ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF GRADUATE STUDIES

THE USE OF DATA MINING TO PREDICT THE LOAN REPAYMENT RISK: THE CASE OF OROMIA CREDIT AND SAVING SHARE COMPANY

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

By

Ketema Feyissa

September, 2018

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF GRADUATE STUDIES

THE USE OF DATA MINING TO PREDICT THE LOAN REPAYMENT RISK: THE CASE OF OROMIA CREDIT AND SAVING SHARE COMPANY

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATES OF ADDIS ABABAUNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

By

Ketema Feyissa

Advisor:

Wondwossen Mulugeta (PhD)

September, 2018

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF GRADUATE STUDIES

THE USE OF DATA MINING TO PREDICT THE LOAN REPAYMENT RISK: THE CASE OF OROMIA CREDIT AND SAVING SHARE COMPANY

By

Ketema Feyissa

September, 2018

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
Wondwossen Mulugeta(PhD)	**Advisor**	----------	----------
Tibebe Beshah (PhD)	**Chairperson**	----------	----------
Million Meshesha (PhD	**Examiner**	----------	----------

DEDICATION

This research work is dedicated to the Almighty GOD who is always help me in all the hard times and challenges of my life.

ACKNOWLEDGEMENTS

I am deeply indebted to all those who in their own way contributed to successful completion of this study. First and foremost I thank the almighty God, to whom all knowledge, wisdom and power belong for sustaining me in good health, sound judgment and strength to move on and complete my master's studies. I am heartily thankful to my advisors, Dr. Wondowssen Mulugeta, for their encouragement, guidance, constructive comments, support and for their help that enabled me to develop an understanding of the subject. I am deeply grateful to Oromia credit and saving share company workers and higher officials, for allow to collect and making the data available for this study. In addition, I would like to thank Mr. Shibiru Olika for their help, during the collection of the data and Mr. Bekele Hordofa for encouragement and support in all my endeavors.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AUC – Area Under Curve
ANN – Artificial Neural Network
ARFF – Attribute Relation File
BLAC – Branch Loan Approval Committee
BM – Branch Manager
CRISP-DM – CRoss-Industry Process for Data Mining
DM – Data Mining
KDP – Knowledge Discovery Process
KDD – Knowledge Discovery in Database
KM – Kilo Meter
MFI – Microfinance Institution
MLP – Multilayer Perception
MSEL – Micro and Small Entrepreneur Loan
M-BIIR – Mobile Money Service
OCSSCO – Oromia Credit and Saving Share Company Organization
ORCSDP – Oromia Rural Credit and Saving Scheme Development Project
OMFI – Oromia Microfinance Institution
OSHO – Oromia self Help Organization
ROC – Receiver Operating Curve
RNN – Recurrent Neural Network
SACO -Saving and Credit Cooperative
SGBL – Solidarity Group Based Loan
SCV – Common Separated Values
WEKA – Waikato Environment for Knowledge
WDEPL – Women Development Entrepreneur Program Loan

ABSTRACT

Data Mining is the process of extracting useful patterns from the huge amount of database and many data mining techniques are used for mining these patterns. Data mining is still a technology of having great expectations to enable the organizations to take more benefit of their huge data bases. Recently, one of the remarkable facts in microfinance institute is the rapid growth of data and this microfinance data is expanding quickly without any advantage to the organization for decision making. The main aim of this research work is utilizing of data mining by developing classification model in order to predict customer loan repayment behavior for loan risk that could help for better decision making and maximize the benefit of the microfinance from organizational datasets.

In this research the applicability of classification data mining techniques to implement customer loan repayment prediction model in Oromia credit and saving share company have been explored within the approach of CRISP-DM process model. After understanding business objectives of the organization, customer profile data are extracted, collected, cleaned, transformed, integrated and finally prepared for experimentation with the classification algorithm to develop a prediction model. The final dataset prepared for experimentation have 147,285 customer profile instances.

The findings of this study revealed all the models built from J48 Decision Tree classifier, Naïve Bayes classifier and Neural Network have high classification accuracy and are generally comparable in predicting customer loan repayment. However, comparison that is based on their performance accuracy suggests that the J48 model performs slightly better in predicting customer loan repayment with classification accuracy of 98.89%.

In this study the following attributes: customer follow up, purpose of loan, distance of customers from microfinance center and amount of loan disburse are the most interesting attributes in determining customer loan repayment prediction. The result of this study used efficiently to model and predict customer loan repayment. Based on the findings of the study, we recommend that microfinance institutions should adopt data mining to enhance their performance. The organizations need to make sure that there is enough data to analyze as well as assure quality of data. Organizations should ensure that the analysts are trained well and deduct the correct information which serves the purposes of the problem in the first place.

Key words: loan repayment risk, Microfinance, Oromia Credit and Saving Share Company Organization

CHAPTER ONE
INTRODUCTION

1.1 Background of the study

In Ethiopia, poverty is pervasive and it is argued that one of its causes is deprived access to credit and other microfinance services to be used for the purpose of working capital as well as investment. Hence, micro credit service is among the areas of priority in the fight against poverty and ensuring sustainable development [1].

In the view of this, the development of microfinance institutions in Ethiopia is a recent phenomenon. The proclamation, which provides for the establishment of microfinance institutions, was issued in July 1996. Since then, various microfinance institutions have legally been registered and started delivering microfinance services[2].

The Ethiopian economy is mainly dependent on agriculture and vulnerable to several internal and external shocks such as frequent draughts, high population growth, low investment, and volatile primary product prices. These and other factors have resulted in declining level of the economy with deteriorating living conditions. Since unemployment is the major problem in the Ethiopian economy, a lot of people were to join the informal sector. This sector is said to have a significant role in the creation of jobs and income generation for a large proportion of the population in Ethiopia. According to a paper compiled by the Ministry of Finance and Economic Development, the number of people earning their livelihood from the informal sector activities and small scale manufacturing industries is eight times larger than those engaged in the medium and large scale industrial establishments [3].But, one of the major challenges in the informal sector was acquiring financial resources. To overcome this problem, microfinance services were indicated as suitable solutions that meet the need of borrowers who need capital in small amount.

The delivery of financial services in Ethiopia has been viewed as an antipoverty tool because it helps the unemployed become employed, thereby increasing their income and consumption and reducing poverty. Improving financial access to the poor can facilitate economic growth by easing liquidity constraints in production, by providing capital to startup new production or adopt new technology and by helping producers assume production risk. As a result, the intervention in

microfinance will have a significant effect on addressing poverty reduction at a macro and micro levels [4].

Access to credit enables poor people to invest in a wide range of assets, better nutrition, improved health, access to schooling, a better roof on their homes and expansion of their small businesses [5]. Access to financial services can be essentially viewed as a fundamental driver of increased household income and resilience. Therefore access or outreach encompasses the ability of Microfinance Institutions (MFIs) to reach the poor and remote people. So, micro financial institutions are performing a key role in economic growth as they are mobilizing savings for productive investments through facilitating role in capital flows towards various sectors of the economy [6, 7]. It is also worth to note that commercial banks in most of the world economies are dominant as a financial institution providing installment loans compared to any other financial institution [11]. The significance of banks in an economy may not be eliminated as they are institutions, which provide liquidity for both lender and borrower [10].Because of this significance lenders have to evaluate the risk, which it face daily while lending [9]. Banks involves continuously in corporate governance to monitor, screen and recovery of loan for better performance of loan [9]. Performance of commercial banks influences economic growth positively. Despite their role in growth of economy, well performing commercial banks helps in economic acceleration while those poorly performing hampers the economic growth and enhance poverty in the country [12]. Hence performance is critical for financial organizations for achieving their objectives. To outreach in finance the low income group in Ethiopia, the proclamation number 40/1996 in which the legal framework that allow the establishment and operation of microfinance institutions was framed. Microfinance credit service has become one of the most prominent instruments in the development programs and strategies of the country.

Oromia Microfinance Institution (OMFI) is one of those MFIs that were established and Operating in the country. This institution is currently engaged in providing financial access to a large number of low income group and unemployed people in the region. Even though their establishment history are short, rising numbers of Micro financial institutions are introducing and expanding their offerings of loan for its customer . Central to the business strategy of every financial service company is the ability to retain existing customer and reach new prospective customers. The Microfinance industry in Oromia is growing rapidly and it has

become more and more important to keep pace with the growth of the industry through technological advancements and innovative ideas to market the organization to the masses. Portfolio of products offered by Microfinance providers has diversified, over the years, attracting more customers than ever. Accumulation of operational data inevitably follows from this growth in industry. There exists an increasing need to convert their data into a corporate asset in order to stay ahead and gain a competitive advantage.

Data mining is adopted to play an important role in these efforts. Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. It is also popularly referred to as knowledge discovery in databases (KDD). Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, etc.

The architecture of a typical data mining system may have the following major components [73]: database, data warehouse, or other information repository; their server which is responsible for fetching the relevant data based on the user's data mining request; knowledge base which is used to guide the search, or evaluate the interestingness of resulting patterns; data mining engine which consists of a set of functional modules for tasks; pattern evaluation module which interacts with the data mining modules and focus the search towards interesting patterns.

Data mining tasks have the following categories [73]:

Class description: It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms.

Association analysis: It is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data.

Classification: It is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data, and can be represented in forms like classification rules, decision trees, etc.

Cluster analysis: Clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to

begin with. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.

Outlier analysis: Outliers are data objects that do not comply with the general behavior of model of the data. Outliers may be detected using statistical tests or using distance measures.

Evolution analysis: It describes and models trends for objects whose behaviors changes over time. It normally includes time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

In the financial area, data mining has been applied successfully in determining the likely eligible candidate for loan disbursement, finding profitable customers, products, characterizing different Product segments [1]. All of these factors are challenging old ways of doing business and forcing Microfinance to consider reinventing themselves to win in the marketplace. In this aspect to find out good customers to disbursing loan is really a challenging issue in the microfinance era.

A gap between knowledge and knowing customer behavior of repayment to provide loan is observed from many studies. The Studies conducted so far shows that the microfinance has used traditional statistical analysis [23], which formulates a hypothesis and test the validity on the data set [23] and also identify their customer background from social information that the customer has in community. While data mining is applied on large dataset and its intention is not to test a hypothesis rather it tries to discover if there are hidden patterns and relationships from a large amount of data with no prior assumption. It further predicts the likelihood of a certain phenomenon from a previous collected and trained data. This research work then aims to know the customers loan repayment behavior by classifying them as standard customer, uncertain customer or loss customer to return the loaned money based on previously stored data and train data. This helps to predict the customer before providing loan.

1.2 Motivation

Data mining has attracted a great attention in the information industry and in society as a whole in recent years, due to wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from market analysis, fraud detection, to production control, disaster management and science exploration. The applications of data mining in microfinance

institutions help for identifications of customer loan repayment behavior and better decision making.

The motivation to select the organization is that microfinance institutions are those institutions that provide small loans typically for working capital for poor who could not get loan from the bank. Microfinance institutions have the advantage for poor in substituting collateral (guarantee) by group guarantees or compulsory savings which makes easy for poor to get access to repeated and larger loans based on repayment performance. These institutions are contributing the way for poor to get loan, use it and generate income which in turn can improve their life. This implies microfinance is contributing for the economic growth of the family which will in turn bring about economic growth of the country. In the view of services provided by microfinance, this institution should be sustainable in loan providing and loan collection. To support microfinance institution, Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets to solve the problems in microfinance and help the managements for batter decision making.

1.3 Statement of the problem

Microfinance has evolved as an approach to economic development intended to benefit low income women and men. It expanded enormously in 1990s [77]. Policy makers, donors, practitioners and academics underline the role of microfinance as a powerful tool for poverty alleviation and economic development. Alleviation of poverty and promotion of economic development can therefore be facilitated through providing credit to the poor. Despite its remarkable achievements, there remained several weaknesses in microfinance that need to be improved to ensure its continuous development and successful implementation. In order to assess the risk of loans repayment, all the mainstream lenders have their own loan repayment risk assessment systems which they have developed, but nowadays, there are many risks related to microfinance loans repayment, for the micro finance and for those who get the loans. The analysis of risk in microfinance loans repayment need understanding what is the meaning of loan repayment risk. These loans are typically unsecured but may also be secured in some cases [1].
According to Hunte [78], default problems destroy lending capacity as the flow of repayment declines, transforming lenders into welfare agencies, instead of a viable financial institution. It incorrectly penalizes creditworthy borrowers whenever the screening mechanism is not efficient.

Lending in Ethiopia has been both a controversial and a difficult matter. On the one hand, customers complain about lack of credit and the excessively high standards set by micro finance. On the other hand, microfinance has been suffering from losses on bad loans [1]. Lending inherently requires that the lender trust the borrower to repay the loan at a later date. For the lender to be able to trust the borrower, the lender must have means of screening out incompetent and untrustworthy borrowers. However, one chooses to put it, the microfinance problem is to distinguish between good and bad customers and its character. By good and bad customer-mean expected return and risk. Moreover, good and bad character refers to the borrower's previous data history.

The review shows that Ethiopian microfinance has trouble distinguishing the good from bad in both customer and character. This is partly due to intrinsic problems in Ethiopia, and partly due to their own methods. In this perspective the review indicates that solutions adopted by micro finance often seemed inefficient from the perspective of a profit-maximizing micro finance. Probably, this reflects both incomplete learning by micro finances about the most effective way to make loans, and internal incentive problems that micro finances have not solved [1].

In addition, the number of transactions in microfinance sector is rapidly growing and huge data volumes are available which represent the customers' behavior and the risks around loan are increased. Therefore, the datasets in the micro finance organization are so huge as mentioned before, and the simple statistical tools would not support to provide significant information and knowledge from these large collections of data as done by the sector professional for the purpose of early warning, planning, preparedness and decision making activities.

Data mining is concerned with finding hidden relationships present in business data to allow businesses to make predictions for future use [8]. Therefore, the primary goal of this research work is to predict the customer loan repayment of Oromia microfinance using application of data mining for better decision making activities.

1.4 Research Questions

Based on the problem discussion above, the purpose of this research work is to gain a better understanding of loan repayment related risks and applications of data mining to solve the problems of loan related risk in the Oromia microfinance.

In order to fulfill this purpose four research questions have been constructed as a foundation for this research work:

1. Which attributes of the customer is the most factor for not return the loan at the right time of agreement signed?
2. Can useful hidden patterns be extracted through data mining for the loan risk to assist the decision makers of the Oromia Microfinance?
3. To what extent the predictive model constructed in the study identify loan risk?

1.5 Objective of the study

1.5.1 General Objective

The main objective of this research work is to apply data mining by developing classification model in order to predict customer loan repayment behavior for loan risk that could help for better decision making and maximize the benefit of the microfinance.

1.5.2 Specific Objectives

In order to achieve the above general objective, the following specific objectives are identified:

- Review related literature in the area of loan risk to understand the customer loan repayment related risk.
- Study specific requirements for customer loan risk analysis.
- Prepare the existing data sets for better use of data mining algorithm and methods.
- Apply data mining to analyze customer data in microfinance.
- Develop a classification model using data mining techniques that classifies customer behavior to predict customer loan repayment behavior.
- Evaluate the applicability and the performance of the model under the study.
- Recommend on what should be done from the finding of the experiments.

1.6 Scope and limitations of the study

The scope of this research is restricted to model building, rule generation, result evaluation and discussion with stakeholders of the organizations and business domain experts, document organization, prototype development for rule generated and recommend on implementation of the models by making discussions within institutional stakeholders. This work will not cover the system implementation except prototype developed due to the need of integration of resources like people, business process and technology.

1.7 Significance of the study

With the continuous changing and development in the credit industry, credit products and services are becoming more and more important in the economy. The increasing demand and increasing competition resulting from new economic environment offer new opportunities, but also put forward new requirements of data mining technologies. Credit granting institutions pursue urgently cost savings and efficiencies. This has led them to use technology in their credit management process. This research work is used to propose best model that used to predict customer loan repay behavior for the organization to integrate into their system helps them easily identify to whom they provide loan or reject.

The result of this research support microfinance loan services in contributing to the smoothing out of peaks and troughs in income and expenditure thereby enabling the poor to cope with unpredictable shocks and emergencies.

The finding of this study can also help to identify problems to make new policy, monitoring and evaluating the activities for the microfinance and different concerned stakeholders.

The outcome of this research shall also be used as a benchmark for microfinance organizations as well as a source of methodological approach for studies dealing on the application of data mining on loan risk analysis as well as other similar areas.

1.8 Thesis Organization

This thesis consists of six chapters, the first chapters consists of background of the study, motivations for the initiation of the study, statement of the problem, research question, objective of the study, scope and limitation of the research, significance of the research. The second chapter dedicated with literature review of data mining: Data mining, knowledge discovery process, knowledge discovery process model, tasks of data mining, types of data mining system, challenges in data mining, data mining tools, applications of data mining in business area, ethics of data mining and related work. The third chapter of this thesis is devoted on research methodology and research design. The CRISP-DM processes model: understanding of the model, understanding of the data, preparation of the data, modeling, evaluation and deployments are discussed. Also data mining algorithms for customer loan repayment predictions: Naïve Bayes, decision tree and neural network and evaluation methods of the model are discussed

under this chapter. Chapter four is concerned with Business understanding; data understanding, data preparation and attribute selection are seen in this chapter. In chapter five, Experimentation: model development, model performance evaluation, model selection and prototype development. Chapter six, conclusion and recommendations are given.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

The development of information technology generated large amount of databases and huge data in various forms in different areas. The data can be simple numerical Figures and text documents, or more complex information like spatial data, multimedia data, and hypertext documents. From these available databases, mining knowledge, regularities or high-level information is essential to support decision making and predict future behavior [13]. To take complete advantage of data stored in files, databases, and other repositories, data retrieval is simply not enough. It requires a technique or powerful tool for analysis and interpretation of such data and that could help in decision-making. These technique or tool is data mining. Data mining is the extraction of hidden predictive and descriptive information from large databases or huge data [17]. It is a powerful technology with great potential to help organizations to focus on the most important information Win their data warehouses. Data mining technique also enables to predict future trends and behaviors, and helps organizations to make proactive knowledge-driven decisions and it can solve the problems that traditionally were too much time consuming like preparing databases for finding hidden patterns, finding predictive information that experts may miss. For this purpose different techniques and algorithms are used to accomplish the tasks of data mining.

This chapter presents a review of literature to provide brief understanding of data mining potential to discover hidden knowledge from huge data in database. It also provides a brief description of knowledge discovery process and knowledge discovery process model; it presents a review of different data mining methods, common tasks of data mining in Business, challenges in data mining and review articles, journals and researches which conducted in related area.

2.2 Data Mining

Data Mining or knowledge discovery in databases can be defined as an activity that extracts some new nontrivial information contained in large databases. According to Chung and Gray [28], data mining on the other hand, is the extraction of valid, novel, potentially useful and understandable correlation and useful patterns in existing data or data warehouse.

The goal of data mining is to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine learning, statistics and database technologies. This new discipline today finds application in a wide and diverse range of business, scientific and engineering scenarios.

The overall knowledge discovery process is outlined as an interactive and iterative process involving more or less the following steps: understanding the application domain, selecting the data, data cleaning and preprocessing, data integration, data reduction and transformation, selecting data mining algorithms, data mining, and interpretation of the results and using the discovered knowledge [29].

According to Silltow [15], in its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge - driven decisions and answer questions that were previously too time- consuming to resolve. Silltow [15] as mentions, the cores of data mining techniques are: association, classification, link analysis, sequence analysis, clustering, prediction, sequential patterns, decision trees, combinations, long- term (memory) processing.

Data mining has got its own challenges and Han [48]mentions these challenges to include: lack individual privacy; issue of data integrity whereby he states that data analysis can only be as good as the data that is being analyzed; the issue of cost; most databases are dynamic as databases usually change continually therefore getting representation data may be a challenge; databases may be huge and there may be difficulty in accessing data.

Newman [16], also argues that a lot of good can come from data mining. He highlights the top benefits of data mining as more money as a result of profits and investments; tapping into new markets; the principle of share and share alike which involves sharing of information that may be useful to similar organizations; learning from the past; help in competitor analysis where data mining helps companies to get information that they can use effectively to stand out from competition.

2.3 Knowledge Discovery Process (KDP)

The aim of researcher under this topic is to provide an overview of knowledge discovery process to obtain different knowledge from descriptions and comprehensive comparison of several main models along with discussions of issues associated with their implementation. In addition before one attempt to extract useful knowledge from a data, it is important to understand the overall approach. Even knowing many algorithms used for data analysis is not sufficient for a successful data mining (DM) researches [32]. Therefore this topic focuses on describing and explaining the process that leads to find new knowledge. The process defines a sequence of steps (with eventual feedback loops) that should be followed to discover knowledge (pattern) in data as shown in the following Figure 2.1.

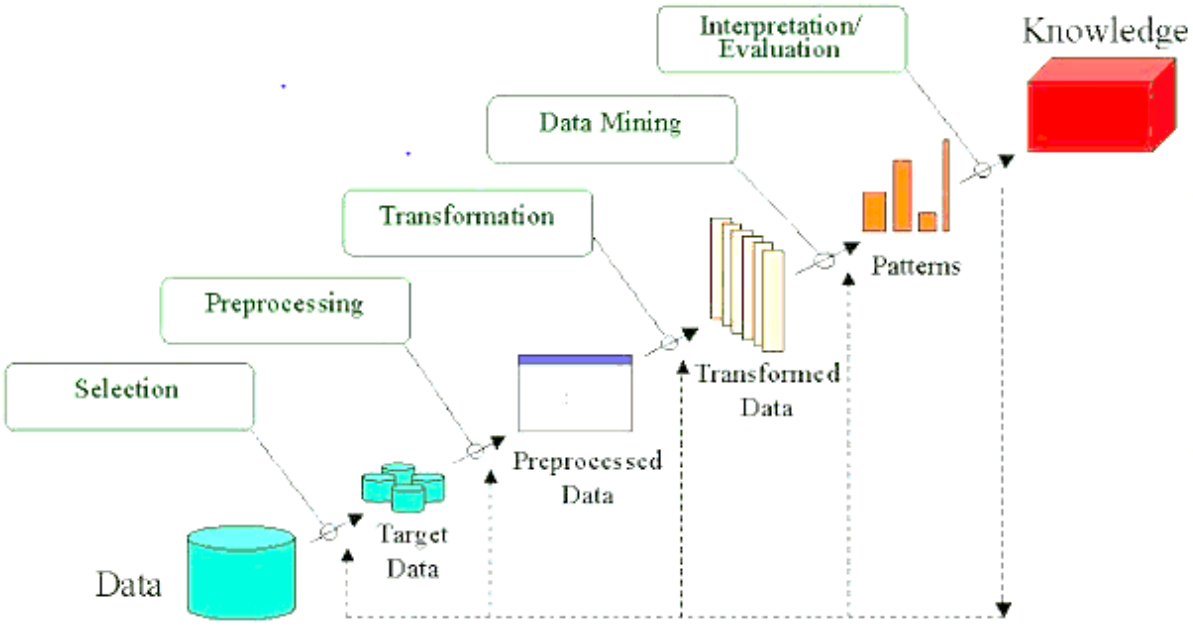

Figure: 2.1 Knowledge discoveries Process [68].

There is some confusion about the term data mining and knowledge discovery, and knowledge discovery in data bases, so it is better to understand the difference between terms. The term knowledge discovery in a data base or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the high level application of particular data mining methods.

Note, however, that many researchers and practitioners use Data mining as a synonym for knowledge discovery; DM is also one step of KDP [32].

Data mining: is defined as an application of specific algorithm for extracting patterns from data. DM is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing [32].

The knowledge discovery process (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to non-database sources of data, although it emphasizes databases as a primary source of data. As shown in Figure 2.1, it consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain.

2.4 Knowledge discovery process Model

The KDP model consists of a set of processing steps to be followed by practitioners' when executing a knowledge discovery project. The model describes procedures that are performed in each of its steps. It is primarily used to plan, work through, and reduce the cost of any given project [34].

Since the 1990s, several different KDPs have been developed. The initial efforts were led by academic research but were quickly followed by industry. The first basic structure of the model was proposed by Fayyad et al. [34], and later improved/modified by others. The process consists of multiple steps that are executed in a sequence. Each subsequent step is initiated upon successful completion of the previous step, and requires the result generated by the previous step as its input. Another common feature of the proposed models is the range of activities covered, which stretches from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results. All the proposed models also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process.

The main differences between the models described below are the number and scope of their specific steps. A common feature of all models is the definition of inputs and outputs. Typical inputs include data in various formats, such as numerical and nominal data stored in databases or flat files; images; video; semi-structured data, such as XML or HTML; etc. The output is the generated new knowledge; usually described in terms of rules, patterns, classification models, associations, trends, statistical analysis, etc. [34].

2.4.1 Academic Research Model

The efforts to establish a KDP model were initiated in academia. In the mid-1990s, when the DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain. The two process models developed in 1996 and 1998 are the nine-step model by Fayyad et al. [34] and the eight-step model by Anand and Buchner [34]. Below we introduce the first of these, which is perceived as the leading research model.

The Fayyad et al. [34], KDP model consists of nine steps, which are outlined as follows:

- Developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.
- Creating a target data set. Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.
- Data cleaning and preprocessing. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.
- Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.
- Choosing the data mining task. Here the data miner matches the goals defined in Step one with a particular DM method, such as classification, regression, clustering, etc.

- Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
- Data mining. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.
- Interpreting mined patterns. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.
- Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge.

2.4.2 Industrial Model (CRISP-DM Process Model)

Industrial Models quickly followed academic efforts. The two representative industrial models are the five-step model by Cabena et al. [34] and the industrial six-step CRISP-DM Model, developed by a large consortium of European companies [34].Starting from the 1990s different organizations were having different models for their data mining tasks. The CRISP-DM (Cross Industry Standard Process for Data Mining) was created for the purpose of having a standard for the processes involving data mining projects. According to Cios et al. [34], the development of CRISP-DM model enjoys strong industrial support and becomes the leading industrial model that the researcher discussed under this topic.

The CRISP-DM process model provides an overview of lifecycles of a data mining project. It contains phases of a project, their respective tasks, and relationships between these tasks. According to this model the life cycle of data mining project consists of six phases shown in Figure2.2. The sequence of the phases is not rigid moving back and forth between the different phases is possible. The outcome of each phase determines which phase has to be performed next. The arrows indicate the most important and frequent dependencies between the phases. The outer circle in the Figure shows the cyclical nature of data mining. It does not end once the solution is deployed. The lessons learned during the process and from the deployed solution can trigger new [37].

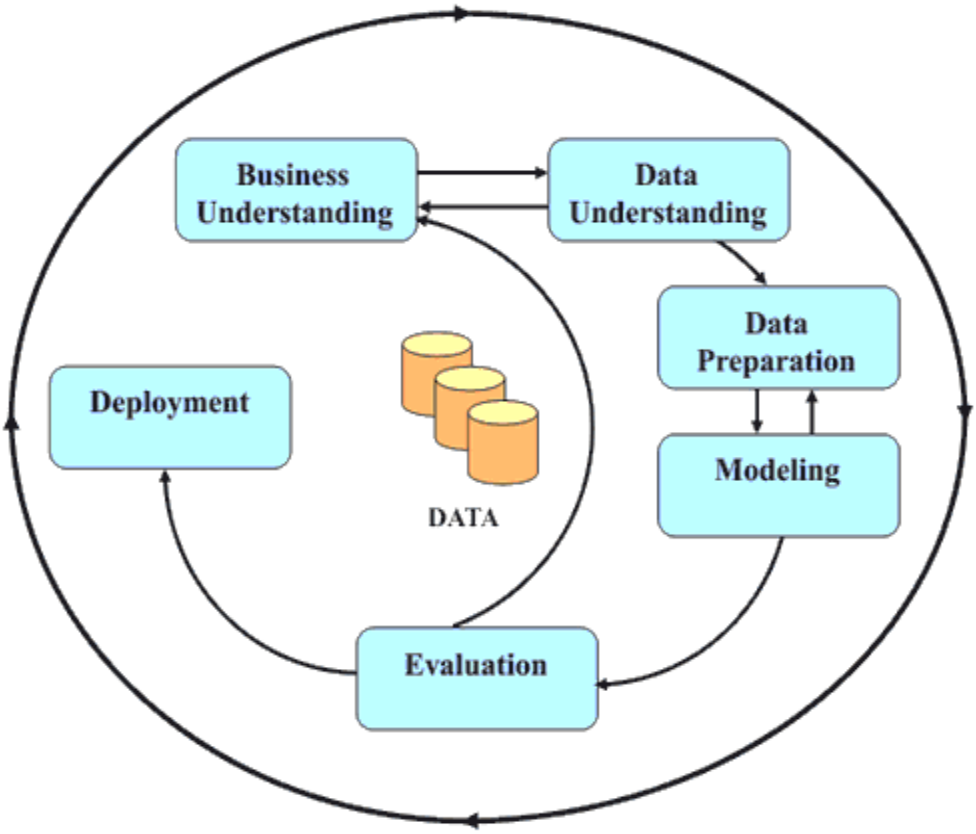

Figure 2.2: The CRISP – DM Process Model [69].

The six phases of CRISP-DM model are explained as follows:

Business Understanding: This is the initial phase which focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data understanding: This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets.

Data preparation: This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data.

Modeling: At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their

parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary.

Evaluation: at this stage the model built before processing the final deployment is thoroughly evaluated and the steps executed to construct the model are reviewed to be sure it properly achieves the business objectives. At the end of this phase a decision on the use of the data mining results should be reached.

Deployment: In this phase now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating a report or as complex as implementing a repeatable KDP.

The model is characterized by an easy-to-understand vocabulary and good documentation. It divides all steps into sub steps that provide all necessary details. It also acknowledges the strong iterative nature of the process, with loops between several of the steps. In general, it is a very successful and extensively applied model, mainly due to its grounding in practical, industrial, real-world knowledge discovery experience [34].

2.4.3 Hybrid Models (a six step KDP)

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of academic and industrial model [40].This is the six-step KDP model shown in Figure 2.3, which is developed by Cios et al (2006).It is developed by adopting CRISP-DM into academic researchers in a way that provide more general, research-oriented description of steps, introduces mining step instead of modeling step, has more detailed feedback mechanisms and the last step is modified to use the discovered knowledge to be applied in other domains [38].

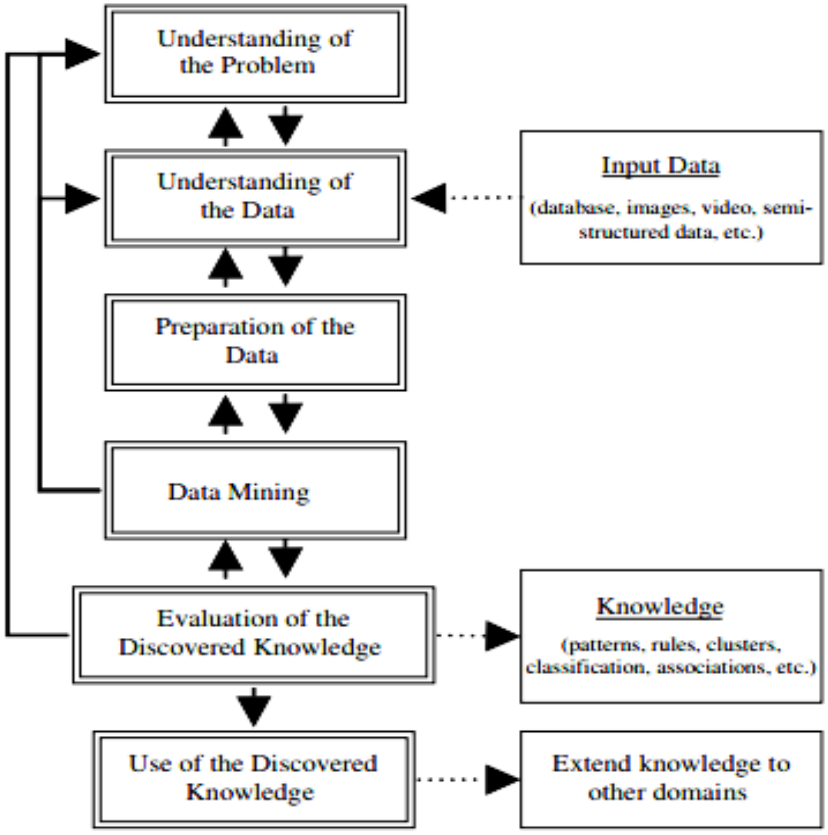

Figure 2.3: The six-steps KDP Model [70].

Understanding of the problem domain: This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

Understanding of the data: This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

Preparation of the data: This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data. The end results are data that meet the specific input requirements for the DM tools selected in first step.

Data mining: Using different data mining methods and algorithms to derive knowledge from the processed data.

Evaluation of the discovered knowledge: includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts. Only approved models are retained.

Use of the discovered knowledge: This is the final step which consists of where and how to use the discovered knowledge. The application area of the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and entire project is documented. Finally, the discovered knowledge is deployed.

2.4.4 Comparisons of Process Model

To understand and interpret the KDP models described above, summary is given in the following table 2.1 to show side by side comparison on information about the domain of origin (academic or industry), the number of steps, and a comparison of steps between the models, notes, and application domains.

Model	Fayyad et al.	Anand and Buchner	Cios et al.	Cabena et al	CRISP-DM
Domain of Origin	Academic	Academic	Hybrid (Academic and Industry)	Industry	Industry
Steps of the model	9	8	6	5	6
List of steps	1. Discovering and understanding the application domain	1. Human resource identification 2. Problem specification	1.Understanding of the problem domain 2.Understanding of the data	1. Business objective determination 2. Data preparation	1. Business understanding 2. Data understanding 3. Data preparation

Model	Fayyad et al.	Anand and Buchner	Cios et al.	Cabena et al	CRISP-DM
	2. Creating target dataset 3. Data cleaning and preprocessing 4. Data reduction and projection 5. Choosing the data mining task 6. Choosing the data mining algorism 7. Data mining 9. Interpreting mined pattern 9.Consolidating discovered knowledge	3. Data prospecting 4. Domain knowledge elicitation 5.Methodology identification 6. Data preprocessing 7. Pattern discovery 8. Post-processing	3. Preparation of the data 4. Data mining 5. Evaluation of the discovered knowledge 6. Use of the discovered knowledge	3. Data mining 4. Analysis of results 5.Assimilation of knowledge	4. Modeling 5. Evaluation 6. Deployment
Note	The most popular and most cited model; provides detailed technical description with respect to data analysis, but lacks business aspect	Provides detailed breakdown of the initial steps of the process; missing step concerned with application of the discovered knowledge and project documentation	Draws from both academic and industrial models and emphasizes iterative aspects; identifies and describes several explicit feedback loops	Business oriented and easy to comprehend by non-data mining specialists; the model definition uses non-data mining jargon	Uses easy to understand vocabulary; has good documentation; divides all steps into sub steps that provide all necessary details
Supporting software reported	Commercial system MineSet™	N/A	N/A	N/A	Commercial system Clementine R
application domains	medicine, engineering, production, e-business, software	Marketing, sales	Medicine, software	Marketing, sales	Medicine, engineering, marketing, sales

Table: 2.1 Comparisons of Process Models [70].

Most models that we have seen follow a similar sequence of steps, while the common steps between the five are domain understanding, data mining, and evaluation of the discovered knowledge. The nine-step model carries out the steps concerning the choice of DM tasks and algorithms late in the process.

The other models do so before preprocessing of the data in order to obtain data that are correctly prepared for the DM step without having to repeat some of the earlier steps. In the case of Fayyad's model, the prepared data may not be suitable for the tool of choice, and thus a loop back to the second, third, or fourth step may be required. The five-step model is very similar to the six-step models, except that it omits the data understanding step. The eight-step model gives a very detailed breakdown of steps in the early phases of the KDP, but it does not allow for a step concerned with applying the discovered knowledge. At the same time, it recognizes the important issue of human resource identification. This consideration is very important for any KDP, and we suggest that this step should be performed in all models. We emphasize that there is no universally "best" KDP model. Each of the models has its strong and weak points based on the application domain and particular objectives. For the purpose of this research work the researcher used CRISP-DM process model.

2.5 Tasks of data mining

Data mining is highly helpful to collect relevant information from various sources of data. So it is highly helpful to achieve particular task. The goal of data mining effort is normally either to create a descriptive mining model or predictive mining model [29]. A descriptive model presents the data in concise form which is essentially a summary of the data points, finds patterns in the data and understands the relations between attributes represented by data. The descriptive model includes the tasks such as Clustering, Association rule, Summarizations, and sequence discovery. The predictive model works by making a prediction about values of data, which uses known results found from different datasets [30]. The predictive data mining model includes classification, prediction, regression and analysis of time series.

2.5.1 Predictive Modeling

Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of

records at large [31]. This approach frequently employs decision tree or neural network-based classification algorithms. The common characteristics of classification tasks are as supervised learning, categories dependent variable and assigning new data to one of a set of well-defined classes. Classification technique is used in customer segmentation, modeling businesses, credit analysis, and many other applications.

Regression: Regression is another Predictive data-mining model is also known as supervised learning technique. This technique analyzes the dependency of some attribute values, which is dependent upon the values of other attributes mainly, present in same item. In the regression techniques target value are known. For instance, we can predict the child's behavior based on family history.

Time Series data analysis: Time-series database uses sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time interval such as hourly, daily, weekly. A sequence database is any database that consist sequence of ordered events, sometimes having concrete notions of time [32].

Prediction: This technique discovers the relationship between independent variables and variable. The prediction is to predict a future state, rather than a current one [32]. Its applications include obtaining forewarning of natural disaster s (flooding, hurricane, snowstorm, etc.), epidemics, stock crashes, etc. As another instance, the sales volume of computers accessories can be forecasted based on the number of computers sold in the past few months.

2.5.2 Descriptive Modeling

Clustering: is a collection of similar data objects to gather. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. Clustering can be considered as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but this method is expensive so clustering can be used as preprocessing approach for attribute subset selection and classification.

Summarization: Summarization is referred as the abstraction or generalization of data. The summarization technique maps data into subsets with simple descriptions. The summarized data

set gives general overview of the data with aggregated information. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation.

Association: The Association technique is used to extract the relationships between attributes and items. In this technique, the presence of one model implies the presence of another model i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of data mining; association rules are useful for analyzing and predicting customer behavior. They also play an important role in shopping basket data analysis, product clustering, and catalog design and store layout. The association rules are also building by programmers can be used to build programs capable of machine learning.

Sequence Discovery: Uncovers correlation among data. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence. Knowing the tasks of data mining is helpful for having knowledge of its task and to select the one that is more related and useful for this researcher work.

2.6 Types of data mining system

Data mining researches generate different varieties of data mining system due to the reason that data mining is the contribution of diversified discipline. Therefore it is necessary to provide a clear classification of the data mining systems to help users identify those that best matches their problems easily. Data mining systems can be categorized on the following classification criteria [39].

Based on the kind of database mined:

Here the types of database are further classified by data model and types of data so that each types use different data mining techniques. When data model is considered as criteria for classification: we can have relational, transactional, object-oriented, or data warehouse mining systems. And when the type of data is taken as criteria for classification the different data mining systems are spatial data, multimedia data, time-series data, text data and World Wide Web data mining systems.

Based on the kind of knowledge mined:

Data mining systems can be categorized according to the kinds of knowledge they mine such as characterization, discrimination, association, classification, clustering, etc. An advanced data mining system should facilitate discovery of knowledge at multiple level of abstraction. That means generalized knowledge (high level), primitive-level knowledge (raw data level) or multiple level knowledge abstraction.

Based on the kind of techniques utilized:

Data mining systems can also be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved example autonomous systems, interactive exploratory systems, query-driven systems, or based on the methods of data analysis employed example database oriented, data warehouse oriented, machine learning, statistics, visualization, pattern recognition, neural networks and so on. Mining knowledge from data is now a day is very important for organization for customer loan repayment prediction and decision making.

2.7 Challenges in Data Mining

High dimensional spare data, uncertain data, incomplete data, Variety and Heterogeneity, Scalability, Speed/Velocity, Accuracy and Trust, privacy preservation, Inter-activeness are the challenges which are faced by data mining [34].

High dimensional spare data: High dimensional spare data significantly deteriorate the reliability of the models derived from the data [35], Common approaches are to employ dimension reduction or future selection [36] to reduce data dimension or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

Uncertain Data: Uncertain data are special type of data reality where each data field is subjected to some random/error distribution. For example, each recording location of GPS system is represented by mean value and variances to indicated expected errors; for uncertain data major challenge is that each data item is represented as sample distributions but not as single value, so most of the existing data mining algorithms cannot be directly applied. Error aware data mining utilizes the mean and variance values with respect to each single data item to build

Naive Bayes model for classification. Similar approaches have also been applied for decision tree or database queries.

Incomplete Data: Incomplete data [35] refers to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values. For example: dropping some sensor node readings to save power for transmission. While most modern data mining algorithms have in-built solutions to handle missing values, data imputation is an established research field that seeks to impute missing values to produce improved models.

Variety and Heterogeneity: - Variety is the characteristic of huge data in data mining. Data is collected from many sources that may generate data on its own or may contribute to it. It means there is variety as well as heterogeneity in the data. These types of data are interconnected, interrelated and inconsistent. Data may be structured which may fit in the database, semi-structured which may partially fit into the database or unstructured may not fit in the database. So mining hidden patterns and knowledge from these heterogeneous data is a challenge before the data scientists.

Scalability: - huge data requires high scalability of its data management and mining tools. This data may contain knowledge and information which may not be possible to collect from conventional data.

Speed/Velocity: - The data mining algorithm must be able to finish the processing within a particular time. The data must be accessed and processed quickly otherwise the results obtained from these will become worthless. The factors which affect the speed of data mining depends include data access time and efficiency of mining algorithms. Parallelism may be included to increase the accessing speed.

Accuracy and Trust: - With an increase in the amount of data, there are many data sources from where data is collected, which may not be verifiable or trustable. Therefore the accuracy and trust of data source becomes an issue, which may propagate to results as well. Therefore data validation and accuracy becomes an important issue for discovery of useful information.

Privacy: - It is an important issue that data must be kept private and invisible to others. Data mining requires personal information in order to produce results. Social media contains all of the

information about an individual and information can be mined from that information and then the privacy disappears. So this is the issue which must be considered by data scientists and tools must be built by taking this consideration into account.

Inter-activeness:- It means feature of the data mining system that allows user interaction by using feedback/guidance. It is an important issue as it allows the users to visualize, evaluate and interpret intermediate and final results.

2.8 Data mining tools

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of one special tool becomes increasingly difficult for each potential user. This decision making process can be supported by criteria for the categorization of data mining tools.

There are different open source data mining tools available to support the tasks of data mining. The tools available in the beginning were invoked from command prompt which requires programming knowledge. Then with contributions of research communities now a days we can get many tools having interactive graphical user interface [41].

The known general-purpose tools for data mining, machine learning and statistics include WEKA, R, Tanagra, KNIME and Orange. From which WEKA and R are the most popular one's [41]. The choice differs with the implementation language, interactive graphical user interface, documentation, stability, performance, visualization of the models.

R is a language and environment for statistical computing and graphics [41]. This tool needs scripting knowledge to manipulate. While WEKA is the best known open-source machine learning and data mining tool [41]. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. It can be accessed through java programming or command line interface or using a WEKA explorer graphical user interface. In this study WEKA 3.8.2 version is used.

2.9 Applications of Data Mining

Various fields use data mining technologies because of fast access of data and valuable information from vast amount of data. Data mining technologies have been applied successfully

in financial sectors like customer segmentation and profitability, high risk loan applicants, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimizing stock portfolios, cash management and forecasting operations, most profitable Credit Card Customers and Cross Selling and so on. This application of DM techniques on financial data can contribute to the solution of financial sector problems and facilitate the decision making process.

Some of the applications areas of data mining technology include the following:

2.9.1 Banking

The banking industry across the world has undergone tremendous changes in the way the business is conducted. With the recent implementation, greater acceptance and usage of electronic banking, the capturing of transactional data has become easier; due to this the volume of such data has grown considerably. It is beyond human capability to analyses this huge amount of raw data and to effectively transform the data into useful knowledge for the organization [22].

Data Mining can help by contributing in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases [21]. By using data mining to analyses patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers will be likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable [22] are some applications of data mining in banking.

2.9.2 Marketing/Retail

One of the most widely used areas of data mining for the banking industry is marketing. The bank's marketing department can use data mining to analyze customer databases. Data mining carry various analyses on collected data to determine the consumer behavior with reference to product, price and distribution channel. The reaction of the customers for the existing and new products can also be known based on which banks will try to promote the product, improve quality of products and service and gain competitive advantage. Bank analysts can also analyze the past trends, determine the present demand and forecast

the customer behavior of various products and services in order to grab more business opportunities and anticipate behavior patterns. Data mining technique also helps to identify profitable customers from non-profitable ones [23]. The data mining techniques can be used to determine that how customers will react to adjustments in interest rates, the risk profile of a customer segment for defaulting on loans [24].

2.9.3 Risk Management

Data mining is widely used for risk management in the banking industry. Bank executives need to know whether the customers they are dealing with are reliable or not. Offering new customers credit cards, extending existing customers lines of credit, and approving loans can be risky decisions for banks if they do not know anything about their customers [22].

Banks provide loan to its customers by verifying the various details relating to the loan such as amount of loan, lending rate, repayment period, type of property mortgaged, demography, income and credit history of the borrower. Customers with bank for longer periods, with high income groups are likely to get loans very easily. Even though, banks are cautious while providing loan, there are chances for loan defaults by customers. Data mining technique helps to distinguish borrowers who repay loans promptly from those who don't [23].

Bank executives by using Data mining technique can also analyze the behavior and reliability of the customers while selling credit cards too. It also helps to analyze whether the customer will make prompt or delay payment if the credit cards are sold to them [23].

Credit scoring, in fact, was one of the earliest financial risk management tools developed. Credit scoring can be valuable to lenders in the banking industry when making lending decisions. Data mining can also derive the credit behavior of individual borrowers with installment, mortgage and credit card loans, using characteristics such as credit history, length of employment and length of residency. A score is thus produced that allows a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a high risk of default. By knowing what the chances of default are for a customer, the bank is in a better position to reduce the risks [22].

2.9.4 Fraud Detection

Another popular area where data mining can be used in the banking industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of data mining more fraudulent actions are being detected and reported. Two different approaches have been developed by financial institutions to detect fraud patterns. In the first approach, a bank taps the data warehouse of a third party and use data mining programs to identify fraud patterns. The bank can then cross-reference those patterns with its own database for signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the bank's own internal information. Most of the banks are using a hybrid approach consisting of data mining for data analysis and internal information [22].

One system that has been successful in detecting fraud is Falcon's 'fraud assessment system'. It is used by nine of the top ten credit card issuing banks. The data mining techniques will help the organization to focus on the ways and means of analyzing the customer data in order to identify the patterns that can lead to frauds [25].

2.9.5 Customer Relationship Management

In the era of cut throat competition the customer is considered as the king. Data mining can be useful in all the three phases of a customer relationship cycle: Customer Acquisition, Increasing value of the customer and Customer retention [26]. Customer acquisition and retention are very important concerns for any industry, especially the banking industry [22].

Today customers have wide range of products and services provided by different banks. Hence, banks have to cater the needs of the customer by providing such products and services which they prefer. This will result in customer loyalty and customer retention.

Data mining techniques helps to analyze the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known which will help the banks to perform better and retain its customers [23].

2.9.6 Customer Segmentation

Customer segmentation is a popular application of data mining with established customers. A segmentation project starts with the definition of the business objectives and ends with the

delivery of differentiated marketing strategies for the segments. There are many different segmentation types based on the specific criteria or attributes used for segmentation. Specifically, customers can be segmented according to their value. The type of segmentation used depends on the specific business objective [76].There are various segmentation types according to the segmentation criteria used. Particularly, customers can be segmented according to their value, socio-demographic and life-stage information, and their behavioral, need/attitudinal, and loyalty characteristics. The type of segmentation used depends on the specific business objective and your target. Different criteria and segmentation methods are appropriate for different situations and business objectives. In behavioral segmentation, customers are grouped by behavioral and usage characteristics. Although behavioral segments can be created with business rules, this approach has inherent disadvantages. It can efficiently handle only a few segmentation fields and its objectivity is questionable as it is based on the personal perceptions of a business expert. Data mining on the other hand can create data-driven behavioral segments. Clustering algorithms can analyze behavioral data, identify the natural groupings of customers, and suggest a solution founded on observed data patterns. Provided the data mining models are properly built, they can uncover groups with distinct profiles and characteristics and lead to rich segmentation schemes with business meaning and value.

2.10 The Ethics of data mining

The use of data, particularly data about people, has serious ethical implications in data mining. We researcher of data mining techniques must act responsibly by making ourselves aware of the ethical issues that surround with particular application. When applied to people, data mining is frequently used to discriminate like who gets the loan, who gets the special offer, and so on.

It is widely accepted that before we make a decision to collect personal information we need to know how it will be used and what it will be used for, what steps will be taken to protect its confidentiality and integrity, what the consequences of supplying or withholding the information are, and any rights of redress they may have.

The potential use of data mining techniques may stretch far beyond what we were conceived when we collect original data. This creates a serious problem. It is necessary to consider under what condition the data we collected and for what purpose we use that data. Surprising things may emerge from data mining. When we presented with data, it is needed to know who is

permitted to have access to it, for what purpose we collected, and what kind of conclusions is legitimate to draw from it. The ethical dimension raises tough questions for those involved in practical data mining. It is necessary to consider the norms of the community that is used to dealing with the kind of data involved, standards that may have evolved over decades or centuries but ones that may not be known to the information specialist. In addition to community standards for the use of data, logical and scientific standards must be adhered to when drawing conclusions from it. If we come up with conclusions, we need to attach caveats to them and back them up with arguments other than purely statistical ones. The points we consider is that data mining is just a tool in the whole process. We are the one who take the results, along with other knowledge, and decide what action to be taken. Of course, anyone who uses advanced technologies should consider the wisdom of what they are doing. So, the conclusion that can be drawn is like when data mining is conducted in a particular domain, we should seriously address the question whether the objective of mining is useful for the mankind and is there anything non-ethical hidden behind the rules extracted from the data mining process.

2.11 Related Works

A number of researches have been conducted using different data mining tools and techniques on various areas within the country and abroad. In this section the most importantly related works on application of data mining are reviewed.

Wakgari [43] has conducted a research on application of data mining techniques for effective customer Relationship management of microfinance. The aim of the researcher was to investigate the potential application of data mining tools and techniques to support customer relationship management of WISDO microfinance. The researcher after applying different data preprocessing use 9550 datasets for the purpose of experimentation. In this research the experiments were compared against each other based on parameters such as the number of leaves, size of the tree and accuracy of the classifier is used. The researcher use J48 algorithm to build classification model. The researcher found experiment six is best because it shows best accuracy 78.5026% and level with a compromised tree size and number of leaves in a better way than any other experiment.

Sousa [47] has conducted a research on credit analysis of credit union using data mining. The focus of the researchers is to develop and evaluate data mining models that classify and predict the behavior of cooperative member's behavior in honoring their obligations. The researcher

used decision tree and artificial neural network to develop the model, after applying data preprocessing on the cooperative datasets by dividing data into training and testing sets. The researcher compared the result which the accuracy of the decision tree in the simulation is 97.07% and 95.58% with the ANN, the decision tree based C4.5 algorithm obtains a result that is statistically similar to that of the model that was based on the MLP artificial neural network. The researchers argue that the knowledge discovery process and the use of the models based on the data mining developed provide the cooperative union with practical advantages. Finally the researchers propose the future studies: using different databases to validate the credit analysis model; using other data mining techniques; using hybrid models, combining different techniques to improve classification and predictive performance; investment analysis, evaluating the type of error and the financial impact that the model has on the cooperative's profitability; and evaluation of discrepant cases, particularly those cases involving the variable "capital", to check for the existence of new patterns and classifications.

Sara Worku [44] has conducted a research on application of data mining technology for credit risk assessment in Addis credit and saving institution. The researcher used WEKA tool to find important variable that risk on the loan provided and to predict the pattern which borrowers' likely bad or good borrower by developing classification model. The researcher collects customer data manually from business plan and legend form from four different branches a total of 4000 customer records. After data collection the researchers applies different preprocessing techniques on the data and build the classification model using J48, PART and NAÏVE BAYES algorithm. After conducting of a total of nine experiments the researcher evaluates the performance of the algorithm and identifies the algorithm that perform better performance accuracy which is 99.825% is PART Rule induction algorithm.

In [46], the target was to conduct a research on developing prediction model of loan risk in banks using data mining. The researchers develop a new model for classifying loan risk in banking sector by using data mining. The model has been built using data from banking sector to predict the status of the loans. For this purpose the researchers use three algorithms J48, BayeNet and Naïve Bayes to build the prediction model by using WEKA application tool, the model has been implemented and tested. The researchers divided 1000 instances of original data set into two groups, training which represents 80% from all data and testing set which represents 20% of the data set. In this research the researchers applied three different classification algorithms and

made a comparison between them according to their accuracy in classifying the data correctly. In this research work is the J48 algorithm performed the accuracy of 78.39%.

Belachew Reganie [45] has conducted a research on application of data mining techniques for customer segmentation and to cluster instances. The researcher after applying different data preprocessing techniques use 1357 records and six attributes for clustering and analysis purpose. The researcher use simple K-means algorithm to cluster the instance and different experiments were conducted with different k-values and seed size. The experiment conducted using segmentation at K=5 and seed size 10 with five clusters selected as best customers segment model in the institution. The researchers also classify the instances into simple groups based on result obtained through clustering. And also the researcher using decision tree classifier J48 algorithm conduct different experiments and build a model by using 10-fold cross validation test mode which registered accuracy of 99.95% is selected as best model for prediction purpose. The researcher believes that, application of other data mining techniques with different algorithms with new tool is potential research area to improve performance of customer relationship management in the institution.

Jia et al. [42] has conducted a research on a comparison of data mining methods in microfinance which focuses on applying data mining technology in developing a loan risk assessment system. In order to produce the result, different algorithms like Decision tree induction algorithm, clustering and Naïve Bayes and the data mining tool called WEKA are used. WEKA is a collection of algorithms for solving real-world data mining problems. Using WEKA, different data mining methods are able to be applied to extract data patterns. Finally, the paper introduces a case study of applying different data mining technologies in developing a loan risk assessment system for a sub-prime lender.

After reviewing the above literatures the researcher was motivated to work on a classification model that Predict customer loan repayment cases based on patterns generated from Oromia credit and saving Share Company collected datasets. This research work different from previous related work in various ways: attribute selection methods and attribute used, the size of data used, there is no research conducted on the organization on customer loan repayment prediction, the data used is the data of group lending, the objective of the research is on customer loan repayment prediction, the methodology used to conduct the research and the algorithm used.

CHAPTER THREE

RESEARCH METHODOLOGY

3. Overview

Data mining is a creative process which requires a number of different skills and knowledge. Currently there is no standard framework in which to carry out data mining projects. This means that the success or failure of a data mining project is highly dependent on the particular person or team carrying it out and successful practice can not necessarily be repeated across the enterprise. Data mining needs a standard approach which will help to translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience.

3.1 Research Methods

Methodology means the steps or procedures that the researcher follows to achieve the objectives stated. It is a road map that shows the direction how the research is going to be done to reach the end. In this research, the researcher aimed to use CRISP-DM process model to achieve the stated objectives. This is because the model has been widely applied for data mining studies. Besides, it is flexible to account for differences i.e. for different business problems and different data. And also it is open-source and industry standard data mining processes model. Accordingly, this study has followed the following methodologies in order to develop classification model and that predict customer loan repayment risk behavior in microfinance historical data.

3.2 Review of related literature

Detail literature review on data mining techniques and assessment will be conducted on works which are related to identify customers loan repayment risk modeling. Various books, journals, magazines, articles, and papers from the internet pertaining to potential of data mining for loan risk analysis and successful data mining applications in prediction and classification of customer loan risk repayment behavior have been reviewed. From the review of this work, best approaches will be used for overall thesis work.

3.3 Research Design

This research is an experimental research because experimental analysis is conducted on microfinance data collected from Oromia microfinance S.C. In this chapter the methods, techniques and tools used to conduct the research are discussed in detail based on the steps of CRISP-DM data mining process model selected to guide the entire process for this research.

As it is discussed in chapter two data mining process models are developed for purely academic purposes and also for industrial purposes. The CRoss Industry Standard Process Model is developed for industrial purpose and the result of this research based on the organization need in future it may be implemented. The process model adopted to undertake this research is the CRISP-DM due to the reason that this model describes each of the knowledge discovery process steps in a better way and it is flexible since it has a feedback mechanism in more steps than the other Process model.

The CRISP-DM model has six steps that are: Understanding of the Business, understanding of the data, preparation of the data, modeling, and evaluation of the Model and Deployment of the model. The research is designed based on steps of this process model and discussed under here.

3.3.1 Understanding of the Business

The Cross-industry standard process for data mining (CRISP-DM) is the dominant process framework for data mining. In the first phase of a data mining project, before we approach data or tools, we define what we are out to accomplish and define the reasons for wanting to achieve this goal.

The business understanding phase includes four tasks [69]:

- **Identifying our business goal**: In this task, the first thing we must do in any project is to find out exactly what we are trying to accomplish. Many data miners have invested time on data analysis, only to find that their management wasn't particularly interested in the issue we were investigating. We must start with clear understanding of a problem that we wants to address, the business goals, Constraints (limitations on what we may do the kinds of solutions that can be used, when the work must be completed, and so on); impact (how the problem and possible solutions fit in the business).

Deliverables for this task include three items (usually brief reports focusing on just the main points):

Background: Explain the business solution that drives the project. This item, like many that follow, amounts only to a few paragraphs.

Business Goals: Define what the organization intends to accomplish with the project. This is usually a boarder goal than we, as a data miner, can accomplish independently.

Business success criteria: Define how the results will be measured. Try to get clearly defined quantitative success criteria. If we must use subjective criteria (hint: terms like gain insight or get a handle on imply subjective criteria), at least get agreement on exactly who will judge whether or not those criteria have been fulfilled.

- **Assessing the situation**: In this task, this is where we get into more detail on the issues associated with our business goals. Now we will go deeper into fact-finding, building out a much fleshier explanation of the issues outlined in the business goals task.

Deliverables for this task include five in depth reports:
Inventory of resources: A list of resources available for the project. These may include people (not just data miners, but also those with expert knowledge of the business problem, data managers, technical support, and others), data, hardware and software.

Requirements, assumptions, and constraints: Requirements will include a schedule for completion, legal and security obligations, and requirements for acceptable finished work. This is the point to verify that we will have access to appropriate data.

Risk and contingencies: identify causes that could delay completion of the project, and prepare a contingency plan for each of them.

Terminology: create a list of business terms and data mining terms that are relevant to our project and write them down in a glossary with definitions, so that everyone involved in the project can have a common understanding of those terms.

Costs and benefits: prepare a cost-benefit analysis for the project. Try to state costs and benefits in money terms. If the benefits don't significantly exceed the costs, stop and reconsider this analysis and our project.

- **Defining data mining goals**: In this task, reaching the business goal often requires action from many people, not just the data miner. So now; we must define our title part within the bigger picture. If the business goal is to reduce customer attrition, for example, our data mining goals might be to identify attrition rates for several customer segments, and develop models to predict which customers are at greatest risk.

Deliverables for this task include two reports:

Data mining goals: define data mining deliverables, such as models, reports, presentations, and processed datasets.

Data mining success criteria: Define the data mining technical criteria necessary to support the business support the business criteria. Try to define these in quantitative terms (such as model accuracy or predictive improvement, identify the person who makes the assessment.

3.3.2 Understanding of the Data

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

The data understanding phase includes the following four tasks. These are described blow [69]:

- **Gathering data**: In this task we just set goals and define a data –mining plan. Every step of the plan depends on having the right data. Better make sure that we really have the right data. Just one deliverable exists for this task: the initial data collection report. In our report, we need to verify that we have acquired the data or at least gain access to the data, tested the data access process, and verified that the data exists. We need to load data into any tools that we will be using for data mining to verify that tools are compatible with the data.

We may do a lot of work to assemble the data we need before we can write the report. First, we will make our plan, as follows:

Outline data requirements: Create a list of the types of data necessary to address the data mining goals. Expand the list with details such as the required time range and data formats.

Verify data availability: Confirm that the required data exists, and that we can use it. If some of the data we want is unavailable, decide how we will address that issue. We consider alternatives such as: substituting with an alternative data source, narrowing the scope of the project and gathering new data.

Define selection criteria: Identify the specific data sources (databases, files, documents and so on.) we will use. Within those sources, specify the tables, fields, and case ranges that are relevant to our project.

Once we have gone through these steps, we must actually obtain the data. At this stage, import the data into the data mining platform we will be using for the project to confirm that it is possible to do so and that we understand the process. In the course of this trial we may discover software or hardware limitations we had not anticipated, such as: Limits on the number of cases or fields, or on the amount of memory we may use, inability to read the data formats of our sources, Difficult dealing with imperfections in the data (for example we might encounter products that won't import or analyze datasets).

Finally, summarize the gathering process in a report. The report should describe our requirements, and explain in some detail exactly what we have gathered and from what sources. Here we confirm that we have actually obtained the data and that it is compatible with our data mining platform. If we have run into difficulties, we will explain what we were and how we have addressed those (using alternative sources, revising plans, changing formats).

- **Describing data:** In this task, now we have data, prepare a general description of what we have. The deliverable for this task is the data description report. In it, we describe the source and formats of the data, the number of cases, the number and descriptions of the fields, and any other general information that may be important. We also make a brief evaluation of the suitability of the data for our data mining goals. For example, verify that the data includes the fields that we expect and need to be there and sufficient cases for analysis.
- **Exploring data:** In this task, we examine the data more closely. For each variable, we look at the range of values and their distributions. We will use simple data manipulation and basic

statistical techniques for further checks into the data. Data exploration supports several purposes like: get familiar with data, spot signs of data quality problems and set the stage for data preparation steps.

The deliverable for this task is the data exploration report. It is the place to document any hypotheses or initial findings that we have developed during data description report, including distributions, summaries, and any signs of data quality problems.

- **Verifying data quality:** In this task, we have the data and we have examined it, and now we have to determine whether it is good enough to support our goals. We will often have some quality problem to address yet still be able to move forward, but at times the data quality is so poor that it cannot support our plan and we will have to look for alternatives. Some of the worst data problems would include: The data we need doesn't exist (did it never exist, or was it discarded? can this data be collected and saved for future use?), it exist, but we can't have it (can this restriction be overcome?), we find severe data quality issues (lots of missing or incorrect values that can't be corrected).

The deliverable for this task is the data quality report. This summarizes the data that we have, minor and major quality issues that we have found, and possible remedies for quality problems or alternatives (such as using an alternative data source). If we are facing any reality serious data quality issues and can't identify an adequate solution, we may have to recommend reconsidering goals or plans.

3.3.3 Preparation of the data

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

The data preparation phase includes five tasks. These are selecting data, cleaning data, constructing data, integrating data and formatting data.

Selecting data: now we will decide which portion of the data that we have is actually going to be used for data mining. The deliverable for this task is the rationale for inclusion and exclusion. In it, we will explain what data will, and will not, be used for further data mining work.

We explain the reasons for including or excluding each part of the data that we have, based on relevance to our goals, data quality, and technical issues- such as limits to the number of fields or rows that our tools can handle, or the suitability of the data formats for our needs.

Cleaning data: is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data[49]. These raise the data quality to the level required by the selected analysis techniques.

The data that we have chosen to use is unlike to be perfectly clean (error free). We will make changes, perhaps tracking down sources to make specific data corrections, excluding some cases or individual cells (items of data), or replacing some items of data with default values or replacements selected by a more sophisticated modeling technique. We may choose to use only subsets of the data for all or some of our data mining work.

The deliverable for this task is the data cleaning report, which documents, in excruciating detail, every decision and action used to clean our data. This report should cover and refer to each data quality problem that was identified in the verify data quality task in the data understanding phase of the process. We report should also address the potential impact on results of the choices we have made during data cleaning.

Constructing data: We to drive some new fields (for example, use the delivery date and the date when a customer placed an order to calculate how long the customer waited to receive an order), aggregate data, or otherwise create a new form of data.

Deliverables for this task include two reports:

Derived attributes: a report that describes what new fields (columns) we have constructed, how we did it, and why.

Generated records: a report that describes what new cases (rows) we have constructed, how you did it, and why.

Integrating data: Our data may now be in several disparate datasets. We will need to merge some or all of those disparate datasets to gather to get ready for the modeling phase. The deliverable for this task is the merged data. (And it would not hurt to document how the merge was performed).

Formatting data: Data often comes to us in formats other than the ones that are most convenient for modeling. (Format changes are usually driven by the design of our tools.) So convert these formats now. The deliverable for this task is our reformatted data. (And a little report describing the changes we have made would be a smart thing to include.)

We should end the data preparation phase of the data mining process with a dataset ready for modeling and a thorough report describing the dataset.

3.3.4 Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

Select modeling Techniques: as the first steps in modeling, select the actual modeling technique to be used. If a tool was selected in business understanding (phase one), this task refers to selecting the specific modeling technique, example building decision tree or neural network.

Generate testing design: Prior to build a model, a procedure needs to be defined to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, if the test design specifies that the dataset should be separated into training and test sets, the model is built on the training set and its quality estimated on the test set.

Build model: The purpose of building models is to use the predictions to make more informed business decisions. The most important goal when building a model is stability, which means that the model should make predictions that will hold true when it's applied to yet unseen data. Regardless of the data mining techniques being used, the basic steps used for building predictive

models are the same. The model set first needs to be split into three components: the training set, the test set and the evaluation set.

Each of these components should be totally separate; that is, they should not have any records that are in common since each set performs a distinct purpose. Models are created using data from the past in order for the model to make predictions about the future. This process is called training the model. In this step, the data mining algorithms find patterns that are of predictive value. Next, the model is refined using the test set. The model needs to be refined to prevent it from memorizing the training set. This step ensures that the model is more general (i.e. stable) and will perform well on unseen data. Next, the performance of the model is estimated using the evaluation set. The evaluation set is the entirely separate and distinct from the training and test sets. The evaluation set (or hold out set) is used to access the expected accuracy of the model when it is applied to data outside the model set. Finally, the model is applied to the score set. The score set is not ore-classified and is not part of the model set used to create the data model. The outcomes for the score set are not known in advance. The final model is applied to the score set to make predictions. The predictive scores will, presumably, be used to make more informed business decisions.

Assess Model: Interpret the models according to your domain knowledge, your data mining success criteria and your desired test design. Judge the success of the application of modeling and discovery techniques technically, and then contact business analysts and domain experts later in order to discuss the data mining results in the business context. This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project.

At this stage we should rank the models and assess them according to the evaluation criteria. We should take the business objectives and business success criteria into account as far as we can here. In most data mining projects a single technique is applied more than once and data mining results are generated with several different techniques.

Model assessment - Summarize the results of this task, list the qualities of our generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.

Revised parameter settings - According to the model assessment, revise parameter settings and tune them for the next modeling run. Iterate model building and assessment until we strongly believe that we have found the best model(s and document all such revisions and assessments.

3.3.5 Evaluation

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why the model is deficient. It compares results with the evaluation criteria defined at the start of the project.

A good way of defining the total outputs of a data mining project is to use the equation:

$$\text{Result} = f(\text{Models}, \text{Findings})$$

we define the total output of the data mining project as not just the models, but also the findings which can be defined as anything (apart from the model) that is important in meeting objectives of the business.

Evaluate Results: previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this chosen model is deficient. Another option of evaluation is to test the models on test application if time and budget permits.

Review process: At this point the resultant model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining project in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of data mining, the review process takes on the form of a quality assurance review.

Determine next step: according to the assessment result and the process review, the analyst decides how to proceed at this stage. The analyst needs to decide whether: to finish the project and move on to deployment, to initiate further iteration or to set up new data mining projects.

3.3.6 Deployment

Creation of the model is generally not the end of the research work. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the client to understand up front what actions will need to carried out to make use of the model created.

Plan deployment: To deploy the data mining results into the business, this task takes the evaluation results and develops a strategy for deployment. If a general procedure was identified to create the relevant models, this procedure is documented here for later deployment.

Plan monitoring and maintenance: Monitoring and maintenance are important issues if the data mining result becomes part of the day to day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessary long periods of incorrect usage of data mining results. To monitor the deployment of the data mining results, the project needs a detailed plan on the monitoring process. This plan takes into account the specific type of deployment.

Produce final report: at the end of the project, the project leader and the team write up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experience (if they have not already been documented as an ongoing activity) or it may be a final and comprehensive presentation of the data mining results.

A review project: assess what went right and went wrong, what was done well and what needs to be improved.

3.4 Data mining Algorithms for Customer loan repayment predictions

Classification algorithms that are mostly used in predictions basing use historical data. Classification is a class prediction technique, which is supervised in nature. This technique possesses a systematic approach to build classification models from an input dataset. The researcher aims to use decision tree classifier, artificial neural network and Naïve Bayes

classifier due to it is supported by most of literatures and the researcher is familiar with it . Each of techniques employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability; i.e. models that accurately predict the class labels of previously unknown records.

In classification, the dataset is divided into two sets, namely the training set (dependent set) and a test set (independent set). The training set consisting of records whose class labels are known. The training set is used to build a classification model, which is subsequently applied to the test set, which consists of records with unknown class labels. The following are classification algorithms that are used in customer loan repayment behavior prediction.

3.4.1 Naïve Bayes Classification Techniques

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive [56] term for the underlying probability model would be ‘independent feature model’. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about ‘4’ inch diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers [56]. Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is

outperformed by more current approaches, such as boosted trees or random forests [57]. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Naïve Bayes algorithm

Naïve Bayes classification algorithm is one of the classification techniques that does not use any rule like that of decision tree. The foundation of Naïve Bayes is the probability theory. The straightforward of calculating probability is to look for the frequent event and classify the unseen instance to the frequent occurring event. Using more complex probability types is better prediction of the unseen event. To obtain the prior probability we divide the frequency of most frequent event by total number of instance. The probability of an event occurring if we know that an attribute has a particular value (or that several variables have particular values) is called the conditional probability. This is also called posterior probability since it calculates the probability after it gains information whereas priori probability calculates before it obtains information.

Given variables:

$X = \{x_1, x_2, x_3 \text{ ---}, x_d\}$, we want to construct the posterior probability for the event

C_j *among* a set of possible outcomes $C = \{c_1, c_2, c_3, \text{---}, c_n\}$.

In a more familiar Language, X is the predictors and C is the set of categorical level present in the dependent variable.

Using Bayes' Rule [48].

$$p(C|x_1, \text{---}, x_d) = \frac{p(C)p(x1,---,xd|C)}{P(x1,---,xd)} \quad \text{---------- ---------- (3.1)}$$

Where $P(C_j|x_1, \text{---}, x_d)$ is the posterior probability of class membership, i.e., the

Probability that X belongs to C_j.

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability:

$p(C, x_1, ---, x_d) = P(C)\, p(x_1|C)\, P(x_2|C,x_1)\, p(x_3|C, x_1,x_2)$ --- $p(x_d|C,x_1,x_2,x_3,---,x_{d-1})$ ---------- (3.2)

The naïve conditional independence assumptions come into play: assume that each feature xi is conditionally statistical independent of every other feature x_j for $j \neq i$.

This means that:

$p(x_i|C, x_j) = p(x_i|C)$

for $i \neq j$, and so the joint model can be expressed as

$p(C|x_1, ---, x_d) = p(C)p(x_1|C)p(x_2|C)P(x_3|C)$ ----- $P(c)\, P(x_{n/}c_n)$

$= p(C) \prod_{i=1}^{d} p(xi|C)$ ---------- ---------- (3.3)

This means that under the above independence assumptions, the conditional distribution over the class variables C can be expressed like this:

$p(C|x_1, --- , x_d) = \frac{1}{Z} P(C) \prod_{i=1}^{d} p(xi|C)$ ---------- (3.4)

Where Z (the evidence) is a scaling factor dependent only on $x_1, ---, x_d$, i.e., a constant if the values of the feature variables are known.

Finally, we can label a new case F with a class level C_j that achieves the highest posterior probability:

Classify $(F_1, ---, F_d) = \text{argmax}\, p(C = c) \prod_{i=1}^{d} p(xi = Fi|C = c)$.c ---------- (3.5)

3.4.2 Decision Tree Classification Techniques

Han and Kamber [58], defined decision tree as a flowchart like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label as shown in the Figure 3.1. The topmost node in a tree is the root node.

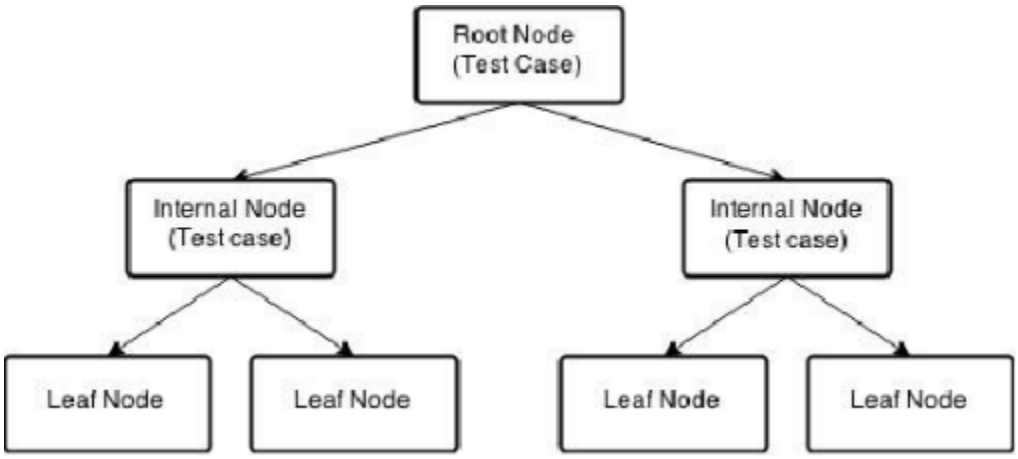

Figure: 3.1 simple decision tree structure

In order to classify an unknown sample, the attribute values are tested against the decision tree. A path is traced from the root to leaf node that holds a class prediction for the sample. At each level of the tree, the appropriate value would be compared and a decision on which direction to go would be made. This means by navigating the decision tree one can assign a value or a class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until leaf node reached.

Attribute selection:

Not all attributes are equally important in classifying given dataset with respect to some target class [17]. Therefore, the issue of which attribute should be considered first, second, third etc. is the basic thing to address in the tree construction step. According to Han and Kamber [17], most decision tree induction tools adopt attribute selection measure known as information gain to determine the precedence of the test attributes among the candidate attributes in the data set.

According to Han and Kamber [17] information gain of an attribute A is compared as:

$$\text{Gain (S, A)} = \text{Entropy(S)} - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \text{Entropy (Si)}$$

With: $\{S_1, \text{----- } S_i \text{ -----}, S_n\}$ = Partition of S according to value of attribute A

n = number of attribute A

$|S_i|$ = number of cases in the partition S_i

|S| = total number of cases in S

Entropy(S) is a measure in information theory that characterizes impurity of a collection S. If the target attributes takes on n different values, then the entropy S relative to this n wise classification is defined as:

$$\text{Entropy (S)} = -\sum_{i=1}^{n} \text{Pilog2Pi}$$

Where pi is the proportion/probability of S belonging to class i. logarithm in base 2 because entropy is a measure of the expected encoding length measured in bits.

Entropy used for:

- When node belongs to only one class, the entropy will become zero.
- When disorder of dataset is high or classes are equally divided then entropy will be maximal
- Helps in making decision as several stages

J48 classification algorithm

J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created to the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. J48 creates a decision node higher up in the tree using expected value of the class.

In Classification the process of building a model of classes from a set of records that contain class labels. J48 Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found [60]. This algorithm generates the rules for the prediction of the

target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable [59]. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Pruning:

Generating a decision to function best with a given of training data set often creates a tree that over-fits the data and is too sensitive on the sample noise. Such decision trees do not perform well with new unseen samples.

We need to prune the tree in such a way to reduce the prediction error rate. Pruning [61] is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The dual goal of pruning is the reduction complexity of the final classifier as well as better predictive accuracy by the reduction of over-fitting and removal of sections of a classifier that may be based on noisy or erroneous data.

3.4.3 Artificial Neural networks classification Techniques

An artificial neural network (ANN), often just called a *neural network* (NN), is a mathematical model or computational model based on biological neural networks, in other words, it is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [49]. ANNs imitate the learning process of the human brain and can process problems involving non-linear and complex data even if the data are imprecise and noisy. These techniques are being successfully applied across an extraordinary range of problem domains.

ANNs have powerful pattern classification and pattern recognition capabilities through learning and generalize from experience. ANNs are non-linear data driven self-adaptive approach as opposed to the traditional model based methods. They are powerful tools for modeling, especially when the underlying data relationship is unknown. ANNs can identify and learn correlated patterns between the input datasets and corresponding target values. After training, ANNs can be used to predict the outcome of new independent input data.

Neural Network Topologies:

Feed forward neural network: is an artificial neural network where inter connections between the units does not form a cycle [52], as shown in the Figure 3.2.

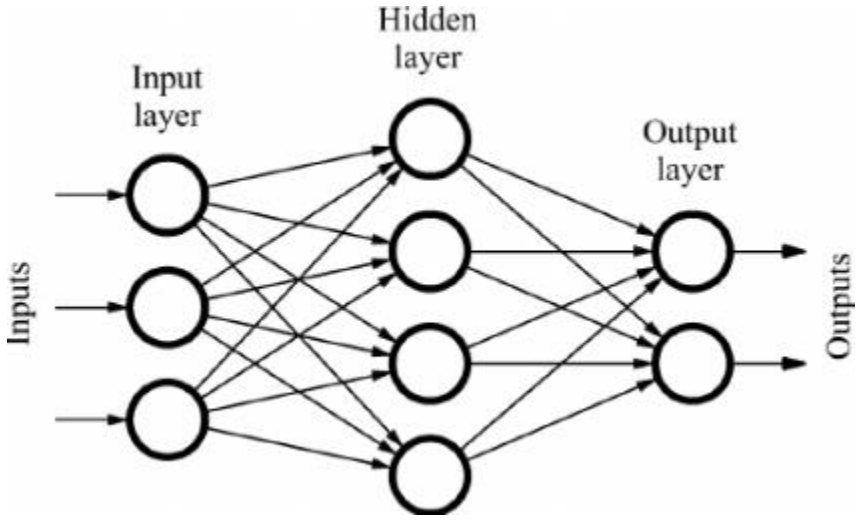

Figure: 3.2 feed forward neural network [52].

The feed forward neural network was the first and simplest type of artificial network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (and to the output nodes. There are no cycles or loops in the network layer [52].

Recurrent network: Recurrent neural networks that do contain feedback connections as shown in the Figure 3.3. Contrary to feed forward networks, recurrent neural networks (RNs) are models with bi-directional data flow. While a feed forward network propagates data linearly from input to output, RNs also propagate data from later processing stages to earlier stages.

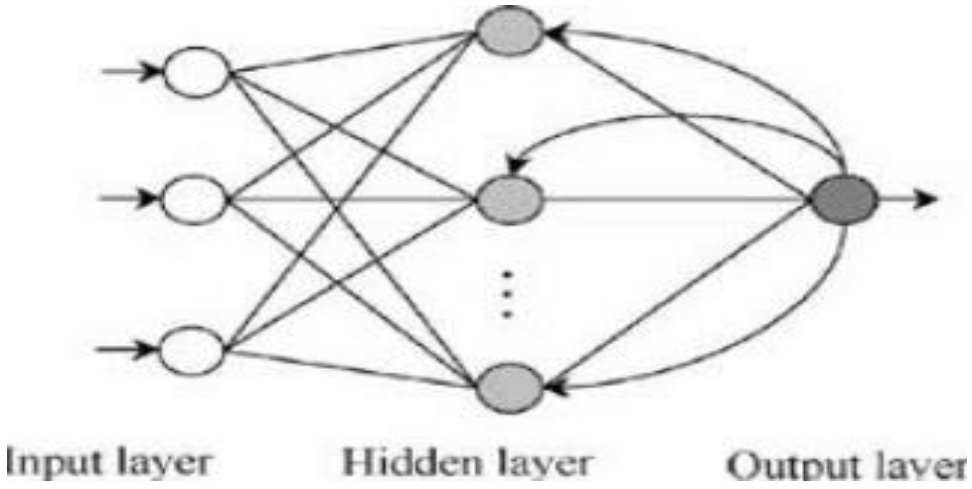

Figure: 3.3 recurrent neural networks [54].

Multilayer perception algorithm

A multilayer perceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training [53, 54]. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable [55].

Multi-layer networks use a variety of learning techniques, the most popular being back-propagation. Here, the output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques, the error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to some state where the error of the calculations is small. In this case the network has learned a certain target function. To adjust weights properly applies a general method for non-linear optimization that is called gradient descent. For this, the network calculates the derivative of the error function with respect to the network weights, and changes the weights such that the error decreases (thus going downhill on the surface of the error function). For this reason, back-propagation can only be applied on networks with differentiable activation functions.

One of the advantages of using neural networks is that they are quite robust with respect to noisy data. Because the network contains many nodes (artificial neurons), with weights assigned

to each connection, the network can learn to work around these uninformative (or even erroneous) examples in the dataset. However, unlike decision trees, which produce intuitive rules that are understandable to non-specialists, neural networks are relatively opaque to human interpretation. Also, neural networks usually require longer training times than decision trees, often extending into several hours.

3.5 Evaluation Methods for Classification Model

Evaluating the performance of a data mining technique is a fundamental aspect of machine learning. Evaluation method is the yardstick to examine the efficient and performance of any model. The evaluation is important for understanding the quality of the model or technique, for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set or model or techniques. There are several criteria for evaluating models for different tasks and other criteria that can be important as well, such as the computational complexity or the comprehensibility of the model. The most widely used measures for evaluating the performance of the techniques used for carrying out different data mining tasks (Classification, association rule mining and clustering) are discussed under here.

Holdout Method and Random Subsampling:

In this method, the given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set. The training set is used to derive the model. The model's accuracy is then estimated with the test set .The estimate is pessimistic because only a portion of the initial data is used to derive the model. Random subsampling is a variation of the holdout method in which the holdout method is repeated k times. The overall accuracy estimates are taken as the average of the accuracy obtained from each iteration.

Cross-Validation:

In K-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or 'folds,' D1, D2,...,Dk, each of approximately equal size. Training and testing is performed k times. In iteration i, partition Di is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets D2, … , Dk collectively serve as the training set to obtain a first model, which is tested on D1; the second iteration is trained on subsets D1, D3, … , Dk and tested on D2; and so on. Unlike the holdout

and random subsampling methods, here each sample is used the same number of times for training and once for testing. For classification, the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data [72].

Leave-one-out: is a special case of k-fold cross-validation where k is set to the number of initial tuples. That is, only one sample is 'left out' at a time for the test set. In stratified cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data [72].

In general, stratified 10-fold cross-validation is recommended for estimating accuracy (even if computation power allows using more folds) due to its relatively low bias and variance.

Bootstrap:

Unlike the accuracy estimation methods just mentioned, the bootstrap method samples the given training tuples uniformly with replacement. That is, each time a tuple is selected, it is equally likely to be selected again and re-added to the training set. For instance, imagine a machine that randomly selects tuples for our training set. In sampling with replacement, the machine is allowed to select the same tuple more than once [72].

Confusion Matrix:

A confusion matrix is a simple performance analysis tool typically used in supervised learning. It is used to represent the test result of a prediction model. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class [71]. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

In the table 3.1 below, a confusion matrix is shown, for which, the various values and related equations are described. Few of these equations are very relevant for performance analysis.

Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	A	B
	Positive	C	D

Table: 3.1 Confusion Matrixes [72]

The entries in the confusion matrix have the following meaning in the context of a data mining problem:

- A is the number of correct predictions that an instance is negative (positive),
- B is the number of incorrect predictions that an instance is positive (negative),
- C is the number of incorrect of predictions that an instance negative (positive),
- D is the number of correct predictions that an instance is positive (negative).

Several standard terms are defined for the two class matrix:

The accuracy (AC): is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$\text{AC} = \frac{A+D}{A+B+C+D}$$

The recall or true positive (TP): rate is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{TP} = \frac{D}{C+D}$$

The false positive (FP): rate is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$\text{FP} = \frac{B}{A+B}$$

The true negative (TN): rate is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$\text{TN} = \frac{A}{A+B}$$

The false negative (FN): rate is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$\text{FN} = \frac{C}{C+D}$$

Finally, precision (P): is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$\text{P} = \frac{D}{B+D}$$

The above concepts for a two-class problem can be extended to a multi class problem by focusing one of the classes as positive at a time and the rest as negative. The average of these parameters like precision, recall etc. for individual classes becomes the final values of the entire model.

Receiver Operating Curve (ROC):

ROC graphs are constructed by plotting the true positive rate against the false positive rate. A number of regions of interest in a ROC graph can be identified. The diagonal line from the bottom left corner to the top right corner denotes random classifier performance, that is, a classification model mapped onto this line produces as many false positive responses as it produces true positive response [71].

An ROC curve demonstrates several things [71]:

- It shows the tradeoffs between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The area under the curve is a measure of text accuracy.

The Figure 3.4 shows three ROC curves representing excellent, good, and worthless tests plotted on the same graph. The accuracy of the test depends on how well the test separates the group being tested into positive and negative cases. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system: A sample plot is shown in Figure 3.4 below [71].

- 0.90-1 = excellent (A)
- 0.80-0.90 = good (B)
- 0.70-0.80 = fair (C)
- 0.60-0.70 = poor (D)
- 0.50-0.60 = fail (F)

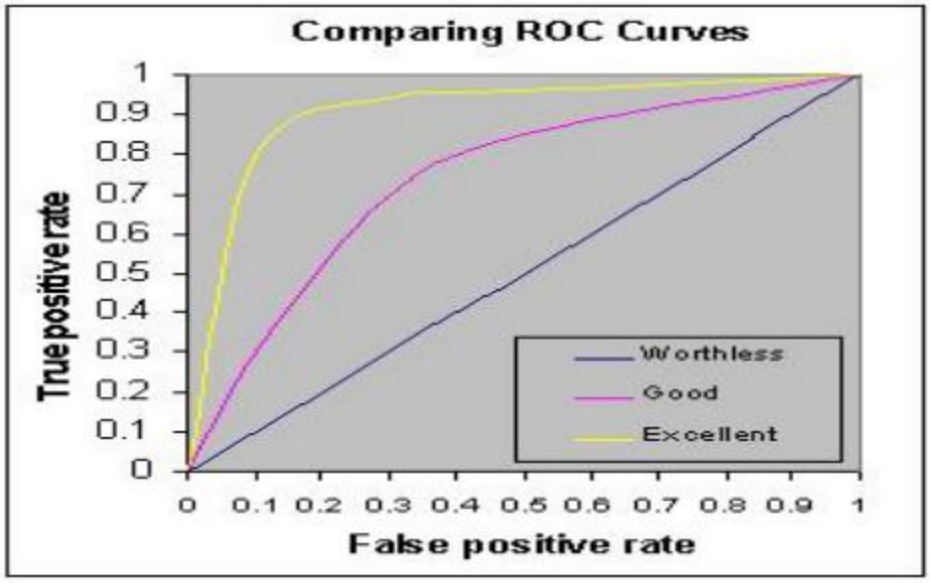

Figure: 3.4 ROC Curve Characteristics [71].

Making predictions has become an essential part of every business enterprise and scientific field of inquiry. Receiver operating characteristic (ROC) curves are useful for assessing the accuracy of predictions.

CHAPTER FOUR

BUSINESS UNDERSTANDING, DATA UNDERSTANDING AND DATA PREPARATION

4.1 Business Understanding

This study is based on a data collected from Oromia credit and Saving Share Company. So, to understand the customer loan process, the researcher has observed closely the work within the organization administration and loan department. In addition, real time observation of the business process was performed to gain an insight how the microfinance institution functions. This used to understand which data should later be analyzed, and how, it is important for data mining researchers to fully understand the business for which we are finding solution.

Therefore, Oromia credit and saving share Company (OCSSCO) is a microfinance institution which provide a reliable sources of financial support and assistance. It is a registered and licensed Microfinance institution operating in Oromia National Regional State. It was initiated on June 1995 as a project under Oromo Self Help Organization (OSHO) with the name of Oromia Rural Credit and Saving Scheme Development Project (ORCSDP) and undertakes its operation under the mandate of the mother organization. OCSSCO was established in 1996 based on the commercial code of Ethiopia and proclamation No. 40/1996 by five shareholders (Oromia National Regional State, Oromo Self Help Organization, Dinsho Enterprise, Oromia Development Association and One Natural Person). OCSSCO registered and commenced its formal operational activities with Head Office in Addis Ababa and four branch offices namely Kuyu, Shashemene, Hetosa and Sinana Dinsho in 1997 [74]. Soon, the operational successes in the four districts and high demand from government and people sides pushed the company to expand its outreaches to all corners of the region. As a result, now the company has been able to make its products accessible in all districts of the region by opening 139 full-fledged branches. OCSSCO has an objective to improve the socio-economic conditions of the low-income people through providing working finance at affordable interest rate and encouraging saving habit of the people.

The main service provided by Oromia credit and saving company is services of credit. The aim of providing credit is to increase the productivity of farmers and pastoralist of the region, operating the job opportunities for graduates and youth who not have jobs and to increase the

benefit of women. This in turn contributes its part for the accomplishment of government policies and strategies that mainly aimed at poverty alleviation.

4.1.2 Mandates of Oromia microfinance

As one of the microfinance institutions in Ethiopia, OCSSCO had been operating under the mandate defined by Proclamation No. 40/1996 in 2009, Proclamation No.40/1996 was replaced by new Microfinance Business Proclamation 626/2009 and the operation of the company presently is bound by this proclamation [74].

Besides the countries rules and regulations of microfinance institutions, the founder of OCSSCO also confers the following mandate that governs the operation of OCSSCO at its establishment.

- Provide credit for small farmers and others engaged in small and micro enterprise activities in cash or in kind.
- Accept savings as well as demand and time deposits.
- Perform transfer of payments which is effective only in Ethiopia.
- Purchase financial instruments such as Treasury bills, the objective of which is to generate income.
- Acquire, maintain and transfer mobile and immobile property including business buildings for carrying on business.
- Provide counseling services to the clients.
- Encourage income generating projects for urban and rural small and micro operators, render managerial, marketing, technical and administrative advice to borrowers and assisting them in obtaining services in those fields, manage fund with the purpose of on lending to peasant farmers and micro entrepreneurs and perform any other relevant activities to achieve the objectives and would help for its success.

4.1.3 Major services of the microfinance

The main service provided by Oromia credit and saving company is credit service, saving service, M-BIIR service, and Fund administration.

Credit Services:

Company has been offering diversified loan products to its customers. Solidarity group based loan, MSE Loan, Business loan, WDEP Loan, General purpose loan, Business Loan, Housing loan and Interest Free Finance are the major loan products of the company.

Solidarity Group Based Loan (SGBL): is a microloan that self-organized groups in rural and urban settings are eligible to borrow through group liability. It targets all segments of low income population but mainly women and unemployed youths.

Micro and Small Enterprise Loan (MSEL): is targets higher education institution graduates as well as other unemployed youths or individuals who establish one of business entities such as Cooperatives, Sole proprietorship, Partnership, Share Company and Private Limited Company.

Business Loan (BL): is a loan type that offered to individual or group business runners. The operators need to present matching collateral for their loan request i.e. obviously legal urban house.

Women Entrepreneurs Development Program Loan (WEDP): is loan intends to enhance women owned individual businesses or enterprises through ensuring financial support. The main targets are individual businesses owned by women and of organized women enterprises that are in operation for at least 6 months.

General Purpose Loan (GPL): is a loan product for permanent employees of government and non-government organizations that borrowed for any personal affairs through presenting equivalent salary guarantee.

Housing Loan (HL): is a housing loan for employees to contribute toward the efforts that have been made by government in reducing the housing problems in towns. HL is provided to permanent government and non-government employees, police forces and government appointee officials who can present required guarantee or collateral.

Interest Free Finance (IFF): is finance services delivered to clients who don't want to borrow with interest due to their religion. The service is offered in consistent with Islamic (Sheria) finance principles. OCSSCO provides interest free finance known as Murabaha Financing which is an asset-based sale transaction used to finance goods.

Saving services:

OCSSCO has currently seven different saving products identified as Regular, Handhura, Sorema, Coin-box, Wadia Saving Account, Fixed Time Deposit and Current Account.

Regular saving: is a non-contractual saving that enables savers to withdraw their deposit at any time they want without any restriction.

Handhura saving: is a saving product that targets parents and other relatives interested to save for their children or children under their guardianship (protection) for future expenditure, school fees, and wealth. In addition to wealth accumulation for future uses, Handhura saving is also used as one of the tools to educate the community including the children about savings and also used to promote saving culture early in their childhood.

Wadia Saving Account: is a voluntary saving type with those clients who don't want an interest for religious purposes can deposit their money for the sake of security. It's deposited and kept separately from conventional system at interest free financial service window. Wadia Saving is subject to withdrawal at any time.

Sebeta Ayo Saving Account: Sebeta Ayo Saving Account is voluntary saving type that targets only women. It offers special interest (higher than the conventional one) to its clients. The objective is to encourage women saving and investment that believed as ground for family & community empowerment.

Sorema saving: is a long term contractual saving designed to mobilize savings from the people during their active age and when they are earning more income that to be deposited with special interest returns and to be withdrawn for investment, creating family or own jobs or to cover future planned or unplanned expenditures.

Coin-box saving: is a saving service that will be delivered to any individual who cannot physically appear to OCSSCO branch office to deposit due to the nature of their business. The major targets are loan customers, petty traders, small shop owners, mini cafeterias, taxi drivers and assistant. OCSSCO rents coin box to customers and help them to deposit at their work place.

Fixed Time Deposit: is a type of deposit, which is deposited for a certain agreed period of time at pre agreed interest-rate that to be paid if only withdrawal is made after the agreed time.

Current Account: is provides current (checking account) services on which the account owner can order payments to the check bearer. But unlike other checks, the existing OCSSCOs' check is used only for internal purpose that to be used for cash order transactions that is to be made between OCSSCOs' branch in which the account is opened and the account owner in the branch.

M-BIRR Services:

OCSSCO provides M-BIRR service which is a mobile money *service* that allows customers to do financial transactions from the convenience of your mobile phone at anytime from anywhere. This service enables customers to: deposit cash at an agent, withdraw cash at an agent, transfer money, buy mobile top up, pay bills, buy goods, repay loans, check balance and get statement. In this there is no need to travel or queue. As long as you have money on your account you can transact from the convenience of your phone.

Fund Administration Services:

OCSSCO administers third part funds. The company and several partners' with common objectives have been jointly providing of micro credit for different purposes like solar energy, biogas, etc.

4.1.4 Loan Policies and Procedures

The institution uses internal credit policies and procedures and strictly followed manuals in various level of credit committees before disbursing loan to customer.

The customer to be eligible, target client of the institution has to fulfill the following conditions [74]:

- The loan applicant age should ranges between 18 to 70 years old depending on product types to take in new clients, but existing clients who are economically productive will be served based on the 5 C's credit appraisal requirements
- She/he should have motivation & capacity to engage in viable income generating activities
- She/he must be ready to accept the operational rules and procedures of the institution
- She/he must be willing to enter into a loan and/or savings deposit agreement
- Every potential borrower client should have a mandatory savings account to be eligible for loan provision of the institution
- She/he must be a member of a solidarity group (for group loan methodology).

- In a given loan cycle, all group members are required to borrow a loan of uniform loan period
- The loan for group business loan ranges from 4 month to 24 months
- Group member size of group business loan minimum five and above
- The applicant should be good character and integrity to be accepted by members of the group she/he belongs to
- Individual borrower applicant accepted as a client, only if she/he can present credible guarantee
- Individual business owners applicant shall submit loan application accompanied by business plan
- Individual applicant should have her/his own business that is active in cash flow activity with sufficient level of income
- The client needs to present business plan for the business activity she/he requests the loan. In case the client in able to prepare the business plan, she/he should be willing to show owns financial records and documents to the SACO of the institution.
- An individual business owner to be eligible for the loan he/she requests should present a personal guarantor(s) with known personalities and sufficient level of income to recover a loan delivered in case of the borrower fails to repay the loan.
- A personal guarantor should bring a letter from her/his organization or municipality to pledge owns salary or house respectively for the debt of the applicant client.
- Every individual person who has been authorized to receive an individual business loan greater than birr 50,000 should present a business license and property ownership certificate (mainly fixed assets)

4.1.5 Lending Methodology

Oromia MFI practices group and individual lending methodologies. The former one is a group lending methodology in which potential clients are required to form their own peer group. Group lending of the institution is offered by social collateral of solidarity group and individual lending delivered with two main categories; lending with guarantee of property and salary of public servants and other licensed institutions through guarantee letters written form their employer organizations.

Oromia MFI utilizes lending methods suitable and acceptable to its clients' needs as well as to its operational activities in order to avoid risks that would endanger its normal tasks and clients business activities

Collateral/Guarantee for Group Business Loan:

The institution uses group guarantee or social collateral for all members in a group and an individual person who qualifies for the amount borrowed by individual client of group business loan. The loan amount up to 15,000 birr will be delivered with only group collateral. However, a client who requests a loan amount above 15,000.00 birr should provide at least one members fixed asset as additional collateral besides the group collateral. The institution records list of borrower's fixed asset like, house, car and license of land for the loan amount greater than 15000 to identify the property of an individual guarantor that helps to claim during arrears follow up.

Collateral for Individual Business Loan:

- The business man client can borrow for the loan size up to 50,000 birr with the appropriate salary guarantee,
- A client who receives a loan greater than 20,000.00 should present business license and certificate of personal property (mainly fixed assets).
- A client should deposit a mandatory savings of 10% before securing the loan and ongoing savings percentage of the respective upfront deposit which will be paid over the loan period on monthly basis in equal installments,
- For all loan sizes, clients should present their asset; house, vehicles ownership certificate (Libra) and company share certificate from respective authority with their own cost and should put the original copy in Oromia MFI office as well.

Loan size of group Business loan:

- The size of group business loan per individual client is largely depends up on the business, services and other activities the clients need and envisaged areas of investments,
- The size of a single loan per individual client increases through time following the improved clients capacity of managing credit money, engagement of clients on relatively higher economic and business activities,

- Minimum and maximum loan size for group business loan clients are Birr 1,000 and 30,000 respectively based on loan cycles and clients repayment capacity,
- The maximum increment amount for the next loan cycle per individual existing clients depends on the loan cycle and client's capacity,

Loan size of individual loan:

- The loan size depends on the salary amount in that the maximum monthly loan repayment can't be more than 1/3 of owns or guarantor's monthly net salary.
- The maximum loan amount delivered for consumption loan secured by salary collateral is calculated by multiplying 1/3 of both borrower's and guarantor's net salary independently with determined loan period.
- An applicant who requests consumption loan using owns or her/his guarantor's fixed asset can secure a loan amount up to birr 100,000 with pledged site plan of a building that should be authenticated by authority of a municipality. To secure this loan amount, the applicant should have additional monthly disposable income used to cover periodic loan repayment.

4.1.6 Loan Approval Procedure

The loan approval committee of different levels of the institution analyzes the appraised loan document and gives their final approval for disbursement of the requested loan.

- Group or individual application should be submitted to the institution
- The applicant of both group and individual lending should be certified,
- A loan approval up to 100,000.00 will be made by Branch Loan Approval Committee (BLAC).
- A loan above 2nd cycle that exceeds 100,000 needs the approval of operation manager of the institution,
- A sub branch/branch checks and submits to head office operation department for analysis as per of the loan processing procedures of the institution and then approves within a one day period.
- The operation department can hold up either to assess further or to appear physically at front line level to evaluate the business of the client either to approve or reject the requisition.

4.1.7 Loan Collection Policy

- The principal and interest loan is divided to the whole loan period and will be paid on monthly base as per of the agreement will made between the branch office of the institution and the clients.
- The principal and interest loan including ongoing Mandatory savings deposit should be made on monthly basis at branch office by cashier of sub-branch or branch,
- When the regular collection involves advance or late repayment of principal and interest, the SACO should inform the casher the status of the client in repayment,
- Collection of monthly installment of principal and interest with ongoing Mandatory savings deposit conducted at office level,
- The SACO plans regular collections on monthly basis that shows details of repayment schedule,
- The Sub branch manager or Branch Manager will keep summary of collections planned in a month,
- Repayment interval is monthly, whereas, if the collection is on a holiday or Sunday, a SACO together with her/ his immediate supervisor arranges by extending one day forward.
- Loan collection shall be made as per of the loan repayment schedule of the institution for individual loan
- Repayment of loan is done depending on the nature of the activity & agreement reached for individual loan
- The SACO is also responsible to provide monthly summaries of collections made from his/her own groups and clients on loan and savings group data that will be checked out and signed by Sub-BM, BM at the end of every month.

At all levels of the institution collection should be made in the order of their priority: penalty, interest and principal, client's savings deposit by recording their passbook & ledger. Any violation of this procedure will lead to disciplinary measure as per of the personnel manual of the institution.

4.1.8 Delinquency management in Oromia MFIs

Delinquency is a deviation of client's expected behavior on loan repayment or fails to comply with contractual agreement she/he entered based on the policies and procedures of the institution. This is due to institutional, client related or external driven causes are common factors reported causes for the incidence of loan delinquency in that the operation staff should know their nature to manage them efficiently. Oromia MFIs monitors and controls any loan which may not be collected on the agreed due date. The seriousness of the arrears is determined by the intention of the client not to pay and its age. Therefore, all arrears should be classified by their ages. The older the age of the default loan, the higher the risk associated with it and the tougher the action that the institution should take.

Main signs of delinquency:

The following signs indicate client delinquency behavior that observed

- Lateness or absenteeism on scheduled repayment day,
- Irregular loan repayment and savings deposit,
- Refusal in tailoring with adjustments,
- Poor group leadership,
- Conflict in the group,
- Refusal to participate in group activities,
- Late disbursement of loans,

Recourse to legal action:

- Branches should send regular reports on collection of loans on monthly basis,
- In particular reports; persistent defaulters and fraudulent staffs should be sent to head office with recommendations.
- Head of operations department should make full report with supporting documents to the Managing Director on cases of fraudulent staff to be handled by legal action.
- Board of directors should be informed on fraudulent cases immediately after confirmation of operation department and internal audit.
- Depending on the required directives of the regulatory body, the institution should report cases like frauds within 15 days period,

- The decision to have recourse of legal action should be taken after careful consideration of cost allocation and the consequences of inaction.

Age Categorization of Loans in Arrears:

The National Bank of Ethiopia categorizes arrears in age as follows based on the number of days of the loan reported in past due. Accordingly, the institution decided to implement the standard arrears age category indicated in the table below:

Ages of arrears in days	**Category**
Up to 90 days	Standard
91 – 180	Substandard
181– 365	Doubtful
Greater than 365 days	Loss

Table: 4.1 Age categorizations of loan in arrears

- The National Bank of Ethiopia directive No.MFI/18/06 states that outstanding loans that are past due for more than 90 days beyond the repayment period stipulated in the loan agreement defined as 'non-performing loans' or arrears should be classified as 'Sub-standard', 'Doubtful', 'Loss'. This is intended to minimize the negative financial impact of loan arrears. It also helps Oromia MFI to take timely action with regard to its loans in arrears.
- Any loan past due for greater than 90 days for all loan term is considered as loan defaulted to be calculated by 25% for provision. For a loan term greater than 180 days for all loan term is considered as loan defaulted to be calculated by 50% for provision. Loans past due for more than 365 days for all loan term should be recommended for write-off the defaulted loan portfolio.
- Restructured non-performing loans shall be categorized, at a minimum as 'substandard' unless equivalent of at least all past due interest is paid by the client at the time of restructuring of the loan and unless 3 (three) consecutive repayments are made by the client in a timely manner in accordance with the restructured terms of the loan.
- The outstanding balance shall consist of the principal loan and all other charges, fees and other amounts which have been capitalized to the outstanding balance.

4.2 Data Understanding

Since the data available, in most cases, is generated from day to day activities in different branch collected for different purposes like for administrative purpose, employee information, loan information etc.; So, the existing data situation should be studied to decide on the relevant aspects of the data and to get understanding of the data nature.

In application of data mining, having data at hand is the prerequisite. This phase includes the initial step of collecting data and understanding the relationship of the data to the data mining problem to be solved by the study. Here the data fields are described and analyzed through discussion with the domain experts.

4.2.1 Data Collection

The data used for this study is collected from the centrally collected data from different branches that stored in centrally managed operational databases of Oromia Credit and saving Share Company. In the database of microfinance there are more than fifteen tables that are integrated in their database; for this research purpose the data is extracted from eight tables (account detail table, product table, CBL table (Detail Transaction Table), Lead table, Branch Detail Table, Loan Detail table, Customer Table, Loan Cycle table). The initial number of records is 149963 that extracted from these different tables. This data covers more information about customer loan related data that taken from different tables. The selection of these data done in consideration of discussion within domain experts, relevancies of the data in relation to research purpose is considered. Then this extracted data is exported to excel worksheet as it extracted from different table for data preprocessing.

The distribution of the initial data that extracted from Database in the categories is shown in the following table 4.2.

Category	Number of instance	Instance ratio	Date
Standard	38,134	25.42%	0 – 90
Substandard	7,844	5.23%	91 – 180
Doubtful	12,523	8.35%	181 -365
Loss	91,462	60.98%	>365
Total	149,963	100%	-

Table: 4.2 Distribution of initially collected data

As it is seen from the table the data seems unbalanced that is standard and doubtful cases are much less in number than standard and loss.

4.2.2 Data Description

The data covers the report of financial and customer data from 2012 – 2017 of different branches of Oromia microfinance. Oromia Microfinance has 139 branches out of this ten branches are connected with main branch with core banking system and the other branch data is reported by softcopy and in main branch entered into the system by data encoder of the microfinance.

Fields from the different tables are selectively taken by consulting with business experts. OCSSCO allows providing data only if it is not confidential information. Therefore, records related to customer identity and addresses like customer account number, customer name, customer capital are not included even in the basic data from the beginning. In addition, records related to premiums are also hidden according to the company rule. From the data restricted by the company, the premium data could be related with the problem to make further analysis on loan repayment cases. Some parameters have null values and others are not related to the problem in hand. By merging all the attributes from the different tables a total of 28 attributes are taken. The data is directly exported to an excel format for preprocessing. The descriptions of the attributes with the relative tables shown below:

When the collected raw data is seen from the point of view of data quality, it is believable because the systems are fully operational and it is real transactional data. But it has redundant attributes and instances. As a result of the relationship of data tables most of the columns found one table are also found in another table. Some attributes has missing values and the tables are a bit complex to understand their relationship

No	Attribute	Data Type	Description
1	Branch code	Numeric	The code given for each branch
2	Customer Name	String	Name of each customer individually
3	Product Code	Numeric	Code give for each product
4	Account Number	Numeric	A number given to every single account that a client holds
5	Name Title	String	Title name given for each product
6	Customer group Name	String	Name given for each group of customer
7	Loan Name Description	String	Description given for loan type
8	Effective Date	Date	Date of loan given

No	Attribute	Data Type	Description
9	Installment start date	Date	First date of loan return in the scheduled loan return intervals
10	Installment amount	Numeric	Amount of money that the customer returns in each scheduled intervals.
11	End of payment	Date	The final date of loan return ended
12	Frequency	Nominal	Each cycle to return loan
13	Number of installment	Nominal	Number of installment that the loans is repay and finalized per each return agreement
14	Over due date	Date	The date at which past due start to count
15	Sex Code	Nominal	Sex of customer
16	Marital Status	Nominal	Status whether the customer is married or not
17	Loan Cycle	Numeric	Loan repetition taken by customer.
18	Date of Birth	Date	Birth date of customer
19	To date	Date	End date for repayment
20	Amount Disburse	Numeric	Amount of money paid out for clients in the form of loan
21	Balance	Numeric	Net balance left on customer when he starts repays the payment
22	Purpose of Loan	String	For what purpose the customer take loan
23	Age	Numeric	Age of customer
24	Residence	Nominal	Place where customer live
25	Pre-loan training	Nominal	Pre-loan training given to customer
26	Distance from MFI	Nominal	How far the customer from MFI center
27	Follow-up	Nominal	Customer looked after loan by MFI
28	Overdue days	Date	Number of days past due date

Table: 4.3 Attribute descriptions.

4.3 Data Preparation (Data preprocessing)

Han and Kamber [58],explained that today's real-world databases are highly susceptible to noisy, missing, and inconsistent data because of their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. So, data preprocessing is an important and prerequisite steps in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Also analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis [64].

If there is much irrelevant and redundant information present or noisy and unreliable data, the knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. The activities during this phase include

data cleaning; attribute selection, normalization, data transformation and aggregation and data formatting etc.

4.3.1 Data Cleaning

Witten and Frank [66], described Data cleaning as a time consuming and labor intensive procedure but one that is absolutely necessary for successful data mining. Data cleaning routines work to cleaning the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Quality of data plays an important role in information oriented organizations, where the knowledge is extracted from data. Consistency, completeness, accuracy, validity and timeliness are the important characteristics of quality data. So, it is important to obtain quality data for knowledge extraction. Data cleaning is an important step in KDD process in order to recognize any inconsistency and incompleteness in the data set and to improve the quality [58].

Under data cleaning there are many activities are done. Some of the activities: in age attribute 1699 (1%) instances has missed values replaced by most frequent values, there is some values of small and capital letter, this values replaced by M and F. In this the same attribute according to the rule of the organization the loan is given for clients' age between 18 and 70; so, the age values below 18 and above 70 instances that contain such values are removed. The overdue data attribute has missed values 1189500 (80%) instances attribute is remove (deleted), the sex attribute which has a missed value of 3921 (3%) instances are removed and for attributes age, sex and residence data inconsistencies, noisy data and outlier were cleaned. Replaced with the same value F and M in sex attributes, u and urban in residence attributes replaced by urban etc.

4.3.2 Data Transformation

Once the data has been assembled and major data problems are fixed, the data must still be transformed for analysis. This involves adding derived fields to bring information to the surface. It may also involve smoothing, aggregation, generalization, normalization, discretization, and attribute construction [58]. On an effort to make the dataset used for this study suitable for the data mining process, few data transformation methods were used. Discretization was used to reduce distinct values of attributes, dimensionality reduction was used to reduce the size of the dataset and attribute selection method was applied to remove weakly relevant attributes.

In order to handle skewed data and dominance of outliers in presentation equal-depth (frequency) partitioning of was adopted. In equal-depth discretization the range is divided into N intervals, each containing approximately same number of samples. The result of the discretization process is applied on continuous value of age attribute and the distance customer far from microfinance center.

Label	Count
18-31	36820
32-45	36824
46-58	36831
59-70	36810

Table: 4.4 Age attribute after discritization

After completing the discretization process distinct values of the age attribute were reduced to 4 from 52 distinct values.

Label	New value	Count
0-5	Near	39506
7-8	Middle	26085
9-12	Far	81694

Table: 4.5 discritization of distance in KM

After completing the discretization process distinct values of the distance from microfinance center attribute were reduced to 3 from 13 distinct values.

4.3.3 Attribute value representation and derivation

In the point of views discussed under these is attribute derivation; the age attribute were derived from today date and date of birth, the full name attribute and Overdue date attribute before selecting the relevant attribute the customer full name is empty and the Overdue date attribute is 80% is not filled. So, the four attribute are reduced before relevant attribute selection. In the value reduction the following attributes values are represented accordingly.

Marital Status		Frequency		Purpose of loan	
Attribute value	Value representation	Attribute value	Value representation	Attribute value	Value representation
Single	1	Weekly	W	Agriculture	1
Married	2	Monthly	M	Trade	2
Windowed	3	Quarterly	Q	Manufacturing	3
Divorced	4	Half yearly	H	Construction	4
		Yearly	Y	Service	5
				Any other	6

Table: 4.6 Attribute value Representation

Under class attribute there are four class values: standard, substandard, doubtful and loss. The distribution of values in these attributes 25.42%, 5.23%, 8.35% and 60.98% and then made discussion with domain experts, they discuss with the researcher it is better if the substandard and doubtful values are merged and named as uncertain and the class value categorized into three class: standard, uncertain and loss.

4.3.4 Attribute selection

As Han and Kamber [58] stated, attribute selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions) and also mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. In order to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes the researcher has used to selection the attribute with the help of domain experts.

Most machine learning algorithms are designed to learn the most appropriate attributes to use for making their decisions [66]. However, algorithms have their own limitations, such as considering insignificance attribute as crucial one and ignoring the most important attribute as irrelevant. Due to this limitation, the attributes for customer loan repayment prediction selection is done in the view of business perspective of the organization, research objectives and from discussion with domain expert out of the following attributes:

No.	Attribute	Data type	Attribute selection
1	Branch code	Numeric	Selected
2	Customer Name	String	Not selected
3	Product code	Numeric	Not selected
4	Account Number	Numeric	Not selected
5	Name Title	String	Not selected
6	Customer group Name	String	Not selected
7	Loan Name description	String	Not selected
8	Effective date	Date	Not selected
9	Installment start date	Date	Not selected
10	Installment amount	Numeric	Selected
11	End of payment	Date	Not selected
12	Frequency	Nominal	Selected
13	Number of installment	Nominal	Selected
14	Overdue date	Date	Not selected
15	Sex Code	Nominal	Selected
16	Marital status	Nominal	Selected
17	Loan cycle	Numeric	Selected
18	Date of birth	Date	Not selected
19	To date	Date	Not selected
20	Amount Disburse	Numeric	Selected
21	Balance	Numeric	Not selected
22	Purpose of loan	String	Selected
23	Age	Numeric	Selected
24	Residence	Nominal	Selected
25	Pre-loan training	Nominal	Selected
26	Distance from MFIs	Nominal	Selected
27	Follow up	Nominal	Selected
28	Overdue days	Date	Selected (Class)

Table: 4.7 Selected attributes

Finally the following 14 attributes are selected for experimentation :Branch code, Installment amount, frequency, Amount disburse, Number of installment, Purpose of loan, Sex code, Marital status, Loan cycle, Age, Residence, Follow up, Distance from MFIs, Pre-loan training then these attribute is ranked using WEKA best first ranker method.

CHAPTER FIVE

EXPERIMENTATION

As the objectives of this study is to predict customer loan repayment behavior before loan disburse using data mining techniques; a classification technique is adopted to develop a predictive model. The models are built with three different supervised machine learning algorithms. These algorithms are Decision tree classifier (J48), Bayesian Classifier (Naïve Bayes) and neural network classifier (multilayer perception) using WEKA 3.8.2 machine learning software. These classifier algorithms are used because it is recommended and commonly used in different literature reviews then do the comparison and selection of the best classifier based on their performance accuracy.

The preprocessing of data done manually and using WEKA preprocessor tool then the final datasets with instance of 147285 and 15 attributes is prepared for experimentation purpose. The class distribution of the final datasets are loss 90165(61.2%), Standard 37679(25.6%) and uncertain 19441(13.2%) respectively.

Six experiments are conducted using original unbalanced class datasets and balanced class datasets, for all experiments two model evaluation techniques are chosen (10 fold cross validation and percentage split with (66%, 70% and 80%) are used. In both model evaluation techniques all attributes are considered.

The performances of the models in this study are evaluated using the standard metrics of accuracy, precision, recall and F-measure which are calculated using the predictive classification table, known as Confusion Matrix. ROC area was also used to compare the performances of the classifiers. These performance evaluation methods are well known measures for evaluation of data mining models for classification.

5.1 Model Building using Decision Tree (J48)

The decision tree algorithms generate models in the form of a tree like-structure, which starts from root attributes and ends with leaf nodes, describing the relationship among attributes and the relative importance of attributes. They represent rules which could easily be understood and interpreted by users, do not require complex data preparation, and perform well for numerical and categorical variables.

This experiment designed to evaluate the performance of a J48 decision tree algorithm pruned tree to predict customer loan repayment behavior before loan disburses and to investigate the effect of algorithms on performance of the model.

Experiment one:

The first experiment is designed to evaluate the performance of a J48 classifier using original unbalanced class dataset to predict customer loan repayment and to evaluate the performance of the model. In this experiment the percentage split with 66%, 70%, 80% training with the rest test and 10 fold cross validation are used. In percentage split of 66%: 49470(98.79%) instances are correctly classified and 607(1.21%) instances are incorrectly classified from 50077 of testing instance, it takes 9.68 second to build the model and the model generates a tree with a size of 1901 and 1564 leaves. Using70%:43632(98.75%) instances are correctly classified and 553(1.25%) instances are incorrectly classified from 44185 of testing instance, it takes 6.85 second to build the model and the model generates a tree with a size of 1901 and 1564 leaves. In 80%:29082(98.73%) instances are correctly classified and 375(1.27%) instances are incorrectly classified from 29457 of testing instance, it takes 6.73 second to build the model and the model generates a tree with a size of 1901 and 1564 leaves. In second case, using 10 fold cross validation model evaluation techniques,145647(98.89%) instances are correctly classified and 1638(1.11%) instances are incorrectly classified from 147285 of instance, it takes5.54 second to build the model and the model generates a tree with a size of 1901 and 1564 leaves.

From the results it can be seen that Decision tree classifier (J48) with 10-fold cross validation has relatively better performance than the three percentages split cases. 145647(98.89) of the instance 147285 are classified correctly to the respective class loss, standard or uncertain.

			Predicted		
Percentage split 66%			Loss	Standard	uncertain
	Actual	Loss	30472	0	259
		Standard	0	12738	0
		Uncertain	348	0	6260
			Predicted		
Percentage split 70%			Loss	Standard	uncertain
	Actual	Loss	26883	0	250
		Standard	0	11193	0
		Uncertain	303	0	5556
			Predicted		
Percentage split 80%			Loss	Standard	uncertain
	Actual	Loss	17947	0	145
		Standard	0	7449	0
		Uncertain	230	0	3686
			Predicted		
10-fold			Loss	Standard	uncertain
	Actual	Loss	89463	0	702
		Standard	0	37679	0
		Uncertain	936	0	18505

Table 5.1: Confusion matrix for experiment one

As shown in table 5.1, the confusion matrix for the percentage split of 66%, 70%, 80% and 10-Fold cross validation model evaluation techniques using decision tree (J48) classifier for target class loss, standard. From the confusion matrix for 66% percentage split 259(0.84%) instances are misclassified as class of uncertain which were actually class of loss and 348(5.27%) instances are misclassified as class loss which were actually uncertain. For percentage split of 70% 250(0.92%) instances are misclassified as uncertain which were actually class of loss and 303(5.17%) instances are misclassified as loss which were actually class of uncertain. For percentage split of 80% 145(0.80%) instances are misclassified as uncertain which were actually loss and 230(5.92%) instances are misclassified as loss which actually uncertain. In 10-fold cross validation 702(0.78%) instances are misclassified as uncertain which were actually class of loss and 936(4.81%) instance are misclassified as class of loss but actually it is class of uncertain.

Algorithm	%Split	Accuracy	TP rate	FP rate	Precision	F-measure	ROC area
J48	66	98.7879	0.988	0.012	0.988	0.988	0.997
	70	98.7484	0.987	0.012	0.987	0.987	0.997
	80	98.727	0.987	0.013	0.987	0.987	0.997
	10-fold	98.8879	0.989	0.011	0.989	0.989	0.997

Table 5.2 Detailed Performance measures for experiment one

The performance measure of the model, we obtained average precision of 0.99 in all cases. From this result, it is very successful model in retrieving relevant values for each class. From the evaluation we obtained, F-measure of 0.99 in all cases, this implies that the model is significantly balanced.

5.2 Model Building Using Naïve Bayes Classifier

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naïve Bayes classification algorithm works based on the following three conditions: The prior probability of a given hypothesis, the probability of the data given that a hypothesis, and the probability of the data itself. Its classification performance is based on the assumption of conditional independence between the attributes.

Experiment two:

The second experiment is designed to evaluate the performance of a Naïve Bayes classifier using original unbalanced dataset to predict customer loan repayment and to evaluate the performance of the model. In this experiment also the percentage split and cross validation model evaluation techniques are used. In percentage split of 66%:45011(89.88%) instances are correctly classified and 5066(10.12%) instances are incorrectly classified from 50077 of testing instance, it takes 0.66 second to build the model. In 70%:39678(89.80%) instances are correctly classified and 4507(10.20%) instances are incorrectly classified from 44185 of testing instance, it takes 0.47 second to build the model. using80%:26423(89.70%) instances are correctly classified and 3034(10.30%) instances are incorrectly classified from 29457 of testing instance, it takes 0.45 second to build the model and 0.7 second to test model. Using 10 fold cross validation: 132412(89.90%) instances are correctly classified and 14873(10.10%) instances are incorrectly classified from 147285 of testing instance, it takes 0.57 second to build the model.

From the results it can be seen that Naive Bayes classifier with 10-fold cross validation has relatively better performance than the other three percentage split cases. 132412(89.90) of the instance 147285 are classified correctly to the respective class loss, standard or uncertain.

Percentage split 66%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	30112	1	618
		Standard	0	12564	174
		Uncertain	4273	0	2335
Percentage split 70%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	26581	1	551
		Standard	0	11039	154
		Uncertain	3801	0	2058
Percentage split 80%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	17711	0	381
		Standard	0	7340	109
		Uncertain	2544	0	1372
10-fold			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	88333	1	1831
		Standard	4	37155	520
		Uncertain	12517	0	6924

Table 5.3 Confusion Matrix for experiment two

As shown in table 5.3, the percentage split of 66%, 70%, 80% and 10-Fold cross validation model evaluation techniques using Naïve Bayes classifier for target class loss, standard and uncertain. From the confusion matrix for 66% percentage split 1 instance is misclassified as standard which is actually class loss, 618 instances are misclassified as class of uncertain which are actually class of loss, resulting total misclassification of 619(2.01%) instances. And 174(1.37%) instances are misclassified as class uncertain which is actually standard and 4273(64.66%) instances are misclassified as class loss which is actually class uncertain. For percentage split of 70%, 1 instance is misclassified as standard which is actually class loss, 551 instances are misclassified as class of uncertain which is actually class of loss, giving total misclassification of 552(2.03%). And 154(1.38%) instances are misclassified as uncertain which is actually class of standard and 3801(64.87%) instances are misclassified as loss which is actually class of uncertain. Using percentage split of 80%, 381(2.11%) instances are misclassified as uncertain which is actually class of loss and 109(1.46%) instances are misclassified as uncertain which is actually standard and 2544(64.96%) instances are misclassified as loss which is actually uncertain.

Using 10-fold cross validation, 1 instance is misclassified as standard which is actually loss class, 1831 instances are misclassified as class of uncertain which is actually class of loss, giving total misclassification of 1832(2.03%) and 4 instances are misclassified as loss which is actually class of standard, 520 instances are misclassified as uncertain which is actually standard class, resulting total misclassification of 524(1.39%) and 12517(64.38%) instance are misclassified as class of loss but actually class of uncertain.

Algorithm	%Split	Accuracy	TP rate	FP rate	Precision	F-measure	ROC area
Naïve Bayes	66	89.8836	0.899	0.138	0.890	0.883	0.983
	70	89.7997	0.898	0.139	0.889	0.882	0.982
	80	89.7002	0.897	0.140	0.888	0.881	0.981
	10-fold	89.9019	0.899	0.137	0.891	0.884	0.983

Table 5.4 Detailed Performance measures for experiment two

From the performance measure, we obtained average precision of 0.89 in all cases; this indicates that the model is very successful in retrieving relevant values for each class. In the case of F-measure, we obtained a result of 0.88 in all cases; this implies that the model is significantly balanced.

5.3 Model Building Using Neural Network

Neural networks produce classification models in the form of a mathematical model, consisting of interconnected computational elements (neurons) and processing information using a connectionist approach to computation. They are used to model complex relationships between inputs and outputs and very often yield very good results.

Experiment Three:

The third experiment is designed to evaluate the performance of a neural network classifier using original unbalanced dataset to predict customer loan repayment and to evaluate the performance of the model. In this experiment the percentage split and cross validation model evaluation techniques are used. In percentage split of 66%: 48472(96.79%) instances are correctly classified and 1605(3.21%) instances are incorrectly classified from 50077 of testing instance, it takes 161.93 second to build the model. In 70%: 42733(96.71%) instances are correctly classified and 1452(3.29%) instances are incorrectly classified from 44185 of testing instance, it takes 194.36 second to build the model. In 80%: 28460(96.62%) instances are correctly classified and 997(3.38%) instances are incorrectly classified from 29457 of testing instance,it takes206.58

second to build the model. Using 10 fold cross validation: 142373(96.67%) instances are correctly classified and 4912(3.33%) instances are incorrectly classified from 147285 of testing instance, it takes 207.03 second to build the model.

From the results it can be seen that neural network classifier with 66% percentage split has relatively better performance than the other three cases. 48472(96.79) of the test data 50077 are classified correctly to the respective class loss, standard or uncertain.

Percentage split 66%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	30368	0	363
		Standard	0	12738	0
		Uncertain	1242	0	5366
Percentage split 70%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	26698	0	435
		Standard	0	11193	0
		Uncertain	1017	0	4842
Percentage split 80%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	17923	0	169
		Standard	0	7449	0
		Uncertain	828	0	3088
10-fold			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	88812	0	1353
		Standard	0	37679	0
		Uncertain	3559	0	15882

Table 5.5 Confusion matrix for experiment three

From the table 5.5, in the percentage split of 66%, 70%, 80% and 10-Fold cross validation model evaluation techniques using the neural network (multilayer precepton) classifier for target class loss, standard and uncertain. From the confusion matrix for 66% percentage split 363(1.18%) instances are misclassified as class of uncertain which were actually class of loss and 1242(18.80%) instances are misclassified as class loss which were actually uncertain. For percentage split of 70%, 435(1.60%) instances are misclassified as uncertain which were actually class of loss and 1017(17.36%) instances are misclassified as loss which were actually class of uncertain. For percentage split of 80% 169(0.93%) instances are misclassified as uncertain which were actually loss and 828(21.14%) instances are misclassified as loss which actually uncertain.

In 10-fold cross validation 1353(1.50%) instances are misclassified as uncertain which were actually class of loss and 3559(18.31%) instance are misclassified as class of loss but actually it is class of uncertain.

Algorithm	%Split	Accuracy	TP rate	FP rate	Precision	F-measure	ROC area
Neural network	66	96.7949	0.968	0.040	0.968	0.967	0.994
	70	96.7138	0.967	0.038	0.967	0.966	0.993
	80	96.6154	0.966	0.046	0.966	0.965	0.992
	10-fold	96.665	0.967	0.040	0.966	0.966	0.993

Table 5.6 Detailed Performance measures for experiment three

Regarding the precision score of the model, average precision 0.97 in all cases, it is very successful model in retrieving relevant values for each class. With F-measure value of 0.97 in all cases, this implies that the model is significantly balanced.

Experiments using Balanced Class:

Class imbalance problem is a hot topic being investigated recently by machine learning and data mining researchers. It can occur when the instances of one class outnumber the instances of other classes. The class have overwhelmed called the majority class while the other called minority class. However, in many applications the class has lower instances are the more interesting and important one. The imbalance problem heightens whenever the class of interest is relatively rare and has small number of instances compared to the majority class. Moreover, the cost of misclassifying the minority class is very high in comparison with the cost of misclassifying the majority class. In classification, a dataset is said to be imbalanced when the number of instances which represents one class is smaller than the ones from other classes. In the above experiment using original unbalanced dataset there is no problems even the data is seems unbalanced. Even though to see the difference of the experiments between dataset with original unbalanced class dataset and by balancing the dataset class using class balancer of WEKA.

Experiment Four (Decision Tree):

The fourth experiment is designed to evaluate the performance of a J48 classifier using balanced dataset class to predict customer loan repayment and to evaluate the performance of the model. In this experiment the percentage split and cross validation model evaluation techniques are used. In percentage split of 66%:49272.90(98.51%) instances are correctly classified and 744.95(1.49%) instances are incorrectly classified from 50077 of testing instance, it takes

6.03second to build the model and the model generated a tree with a size of 2095 and 1705 leaves, in 70%: 43630.54(98.59%) instances are correctly classified and 623.61(1.41%) instances are incorrectly classified from 44154.15 of testing instance, it takes 5.4 second to build the model and the model generated a tree with a size of 2095 and 1705 leaves, in 80%: 29030.75(98.59%) instances are correctly classified and 415.48(1.41%) instances are incorrectly classified from 29446.23 of testing instance, it takes 5.5 second to build the model and the model generated a tree with a size of 2095 and 1705 leaves. In second case using 10 fold cross validation: 145356.21(98.69%) instances are correctly classified and 1928.79(1.31%) instances are incorrectly classified from 147285 of testing instance, it takes 5.43 second to build the model and the model generated a tree with a size of 2095 and 1705 leaves.

From the results it can be seen that Decision tree classifier (J48) with 10-fold cross validation has relatively better performance than the other three cases. 145356.21(98.69%) of the test data 29446.23 are classified correctly to the respective class loss, standard or uncertain.

Percentage split 66%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	16356.83	0	376.25
		Standard	0	16597.36	0
		Uncertain	368.7	0	16318.7
Percentage split 70%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	14476.12	0	297.84
		Standard	0	14584.26	0
		Uncertain	325.77	0	14470.16
Percentage split 80%			Predicted		
			Loss	Standard	uncertain
	Actual	Loss	9665.45	0	185.68
		Standard	0	9705.9	0
		Uncertain	229.81	0	9659.4
10-fold			Predicted		
			Loss	Standard	Uncertain
	Actual	Loss	48133.41	0	961.59
		Standard	0	49095	0
		Uncertain	967.2	0	48127.8

Table 5.7 Confusion matrix for experiment Four

As shown in table 5.7, for the percentage split of 66%, 70%, 80% and 10-Fold cross validation model evaluation techniques using decision tree (J48) classifier for target class loss, standard and

uncertain. From the confusion matrix for 66% percentage split 376.25(2.25%) instance are misclassified as uncertain which were actually class loss and 368.7(2.21%) instances are misclassified as class of loss which were actually class of uncertain. For percentage split of 70% 297.84(2.02%) instance are misclassified as uncertain which were actually class loss and 325.77(2.20%) instances are misclassified as class of loss which were actually class of uncertain. For percentage split of 80% 185.68(1.88%) instances are misclassified as uncertain which were actually loss and 229.81(2.32%) instances are misclassified as loss which were actually uncertain. In 10-fold cross validation 961.59 (1.96%) instances are misclassified as uncertain which were actually class loss and 967.2(1.97%) instances are misclassified as class of loss which were actually class of uncertain.

Algorithm	%Split	Accuracy	TP rate	FP rate	Precision	F-measure	ROC area
J48	66	98.5106	0.985	0.007	0.985	0.985	0.995
	70	98.5877	0.986	0.007	0.986	0.986	0.995
	80	98.5889	0.986	0.007	0.986	0.986	0.994
	10-fold	98.6904	0.987	0.007	0.987	0.987	0.995

Table 5.8 Detailed Performance measures for experiment Four

Regarding the precision score of the model, average precision 0.99 in all cases, it is very successful model in retrieving relevant values for each class. With F-measure value of 0.99 in all cases, it can be calculated that precision and recall of the model are significantly balanced.

Experiment Five:

The fifth experiment is designed to evaluate the performance of a Naïve Bayes classifier using balanced dataset to predict customer loan repayment and to evaluate the performance of the model. In this experiment the percentage split and cross validation model evaluation techniques are used. In percentage split of 66%: 45011(89.88%) instances are correctly classified and 5066(10.12%) instances are incorrectly classified from 50077 of testing instance, it takes 0.89 second to build the model, in 70%: 39678(89.80%) instances are correctly classified and 4507(10.20%) instances are incorrectly classified from 44185 of testing instance, it takes 0.45 second to build the model, in 80%: 26423(89.70%) instances are correctly classified and 3034(10.30%) instances are incorrectly classified from 29457 of testing instance, it takes 0.35 second to build the model. In second case using 10 fold cross validation: 132412(89.90%)

instances are correctly classified and 14873(10.10%) instances are incorrectly classified from 147285 of testing instance, it takes 0.39 second to build the model.

From the results obtained, it can be seen that Bayes classifier with 10-fold cross validation has relatively better performance than the other three cases. 132412(89.90%) of the instance 147285 are classified correctly to the respective class loss, standard or uncertain.

			Predicted		
Percentage split 66%			Loss	Standard	Uncertain
	Actual	Loss	30112	1	618
		Standard	0	12564	174
		Uncertain	4273	0	2335
			Predicted		
Percentage split 70%			Loss	Standard	Uncertain
	Actual	Loss	26581	1	551
		Standard	0	11039	154
		Uncertain	3801	0	2058
			Predicted		
Percentage split 80%			Loss	Standard	Uncertain
	Actual	Loss	17711	0	381
		Standard	0	7340	109
		Uncertain	2544	0	1372
			Predicted		
10-fold			Loss	Standard	Uncertain
	Actual	Loss	88333	1	1831
		Standard	4	37155	520
		Uncertain	12517	0	6924

Table 5.9 Confusion matrix for experiment Five

As depicted in the confusion matrix in a table 5.9 for the percentage split of 66%, 70%, 80% and 10-Fold cross validation using Naïve Bayes classifier for target class loss, standard and uncertain. From the confusion matrix for 66% percentage split 1 instance are misclassified as standard which were actually class loss, 618 instances are misclassified as class of uncertain which were actually class of loss, giving total misclassification of 619(2.01%) and 174(1.37%) instances are misclassified as class uncertain which were actually standard and 4273(64.66%) instances are misclassified as class loss which were actually class uncertain. For percentage split of 70% 1 instance are misclassified as standard which were actually class loss, 551 instances are misclassified as class of uncertain which were actually class of loss, giving total misclassification of 552(2.03%) and 154(1.38%) instances are misclassified as uncertain which were actually class

of standard and 3801(64.87%) instances are misclassified as loss which were actually class of uncertain. For percentage split of 80% 381(2.11%) instances are misclassified as uncertain which were actually loss and 109(1.46%) instances are misclassified as uncertain which were actually standard and 2544(64.96%) instances are misclassified as loss which actually uncertain. In 10-fold cross validation 1 instance are misclassified as standard which were actually class loss, 1831 instances are misclassified as class of uncertain which were actually class of loss, giving total misclassification of 1832(2.03%) and 4 instances are misclassified as loss which were actually class of standard, 520 instances are misclassified as uncertain which were actually standard, giving total misclassification of 524(1.39%) and 12517(64.38%) instance are misclassified as class of loss but actually it is class of uncertain.

Algorithm	%Split	Accuracy	TP rate	FP rate	Precision	F-measure	ROC area
Naïve Bayes	66	89.8836	0.899	0.138	0.890	0.883	0.983
	70	89.7997	0.898	0.139	0.889	0.882	0.982
	80	89.7002	0.897	0.140	0.888	0.881	0.981
	10-fold	89.9019	0.899	0.137	0.891	0.884	0.983

Table 5.10 Detailed Performance measures for experiment Five

Regarding the precision score of the model, average precision 0.89 in all cases, it is very successful model in retrieving relevant values for each class. With F-measure value of 0.88 in all cases, it can be calculated that precision and recall of the model are significantly balanced

Experiment Six:

The sixth experiment is designed to evaluate the performance of a neural network classifier using balanced dataset to predict customer loan repayment and to evaluate the performance of the model. In this experiment the percentage split and cross validation are used. In percentage split of 66%: 48472(96.79%) instances are correctly classified and 1605(3.29%) instances are incorrectly classified from 50077 of testing instance, it takes 236.8 second to build the model, in 70%: 42733(96.71%) instances are correctly classified and 1452(3.29%) instances are incorrectly classified from 44185 of testing instance, it takes 231.99 second to build the model, in 80%: 28464(96.63%) instances are correctly classified and 993(3.34%) instances are incorrectly classified from 29457 of testing instance, it takes 231.46 second to build the model. In second case using 10 fold cross validation: 142538(96.78%) instances are correctly classified and 4747(3.22%) instances are incorrectly classified from 147285 of testing instance, it takes 231.6 second to build the model.

From the results it can be seen that neural network classifier with 66% percentage split has relatively better performance than the other three cases. 48472(96.79%) of the instance 50077 are classified correctly to the respective class loss, standard or uncertain.

Percentage split 66%			Predicted		
			Loss	Standard	Uncertain
	Actual	Loss	30356	0	375
		Standard	0	12738	0
		Uncertain	1230	0	5378
Percentage split 70%			Predicted		
			Loss	Standard	Uncertain
	Actual	Loss	26673	0	460
		Standard	0	11193	0
		Uncertain	992	0	4867
Percentage split 80%			Predicted		
			Loss	Standard	Uncertain
	Actual	Loss	17909	0	183
		Standard	0	7449	0
		Uncertain	810	0	3106
10-fold			Predicted		
			Loss	Standard	Uncertain
	Actual	Loss	88843	0	1322
		Standard	0	37679	0
		Uncertain	3425	0	16016

Table 5.11 Confusion matrix for experiment Six

As shown in table 5.11, for the percentage split of 66%, 70%, 80% and 10-Fold cross validation model evaluation techniques using neural network classifier for target class loss, standard and uncertain. From the confusion matrix for 66% percentage split 375(1.22%) instance are misclassified as uncertain which were actually class loss and 1230(18.61%) instances are misclassified as class of loss which were actually class of uncertain. For percentage split of 70% 460(1.70%) instance are misclassified as uncertain which were actually class loss and 992(16.93%) instances are misclassified as class of loss which were actually class of uncertain. For percentage split of 80% 183(1.01%) instances are misclassified as uncertain which were actually loss and 810(20.68%) instances are misclassified as loss which were actually uncertain. In 10-fold cross validation 1322(1.47%) instance are misclassified as uncertain which were actually class loss and 3425(17.62%) instances are misclassified as class of loss which were actually class of uncertain.

Algorithm	%Split	Accuracy	TP rate	FP rate	Precision	F-measure	ROC area
Neural network	66	96.7949	0.968	0.040	0.968	0.967	0.994
	70	96.7138	0.967	0.037	0.967	0.967	0.994
	80	96.629	0.966	0.045	0.966	0.965	0.992
	10-fold	96.777	0.968	0.038	0.967	0.967	0.994

Table 5.12 Detailed Performance measures for experiment Six

Regarding the precision score of the model, average precision 0.97 in all cases, it is very successful model in retrieving relevant values for each class. With F-measure value of 0.97 in all cases, it can be calculated that precision and recall of the model are significantly balanced.

Classification accuracy can be alternatively seen by calculating absolute error, square error, mean square error, relative absolute error and relative square error which can be taken directly from the result of the WEKA classification process. A screen shot of the results are attached as an appendix 2, 3, 4 as a sample for decision tree 10-fold cross validation.

5.4 Comparison of the algorithm

In any branch of science, it is almost a common requirement that performance of various model have to be compared with each other to understand the suitability of a model to a given problem. In this section also after performing the experiments comparing the model and selecting the best model is the task to be done. Basically, the experiments were conducted on all attributes that selected by discussion with domain experts and purpose of the research. The models were compared using different performance measures like correctly classified instance, incorrectly classified instance, ROC Area and execution time (time taken to build the model).

Percentage split	**Model**	**Accuracy**	**TP Rate**	**FN Rate**	**Precision**	**F-Measure**	**ROC Area**	**Time (sec.)**
66%	Decision tree	98.79%	0.988	0.012	0.988	0.988	0.997	9.68
	Naïve Bayes	89.88%	0.899	0.138	0.890	0.883	0.983	0.66
	Neural Network	96.79%	0.968	0.040	0.968	0.967	0.994	161.93
70%	Decision tree	98.75%	0.987	0.012	0.987	0.987	0.997	6.85
	Naïve Bayes	89.80%	0.898	0.139	0.889	0.882	0.982	0.47
	Neural Network	96.71%	0.967	0.038	0.967	0.966	0.993	194.36
80%	Decision tree	98.73%	0.987	0.013	0.987	0.987	0.997	6.73
	Naïve Bayes	89.70%	0.897	0.140	0.888	0.881	0.981	0.45
	Neural Network	96.62%	0.966	0.046	0.966	0.965	0.992	206.58
10-fold	Decision tree	98.89%	0.989	0.011	0.989	0.989	0.997	5.54
	Naïve Bayes	89.90%	0.899	0.137	0.891	0.884	0.983	0.57
	Neural Network	96.67%	0.967	0.040	0.966	0.966	0.993	207.03

Table 5.13 comparison of the algorithm using original unbalanced datasets

As shown in table 5.13 J48 decision tree has scored highest classification accuracy in percentage splits of (66%, 70%, and 80%) and 10-fold cross validations. Also a model built using J48 scored the highest TP rate than neural network and Naïve Bayes. In regard of the ROC Area, looking the area under the curve (AUC) as an indicator for the quality of separation, Table 5.13 confirms decision tree classifier are the most accurate classifiers. In addition to the above matrices the time taken to build model can be used as a measure of performance of classification model. The time taken to build the model using Naïve Bayes model is shorter than J48 and neural network. Time is recognized when the data becomes larger and larger but still the difference is not significant. One important thing observed here is that all the models are better in their performance.

Percentage split	Model	Accuracy	TP Rate	FP Rate	Precision	F-Measure	ROC Area	Time (sec.)
66%	Decision tree	98.51%	0.985	0.007	0.985	0.985	0.995	6.03
	Naïve Bayes	89.88%	0.899	0.138	0.890	0.883	0.983	0.89
	Neural Network	96.79%	0.968	0.040	0.968	0.967	0.994	236.8
70%	Decision tree	98.59%	0.986	0.007	0.986	0.986	0.995	5.4
	Naïve Bayes	89.80%	0.898	0.139	0.889	0.882	0.982	0.45
	Neural Network	96.71%	0.967	0.037	0.967	0.967	0.994	231.99
80%	Decision tree	98.59%	0.986	0.007	0.986	0.986	0.994	5.5
	Naïve Bayes	89.70%	0.897	0.140	0.888	0.881	0.981	0.35
	Neural Network	96.63%	0.966	0.045	0.966	0.965	0.992	231.46
10-fold	Decision tree	98.69%	0.987	0.007	0.987	0.987	0.995	5.43
	Naïve Bayes	89.90%	0.899	0.137	0.891	0.884	0.983	0.39
	Neural Network	96.78%	0.968	0.038	0.967	0.967	0.994	231.6

Table 5.14 comparison of the algorithm using balanced datasets

As shown in table 5.14, J48 decision tree has scored highest classification accuracy in percentage splits of (66%, 70%, and 80%) and 10-fold cross validations. Also a model built using J48 scored the highest TP rate than neural network and Naïve Bayes. In regard of the ROC Area, looking the area under the curve (AUC) as an indicator for the quality of separation, Table 5.14confirms decision tree classifier are the most accurate classifiers. In addition to the above matrices the time taken to build model can be used as a measure of performance of classification model. The time taken to build the model using Naïve model is shorter than J48 and neural network. Time is recognized when the data becomes larger and larger but still the difference is not significant. One important thing observed here is that all the models are better in their performance.

5.5 Specific Rule Extraction

The model Developed with J48 classifier with original unbalanced datasets was selected as the best model for this study and this model generated significant rules that are useful for prediction of customer loan repayment. These rules are selected from decision tree generated as shown in appendix 1 that converts most of the instances in the dataset. After selecting that rules the researcher turned to the domain expert for discussion.

Rule: 1 If follow up = Yes then the loan repayment = Standard (37679.0)

This rule show, if there is follow up is yes, gave a correct result for 37679 of the 37679 cases that it covers; thus its success fraction is 37679 (100%). This rule is a strong rule for predicting customer loan repayment.

Rule: 2 If follow up = No and Purpose of loan =1 and distance from MFI = far and amount disburse <= 1500 and branch code = 16 then loan repayment = Loss (89.0/1.0)

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is far from microfinance center and the amount he/she asked is <=1500 and the microfinance branch code is 16 then the loan repayment is loss (89.0/1.0)

Rule: 3 If follow up = No and purpose of loan = 1 Distance from MFI= far and amount disburse > 15,000 then loan repayment= loss (1906.0/155.0)

If there is no customer follow up, the purpose of loan is agriculture, the customer is far from microfinance center and the amount he/she asked is >15,000, then the loan repayment is loss (1906.0/155.0), that is the success fraction is 92.48% out of 2061 cases.

Rule: 4 If follow up = no and purpose of loan = 5 and Distance from MFI = far then the loan repayment = loss (78.0/14.0)

This rule show, if there is no customers follow up, the purpose of loan is different services, the customer is far from microfinance center then the loan repayment is loss (78.0/14.0) that is the success fraction is 84.78% out of 92 cases.

Rule: 5 If follow up = no and loan purpose =5 and Distance from MFI = middle then loan repayment = uncertain (88.0/4.0)

This rule show, if there is no customer follow up, the purpose of loan is different services, the customer is middle from microfinance center then the loan repayment is uncertain (88.0/4.0) that is the success fraction is 95.65% out of 92 cases.

Rule: 6 If follow up = no and purpose of loan = 4 and distance from MFI = middle then loan repayment = uncertain (2659.0/4.0)

This rule show, if there is no customer follow up, the purpose of loan is construction, the customer is middle from microfinance center then the loan repayment is uncertain (2659.0/4.0)that is the success fraction is 0.99.85% out of 2663 cases.

Rule: 7 If follow up = no and purpose of loan = 3 and distance from MFI = Middle the loan repayment is uncertain (1255.0/18.0)

This rule show, if there is no customer follow up, the purpose of loan is manufacturing, the customer is in the middle of microfinance center then the loan repayment is uncertain (1255.0/18.0)that is the success fraction is 98.59% out of 1273 cases.

Rule: 8 If follow up = no and loan purpose = 1 and distance from MFI = near and number of installment <= 9 then loan repayment = loss (9354.0/40.0)

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is near to microfinance center and number of installment <=9 then loan repayment is loss (9354.0/40.0) that is the success fraction is 99.57 % out of 9394 cases.

Rule: 9 If follow up = no and purpose of loan = 2 and distance from MFI = far and number of installment <= 1 and Installment amount > 4590.6 then loan repayment = uncertain (269.0/3.0)

This rule show, if there is no customer follow up, the purpose of loan is trade, the customer is far from microfinance center, number of installment is <= 1 and installment amount is >4590.6 then the loan repayment is uncertain (269.0/3.0)that is the success fraction is 98.90% out of 272 cases.

Rule: 10 If follow up = no and purpose of loan = 2 and distance from MFI = far and number of installment <= 1 and installment amount <= 4590.6 and amount disburse <= 3850 then loan repayment = uncertain (58.0/6.0)

This rule show, if there is no customer follow up, the purpose of loan is 2, the customer is far from microfinance center, number of installment is <=1, installment amount is <= 4590.6 and amount disburse is <= 3850 then loan repayment is uncertain (58.0/6.0)that is the success fraction is 90.63% out of 64 cases.

Rule: 11 If follow up = no and loan purpose = 1 and distance from MFI = middle and branch code = 320216 and marital status = 2 then loan repayment = loss (385.0/16.0)

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is middle from microfinance center, branch code is 320216 and marital status is 2 then loan repayment is loss (1349.0/1.0) that is the success fraction is 99.93% out of 1350 cases.

Rule: 12 If follow up = no and purpose of loan =1 and distance from MFI = far and amount disburse <= 15,000 and branch code = 20216 then loan repayment = loss (1349.0/1.0)

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is far from microfinance center, amount disburse is <= 15,000 and branch code is 20216 then loan repayment is loss (1349.0/1.0)that is the success fraction is 99.93% out of 1350 cases.

Rule: 13 If follow up = no and Purpose of loan = 1 and Distance from MFI = far and Amount Disburse <= 15,000 and Branch code = 170 and Installment amount <= 4927.35 and Marital status = 1 then loan repayment = loss (106.0/1.0)

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is far from microfinance center, amount disburse is <= 15,000, branch code is 170, installment amount is <= 4927.35 and marital status is 1 then loan repayment is loss (106.0/1.0)that is the success fraction is 99.07% out of 107 cases.

Rule: 14 Follow up = no and purpose of loan = 1 and Distance from MFI = far and Amount disburse <= 15,000 Branch code = 130 and Age = 32 – 45 and Marital status = 3 then loan repayment = loss (345.0/1.0)

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is far from microfinance center, branch code is 130, the age is between 32 and 45, and marital

status is 3 then loan repayment is loss (345.0/1.0) that is the success fraction is 99.71% out of 346 cases.

Rule: 15 If follow up = no and purpose of loan = 1 and distance from MFI = far and Amount disburse > 15,000 then loan repayment = loss (1906.0/155.0)that is the success fraction is 92.48% out of 2061 cases.

This rule show, if there is no customer follow up, the purpose of loan is agriculture, the customer is far from microfinance center and amount disburse is >15,000 then the loan repayment is loss (1906.0/155.0)that is the success fraction is 92.48% out of 2061 cases.

Rule: 16 If follow up = no and purpose of loan = 6 and Distance from MFI = near and branch code = 1320 then loan repayment = Uncertain (123.0/1.0)

This rule show, if there is no customer follow up, the purpose of loan is others, the customer is near to microfinance center and branch code is 1320 then loan repayment is uncertain (123.0/1.0) that is the success fraction is 99.19% out of 124 cases.

Rule: 17 If Follow up = no and purpose of loan = 6 and distance from MFI = middle and branch code = 32017 then loan repayment = uncertain (677.0/1.0)

This rule show, if there is no customer follow up, the purpose of loan is other, the customer is in middle of microfinance center and branch code is 32017 then the loan repayment is uncertain (677.0/1.0)that is the success fraction is 99.85% out of 678 cases.

Rule: 18 If follow up = no and purpose of loan = 4 and distance from MFI = middle then loan repayment = uncertain (2658.0/4.0) that is the success fraction is 99.85% out of 2662 cases.

This rule show, if there is no customer follow up, the purpose of loan is construction and the customer is in middle of microfinance center then customer loan repayment is uncertain (2658.0/4.0)that is the success fraction is 99.85% out of 2662 cases.

Rule: 19 If follow up = no and purpose of loan = 3 and distance from MFI = middle then loan repayment = uncertain (1255.0.0/18.0)

This rule show, if there is no customer follow up, the purpose of loan is manufacturing and the customer is in the middle of microfinance center then customer loan repayment is uncertain (1255.0/18.0) that is the success fraction is 98.59% out of 1273 cases.

Rule: 20 If follow up = no and purpose of loan = 1 and distance from MFI = middle then loan repayment = loss (657.0.0/1.0)

This rule show, if there is no customers follow up, the purpose of loan is agriculture and the customer is in middle of microfinance center then loan repayment is loss (657.0/1.0) that is the success fraction is 99.85% out of 658 cases.

5.6 Evaluation

As objectives of this research and research question, data mining goals are defined based on the problems in microfinance for classification and prediction of customer who return loan or not return before loan disburse. The goal are evaluated against the selected model built with J48 decision tree algorithm using 10-fold cross validation successfully met the objectives of data mining goals. Significant rules that are useful for predicting customers' loan repayment before loan disburse are extracted from the dataset. The domain experts during the discussions confirmed that most of the rules generated by the model are important in prediction of customers' loan repayment. Also made the discussion on the attributes used in this research work and out of fourteen attribute and one class attribute used, five most determinant attributes are identified by discussion with domain experts that are highly relevant in predicting customer loan repayment from customer datasets. These attributes are Customer follow up, Branch code (customer location), distance of customer from microfinance, purpose of loan, customer residence are found to be the most determinant attributes of customer loan repayment prediction.

From information gathered during interview and group discussion with domain experts, the microfinance use as one criterion to screen the customer to provide loan, the customer should be the permanent resident person in the kebele and should have social acceptance in the community. Also the loan is group loan, the borrowers organized in a group of minimum five and above and also each group member should know the background of each other to form a group loan member. The company screen the customer based on the information gained from each group members to allow the loan. Therefore, by integrating into their system the generated rules, it is

possible to screen customer who repay the loan from the one that not and also possible to manage the customer after loan disburse to collect the loan in the give time of loan collection. In this study data mining techniques have revealed an important in customer loan repayment prediction and support the microfinance in screening customers before loan disburse and also help managements for easy and quick decision making.

5.7 Deployment

The purpose of the data mining process is to increase the knowledge gained from the data stored. And, deployment is the last step of data mining process, which means using the data mining result of classification and prediction. The knowledge gained from data need to be organized and presented in a way that the organization can understand and use it for successful screen customers to integrate this rule into their system. To make this result applicable, integration of resources like people, business processes, and technology, are required. Moreover, the integration of resource is based on the information or result obtained from the classification model and some modification based on the interest of the organization. In this research work different models are developed using three classification algorithms and the result of the model are evaluated using performance evaluations. Finally the best model is selected. The rules are generated using decision tree original unbalanced datasets using selected model and the result of the research is discussed with domain experts. Based on discussion with domain experts this model is useful for the Oromia credit and saving Share Company to predict the customer loan repayment before loan disburse. To evaluate the use of the model in addition to the response of domain experts, we compare the previous systems of the organization: there is no system that support the organization to identify the customer loan repayment prediction before loan disburse but they use manual system to identify the customer that means they collect information from each group members about the customer ability to repay loan, such way leads the organization to wrong decision and also it takes their time to make decision. Besides domain experts, based on the result obtained the researcher discussed with the organizations stakeholders on the advantages of the result and ways of its implementation. Also even if the purpose of the study is academic purpose and the use of model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the Organization can use it. Prototype development is needed to show that the data mining model developed could be deployed. The researcher developed the prototype using python programming language for

twenty selected rules of the model as shown in figure 5.1 and sample of the code is given in the appendix part.

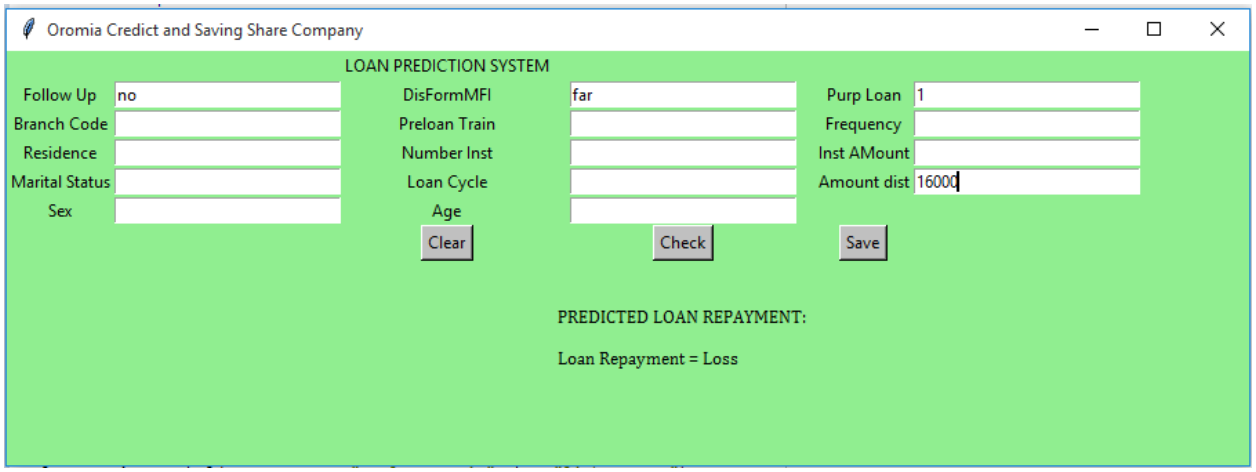

Figure 5.1 Screen shoot of prototype developed for the rule generated

Finally we recommend the organization after integrating the necessary adjustments by group of domain experts the result of this study can be deployed for customer loan repayment identification decision making and successful customer screening purpose in the organization and the organization will be beneficiary from the research result.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

The objective of this study is, to develop classification and predictive model that predict the loan repayment of customer using data mining techniques from Oromia Credit and Saving Share Company customer information datasets, evaluate data mining models in classifying and predicting the behavior of customer's in respect of their obligations, compare the model and select the best model based on its performance.

The review shows that in Ethiopia most microfinance has trouble in distinguishing the customer who returns the loan from the one who not return. This is partly due to fundamental problems in microfinance, and partly due to their own methods to screen the customers. In this perspective the review indicates that solutions adopted by microfinance often seemed inefficient from the perspective of a profit-maximizing microfinance by returning their loan. So, applying data mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated datasets to solve the problems in microfinance and enhance the decision making of microfinance. In the view of this, the ability to predict customer loan repayment is very important in microfinance sectors before loan disburses. In Oromia credit and saving share company, customers' loan repayment ability is based upon diverse factors, out of these the following features are selected with the help of domain expert and used: customers follow up after loan disburse, purpose of loan, Branch code (location of the customer), distance of customer from loan providing microfinance center, giving training before providing loan, frequency (interval in which customer return loan), customer residence, number of installment, installment amount, marital status of customer, loan cycle, amount disburse, age, sex of the customer are selected for data mining purpose in this research.

In this study the process of data preparation and modeling followed the CRISP-DM steps suggested in literature and methodology parts: Business understanding, data understanding, Data preparation (data selection, data pre-processing and cleaning, data transformation, data mining, interpretation, and validation of the results), Modeling, evaluation and deployment are used. The data is divided into training and testing datasets.

As discussed in chapter three no single algorithm or technique works best across all types of datasets problems. The choice of algorithms are governed by the important aspects of datasets being used, the problem area, research objective, and data preprocessing techniques involved, performance evaluation criteria, security, privacy, data integrity issues and the support of literatures are considered. As a result, in this study we use three algorithms: decision tree (J48), Naïve Bayes and Neural network using WEKA 3.8.2 machine learning software.

The model is built on the preprocessed microfinance customer datasets. The performances of the model are evaluated using the standard metrics of accuracy, precision, recall and F-measure. Percentage split and 10-fold cross validation model evaluation techniques is used for training and testing the model. All models that built performed well in predicting customer loan repayment behavior. The most effective model to predict the customer loan repayment behavior is J48 decision tree classifier using 10-fold cross validation model evaluation techniques implemented on all selected attributes with classification accuracy of 98.89. Finally, on the result of the experiment we discussed with domain experts and they confirm the rules generated are useful for customer loan repayment prediction. And also they confirm the following features: customer follow up, distance of the customer from loan providing microfinance center, purpose of loan and branch code that indicates the location area of the customer are the most determinant factor and ranked attributes that identified by using gain ratio attribute evaluator using best first method supported by domain expert.

The outcome of the study is highly useful for the microfinance institutes in predicting customer loan repayment behavior or revising existing customer loan repayment prediction that they use based on the information collected from the group of the customer before loan disburse.

6.2 Recommendations

Data mining has applied in many business sectors, particularly in microfinance data to support decision making. In this study data mining classification and prediction techniques are applied on Oromia credit and saving share company datasets and a good performance was achieved in this technique. The researcher recommends the following points based on the outcome of the research:

- Customer loan repayment prediction has many benefits for both lenders and borrowers. Customer loan repayment prediction helps to increase the speed and consistency of the loan application process and allows the automation of the lending process. Also, it greatly reduces the need for human intervention on customer screening evaluation and the cost of delivering credit. Therefore, the result of this study shows the possibility of applications of data mining techniques for loan prediction in microfinance. As a result we recommend the organization to implement the effective data mining techniques by integrating into their system using the result of this study.
- The researcher believes that there are other important features from the interview and discussions with domain experts' information obtained that not included in this study. Features for predictions of customer loan repayment. We recommend the organization to record customer information properly; also it is better to record other new customer features that contribute good values in prediction of customer loan repayment like: Saving habit of borrowers, educational background of customer, perception of borrowers on loan repayment period, source of income of the customer, past business experience of the customer, business type and family size of the customer.
- From the view point of the microfinance institutions as the researchers understand from the Mission of microfinance and different literatures, microfinance Provide affordable, innovative and customers' responsive financial services to rural and urban economically active people and for poor who has no guarantee to take loan from bank, to improve their life and income. We recommend the higher officials of the microfinance, to make a policy for awareness of society by using the result generated in data mining and other useful information, how they become active participant to improve their livelihood by repaying the loan and become beneficiary from the microfinance.

- The researcher recommends that the loan officers should follow up the customer after loan is provided and also providing advice, to use the loan for intended purpose that makes them more productive.
- We recommend that other researchers to conduct the research using other features like: Saving habit of borrowers, educational background of customer, perception of borrowers on repayment period, source of income of the customer, past business experience of the customer, business type, family size and the purpose of saving.
- The techniques employed in this study were decision tree, Naïve Bayes and neural network algorithm. Even though an encouraging result was obtained, using other types of techniques with changing different parameter might perform better therefore; we recommended other researchers to test with other types of techniques like: support vector machine or using other machine learning software.

Reference

[1].*Mengistu K., Mengistu U., Nigussie D., Endrias G., Mohammadamin H., Temesgen K., &Yemisrach G.(eds.), 2013. Proceedings of the National Conference on 'Loan and Saving: The Role in Ethiopian Socioeconomic Development', 15-16 February 2013, Haramaya, Ethiopia.*

[2].*WoldayAmha (2000). Networking Microfinance Activities in Ethiopia: Challenges and Prospects.Occasional Paper No. 1 .AEMFI. Addis Ababa, Ethiopia*

[3].*MoFED (Ministry of Finance and Economic Development), (2002).Sustainable developmentand poverty reduction program. Addis Ababa, Ethiopia.*

[4].*WaldayAmha (2001). Poverty Assessment in Ethiopia: The experience of Microfinance in poverty Reduction. July, 2001 Addis Ababa, Ethiopia*

[5]. *World Bank, 2008. Finance for All: Policies and pitfall in Expanding Access. A World Bank Policy Research Report – 2008, The World Bank, Washington DC.*

[6].*Shanmugan, B., Bourke, P. (1990).The Management of Financial Institutions: selected Readings. Addison: Wesley Publishing Company.*

[7].*Sufian, F., Parman S. (2009), Specialization and other determinants of non-commercial bankinancial institutions' profitability: empirical evidence from Malaysia. Studies in Economics and Finance, 26(2), 113 - 128.*

[8]Xiaohua Hu, (2005) A Data Mining Approach for Retailing Bank Customer Attrition Analysis. Applied Intelligence. Vol. 22, pp. 47–60.

[9]. *Mohammad, A. (2014), "What influences banks' lending in sub-Saharan Africa. Journal of Emerging Market Finance, 13(1), 1-42.*

[10].*Kashyap, A.K., Rajan, R., Stein, J.C. (1999), 'Banks as Liquidity Providers: An Explanation for the Co-existence of Lending and DepositTaking', NBER Working Paper Series, No. 6962.*

[11].*Greuning, H., Bratanovic, S.B. (2003), Analyzing and Managing Banking Risk: a Framewor for Assessing Corporate Governance and Financial Risk. 2nded. Washington, DC: The World Bank.*

[12]. *Barth, J.R., Caprio, G. Jr., Levine, R. (2004), Bank regulation and supervision: what works best? Journal of Financial Intermediation 13, 205 - 248*

[13].*Ghosh A, Nath B, Multi-objective rule mining using genetic algorithms, Information Sciences 163 (2004) 123–133; 2004.*

[14]. *Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data miningto knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G.Piatetsky-Shapiro, P. Smyth,and R. Ut.*

[15].*Silltow J., (2006).Data mining tools and techniques. United Kingdom*

[16]. *Newman D., (2013).Introduction to Data Mining. California: Irvine*

[17]. *Han, &Kamber J. M. (2001). Data mining: concepts and techniques. Morgan Kaufmann.*

[18].*Klosgen, W. Zytkow, J (2002) ''Handbook of Data Mining and Knowledge Discovery'', Oxford University Press*

[19].*Adrians P, Zantinge D. (1996). Data mining. Addison-Wesley Longman, England.*

[20]. A Nikoukar, IS Amiri, J Al (2015). International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2015, ISSN (Online): 2320-9801; ISSN (Print): 2320-9798

[21].VivekBhambri"Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011

[22].*MadanLalBhasin, "Data Mining: A Competitive Tool in the Banking and Retail Industries", The Chartered Accountant October 2006*

[23]. *B. Desai and Anita Desai, "The Role of Data mining in Banking Sector", IBA Bulletin, 2004.*

[24].S.S.Kaptan, "New Concepts in Banking", Sarup and Sons, Edition, 2002

[25]. *S. S. Kaptan, N S Chobey, "Indian Banking in Electronic Era", Sarup and Sons, Edition 2002.*

[26].*RajanishDass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad.*

[27]. *M Zaman, Predictive analytics; the future of business intelligence www.mahmoudyoussef.com*

[28]. *Chung, H. M., Gray, P. (1999), "Special Section: Data Mining". Journal of Management InformationSystems, (16:1),11-17.*

[29]. *Fayyad, U. Piatetsky-Shapiro, G. & Smyth, P. (1996), "From data mining to knowledge Discovery in database", American Association for Artificial Intelligence Press, Cambridge.*

[30].*Pradnya P. Sondwale, "Overview of Predictive and Descriptive Data Mining Techniques" IJARCSSE, Volume 5, Issue 4, April 2015*

[31]. *.Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013.*

[32].*BrijeshKumarBaradwaj, Saurabh Pal" Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011*

[33]. *Wolfgang Karl Hardle," Time Series Data Mining Methods: A Review", Berlin, March 25, 2015.*

[34].*Cios, K.J.:Pedrycz, W.Swiniarski, R.W, Data Mining A Knowledge Discovery Approach Kurgan, L. 2007, XV, 606 p., Hardcover ISBN:978-0-387-33333-5; http://www.springer.com/978-0-387-33333-5*

[35].*X.Wu, X.Zhu, Gong-Qing.Wu and W.Ding ," Data mining with big data". IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, Jan. 2014.*

[36].*H.Wang, G.Nie and K.Fu ," Distributed data mining based on semantic web and grid".IEEEInternational Conference on Computational Intelligence and Natural Computing, 2009.*

[37]. *Pete Chapman, Julian Clinton, Randy Kerber , Thomas Khabaza, Thomas Reinartz , Colin Shearer andRüdiger Wirth, R(2000).CRISP-DM 1.0 step by step data mining guide.*

[38].*Cios, K Witold, P Roman, S and Kurgan A. (2007). Data Mining A Knowledge Discovery Approach, Springer.*

[39].*Deshpande, S.P., &Thakare, D.V. (2010). Data mining system and applications: A review. International Journal of Distributed and Parallel System (IJDPS), Vol.1 No.1, 32-44.*

[40].*Maimon, o., &Rokach, L. (Eds.).(2005). Data mining and Knowledge discovery handbook (V0l.2). New York: Springer.*

[41].*Zupan, B., &Demsar, J. (2008). Open-source tools for data mining. Clinics in laboratory medicine, 28(1), 37-54.*

[42]. *Wu Jia, Vadera Sunil, and Dayson Karl, "A Comparison of Data Mining Methods in Microfinance", University of Salford.*

[43].*WakgariDibaba,"Application of Data Mining Techniques for Effective Customer Relationship Management of Microfinances: The Case of Wisdom Microfinance", Unpublished Master's project, Addis Ababa University, 2009.*

[44]. *Sara Worku, "Applications of Data mining Technology for Credit Risk Assessment in Addis Credit and saving institution",Unpublished Master's project, Addis Ababa University, 2016.*

[45].*BelachewRegane, "Application of Data Mining Techniques for Customers Segmentation and Prediction: The Case of BuusaaGonofa Microfinance Institution" Unpublished Master's project, Addis Ababa University, 2013.*

[46].*AboobydaJafar Hamid and Tarig Mohammed Ahmed, "Developing Prediction Model of Loan risk in Banks Using Data Mining",University Khartoum, Sudan*

[47]. *Marcos de Moraes Sousa and Reginaldo Santana Figueiredo, "Credit Analysis Using Data Mining: Application in the Case of a Credit Union",Universidade Federal de Goiás, Goiás, Brasil*

[48].*Jiawei* Han *and MichelineKamber. Data Mining: Concepts and Techniques, 2nd edition.Morgan Kaufmann Publishers, March* 2006. *ISBN 1-55860-901-6.*

[49]. *Wu, S. (2013). "A review on coarse warranty data and analysis", Reliability Engineering and System, 114: 1–11, doi:10.1016/j.ress.2012.12.021*

[50].*Dileep B. Desai, Dr. R.V.Kulkarni "A Review: Application of Data Mining Tools in CRM for Selected Banks", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2), 2011, 199 – 201.*

[51].*Nashaat El-KhamisyMohamed , Ahmed ShawkyMorsi El-Bhrawy: "Artificial Neural Networks in Data Mining", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 55-59: www.iosrjournals.org*

[52]. *Zell, Andreas (1994). Simulation NeuronalerNetze [Simulation of Neural Networks] (in German) (1st ed.). Addison-Wesley.p. 73. ISBN 3-89319-554-8*

[53]. *Rosenblatt, Frank. x. Principles of Neuro dynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961*

[54].*Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart,James L. McClelland, and the PDP research group .(editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.*

[55].*Cybenko, G. 1989. Approximation by super positions of a sigmoidal function Mathematics of Control, Signals, and Systems, 2(4), 303–314.*

[56]. *Harry Zhang "The Optimality of Naive Bayes". FLAIRS2004 conference. (available online: PDF (http://www.cs.unb.ca/profs/hzhang/ publications/FLAIRS04ZhangH.pdf))*

[57]. *Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms".Proceedings of the 23rd international conference on Machine learning, 2006. (available online PDF (http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.122.5901& rep=rep1&type=pdf))*

[58]. *Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco*

[59].*Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.6, no., pp.161,165, 8-10 April 2011*

[60].*Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees."Image Processing Division, National Institute for Space Research—INPE*

[61].*Arbres de décision, Ingénierie des connaissances (Master 2 ISC)*

[62] .*Jiawei Han, M. Kamber and J. Pei (2012).Data mining: concepts and techniques. Third edition, Morgan Kaufmann Publishers is an imprint of Elsevier*

[63]. *C. Velayutham and K. Thangavel, "Unsupervised Quick Reduct Algorithm Using Rough Set Theory", Journal of Electronic Science and Technology, Vol. 9, No. 3, September 2011*

[64]. *Pyle, D., 1999 Data Preparation for Data Mining Morgan Kaufmann Publisher.*

[65].*Zubair K. (2014). A survey of data mining: Concepts with applications and its future scope: International of computer science trends and technology Vol.2 ISS3*

[66]. *Witten H. Ian and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann Publishers, San Francisco*

[67].*Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse, (2014): WEKA Manual for version 3.8.2*

[68]. *Ana Azevedo and M.F. Santos (2008), KDD, SEMMA AND CRISP-DM: a parallel overview*

[69]. http://www.crisp-dm.org/

[70].Pal, N.R., Jain, L.C., (Eds.) 2005. Advanced Techniques in Knowledge Discovery and Data Mining, SpringerVerlag)

[71]. VladislavPyzhov and StanislavPyzhov (2017), Comparison of methods of data mining Techniques for the predictive accuracy, online at https://mpra.ub.uni-muenchen.de/79326/ *MPRA Paper No. 79326, posted 27 May 2017 04:43 UTC*

[72]. B. Kiranmai, Dr. A. Damodaram (2014). 'A Review on Evaluation Measures for Data Mining Tasks'. International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 3 Issue 7 July, 2014 Page No. 7217-7220

[73] Jiawei Han, MichelineKamber(2001). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

[74] Foundation of Organization andBackground Document, BPR document, Operational manual of OromiaCredit and Saving Share Company, 2017

[75] Chaia, A. J. (2003). Modelos de gestão de risco de crédito e suaaplicabilidadeao mercadobrasileiro. Dissertação de Mestrado. FEA/USP.

[76]. K. Tsiptsis and A. Chorianopoulos, "Data Mining Techniques in CRM: Inside Customer Segmentation", John Wiley and Sons, Ltd., Publication, 2009.

[77]. Ledgerwood*, J. 1999.'Microfinance Handbook': An Institutional and Financial Perspective. Washington: the World Bank.*

[78]. Hunte, C. K. 1996. 'Controlling Loan Default and improving the lending Technology in Credit Institutions', Savings and Development. Vol. xx, No.1, available at <www.gdrc.org/icm/grameen-ref.html>, viewed on March 07, 2011.

APPENDICES

Appendix 1: Sample rules generated from Decision tree J48 Classifier

```
J48 pruned tree
------------------

follow_up = no
|   Purp_Loan = 1
|   |   DisFromMFI = far
|   |   |   Amount_Disb <= 15000
|   |   |   |   Branch Code = 13: Loss (79.0/4.0)
|   |   |   |   Branch Code = 16: Loss (89.0/1.0)
|   |   |   |   Branch Code = 17: Loss (52.0/1.0)
|   |   |   |   Branch Code = 20: Loss (70.0)
|   |   |   |   Branch Code = 130
|   |   |   |   |   Age = 32-45
|   |   |   |   |   |   Marital_stus = 1
|   |   |   |   |   |   |   Inst_ Amount <= 291.47: Loss (24.0)
|   |   |   |   |   |   |   Inst_ Amount > 291.47
|   |   |   |   |   |   |   |   Amount_Disb <= 5809.51: Uncertain (8.0/1.0)
|   |   |   |   |   |   |   |   Amount_Disb > 5809.51: Loss (8.0)
|   |   |   |   |   |   Marital_stus = 2: Loss (6.0)
|   |   |   |   |   |   Marital_stus = 3: Loss (345.0/1.0)
|   |   |   |   |   |   Marital_stus = 4: Loss (0.0)
|   |   |   |   |   Age = 18-31
|   |   |   |   |   |   Loan_cyc = 0: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 1
|   |   |   |   |   |   |   Inst_ Amount <= 282.73
|   |   |   |   |   |   |   |   Sex = F
|   |   |   |   |   |   |   |   |   Amount_Disb <= 2400: Loss (4.0)
|   |   |   |   |   |   |   |   |   Amount_Disb > 2400: Uncertain (5.0)
|   |   |   |   |   |   |   |   Sex = M: Loss (6.0/1.0)
|   |   |   |   |   |   |   Inst_ Amount > 282.73: Loss (6.0/1.0)
|   |   |   |   |   |   Loan_cyc = 2: Uncertain (5.0/1.0)
|   |   |   |   |   |   Loan_cyc = 3: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 4: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 5: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 6: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 7: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 8: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 9: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 10: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 11: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 12: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 13: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 14: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 90: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 150: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 180: Loss (0.0)
|   |   |   |   |   |   Loan_cyc = 210: Loss (0.0)
|   |   |   |   |   Age = 45-58: Loss (10.0/3.0)
|   |   |   |   |   Age = 59-70: Loss (0.0)
|   |   |   |   Branch Code = 133: Loss (20.0/1.0)
|   |   |   |   Branch Code = 134
|   |   |   |   |   Sex = F
```

Appendix 2: Sample taken from WEKA classification result Decision tree 10-fold cross validation

```
Correctly Classified Instances      145647               98.8879 %
Incorrectly Classified Instances      1638                1.1121 %
Kappa statistic                          0.9795
Mean absolute error                      0.0112
Root mean squared error                  0.0807
Relative absolute error                  3.0847 %
Root relative squared error             18.9865 %
Total Number of Instances           147285

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.992    0.016    0.990      0.992    0.991      0.977    0.996     0.996     Loss
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Standard
                 0.952    0.005    0.963      0.952    0.958      0.951    0.992     0.966     Uncertain
Weighted Avg.    0.989    0.011    0.989      0.989    0.989      0.979    0.997     0.993

=== Confusion Matrix ===

     a     b     c   <-- classified as
 89463     0   702 |     a = Loss
     0 37679     0 |     b = Standard
   936     0 18505 |     c = Uncertain
```

Appendix 3: Sample taken from WEKA classification result Naïve Bayes 10-fold cross validation

```
Correctly Classified Instances      132412               89.9019 %
Incorrectly Classified Instances     14873               10.0981 %
Kappa statistic                          0.8012
Mean absolute error                      0.0661
Root mean squared error                  0.2428
Relative absolute error                 18.2734 %
Root relative squared error             57.0948 %
Total Number of Instances           147285

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980    0.219    0.876      0.980    0.925      0.798    0.980     0.990     Loss
                 0.986    0.000    1.000      0.986    0.993      0.991    0.996     0.997     Standard
                 0.356    0.018    0.747      0.356    0.482      0.471    0.970     0.723     Uncertain
Weighted Avg.    0.899    0.137    0.891      0.899    0.884      0.804    0.983     0.956

=== Confusion Matrix ===

     a     b     c   <-- classified as
 88333     1  1831 |     a = Loss
     4 37155   520 |     b = Standard
 12517     0  6924 |     c = Uncertain
```

Appendix 4: Sample taken from WEKA classification result neural network 10-fold cross validation

```
Correctly Classified Instances      144218               97.9176 %
Incorrectly Classified Instances      3067                2.0824 %
Kappa statistic                          0.9613
Mean absolute error                      0.016
Root mean squared error                  0.1134
Relative absolute error                  4.4258 %
Root relative squared error             26.6681 %
Total Number of Instances           147285

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.991    0.039    0.976      0.991    0.983      0.956    0.991     0.990     Loss
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Standard
                 0.885    0.007    0.953      0.885    0.918      0.907    0.984     0.945     Uncertain
Weighted Avg.    0.979    0.025    0.979      0.979    0.979      0.961    0.992     0.987

=== Confusion Matrix ===

     a     b     c   <-- classified as
 89325     0   840 |     a = Loss
     0 37679     0 |     b = Standard
  2227     0 17214 |     c = Uncertain
```

Appendix 5: sample Code for implementing rules

```
def check():
    followup = Follow_Up_field.get()
    purposeOfLoan =  Purp_Loan_field.get()
    distFromMFI = DisFormMFI_field.get()
    amountDisburse = Amount_dist_field.get()
    branchCode = Branch_Code_field.get()
    numbOfinstallment = Number_Inst_field.get()
    installmentamount = Inst_AMount_field.get()
    maritalStatus = Marital_Status_field.get()
    age = Age_field.get()
    print(followup)
    print(purposeOfLoan)
    print(distFromMFI)
    print(branchCode)
    print(amountDisburse)

    if distFromMFI == '10':
        print('true')
    else:
        print('false')

    standard ="Loan Repayment = Standard"
    uncertain = "Loan Repayment = Uncertain"
    loss = "Loan Repayment = Loss"
    unknown = "Invalid Input"
    #Rule: 1 If follow up = Yes then the loan repayment = Standard (37679.0)
    if followup == "yes":
        res.configure(text = "\n\nPREDICTED LOAN REPAYMENT: \n\n" +  str(standard))
    repayment = standard
    #return repayment
#Rule: 2 If follow up = No and Purpose of loan =1 and distance from MFI = far and amount disburse <= 1500 and branch code = 16 then loan repayme
elif followup == "no" and int(purposeOfLoan) == 1 and distFromMFI == "far" and float(amountDisburse) <= 1500 and int(branchCode) == 16:
    res.configure(text = "\n\nPREDICTED LOAN REPAYMENT: \n\n" +  str(loss))
    repayment = loss
    #return repayment
#Rule: 3 If follow up = No and purpose of loan = 1 Distance from MFI= far and amount disburse > 15,000 then loan repayment= loss (1906.0/155.0)
elif(followup == "no" and
   int(purposeOfLoan) == 1 and
   distFromMFI == "far" and
   float(amountDisburse) > 15000):
   res.configure(text = "\n\nPREDICTED LOAN REPAYMENT: \n\n" +  str(loss))
   repayment = loss
   #return repayment
#Rule: 4 If follow up = no and purpose of loan = 5 and Distance from MFI = far then the loan repayment = loss (78.0/14.0)
elif(followup == "no" and
    int(purposeOfLoan) == 5 and
    distFromMFI == "far"):
    res.configure(text = "\n\nPREDICTED LOAN REPAYMENT: \n\n" +  str(loss))
    repayment = loss
    #return loss
#Rule: 5 If follow up = no and loan purpose =5 and Distance from MFI = middle then loan repayment = uncertain (88.0/4.0)
elif(followup == "no" and
    int(purposeOfLoan) == 5 and
    distFromMFI == "middle"):
    res.configure(text = "\n\nPREDICTED LOAN REPAYMENT: \n\n" +  str(uncertain))
    repayment = uncertain
#Rule: 6 If follow up = no and purpose of loan = 4 and distance from MFI = middle then loan repayment = uncertain (2659.0/4.0)
elif(followup == "no" and
    int(purposeOfLoan) == 4 and
    distFromMFI == "middle"):
    res.configure(text = "\n\nPREDICTED LOAN REPAYMENT: \n\n" +  str(uncertain))
    repayment = uncertain
    #return uncertain
#Rule: 7 If follow up = no and purpose of loan = 3 and distance from MFI = Middle the loan repayment = uncertain (1255.0/18.0)
elif(followup == "no" and
    int(purposeOfLoan) == 3 and
```

Appendix 6: Interview questionnaires

1. Is there a means that the organization uses to identify the loan repayment capacity of the customers who take the loan?
2. What is the best feature to consider in passing loan decision by the company?
3. List some factors that makes your customers influence to pay back their loan according to their agreement and obligations.
4. How the MFIs identify that the borrowers use the loan for intended purpose of loan?
5. How often follow up your customers that they loaned money is applicable for its purpose the take?
6. What are the criteria that used to approve the loan?
7. If the customer starts to fall behind in his repayments what actions should the organization take?

DECLARATION

I declare that this research work is my original work and has not been submitted for any academic qualification in this or any other University for examination. Where other sources of information have been used, they have been acknowledged.

Ketema Feyissa

Signature: ----------

Date: ----------

This Thesis has been submitted for examination with my approval as a university advisor.

Wondwossen Mulugeta (PhD)

Signature: ----------

Date: ----------