, 2 &3

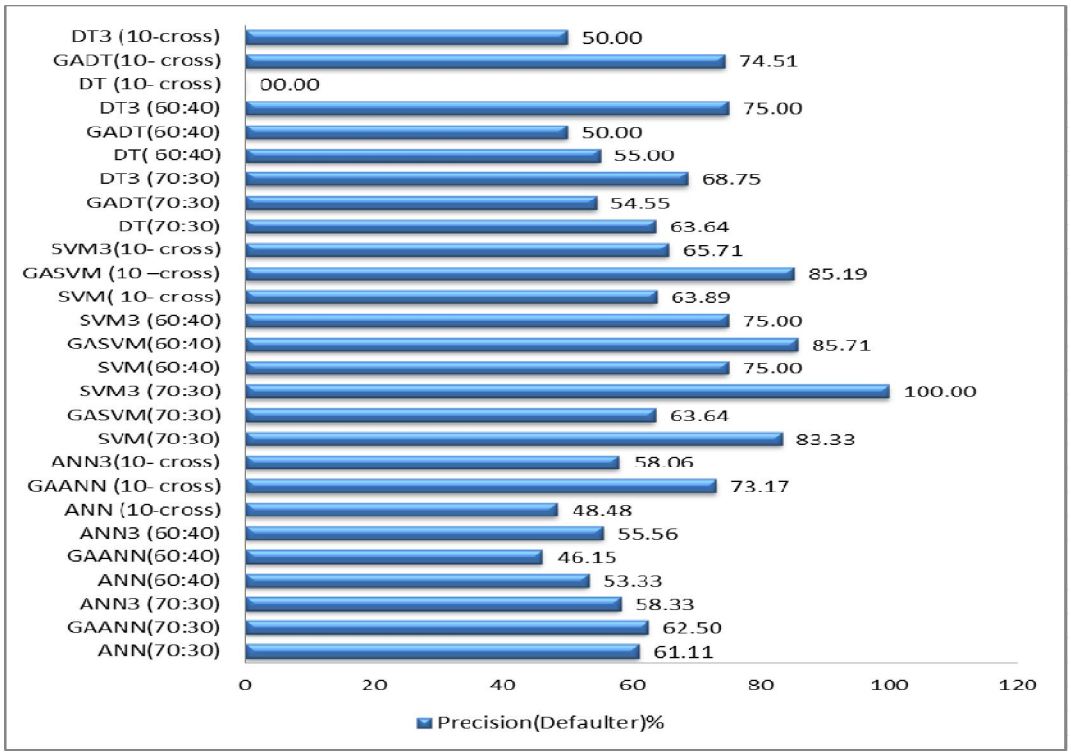

Figure 6.8 Results of Precision (defaulter) for the SCD2Models of Stages 1, 2 &3

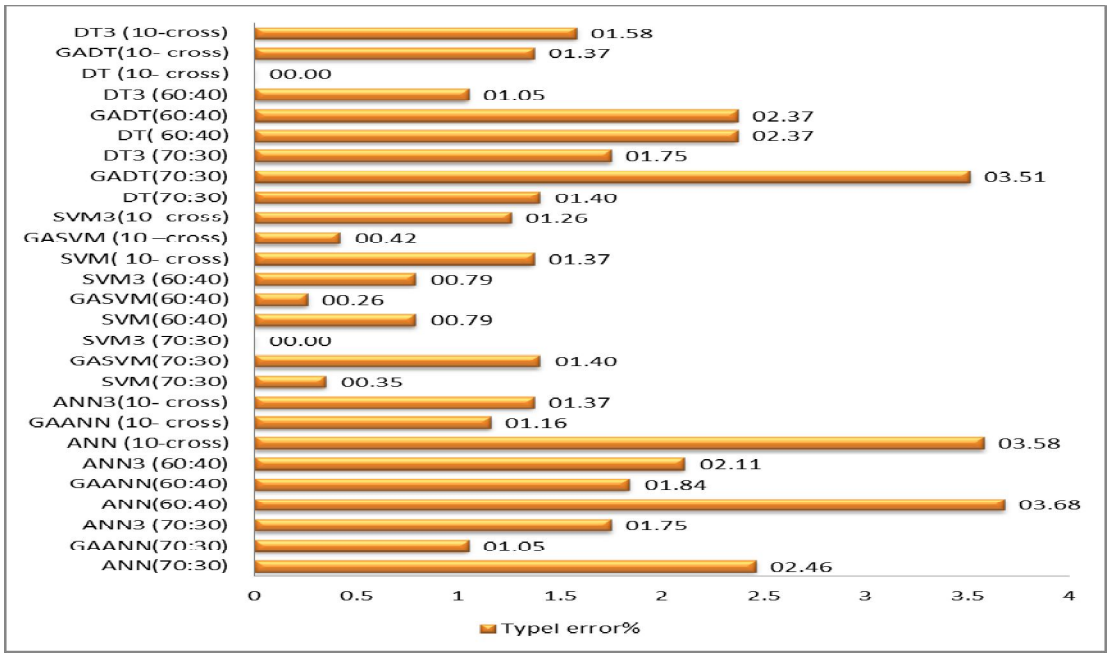

Figure 6.9 Results of Type I Error for the SCD2Models of Stages 1, 2 &3

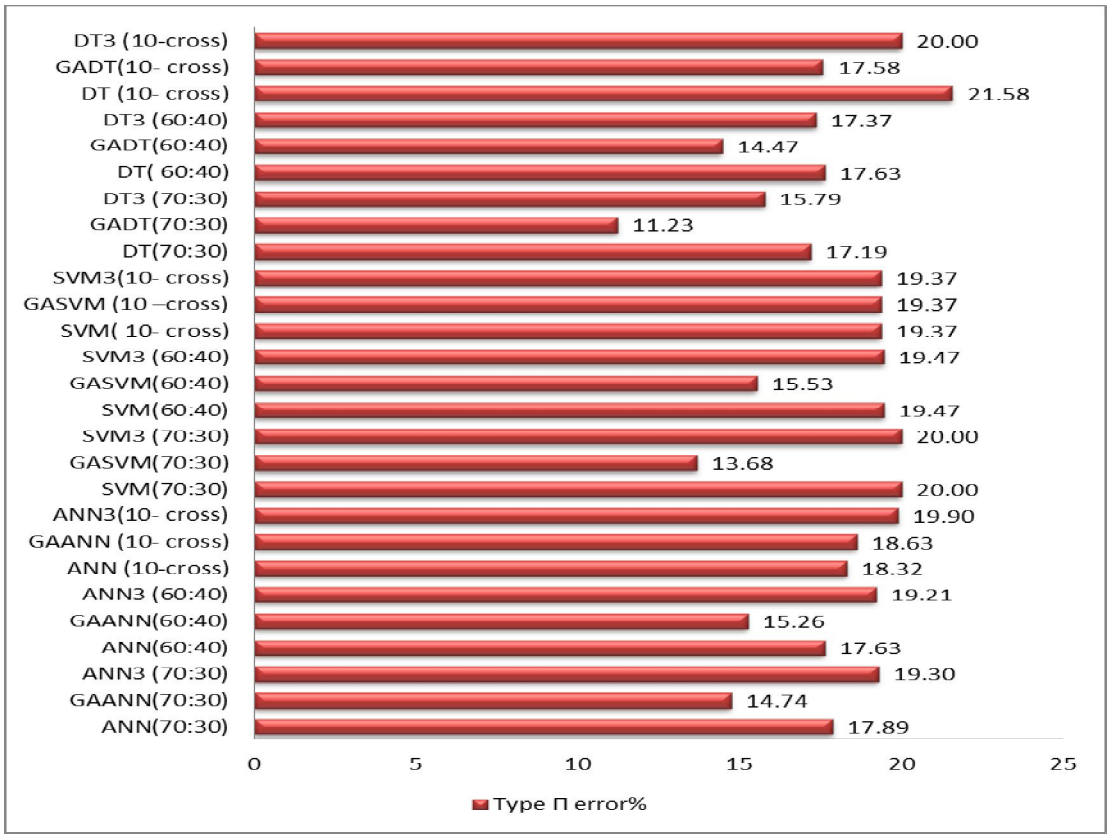

Figure 6.10 Results of Type II Error for the SCD2Models of Stages 1, 2 &3

6.4.4.3Comparisons between Stages 1, 2 and 3 Experiments Resulting Models and Discussion for the German Dataset

The first part of this section conducts the comparisons of stage 2 models to the stage 1 models as follows:

- Like the two Sudanese credit datasets in terms of accuracy and precision (GOOD), all models of stage 2 outperformed all corresponding models of stage 1. See figures 6.11and 6.12.
- Except for GAANN (70:30) and GASVM (70:30), the remainder models of stage 2 got better Precision (BAD) than corresponding models of stage 1. See figure 6.13. GAANN (70:30) was slightly worse than ANN (70:30).
- Except for GASVM (70:30), GADT (70:30) and GADT (60:40) models (these models are arranged according to the increase in the Type I error rate), the rest of the models of stage2 got lower Type I error rates than the corresponding models of stage 1. See figure 6.14.
    - Except for GAANN (10-cross), the remainder models of stage 2 got lower Type Π error rates than the corresponding models in stage 1. GAANN (10-cross) is slightly worse than ANN (10-cross) in terms of this measure. See figure 6.15.

The second part of this section conducts the comparisons of stage 2 models to the stages 3 as follows:

- The resulting models from stage 2 experiments for the German dataset outperformed all other corresponding models of stage 3 in terms of accuracy. See figures 6.11.
- Except for GASVM (10-cross), all other models of stage 2 outperformed the corresponding models of stage 3 in terms precision (GOOD). See figures 6.12.
- In terms of precision (BAD), it was observed that (see figure 6.13):
  - Five models(GAANN (10-cross), GASVM (70:30), GASVM (60:40), GASVM (10-cross) and GADT (10-cross)) of stage2 outperform all other corresponding models of stage 3.
  - Four modelsGAANN (70:30), (GAANN (60:40), GADT(70:30) and GADT(60:40)) of stage 2 were slightly worse than the corresponding models of stage 3.
- In terms of Type I error, it was observed that (see figure 6.14):
  - Five models GAANN (60:40), GAANN (10-cross), GASVM(60:40) , GASVM(10-cross)), (GADT (60:40)) of stage2 outperformed all other corresponding models of stage 3.
  - Four models GAANN (70:30), GASVM (70:30), GADT (70:30) and GADT (10-cross) of stage 2 were worse than the corresponding models of stage 3.
- In terms of Type Π error it was observed that (see figure 6.15):
  - Seven models ((GAANN (70:30), GAANN (60:40), GAANN (10-cross), GASVM (70:30), GASVM (60:40),

GADT(70:30) and GADT(10-cross)) of stage2 outperform all other corresponding models of stage 3.

- One model GASVM (10-cross) in stage 2 got worse than the corresponding model in stage 3.
- One model (GADT (60:40)) in stage 2 obtained the same Type Π error rate as the corresponding model in stage 3.

The third part of this section conducts the comparisons of stage 1 models to the stages 3 as follows:

- Eight resulting models from stage 3 experiments for this dataset (all three ANN , all three DT and two SVM models) outperformed all other corresponding models in stage1 in terms of accuracy. Only one model in stage 3 (SVM3(10-cross) achieved the same accuracy as the corresponding model in stage 1. See figure 6.11.
- In terms of precision (GOOD), SVM3 (70:30), SVM3 (60:40), SVM3 (10-cross), DT3(70:30), DT3(60:40) and DT3(10-cross) in stage 3 outperformed the corresponding models in stage1. All ANN models (ANN3 (70:30), ANN3 (60:40) and ANN3 (10-cross)) in stage 3 were slightly worse than the corresponding models in stage1. See figure 6.12.
- In terms of precision (BAD), all ANN models (ANN3 (70:30), ANN3(60:40) and ANN3(10-cross)) and all DT models (DT3(70:30), DT3(60:40) and DT3(10-cross)) of stage 3 outperformed the corresponding models of stage1. All SVM models (SVM3 (70:30), SVM3 (60:40) and SVM3

(10-cross)) of stage 3 were worse than the corresponding models of stage1. See figure 6.13.

- In terms of Type I error all ANN models (ANN3 (70:30), ANN3(60:40) and ANN3(10-cross)) and two DT models (DT3(70:30) and DT3(10-cross)) of stage 3 were better than the corresponding models in stage1. DT3 (60:40) and all SVM models (SVM3 (70:30), SVM3 (60:40) and SVM3 (10-cross)) in stage 3 were worse than the corresponding models in stage1. See figure 6.14.
- In terms of Type Π error, all SVM models (SVM3 (70:30), SVM3(60:40) and SVM3(10-cross)) and all DT models (DT3(70:30), DT3(60:40) and DT3(10-cross)) in stage 3 outperformed the corresponding models in stage1. While two ANN models (ANN3 (70:30) and ANN3 (10-cross)) were worse than the corresponding models in stage1, one ANN model (ANN3 (60:40)) obtained the same Type Π error rate as the corresponding model in stage 1. See figure 6.15.

The concluding remarks for these comparisons are as follows:

1. From the first and second parts of these comparisons, it is clear that applying GA as a feature selection technique to single techniques in stage 1 leads to:
    - Superiority of the all models of stage 2 in terms of accuracy to other resulting models from stage 1 and stage 3.

- Superiority of the seven models of stage 2 in terms of precision (GOOD) to other resulting models from stage 1 and stage 3.
- Superiority of the four models of stage 2 in terms of precision (BAD) and Type I error to other resulting models from stage 1 and stage 3.
- Superiority of the six models of stage 2 in terms of Type Π error to other resulting models from stage 1 and stage 3. .
- These results reveal that combining GA with the ANN, SVM and DT is more beneficial than applying these techniques individually for this dataset.

2. The concluding remarks from the third part of these comparisons are as follows :

– Superiority of the eight models of stage 3 in terms of accuracy to other resulting (corresponding) models from stage 1(slightly improvement). (These eight models were outperformed by the corresponding models of stage 2 in terms of accuracy).

– Superiority of the six models of stage 3 in terms of precision (GOOD) to other resulting (corresponding) models from stage 1. (These six models were outperformed by the corresponding models of stage 2 in terms of precision (GOOD)).

– Superiority of the six models of stage 3 in terms of precision (BAD) to other resulting (corresponding) models from stage 1. (Four models (ANN3 (70:30), ANN3 (60:40), DT3 (70:30), DT3 (60:40) and DT3(60:40)) among these six models got higher precision (BAD) than the corresponding model in stage 2).

- Superiority of the five models of stage 3 in terms of Type I error to other resulting (corresponding) models from stage 1. (Three models ANN3 (70:30), DT (70:30) and DT (10-cross) among these four models were better than the corresponding models of stage 2).
- Superiority of six models of stage 3 in terms of Type Π error to other resulting (corresponding) models from stage 1. (Five models among these models were outperformed by the corresponding models of stage 2 and one model achieved the same Type Π rate as the corresponding model of stage 2).
- In terms of all measures, the improvements in the models of stage 3 were only slight improvements in most cases. See figures 6.7-6.10.
- It is notably that there is no model in stage3 which can compete with the models in stage2.

3. Final concluding remarks for all comparisons for all experiments of all stages :

- These experiments indicate that combining GA to the individual techniques and applying these hybrid models to the original dataset is better than applying individual techniques to the reduced dataset.
- GASVM (70:30) outperformed all other models of all stages it terms of accuracy. GASVM (60:40) ranked at second in terms of this measure indicator.
- GAANN(70:30) model outperformed all other models of all stages it terms of precision (GOOD). GASVM (70:30) ranked at second in terms of this measure indicator.

– GASVM (10-cross) outperformed all other models of all stages it terms of precision (BAD). GASVM (60:40) ranked at second in terms of this measure indicator.

– GASVM (10-cross) outperformed all other models of all stages it terms of Type I error. SVM (70:30) ranked at second in terms of these measure indicator.

– GAANN (70:30) outperformed all other models of all stages it terms of Type Π error. GASVM (70:30) ranked at second in terms of these measure indicators.

– GAANN (70:30), GASVM (70:30), GASVM (60:40) can be chosen to be the optimal models for the German credit dataset.

– It is notable to mention that, each model of these models showed superiority in terms of some measure indicators but at the same time they slightly declined in terms of other indicators.

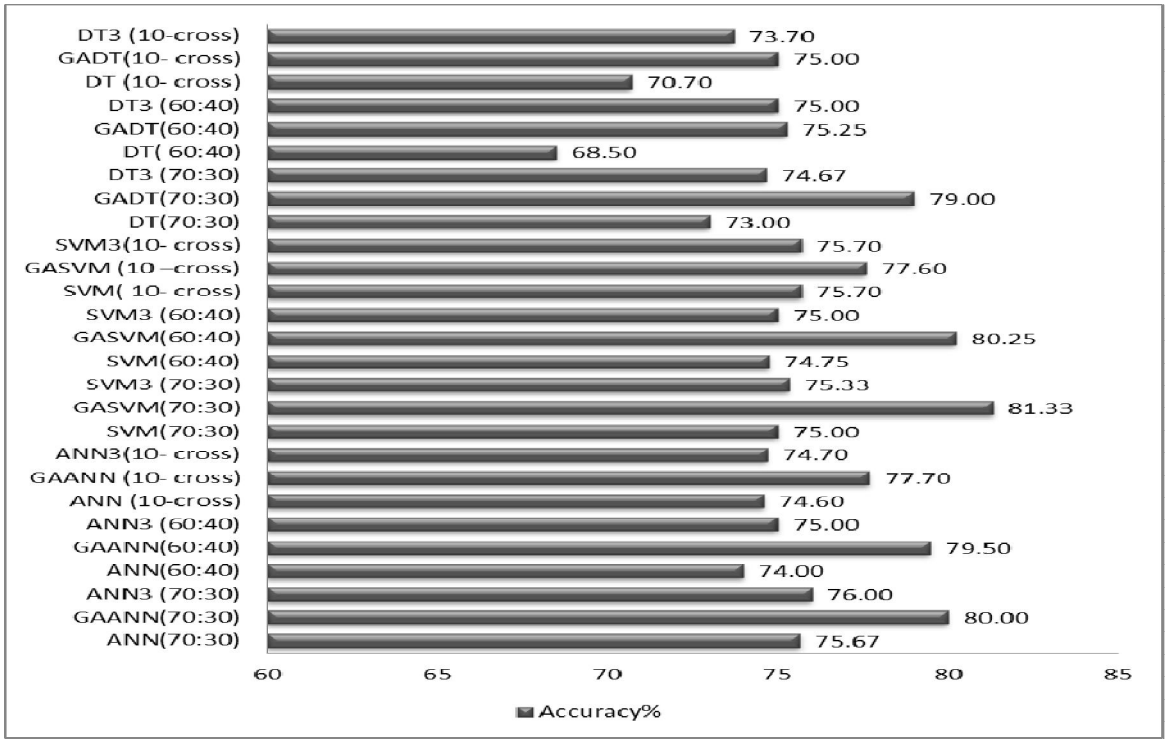

Figure 6.11 Results of Accuracy for the German Dataset Models of Stage 1, 2 &3

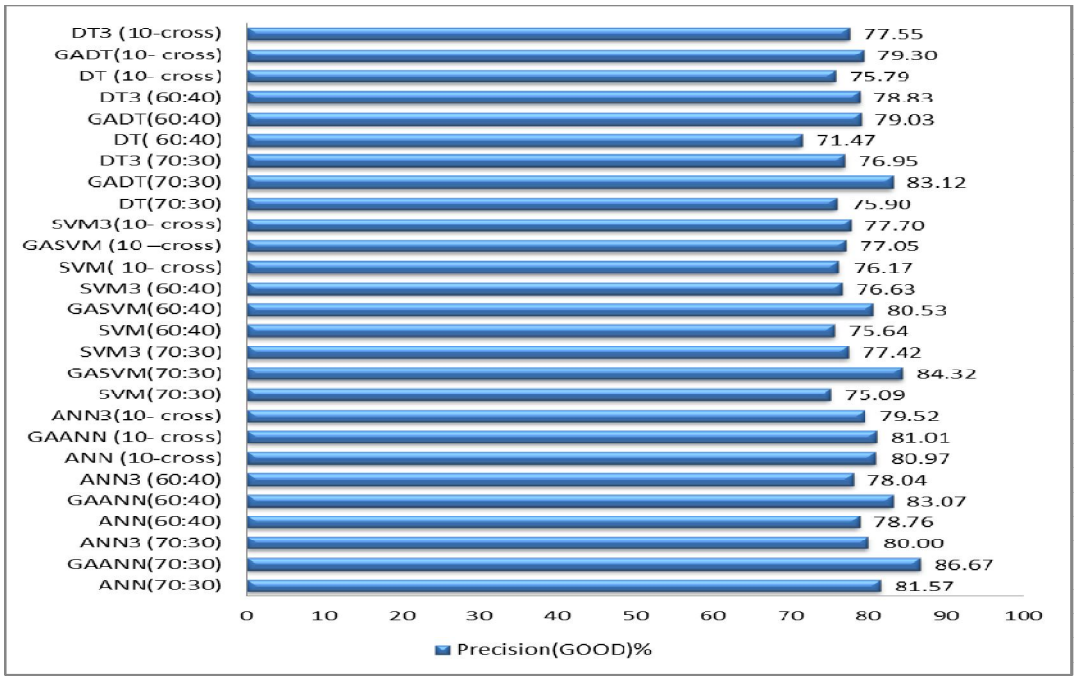

Figure 6.12 Results of Precision (GOOD) for the German Dataset Models of Stages 1, 2&3

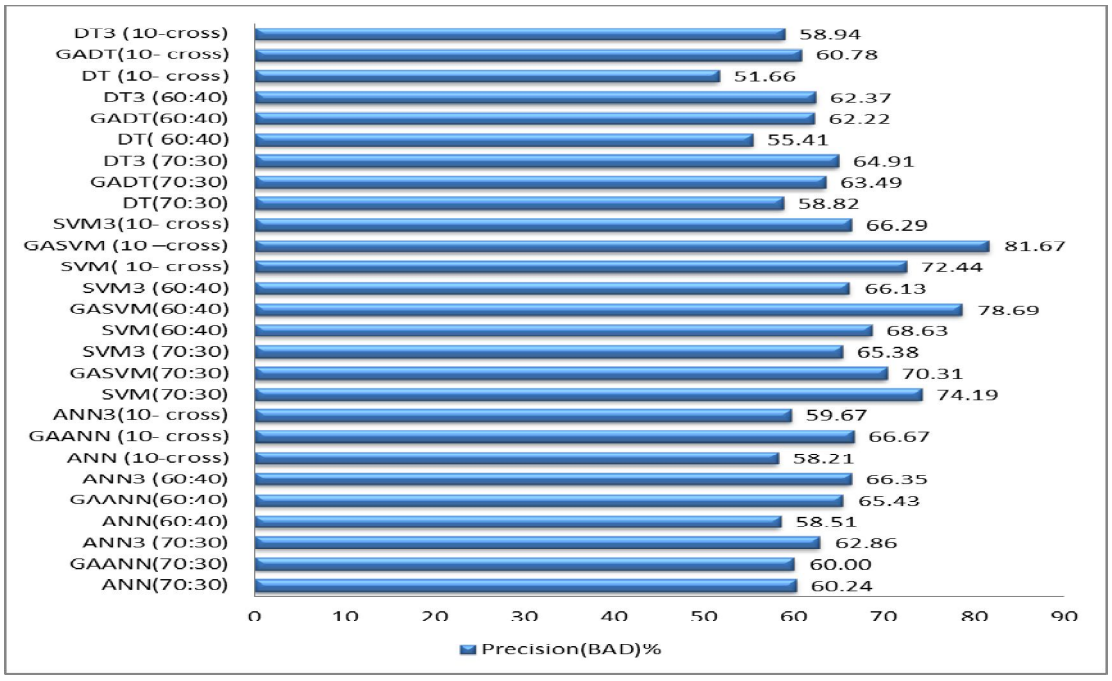

Figure 6.13 Results of Precision (BAD) for the German Dataset Models of Stages 1, 2 &3

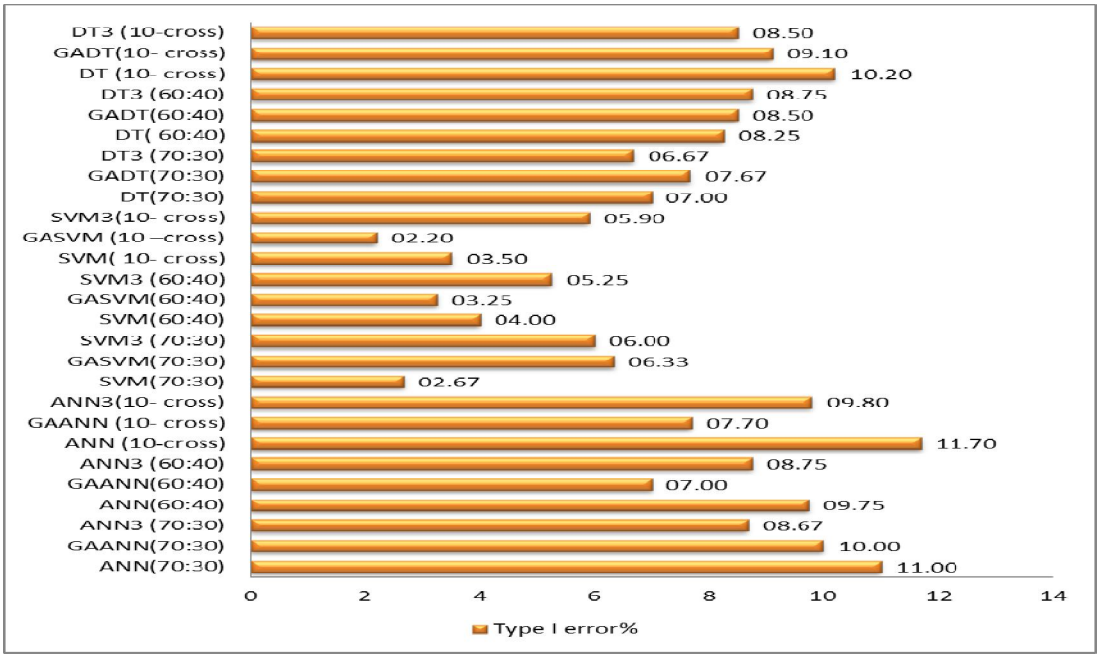

Figure 6.14 Results of Type IError for the German Dataset Models of Stage 1, 2 &3

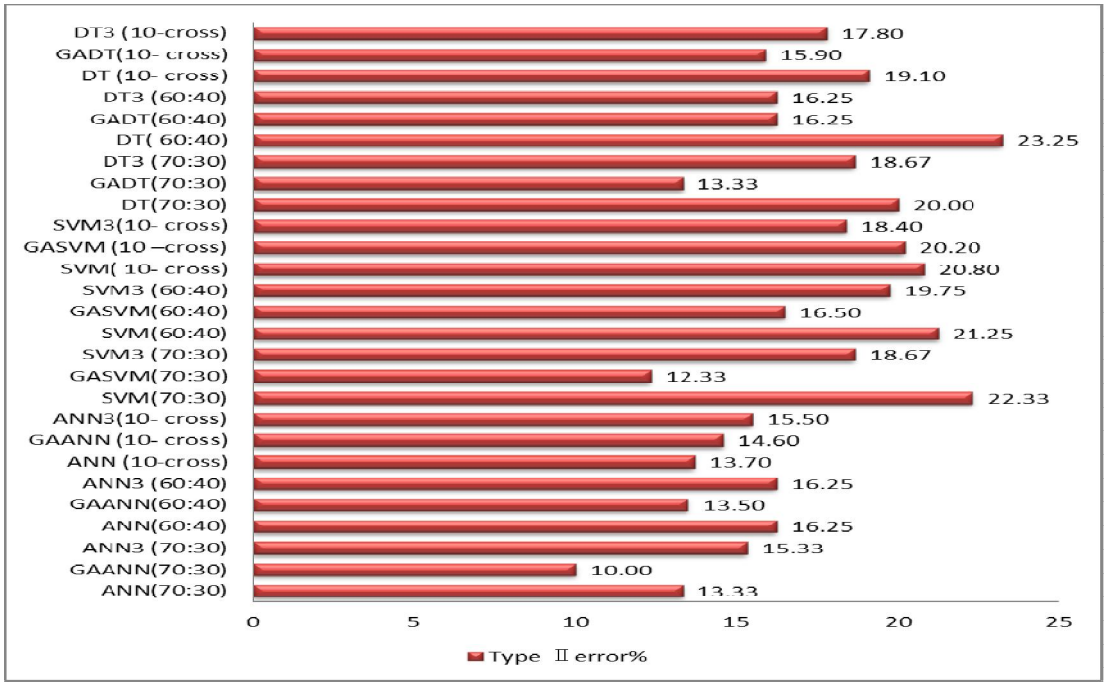

Figure 6.15 Results of Type II Error for the German Dataset Models of Stages 1, 2 &3

6.5 Summary

According to the above experimental results, the following conclusions can be drawn :

1. For all datasets, combining GA as a wrapper-feature selection technique with ANN, SVM and DT classification techniques is more beneficial than applying these techniques individually.
2. The improvements brought by the techniques were different from one dataset to another. For example most DTs models especially for two Sudanese datasets got substantial improvement as the result of combining GA with them. This result is expected because the accuracy of DT technique is affected by the redundancy and irrelevance of the data attributes [121]. See figures 6.1-6.15
3. Except for two models(out of nine models)of SCD1 and SCD2, applying individual techniques to the reduced datasets does not bring a significant improvement in terms of measure indicators compared to the resulting models of stage1.
4. The type of validation used slightly affected the performance of the technique for each dataset. For examples:
   - GAANN(70:30), GAANN(60:40) and GAANN(10-cross) achieved accuracies 80% ,79.50% and 77.70 % respectively for SCD1.
   - SVM (70:30), SVM (60:40) and SVM(10-cross) achieved Type I error rates 00.35% ,00.79% and 1.37 % respectively for SCD2.
   - DT3 (70:30), DT3 (60:40) and DT3 (10-cross) achieved precisions (GOOD) 76.95%, 78.83 and 77.55% respectively for the German dataset.

5. Finally, the answer for the question at the start of these comparisons is that: for all datasets, it is efficient to combine identified single techniques with GA as a feature selection technique to develop the proposed CSMs more than applying these classification techniques to the original datasets, or applying these single techniques to the reduced datasets.

CHAPTER SEVEN

7. Conclusions and Recommendations

7.1Conclusions

7.1.1 Summary of the Thesis

This research introduces the concept of CSMs. These models have been applied by many researchers to improve the process of assessing credit worthiness by differentiating between prospective loans on the basis of the likelihood of repayment. Thus, CS is a very typical DM classification problem.

At the outset of this research, the main aim of this research is identified in Chapter One wasto enhance the process of loan granting in Sudanese commercial banks by developing and introducing CSM(s) to evaluate their personal loans using DM classification techniques. This aim led the researcher to conduct an extensive literature review on the most widely used DM classification techniques in CS (presented in Chapter Three), thereby identifying the gaps in the literature (presented at the end of Chapter Three). The vital fact from this survey is that, till now there is no best technique for CS problems for all situations and datasets. By the end of this survey, one of the objectives of this research was obtained. This objective is reviewing the different DM techniques which are applied to CS problem and address their advantages, shortcomings and their potential influence in improving the loan granting process.

CS problem as mentioned above is modeled as a DM classification problem. The first step in DM process is problem identification. Therefore, in this research Sudanese banking sector which is the domain of this research was studied. This stage of this research includes surveying and interviewing loan officers in 10 banks. As the results of these surveys and interviews, one objective of this research was fulfilled. This objective is identifying the currently used credit risk evaluation systems used in the Sudanese banking (in Chapter Five).The advantages and disadvantages for these systems are also identified.

Like many developing countries, in Sudan credit agencies and credit bureaus do not exist and thus the relevant and trusted data on credit behavior of loan applicants are not currently available. Financial organizations have not built credit datasets from the performing and non-performing loans in the past. Hence, obtaining a credit dataset was a real challenge.

Thereafter, CS readiness test was applied to these ten banks. Only two banks passed this test.

Two credit datasets for these two Sudanese banks were built. Preprocess tasks were made to build these datasets as follows: Identification and collection of the relevant data(under direction of expertise loan officers), data integration, Missing values manipulation, numerical attributes normalization, outliers removing, transformation of the categorical attributes to numerical (to gain numerical version of a dataset) and instance labeling.

After the completions of these preprocess tasks two additional objectives of this research were achieved. These objectives are investigating loan

variables used in the loan granting decision-making process in the Sudanese banks and building high quality Sudanese credit dataset(s) (in Chapter Five).

The first Sudanese Credit Dataset1 (SCD1) was provided by Agricultural Bank of Sudan and the second one (SCD2) was provided by Al Salam Commercial Bank. In addition to these two datasets, German credit dataset was also employed in this research as a benchmarking dataset.

According to the concluding points from the literature survey, numerous classification techniques have been adopted for developing the proposed CSMs in this research. These techniques are ANN, SVM and DT. In addition to these classification techniques GA was applied as the wrapper-based feature selection technique.

RapidMiner 5.3.007was chosen as the main software package to simulate these proposed CSMs.

Experiments of this research were conducted in three stages for each dataset (stage 1, stage 2 and stage 3). In all stages two validation methods holdout (with two splitting ratios) and k-fold cross validation (K=10) were employed. Three sampling types were tested (linear, shuffled and stratified) for each experiment. The sample type which achieved the best accuracy for the given model was chosen.

ANN learned by means of a feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron). For all ANN experiments in this research, networks with one and two hidden layers and learning rate of 0.2 and 0.3 were tested with different validations and sampling types. For all

SVM experiments in this research four kernel types (Dot, Radial, Polynomial and Anova) were tested with different validations and sampling types. For all DT experiments all split criteria (Information gain, Gain ratio, Gini index and Accuracy) were tested with different validations and sampling types.

In the experiments of stage1 three identified classification techniques were applied to each dataset individually. In experiments of stage2, GA was combined with the identified techniques as a feature selection technique. In stage3 all techniques of stage1 were applied to intersected reduced datasets (As a result of using GA in stage 2, tables of attributes and their weights were produced. By using these tables new reduced sets of features were identified for each dataset).

All experiments in these different stages were repeated for each dataset. In each stage of these experiments nine CSMs were developed for each dataset. Hence for all stages 27 CSMs were developed for each dataset. The classification performances of proposed CSMs were identified using Confusion Matrix. Five evaluation measures were identified to evaluate and compare the proposed CSMs. These measures were accuracy, precision (defaulter), precision (Non-defaulter), Type I and Type Π errors.

Different comparisons between the proposed models were conducted and discussed. The resulting models of each of the three stages experiments for each dataset were compared and discussed. For these comparisons optimal model(s) were identified for each dataset in each stage.

Lastly, the resulting models of the stage 1, stage 2 and stage 3 experiments for each dataset were compared and discussed. The reason behind these last comparisons is to answer the question: Is it efficient to develop the proposed CSMs by applying the identified classification techniques individually to the original datasets, or to combine these techniques with GA as a feature selection technique or applying these single techniques to the reduced datasets?

7.1.2 Findings of the Thesis

Findings of this thesis can be summarized as follows:

1. For all datasets, combining GA as a wrapper-feature selection technique with ANN, SVM and DT classification techniques is more beneficial than applying these techniques individually. This result agrees with the prior findings in literature. The improvements brought by the techniques were different from one dataset to another.
2. Different sets of optimal models are identified for each dataset as follows:

– GAANN (70:30), GADT (70:30) and SVM3 (60:40) are chosen to be the optimal models for the SCD1.

– GADT (70:30) is strongly recommended to be the optimal model for SCD2. GASVM (70:30) and SVM3 (70:30) can be also chosen as the second optimal models for this dataset.

– GAANN (70:30), GASVM (70:30), GASVM (60:40) can be chosen to be the optimal models for the German credit dataset.

3. The experiments carried out in this research show that, the performance of each technique heavily depends on the nature of datasets.
4. The type of validation used slightly affected the performance of the techniques for each dataset.
5. Finally, the research concludes that, it is efficient to combine identified single techniques with GA as a feature selection technique to develop the proposed CSMs more than applying these classification techniques to the original datasets, or applying these single techniques to the reduced datasets.

7.2 Recommendations for Future Research

1. This is a first attempt to build CSMs for Sudanese credit dataset. Hence, further studies are needed and the credit dataset has to be extended and collected from more than two banking sectors. Further studies are also needed to investigate if it is useful to separate these datasets according to their types of loans and financing modes or considers all types of loans equal irrespective of the loan conditions. Furthermore all the factors taken into account in the decision making of granting a loan have to be recorded and added to the credit datasets e.g. the behaviour of borrowers in previous loans with the crediting bank and other banks.
2. In order to evaluate these CS proposed models in this research, they have to be implemented in real life and have the results of the models compared with the loan officers' decisions.

Therefore, user friendly interfaces have to be designed and connected with the proposed CSMs. The proposed CSS submit screen is presented in Figure 7.1.

3. The results obtained in this research can be further improved by using other artificial intelligence techniques of classifications such as Case-Base reasoning, rough sets, genetic programming and fuzzy systemsas well as using other alternatives of the hybrid models and ensemble techniques.

4. Explanation of loan granting decision is important for bankers and consumers. Hence, transparency is of special importance to CSMs. ANN and SVM are criticized for their poor explanation capability, specifically when applied to CSSs because the reasoning of their decision is not available. In the literature there are many treatments to solve this problem for these models. Therefore, these treatments have to be addressed in future research.
5. There are many alternatives of feature selection techniques such as principal component analysis (PCA),stepwise forward selection, Stepwise backward elimination and combination of forward selection and backward elimination. In addition to these techniques, particle swarm optimization (PSO) and rough sets (RS) are recently applied as feature selection techniques. These techniques have to be applied to these datasets and compared with GA.
6. The problem of imbalanced datasets is not discussed in this research. Using one of sampling techniques to balance the datasets of this

research may lead to enhance the performance of the proposed CSMs. Mining imbalanced datasets opens a front of interesting problems and research directions in context of CS problems.

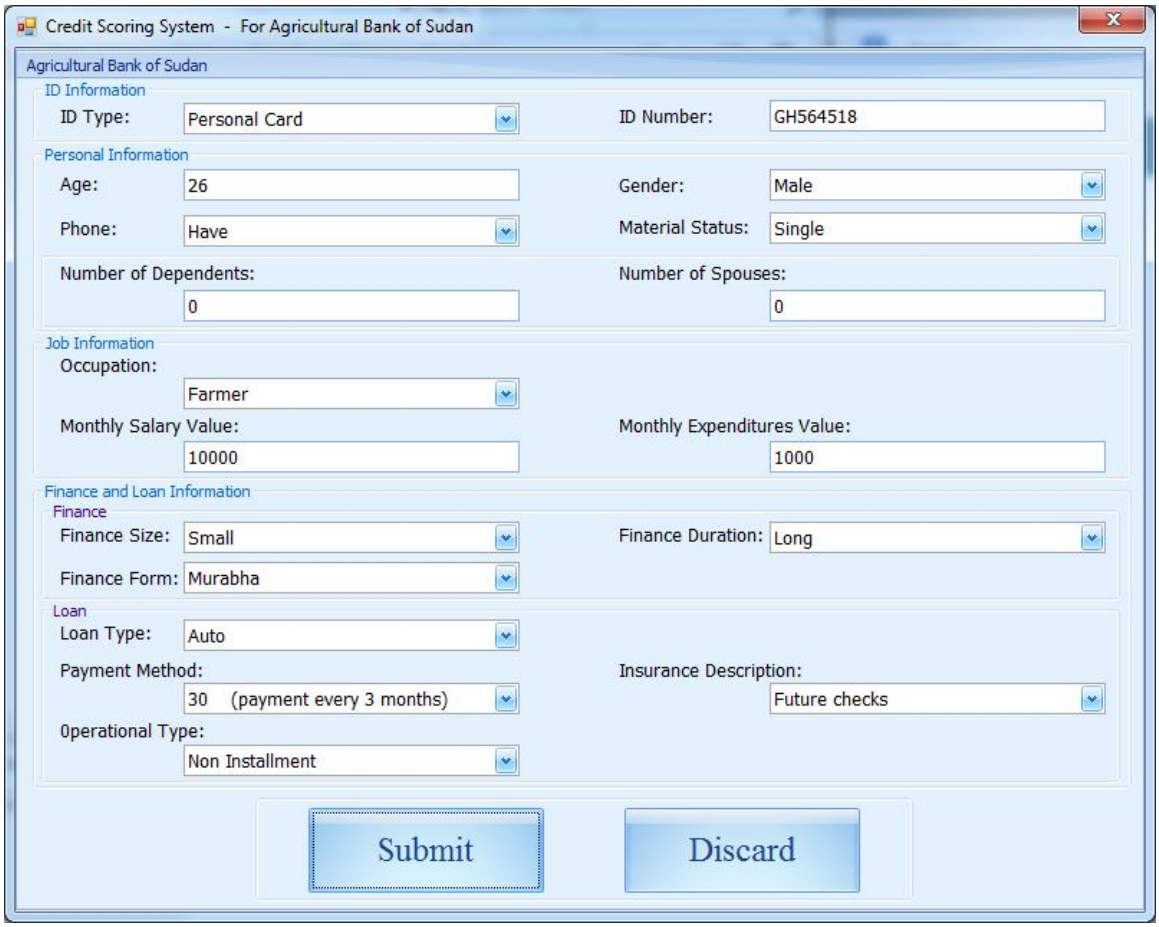

Figure 7.1Credit Scoring System Submit Screen

References

[1] *Business Dictionary*, http://www.businessdictionary.com/.

[2] *http://rapid-i.com*.

[3] *International Convergence of Capital Measurement and Capital Standards,A Revised Framework*, in B. C. o. B. Supervision, ed., Bank for International Settlements, 2004.

[4] *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability* Board of Governors of the Federal Reserve System, 2007.

[5] A I MARQUES, V GARCíA and J. S. SáNCHEZ, *A literature review on the application of evolutionary computing to credit scoring*, Journal of the Operational Research Society advance online publication 7 (2012).

[6] H. ABDOU, *Genetic programming for credit scoring: The case of Egyptian public sector banks*, Expert Systems with Applications 36 (2009), pp. 11402–11417.

[7] H. ABDOU and J. POINTON, *Credit Scoring , Statistical Techniques and Evaluation Criteria : A Review of the Literature*, Intelli gent Systems in Accounting, Finance & Management., 18 (2011), pp. 59-88.

[8] H. ABDOU, J. POINTON and A. EL-MASRY, *Neural nets versus conventional techniques in credit scoring in Egyptian banking*, Expert Systems with Applications, 35 (2008), pp. 1275-1292.

[9] L. ADEL , R. N. AINON; and T. Y. WAH, *Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier*, Maejo Int. J. Sci. Technol. , 4 (2010), pp. 136-158.

[10] R. K. AGGARWAL and T. YOUSEF, *Islamic Banks and Investment Financing*, Journal of Money, 32 (2000), pp. 93-120.

[11] S. AKKOC, *An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data*, European Journal of Operational Research, 222 (2012), pp. 168-178.

[12] R. ANDERSON, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, USA,New York, 2007.

[13] A. ARZY SOLTAN and M. MEHRABIOUN MOHAMMADI, *A hybrid model using decision tree and neural network for credit scoring problem*, Management Science Letters, 2 (2012), pp. 1683-1688.

[14] A. ASUNCION and D. J. NEWMAN, *UCI machine learning repository*, http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA:University of California, School of Information and Computer Science 2007.

[15] B. BAESENS, R. SETIONO, C. MUES, J. VANTHIENEN and K. RIDGE, *Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation.*, Management Science, 49 (2003), pp. 312-329.

[16] B. BAESENS, T. VAN GESTEL, S. VIAENE, M. STEPANOVA, J. SUYKENS and J. VANTHIENEN, *Benchmarking state-of-the-art classification algorithms for credit scoring*, Journal of the Operational Research Society, 54 (2003), pp. 627-635.

[17] A. C. BAHNSEN and A. M. GONZALEZ, *Evolutionary algorithms for selecting the architecture of a MLP Neural Network: A Credit Scoring Case, 11th IEEE International Conference on Data Mining Workshops*, Vancouver, 2011.

[18] A. BAHRAMMIRZAEE, *A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems*, Neural Computing and Applications, 19 (2010), pp. 1165-1195.

[19] T. BELLOTTI and J. CROOK, *Support vector machines for credit scoring and discovery of significant features*, Expert Systems with Applications, 36 (2009), pp. 3302-3308.
[20] A. BLANCO, R. PINO-MEJíAS, J. LARA and S. RAYO, *Credit scoring models for the microfinance industry using neural networks: Evidence from Peru*, Expert Systems with Applications, 40 (2013), pp. 356-364.
[21] V. BUMACOV and A. ASHTA, *The conceptual framework of credit scoring from its origins to microfinance*, Burgundy School of Business, 2011.
[22] D. CAIRE and R. KOSSMANN, *Credit Scoring : Is It Right for Your Bank ?*, Bankok Consulting 2003, pp. 1-12.
[23] S. CHATTOPADHYAY, S. BANERJEE, F. A. RABHI and U. R. ACHARYA, *A Case-Based Reasoning system for complex medical diagnosis*, Expert Systems, 30 (2013), pp. 12-20.
[24] H.-C. CHEN and Y.-C. CHEN, *A comparative study of discrimination methods for credit scoring., 40th International Conference on Computers & Industrial Engineering (CIE-40)*, IEEE, Awaji City, Japan, 2010.
[25] W. CHEN, G. XIANG, Y. LIU and K. WANG, *Credit risk Evaluation by hybrid data mining technique*, Systems Engineering Procedia, 3 (2012), pp. 194-200.
[26] B.-W. CHI and C.-C. HSU, *A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model*, Expert Systems with Applications, 39 (2012), pp. 2650-2661.
[27] J.-Y. CHIUA, Y. YAN, G. XUEDONGB and R.-C. CHEN, *A New Method for Estimating Bank Credit Risk, International Conference on Technologies and Applications of Artificial Intelligence(TAAI)*, IEEE, Taiwan, 2010, pp. 503-507.
[28] S. CHO, H. HONG and B.-C. HA, *A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction*, Expert Systems with Applications, 37 (2010), pp. 3482-3488.
[29] C.-L. CHUANG and R.-H. LIN, *Constructing a reassigning credit scoring model*, Expert Systems with Applications, 36 (2009), pp. 1685-1694.
[30] A. CORREA B. and A. GONZALEZ M., *Evolutionary algorithms for selecting the architecture of a MLP Neural Network: A Credit Scoring Case, 11th International Conference on Data Mining Workshops*, IEEE, Vancouver, 2011, pp. 725-732.
[31] M. DELLAERT and C. DEKONINCK, *Challenges and opportunities for retail and corporate banks in a fast changing and globalizing world*, Management Centre Europe (MCE), 2012.
[32] C. J. DESAI V., AND OVERSTREET G., *Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms*, Computer Journal of Mathematics Applied in Business and Industry, 8 (1997), pp. 324-346.
[33] Y. DONG, *A CASE BASED REASONING SYSTEM FOR CUSTOMER CREDIT SCORING : COMPARATIVE STUDY OF SIMILARITY MEASURES, The 51st Annual Meeting of the International Society for the Systems Sciences*, Curran Associates, Inc., Tokyo, Japan, 2007, pp. 910-922.
[34] M. DOUMPOS, K. KOSMIDOU, G. BAOURAKIS and C. ZOPOUNIDIS, *Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis*, European Journal of Operational Research, 138 (2002), pp. 392-412.
[35] D. DUKI, G. DUKI and L. KVESI, *A Credit Scoring Decision Support System, 33rd Int. Conf. on Information Technology Interfaces*, Cavtat, Croatia, 2011.
[36] D. DURAND, *Risk Elements in Consumer Instalment Financing*, in National Bureauof Economic Research, ed., USA.New York., 1941.

[37] H. T. ELSHOUSH and I. M. OSMAN, *Alert correlation in collaborative intelligent intrusion detection systems—A survey*, Applied Soft Computing, 11 (2011), pp. 4349-4365.
[38] A. B. EMEL, M. ORAL, A. REISMAN and R. YOLALAN, *A credit scoring approach for the commercial banking sector*, Socio-Economic Planning Sciences, 37 (2003), pp. 103–123.
[39] U. FAYYAD, G. PIATETSKY-SHAPIRO and P. SMYTH, *From data mining to knowledge discovery in databases*, AI magazine (1996), pp. 37-54.
[40] S. FINLAY, *Are we modelling the right thing? The impact of incorrect problem specification in credit scoring*, EXPERT SYSTEMS WITH APPLICATIONS, 36 (2009).
[41] S. FINLAY, *Multiple classifier architectures and their application to credit risk assessment*, European Journal of Operational Research, 210 (2011), pp. 368-378.
[42] J. M. GANG WANG *A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine*, Expert Systems with Applications, 39 (2012), pp. 5325–5331.
[43] V. GARCíA, A. I. MARQUES and J. S. SáNCHEZ, *On the use of data filtering techniques for credit risk prediction with instance-based models*, Expert Systems with Applications, 39 (2012), pp. 13267-13276.
[44] A. GHODSELAHI and A. AMIRMADHI, *Application of Artificial Intelligence Techniques for Credit Risk Evaluation*, International Journal of Modeling and Optimization, 1 (2011), pp. 243-249.
[45] P. GIUDICI and S. FIGINI, *Applied Data Mining for Business and Industry*, John Wiley & Sons Ltd, 2009.
[46] J. HAN and M. KAMBER, *Data Mining:Concepts and Techniques* Morgan Kaufmann, 2006.
[47] J. HAN, M. KAMBER and J. PEI, *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers is an imprint of Elsevier., 225Wyman Street,Waltham, MA 02451, USA, 2012.
[48] A. M. M. HASSEN, *Measurement and Disclosure and their Role in Human Resources Cost Accounting in Banking Sector – Applied and Field Study on a Sample of Banks Sudan, Accounting* Sudan University for Science and Technology, 2013.
[49] F. HOFFMANN, BAESENS, B., MARTENS, J., PUT, F. AND VANTHIENEN, J, *Comparing a Genetic Fuzzy and a Neurofuzzy Classifier for Credit Scoring*, International Journal of Intelligent Systems, 17 (2002), pp. 1067-1083.
[50] N. HSIEH, *Hybrid mining approach in the design of credit scoring models*, Expert Systems with Applications, 28 (2005), pp. 655-665.
[51] C.-L. HUANG, M.-C. CHEN and C.-J. WANG, *Credit scoring with a data mining approach based on support vector machines*, Expert Systems with Applications, 33 (2007), pp. 847-856.
[52] J.-J. HUANG, G.-H. TZENG and C.-S. ONG, *Two-stage genetic programming (2SGP) for the credit scoring model*, Applied Mathematics and Computation, 174 (2006), pp. 1039-1053.
[53] J. P. HUSSEIN A. ABDOU, *Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature*, Intelligent Systems in Accounting,Finance & Management 18 (2-3) (2011), pp. 59-88.
[54] J. P. HUSSEIN ABDOU , AHMED EL-MASRY, *Neural nets versus conventional techniques in credit scoring in Egyptian banking.*, Expert Systems with Applications 35 (2008), pp. 1275–1292.
[55] M. ISMAIL, *THE PRACTICE OF ISLAMIC BANKING SYSTEM IN SUDAN*, Journal of Economic Cooperation, 4 (2005), pp. 27-50.
[56] E. KAMBAL, I. OSMAN, M. TAHA, N. MOHAMMED and S. MOHAMMED, *Credit scoring using data mining techniques with particular reference to Sudanese banks*,

International Conference on Computing, Electrical and Electronic Engineering (Icceee), IEEE, Khartoum-Sudan 2013, pp. 378-383.

[57] S. KAPOOR, *How to design a good banking system?*, Re-deine.org http://www.re-define.org, 2010.

[58] A. KHASHMAN, *Credit risk evaluation using neural networks: Emotional versus conventional models*, Applied Soft Computing, 11 (2011), pp. 5477-5484.

[59] A. KHASHMAN, *Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes*, Expert Systems with Applications, 37 (2010), pp. 6233-6239.

[60] M. F. KIANI and A. L. R. LOGIT, *A New Hybrid Method for Credit Scoring Based on Clustering and Support Vector Machine (ClsSVM)*, *2nd IEEE International Conference on Information and Financial Engineering (ICIFE)*, IEEE, Chongqing, China 2010, pp. 585-589.

[61] E. M. KNOX and R. T. NG, *Algorithms for Mining Datasets Outliers in Large Datasets*, *In Proc. of the VLDB Conference*, New York, USA, 1998, pp. 392-403.

[62] H. C. KOH and W. C. TAN, *A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques*, International Journal of Business and Information, 1 (2006), pp. 96-118.

[63] S. B. KOTSIANTIS, I. D. ZAHARAKIS and P. E. PINTELAS, *Machine learning: a review of classification and combining techniques*, Artificial Intelligence Review, 26 (2007), pp. 159-190.

[64] A. LAHSASNA, R. N. AINON and T. Y. WAH, *Credit Scoring Models Using Soft Computing Methods : A Survey*, The International Arab Journal of Information Technology, 7 (2010).

[65] A. LAHSASNA, R. N. AINON and T. Y. WAH, *Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier*, Maejo International Journal of Science and Technology, 4 (2010), pp. 136-158.

[66] A. LAHSASNA, R. N. AINON and T. Y. WAH, *Intelligent Credit Scoring Model using Soft Computing Approach*, *International Conference on Computer and Communication Engineering* IEEE, Kuala Lumpur, Malaysia, 2008, pp. 396-402.

[67] M.-C. LEE, *Enterprise Credit Risk Evaluation models : A Review of Current Research Trends*, International Journal of Computer Applications, 44 (2012), pp. 37-44.

[68] T.-S. LEE, C.-C. CHIU, Y.-C. CHOU and C.-J. LU, *Mining the customer credit using classification and regression tree and multivariate adaptive regression splines*, Computational Statistics & Data Analysis, 50 (2006), pp. 1113-1130.

[69] T.-S. LEE, C.-C. CHIU, C.-J. LU and I.-F. CHEN, *Credit Scoring Using the Hybrid Neural Discriminant Technique*, Expert Systems with Applications, 23 (2002), pp. 245-254.

[70] W. LI and J. LIAO, *An Empirical Study on Credit Scoring Model for Credit Card by Using Data Mining Technology*, *Seventh International Conference on Computational Intelligence and Security*, IEEE, Sanya, Hainan, China 2011, pp. 1279-1282.

[71] M. K. LIM and S. Y. SOHN, *Cluster-based dynamic scoring model*, Expert Systems with Applications, 32 (2007), pp. 427-431.

[72] H. LIU, S. MEMBER, L. YU and S. MEMBER, *Toward Integrating Feature Selection Algorithms for Classification and Clustering*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 17 (2005), pp. 491-502.

[73] Y. LIU, *A framework of data mining application process for credit scoring Institut für Wirtschaftsinformatik, Abteilung Wirtschaftsinformatik II, Georg-August-Universität Göttingen*, 2002.

[74] Y. LIU., *New Issues in Credit Scoring Application*, Arbeitsbericht Nr. 16/2001 (2001).

[75] M. S. LONG, *Credit Scoring Development for Optimal Credit Extension and Management Control*, *College on Industrial Management,Georgia Institute of Technology*, Purdue University, Atlanta Georgia, 1973.

[76] M. MAHMOUD, N. ALGADI and A. ALI, *Expert System for Banking Credit Decision*, *2008 International Conference on Computer Science and Information Technology*, IEEE, Singapore, 2008, pp. 813-819.

[77] A. MARCANO-CEDEñO, A. MARIN-DE-LA-BARCENA, J. JIMENEZ-TRILLO, J. A. PIñUELA and D. ANDINA, *Artificial metaplasticity neural network applied to credit scoring.*, International journal of neural systems, 21 (2011), pp. 1-7.

[78] A. I. MARQUES, V. GARCíA and J. S. SáNCHEZ, *Exploring the behaviour of base classifiers in credit scoring ensembles*, Expert Systems with Applications, 39 (2012), pp. 10244-10250.

[79] D. MARTENS, J. HUYSMANS, R. SETIONO, J. VANTHIENEN and B. BAESENS, *Rule Extraction from Support Vector Machines : An Overview of Issues and Application in Credit Scoring*, Studies in Computational Intelligence (SCI), 63 (2008), pp. 33-63.

[80] L. J. MESTER, *What's the Point of Credit Scoring? 1997.*, FEDERAL RESERVE BANK OF PHILADELPHIA, 1997.

[81] A. MUKHERJEE, *credit scoring model using data mining techniques-a pragmatic approach.* , VDM Verlag Dr. Müller, Germany, 2010.

[82] M. NEGNEVITSKY, *Artificial Intelligence: A Guide to Intelligent Systems*, Pearson Education, 2011.

[83] N. NWULU, S. OROJA and M. ILKAN, *Credit Scoring Using Soft Computing Schemes : A Comparison between Support Vector Machines and Artificial Neural Networks*, *nternational Conference, DEIS* . Springer-Verlag Berlin Heidelberg, London, UK, 2011, pp. 275-286.

[84] C. ONG, J. HUANG and G. TZENG, *Building credit scoring models using genetic programming*, Expert Systems with Applications, 29 (2005), pp. 41-47.

[85] S. ORESKI, D. ORESKI and G. ORESKI, *Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment*, Expert Systems with Applications, 39 (2012), pp. 12605-12617.

[86] V. PACELLI, *An Artificial Neural Network Approach for Credit Risk Management*, Journal of Intelligent Learning Systems and Applications, 03 (2011), pp. 103-112.

[87] R. S. RAGHAVAN, *Risk Management in Banks*, *www.icai.org*, www.icai.org, 2003, pp. 841-851.

[88] B. K. RAGHAVENDRA and B. JAY, SIMHA, , *Evaluation of Feature Selection Methods for Predictive Modeling Using Neural Networks in Credits Scoring*, Int. J. Advanced Networking and Applications, 718 (2010), pp. 714-718.

[89] S. RAMASWAMY, R. RASTOGI and K. SHIM, *Efficient algorithms for mining outliers from large data sets*, *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, ACM SIGMOD Record, New York, NY, USA, 2000, pp. 427-438.

[90] A. B. RUSSELL, *What Gave Rise to the Credit Crisis*, *UICIFD E-Book*, http://blogs.law.uiowa.edu/ebook/sites/default/files/Part_5_1.pdf, 2010, pp. 1-19.

[91] R. SINGH and R. R. AGGARWAL, *Comparative Evaluation of Predictive Modeling Techniques on Credit Card Data*, International Journal of Computer Theory and Engineering, 3 (2011), pp. 1-6.

[92] D. O. STJEPAN ORESKI , GORAN ORESKI *Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment*, Expert Systems with Applications 39 (2012), pp. 12605–12617.

[93] M. ŠUŠTERŠI, D. MRAMOR and J. ZUPAN, *Consumer Credit Scoring Models With Limited Data*, Expert Systems with Applications, 36 (2009), pp. 4736-4744.

[94] B. TANG and S. QIU, *A new Credit Scoring Method Based on Improved Fuzzy Support Vector Machine*, *Computer Science and Automation Engineering (CSAE)*, IEEE, China, 2012, pp. 73-75.

[95] L. C. THOMAS, D. B. EDELMAN and L. N. CROOK, *Credit Scoring and Its Applications*, Society for Industrial and Applied Mathematics, 2002.

[96] Y. TIAN, Y. SHI and X. LIU, *Recent advances on support vector machines research*, Technological and Economic Development of Economy, 18 (2012), pp. 5-33.

[97] C.-F. TSAI and M.-L. CHEN, *Credit rating by hybrid machine learning techniques*, Applied Soft Computing, 10 (2010), pp. 374-380.

[98] H.-E. TSENG, C.-C. CHANG and S.-H. CHANG, *Applying case-based reasoning for product configuration in mass customization environments*, Expert Systems with Applications, 29 (2005), pp. 913-925.

[99] B. TWALA, *Multiple classifier application to credit risk assessment*, Expert Systems with Applications, 37 (2010), pp. 3326-3336.

[100] S. VUKOVIC, B. DELIBASIC, A. UZELAC and M. SUKNOVIC, *A case-based reasoning model that uses preference theory functions for credit scoring*, Expert Systems with Applications, 39 (2012), pp. 8389-8395.

[101] S. VUKOVIC, B. DELIBASIC, A. UZELAC and M. SUKNOVIC, *A case-based reasoning model that uses preference theory functions for credit scoring*, Expert Systems with Applications 39 (2012), pp. 8389–8395.

[102] B. WAH, S. HUAT, N. HUSELINA and M. HUSAIN, *Using data mining to improve assessment of credit worthiness via credit scoring models*, Expert Systems With Applications, 38 (2011), pp. 13274-13283.

[103] Y. B. WAH and I. R. IBRAHIM, *Using Data Mining Predictive Models to Classify Credit Card Applicants*, *6th International Conference on Advanced Information Management and Service (IMS)*, IEEE, Seoul, Korea (South), 2010, pp. 394-398.

[104] R. WALL, P. CUNNINGHAM, P. WALSH and S. BYRNE, *Explaining the output of ensembles in medical decision support on a case by case basis*, Artificial Intelligence in Medicine, 28 (2003), pp. 191-206.

[105] G. WANG, J. HAO, J. MA and H. JIANG, *A comparative assessment of ensemble learning for credit scoring*, Expert Systems with Applications, 38 (2011), pp. 223-230.

[106] G. WANG, J. MA, L. HUANG and K. XU, *Two credit scoring models based on dual strategy ensemble trees*, Knowledge-Based Systems, 26 (2012), pp. 61-68.

[107] J. WANG, K. GUO and S. WANG, *Rough set and Tabu search based feature selection for credit scoring*, Procedia Computer Science, 1 (2010), pp. 2425-2432.

[108] Q. WANG, K. K. LAI, D. NIU and Q. ZHANG, *A Triple Artificial Neural Network Model Based on Case Based Reasoning for Credit Risk Assessment*, *Fifth International Conference on Business Intelligence and Financial Engineering*, IEEE, China, 2012, pp. 10-14.

[109] Y. WANG, S. WANG and K. K. LAI, *A New Fuzzy Support Vector Machine to Evaluate Credit Risk*, IEEE TRANSACTIONS ON FUZZY SYSTEMS, 13 (2005), pp. 820-831.

[110] D. WEST, *Neural network credit scoring models*, Computers & Operations Research 27 (2000), pp. 1131-1152.

[111] W. XU, S. ZHOU, D. DUAN and Y. CHEN, *A Support Vector Machine Based Method for Credit Risk Assessment*, *7th International Conference on E-Business Engineering*, IEEE, Shanghai, China, 2010, pp. 50-55.

[112] P. YAO, *Fuzzy Rough Set and Information Entropy Based Feature Selection for Credit Scoring*, *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Tianjin, China, 2009, pp. 247-251.

[113] P. YAO, *Hybrid Fuzzy SVM Model Using CART and MARS for Credit Scoring*, *International Conference on Intelligent Human-Machine Systems and Cybernetics*, IEEE, Hangzhou, Zhejiang, China, 2009, pp. 392-395.

[114] I.-C. YEH and C.-H. LIEN, *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*, Expert Systems with Applications, 36 (2009), pp. 2473-2480.

[115] H. YU, X. HUANG, X. HU and H. CAI, *A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation*, *International Conference on Management of e-Commerce and e-Government*, IEEE, Chengdu, China, 2010, pp. 35-38.

[116] L. YUN, Q.-Y. CAO and H. ZHANG, *Application of the PSO-SVM Model for Credit Scoring*, *Seventh International Conference on Computational Intelligence and Security*, IEEE, Sanya,Hainan,China, 2011, pp. 47-51.

[117] D. ZHANG, M. HIFI, Q. CHEN and W. YE, *A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines*, *Fourth International Conference on Natural Computation*, IEEE, Jinan, China, 2008, pp. 8-12.

[118] D. ZHANG, H. HUANG, Q. CHEN and Y. JIANG, *A Comparison Study of Credit Scoring Models*, *Third International Conference on Natural Computation (ICNC)*, IEEE, Haikou, Hainan, China, 2007, pp. 7-10.

[119] D. ZHANG, S. C. H. LEUNG and Z. YE, *A Decision Tree Scoring Model Based on Genetic Algorithm and K-Means Algorithm*, *Third International Conference on Convergence and Hybrid Information Technology(ICCIT)*, IEEE, Busan, Korea, 2008, pp. 1043-1047.

[120] L. ZHANG and X. HUI, *Application of Support Vector Machines, 6th International Symposium of Neural Networks (ISNN 2009*, Springer-Verlag, China, 2009, pp. 283-290.

[121] W. ZHONG-YIN, *Credit valuation modeling and evaluation using data mining for consumer loan market with skewed data*, *Department of Industrial Management*, National Taiwan University of Science and Technology, 2007.

[122] J. ZHOU and T. BAI, *Credit Risk Assessment Using Rough Set Theory and GA-Based SVM*, *The 3rd International Conference on Grid and Pervasive Computing - Workshops*, IEEE, Kunming, China, 2008, pp. 320-325.

[123] J. ZURADA, *Does Feature Reduction Help Improve the Classification Accuracy Rates ?* , Review of Business Information Systems, 14 (2010), pp. 35-40.

[124] J. ZURADA, *Using Memory-Based Reasoning For Predicting Default Rates On Consumer Loans*, Review of Business Information Systems, 11 (2007), pp. 1-16.

[125] J. ZURADA and N. KUNENE, *Performance Assessment of Data Mining Methods for Loan Granting Decisions : A Preliminary Study*, *10th International Conference, ICAISC* springer Zakopane, Poland, 2010, pp. 495-502.

Appendix A

Islamic Financing Modes

The following modes are the most popular Islamic financing modes:

- Al-Murabahah mode: Is the popular mode used by a large number of Islamic banks. It is the main mode used to finance trade sector in Sudan[55]. In this mode the bank purchases an asset on behalf of an entrepreneur. The bank resells the asset to the entrepreneur at a predetermined price that covers the original cost and a negotiated profit margin. The murabahah payment is always deferred and made in lump sum or in installments. Ownership resides with the bank until all payments are made[10].

- Al-Musharaka mode: is the second mode that is widely used Islamic banks[55].In this mode the entrepreneur and the bank jointly supply the capital and manage the project. Losses are borne in proportion to the contribution of capital while profit is divided in a ratio agreed upon freely in advance[10]. Al-musharakah is also called al-muzar'ah in case the bank provides the land and machinery to the farmer[55].

- Al-Mudarabah mode: in this mode the bank acts as the creditor (rabb-almal) and the client as an entrepreneur [55](mudarib). The bank provides capital and the client provides efforts, expertise and complete control over the project. In case of a loss, the bank gains no return on its investment and the client receives no compensation for her (his) effort. In case of a gain, profit is split according to a negotiated equity percentage[10].

• Bai’ al_salam[55]: Is type of selling where the delivery product is deferred while payment of the price is introduced. Salam is limited to commodities whose quality and quantity can be fully prescribed at the time of the contract is assigned.

• Al-qardalhasan (profit free loan): In this mode the bank provides loan but obtains no profit. Repayments of this loan are made according to agreement between two parties[55].

• Ijara financing: in this mode the bank purchases the asset and leases it to entrepreneur at an agreed rate, for a defined period. The ownership of the asset either remains with the bank or is gradually transferred to the entrepreneur in a rent-to-own contract[55].

Appendix B

Parts of Sudanese Credit Datasets

Sudanese Credit Datasets1

Phone	IDType	Gender	age	Occupation	MaritalStatus	NumberOfChildren	NumberOfSpouses	MonthlySalaryValue	MonthlyExpendituresValue	finance size	finance duration	payment method	finance form	loan type	insurance description	0prational type	status
holder	PersonalCard	female	38	other	divorce	0	0	0.014393852	0.020134228	small	medium	1	murabha	automatic Agriculture	Mortgage mechanics	Non-Instalment	nondefaulter
holder	PersonalCard	female	48	other	divorce	3	0	0.014393852	0.020134228	micro	short	1	murabha	Traditional Agriculture	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	50	farmer	single	17	2	0.04941539	0.194630872	micro	short	1	murabha	automatic Agriculture	Mortgage mechanics	Non-Instalment	defaulter
holder	PersonalCard	male	32	lawyer	single	6	1	0.029403083	0.020134228	micro	short	30	murabha	Traditional Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	49	pensioner	single	8	1	0.119458467	0.060402685	micro	short	1	murabha	Irrigated agriculture	Personal guarantee	Non-Instalment	defaulter
unholder	PersonalCard	male	59	Free business	single	6	1	0.046913852	0.073825503	micro	long	360	murabha	Irrigated agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	42	Free business	single	0	1	0.026401237	0.023489933	micro	long	1	salam	automatic Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	50	farmer	single	9	2	0.024400006	0.026845638	micro	short	1	murabha	Transport sector	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	45	merchant	single	6	1	0.059421544	0.043624161	micro	short	1	salam	Irrigated agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	61	other	single	2	2	0.059421544	0.053691275	micro	medium	90	murabha	Transport sector	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	31	other	single	5	1	0.074430775	0.023489933	micro	short	1	murabha	Traditional Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	38	employee	single	1	1	0.048915083	0.048657718	micro	short	1	salam	Traditional Agriculture	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	45	farmer	single	9	2	0.024400006	0.026845638	micro	short	1	murabha	Local trade	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	28	other	single	1	2	0.059421544	0.043624161	micro	short	1	murabha	automatic Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	34	merchant	single	6	1	0.059421544	0.060402685	micro	short	30	murabha	Traditional Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	46	other	single	9	1	0.089440006	0.093959732	micro	medium	360	murabha	automatic Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	31	farmer	single	2	1	0.03440616	0.020134228	small	long	360	murabha	automatic Agriculture	Mortgage mechanics	Instalment	nondefaulter
holder	PersonalCard	male	50	merchant	single	11	1	0.024400006	0.026845638	micro	short	30	murabha	Traditional Agriculture	Adoption confidence	Instalment	nondefaulter
holder	PersonalCard	male	35	employee	single	5	1	0.159383021	0.053691275	micro	short	1	salam	Irrigated agriculture	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	51	other	single	8	3	0.024179871	0.021006711	micro	short	30	murabha	Traditional Agriculture	Personal guarantee	Instalment	nondefaulter

holder	PersonalCard	male	39	other	single	3	2	0.011392006	0.006711409	micro	short	180	murabha	Local trade	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	52	other	single	5	1	0.119458467	0.053691275	micro	short	90	murabha	Traditional Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	46	other	single	11	2	0.059421544	0.053691275	small	long	360	murabha	automatic Agriculture	Futures checks	Instalment	nondefaulter
holder	PersonalCard	male	63	teacher	single	7	2	0.041910775	0.038255034	micro	short	1	murabha	Irrigated agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	45	teacher	single	3	1	0.024400006	0.016778523	micro	medium	90	murabha	Irrigated agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	22	employee	single	4	1	0.04941539	0.040268456	small	short	1	murabha	Traditional Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	40	other	single	3	1	0.035406775	0.033557047	micro	short	1	murabha	automatic Agriculture	Mortgage mechanics	Non-Instalment	defaulter
holder	PersonalCard	male	26	other	single	5	1	0.019977286	0.016308725	micro	short	1	murabha	Transport sector	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	45	merchant	single	3	1	0.064424621	0.040268456	small	medium	360	murabha	Irrigated agriculture	other	Instalment	nondefaulter
holder	PersonalCard	male	50	other	single	1	1	0.039409237	0.023489933	micro	short	1	murabha	Irrigated agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	30	worker	single	2	1	0.024400006	0.016778523	micro	short	1	salam	Irrigated agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	27	farmer	single	4	1	0.014393852	0.016778523	micro	short	1	salam	automatic Agriculture	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	38	other	single	3	1	0.019396929	0.006711409	micro	short	1	murabha	Traditional Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	44	other	single	4	1	0.04741416	0.046979866	micro	short	1	murabha	Local trade	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	29	teacher	single	2	1	0.089440006	0.060402685	micro	short	360	murabha	automatic Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	38	employee	single	8	1	0.119458467	0.043624161	micro	short	1	murabha	Local trade	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	48	employee	single	2	1	0.029403083	0.033557047	small	short	180	murabha	automatic Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	51	Free business	single	13	1	0.029403083	0.026845638	micro	short	1	salam	automatic Agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	27	teacher	single	4	1	0.039409237	0.020134228	micro	short	360	murabha	automatic Agriculture	Futures checks	Instalment	defaulter
holder	PersonalCard	male	32	merchant	single	9	2	0.029403083	0.036912752	micro	medium	90	murabha	Traditional Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	38	merchant	single	6	1	0.064424621	0.023489933	micro	long	360	salam	automatic Agriculture	Personal guarantee	Instalment	defaulter
unholder	PersonalCard	male	48	Free business	single	3	1	0.099446159	0.023489933	micro	medium	90	murabha	Traditional Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	54	farmer	single	3	1	0.089440006	0.053691275	micro	short	1	salam	Irrigated agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	57	Free business	single	5	1	0.024400006	0.023489933	small	long	180	murabha	automatic Agriculture	Personal guarantee	Instalment	nondefaulter
holder	PersonalCard	male	38	Free business	single	9	2	0.004387698	0.020134228	micro	short	1	murabha	automatic Agriculture	Personal guarantee	Non-Instalment	nondefaulter
holder	PersonalCard	male	39	Free business	single	16	1	0.016895391	0.033557047	normal	short	30	murabha	automatic Agriculture	other	Instalment	nondefaulter
holder	PersonalCard	male	40	merchant	single	1	1	0.041910775	0.030872483	small	long	360	murabha	automatic Agriculture	Futures checks	Instalment	nondefaulter
holder	PersonalCard	male	50	merchant	single	6	1	0.099446159	0.026845638	micro	short	180	murabha	Irrigated agriculture	Personal guarantee	Instalment	defaulter
holder	PersonalCard	male	19	farmer	single	9	1	0.059421544	0.026845638	micro	short	1	salam	Traditional Agriculture	Personal guarantee	Instalment	defaulter

Sudanese Credit Dataset 2

Gender	Age	MaritalStatus	#fChildren	#Spouses	Occupation	Phone	IdType	Approved Credit Amount	Profit Margin	Periodical instalment Amount	Finance Duration	Periodicity of Payments	Purpose of Credit	Sector	Gurantee Type	Finance Size	Status
0	41	2	4	1	1	1	1	0.016947668	0.00794503	0.022057824	3	1	4	3	1	2	Non-Defaulter
0	26	1	0	0	1	1	2	0.006233122	0.002139821	0.007318621	3	1	4	3	2	2	Non-Defaulter
0	50	2	0	1	13	1	2	0.001748777	0.000334819	0.003738614	2	1	4	3	2	1	Defaulter
0	40	1	0	0	13	1	2	0.000890446	0.000376242	0.004044501	2	1	4	3	3	1	Non-Defaulter
0	36	2	1	1	13	1	1	0.000890446	0.000353357	0.003999184	2	1	4	3	3	1	Non-Defaulter
0	38	2	2	1	2	1	3	0.004924458	0.000894377	0.011057235	2	1	4	3	2	2	Non-Defaulter
0	34	2	0	1	13	1	1	0.009164295	0.002436421	0.014467304	2	1	4	3	2	2	Non-Defaulter
0	53	2	0	1	13	1	1	0.012872053	0.003377942	0.020562378	2	1	4	3	2	2	Non-Defaulter
1	26	1	0	0	13	1	2	0.004935406	0.001724673	0.00563058	3	1	4	3	2	2	Non-Defaulter
0	49	2	0	1	13	1	1	0.004028903	0.001067164	0.005936466	2	1	4	3	2	2	Non-Defaulter
1	49	2	4	1	13	1	1	0.000890446	0.000353357	0.003999184	2	1	4	3	3	1	Non-Defaulter
0	28	1	0	0	13	1	1	0.004466827	0.000702365	0.008961345	2	1	4	3	2	2	Non-Defaulter
0	45	2	0	1	13	1	3	0.000890446	0.000376242	0.004044501	2	1	4	3	3	1	Non-Defaulter
0	45	2	0	1	1	1	1	0.003771258	0.001005373	0.005517288	2	1	4	3	2	2	Non-Defaulter
0	43	2	0	1	13	0	3	0.003681483	0.001043821	0.005460643	2	1	4	3	2	2	Non-Defaulter
0	28	2	0	1	13	0	3	0.004538355	0.001261465	0.006865455	2	1	4	3	2	2	Non-Defaulter
0	52	1	0	0	13	1	3	0.006752792	0.001204021	0.015237685	2	1	4	3	2	2	Non-Defaulter
1	34	1	0	0	13	1	3	0.013736224	0.004539625	0.017073005	3	1	4	3	2	2	Non-Defaulter
0	42	2	0	1	13	1	1	0.005999562	0.002188797	0.007148684	3	1	4	3	2	2	Non-Defaulter
1	33	1	0	0	13	1	1	0.000890446	0.000376242	0.004044501	2	1	4	3	3	1	Non-Defaulter
1	30	1	0	0	13	1	3	0.003809941	0.002065214	0.007012734	3	1	4	3	2	2	Non-Defaulter

0	52	2	0	1	13	1	3	0.000890446	0.000547886	0.002424435	2	1	4	3	3	1	Defaulter
0	41	2	0	1	13	1	3	0.00745931	0.00268313	0.009074636	3	1	4	3	2	2	Non-Defaulter
0	27	1	0	0	13	0	3	0.005065324	0.001395346	0.007737799	2	1	4	3	2	2	Non-Defaulter
0	44	2	3	1	13	1	1	0.009822641	0.002603487	0.015543572	2	1	4	3	2	2	Non-Defaulter
0	45	2	3	1	13	1	1	0.01173272	0.003088666	0.018681742	2	1	4	3	2	2	Non-Defaulter
0	28	1	0	0	13	1	1	0.000890446	0.000376242	0.004044501	2	1	4	3	3	1	Non-Defaulter
1	44	2	0	1	13	1	1	0.014028173	0.007688252	0.019010287	3	1	4	3	1	2	Non-Defaulter
0	34	1	0	0	13	0	1	0.006729436	0.001509089	0.008462863	2	1	4	3	2	2	Non-Defaulter
0	61	2	0	1	13	1	3	0.020597037	0.004444191	0.059047446	2	1	4	3	4	2	Non-Defaulter
0	43	2	3	1	13	1	1	0.000890446	0.000376242	0.004044501	2	1	4	3	3	1	Non-Defaulter
0	53	2	2	1	13	1	1	0.003372017	0.00043689	0.009969638	2	1	4	3	2	2	Non-Defaulter
0	36	2	0	1	13	0	3	0.0040435	0.001439516	0.004475008	3	1	4	3	2	2	Non-Defaulter
0	48	2	0	1	13	1	3	0.014466097	0.003500152	0.020845607	3	1	4	3	2	2	Non-Defaulter
0	48	2	0	1	13	1	3	0.001184877	0.000211694	0.002594372	2	1	4	3	3	1	Non-Defaulter
0	52	2	7	1	13	1	1	0.003779432	0.000972646	0.005302035	2	1	4	3	2	2	Non-Defaulter
0	25	1	0	0	13	1	2	0.01029122	0.002722493	0.016313953	2	1	4	3	2	2	Non-Defaulter
1	33	1	0	0	13	1	3	0.003605576	0.000681539	0.003115512	2	1	4	3	2	2	Non-Defaulter
0	51	2	0	1	13	1	1	0.019429239	0.004667328	0.028220873	3	1	4	3	2	2	Non-Defaulter
0	53	2	0	1	13	1	1	0.024246405	0.005800174	0.035392215	3	1	4	3	2	2	Non-Defaulter
0	29	1	0	0	13	1	3	0.004155901	0.001816216	0.00601577	3	1	4	3	2	2	Non-Defaulter
0	56	2	0	1	13	1	1	0.009204292	0.003272439	0.011374451	3	1	4	3	2	2	Non-Defaulter
0	55	2	0	1	13	1	3	0.000890446	0.000376242	0.004044501	2	1	4	3	3	1	Non-Defaulter
0	29	1	0	0	3	1	1	0.000890446	0.000353357	0.003999184	2	1	4	3	3	1	Defaulter
1	33	2	4	1	13	1	1	0.004977739	0.001738404	0.005687225	3	1	4	3	2	2	Non-Defaulter
0	29	1	0	1	4	1	1	0.003517991	0.000925501	0.004554312	3	1	4	3	2	2	Non-Defaulter
0	40	2	0	1	13	1	3	0.006037807	0.001284808	0.010909956	2	1	4	3	2	2	Non-Defaulter
1	37	2	0	1	13	1	1	0.005495073	0.00190364	0.006355644	3	1	4	3	2	2	Non-Defaulter
0	33	1	0	0	13	1	3	0.000789577	0.000326351	0.003613994	2	1	4	3	3	1	Defaulter