ST. MARY'S UNIVERSITY FACULTY OF INFORMATICS

DEPARTMENT OF COMPUTER SCIENCE

PREDICTING BANK CREDIT RISK USING DATA MINING TECHNIQUE: THE CASE OF BANK OF ABYSSINIA

BY TENINET BELAY ALEMU

ID SGS/0173/2009B

ADVISOR(S): MILLION (PHD)

JUNE, 2019

ST. MARY'S UNIVERSITY FACULTY OF INFORMATICS

DEPARTMENT OF COMPUTER SCIENCE

PREDICTING BANK CREDIT RISK USING DATA MINING TECHNIQUE: THE CASE OF BANK OF ABYSSINIA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ST.MARY'S UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

BY TENINET BELAY ALEMU

ID SGS/0173/2009B

ADVISOR(S): MILLION MESHESH (PHD)

June, 2019

ST. MARY'S UNIVERSITY FACULTY OF INFORMATICS

DEPARTMENT OF COMPUTER SCIENCE

PREDICTING BANK CREDIT RISK USING DATA MINING TECHNIQUE:

THE CASE OF BANK OF ABYSSINIA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ST.MARY'S UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

By

TENINET BELAY ALEMU

Name and Signature of Members of the Examining Board

Name **Date** **Signature**

1. Chairman, Examining Board ________________ ____________ ____________
2. Dr. Million Meshesha , Advisor ________________ ____________ ____________
3. External Examiner ________________ ____________ ____________

DECLARATION

The thesis is my original, has not been presented for a degree in any other university and that all sources of material used for the thesis have been duly acknowledged.

TENINET BELAY ALEMU
June, 2019

Place: Addis Ababa

Date of Submission: ___________________

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

Million Meshesha (PhD) ____________________ ______________

Advisor's Name Signature Date

ACKNOWLEDGMENTS

First and foremost, I would like to thank God Almighty for making it possible for me to come this far in my studies.

Secondly, my deepest gratitude goes to my advisor Dr. Million Meshesha for his invaluable guidance, support and encouragement in carrying out my thesis work ,I am really grateful for his constant supervision and constructive comments up to the submission of my thesis. As my advisor, his observations, guidance and comments help me to find the right direction of the work to move forward and to complete the thesis.

Also, I want to thank to Ato Tewodros Haile, Manager of Bank of Abyssinia, portfolio Division, for his unreserved help and advice during the data collection and domain expert Discussion and all staffs who have working on Banks portfolio and Credit department for their unreserved support and comment on data collection and attribute selection.

My deepest thank goes to my family for their emotional support and encouragement. I also want to thank all the people who have directly or indirectly helped me throughout the course of this thesis work.

I would like to dedicate the work to my lovely Mother, Tiruwork Bekalu for her unreserved love and support for the accomplishment of my Thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATION

ARFF- Attribute-Relation File Format

BOA: Bank of Abyssinia

CRA: Credit Risk Assessment

BCRPM: Banking Credit Risk Prediction Model Prototype

CRISP-DM: Cross-Industry Standard Process for Data Mining

CSV: Comma Separated Value

KDD: Knowledge Discovery in Databases

LC: Letter of Credit

NBE: National Bank of Ethiopia

NPL: Non-performing Loan

SEMMA: Sample, Explore, Modify, Model, and Assess

SPSS: ststistical package for social scinces

SQL: Structured Query Language

WEKA: Waikato Environment for Knowledge Analysis

ABSTRACT

Credit facilities and investments are the cornerstones of the growing economy of Ethiopia. Bank of Abyssinia being one of the former private banks has played its own role in the economy by rendering loan facilities to the individuals and companies which are running business in various sectors. The bank uses internal and National bank credit policies, procedures and strictly followed manuals in various levels of credit committees before disbursing loan to customers. However, there are total defaulters and inconsistent loan repaying customers which declines the profitability of the bank in particular and threatens the growing economy of the country in general. While fueling the sprinting economy in the country, minimizing the possible defaulters is the prime concern of the bank. It is there for the main objective of this study is to apply data mining to predict banking credit risk in Bank of Abyssinia S.C.

Identifying customers and contracts which are more likely to be inconsistent loan payers or defaulters is an important issue .This data mining research has been carried out to identify trends of Low risky and High risky or NLP(non-performing loan)patterns from the historic data and build predictive model to assist the management of the bank.

For conducting experiment a six-step hybrid Knowledge Discovery Process model is used. The required data was collected from the Portfolio and Credit department of the Bank and pre-processed the data for mining using Weka software. The researcher used three data mining algorithms (J48 Decision Trees, JRip rules induction and Naïve Bayes) to develop the predictive model.

The results indicated that J48 decision tree is the best predictor with **97.0167**%

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Ethiopia is one of the poorly developed countries in terms of infrastructure and services [17]. The banking sector is one of the cornerstones which play its own role in the development of the country. The number of private banks in the country is increasing from year to year. The presence of many banks in the country has created aggressive competition in the traditional brick and mortar based market. All private and governmental banks have partially or fully automated their banking operations as per the national bank directive to implement standard core banking solution.

Many of these private and governmental banks have a huge amount of data which is used for statement, auditors' verification of transactions and functional level reporting purposes. In order to discover the set of critical success factors that will help banks reach their strategic goals and remain in the competition, they need to move beyond standard business reporting and sales forecasting. They should learn from their abundant historical data, by applying data mining and predictive analytics to extract actionable intelligent insights and quantifiable predictions. These insights can support the decision of management, auditors and other clerical staffs in the pillar activities of the bank like credit risk assessment (CRA) to grant loan to a customer.

According to Chamatkar et al. [18], Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, and the web. It uses statistical and pattern matching techniques. The concern in data mining are noisy data, missing values, static data, sparse data, dynamic data, relevance, interestingness, heterogeneity, algorithmic efficiency and the size and complexity of data. The data we have is often vast, and noisy, meaning that it's imprecise and the data structure is complex. This is where a purely statistical technique would not succeed, because of its vastness, so data mining is a solution. Data mining has become a popular tool for analyzing large datasets. The efficient database management systems have been very important assets for management of a large corpus of data, especially for effective and efficient retrieval of particular information from a large collection whenever needed.

The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Information retrieval is simply not enough anymore for decision-making because information must be extracted and available [23].

The information or knowledge extracted so can be used for any of the following applications [28]; market Analysis, Fraud Detection, Customer Retention, Production Control, and Science Exploration.

According to Baseer [33], since bank is service industry, maintaining strong and effective customer resource management (CRM) is critical issue and to do this using data mining tool is helpful for banks to understand their existing and future customer so that they deliver tailor made product and service to their customer [33].

Data mining is a powerful technology with great potential to help banking firms focus on the most important information in the data they have collected about the behavior of their existing and potential Creditors. Data mining assists Banking sector in predicting Creditors status nonpayment loan, medical coverage and predicting creditor's loan payment pattern [33].

Data mining is applied in the banking industry and tremendous competitive advantages accrue to those industries who have implemented it successfully. Some of the areas data mining can be applied to banking industry are in identifying risk factors that predict profits, claims and losses, creditor level analysis, marketing and sales analysis, developing new product lines, financial analysis ,estimating outstanding claims provision and detecting fraud.[41].

The aim of this study is to assess the applicability of data mining techniques in banking industry to build models that can predict Creditors loan repayment based on the historical data.

1.2 BANK OF ABYSSINIA

The present-day Bank Of Abyssinia was established on February 15, 1996 (90 years to the day after the first but defunct private bank was established in 1906 during Emperor Menelik II) in accordance with 1960 Ethiopian commercial code and the Licensing and Supervision of Banking Business Proclamation No. 84/1994. BoA started its operation with an authorized and paid up capital of Birr 50 million, and Birr 17.8 million respectively, and with only 131 shareholders [27]. In two decades since its establishment BoA has registered a significant growth in paid up capital and total asset. It also attracted many professional staff members, valuable shareholders and large customers from all walks of life. This performance indicates public confidence in the Bank and reliability and satisfaction in its services. Currently, employing the state-of-art banking technology, the Bank provides excellence domestic, international and special banking services to its esteemed and valuable customers. It also strives to serve all economic and services sectors via its ever increasing branch networks throughout the country. Currently the bank has more than 5825 staff and more than 340 branches and more than

1,012,177. In addition, an authorized and paid up capital of BoA as at 30 June 2018 is Birr 4.24 billion and Birr 2.56 billion, respectively, a total deposit balance of Birr 25.9 Billion and a total loans and advances of Birr 17.99 billion, which in effect enhance the risk absorbing and the lending capacity of the Bank. Bank of Abyssinia‘s vision is to be the bank of choice for customers, employees and shareholders. Its mission is to provide customer-focused financial services through competent, motivated employees and modern technology in order to maximize value to all stakeholders [27].

Bank of Abyssinia S.C being one of the technologically rich banks in the country, has introduced many channels like SMS, Mobile banking, Agent banking, internet, ATM and POS banking services. It has been using legacy system application since its inception and is using core banking solutions since 2011 [26]. These years of day-to-day transactions created terabytes of data which are collected, generated, printed, and stored only for the sake of customer statements, reports for lower level functional managers and auditors.

Bank of Abyssinia Core banking system includes the loan module, where credit contract detailed information is booked, and a huge number of credit related transactions like repayment, daily interest accruals, status changes, full liquidations and contract amendment information is captured over several years.

The analysis of these data will leads to a better understanding of the customer’s profile and attached contract, thus supporting the offer of new products or services and identification of risky disbursements. These data usually hold valuable information like trends and patterns, which can be employed to improve credit assessment. The bulky nature of the data makes its manual analysis an impossible task. In many cases, several related features need to be simultaneously considered in order to accurately model credit contract behavior.

Credit facility is the corner stone of Bank of Abyssinia existence and profitability. As a result, the automatic extraction of useful credit knowledge from its historic data will be an input to the credit risk grading process. Hence, in order to assist the management of the bank to make effective decisions in reducing risk of defaulters and focus on risk free sectors and customers, implementing data mining will play a great role.

1.3 STATEMENT OF THE PROBLEM

Since the enactment of the proclamation for the licensing and supervision of banks, the numbers of competent and powerful banks with the state of the art technology are increasing every year. Despite the number of banks and potential customers, the banking services are so limited and nearly homogenous which may be attributed to several factors. The presence of many private banks and homogenous service strategy has created a good opportunity for credit customers to identify the structured requirement of banks easily. These customers will apply credit for different banks, some with misleading formalities at different times. This environment has created a brutal competition between banks and kept them struggling in cash collection and in rendering risk free credit facility.

Management and measurement of risk is at the core of every financial institution. Today's major challenge in the banking and insurance world is therefore the implementation of risk management systems in order to identify, measure, and control business exposure. Here credit and market risk present the central challenge. One can observe a major change in the area of how to measure and deal with them, based on the advent of advanced database and data mining technology [19].

Loan (credit facility) is one of the major services that contribute to the lion share of profitability of any bank in Ethiopia. In order to lend money to customers, banks need to collect more cash from customers and other various means. Currently each and every bank is extending its effort in deposit mobilization by various methods like branch expansion, increased interest rate for fixed time depositors and normal depositors. This hard fetched deposit is later given as a credit to various customers. But there will be inconsistent loan repayments, defaulters and corruptions related to various sector credit facility customers [28]. Therefore, Banks are in need of loan repayment prediction and customer credit analysis from their historic credit facility data. In order to alleviate these stated problems and find new, applicable and interesting classifications and predictions for loan facilitation and decision making the banks need to look deep into their historic data.

The credit risk assessment of customers involves structured and unstructured management decision elements [29]. The structured decisions are those where the processes necessary for the granting of loan are known beforehand and several computational tools to support the decisions are available. For non- structured decisions, only the managers' intuition and experience are used. Specialists may support these managers, but the final decisions involve a substantial amount of subjective elements. Data mining comes here into picture to assist the unstructured decision of management in predicting and execution of the necessary follow up procedures.

The traditional approach of credit risk assessment employed by financial managers largely depends on their previous experience and does not follow the procedures defined by their institutions. The large amount makes its manual analysis an impossible task. In many cases, several related features need to be simultaneously considered in order to accurately model credit user behavior. These needs for automatic extraction of useful knowledge from a large amount of data and in this regard data mining assume a very important role in credit risk assessment by allowing the replacement of general risk assessment by careful analysis of each loan commitment.

A number of researches have been conducted in credit risk assessment using data mining technique for banks in Ethiopia, Askale Worku [22], works on applying data mining technology for credit risk assessment for Dashen bank, Mertework Shawel [42], on the other hand works on applying data mining technology in supporting credit risk assessment in case of Nib international bank, Mengestu Tesefay [43], further applied credit risk assessment in case of united bank. And their result show that the problem in credit risk assessment can be solved by the use of data mining technology.

Therefore, this research try to address the following basic and general questions:-

✓ What are the suitable attributes to apply data mining for credit risk assessment?

✓ What are the major tasks for improving the quality of data sets?

✓ Which data mining techniques can best be used for the credit facility (loan) risk assessment area?

✓ What are the interesting patterns and relationships for the risky and risk free credit contracts of Bank of Abyssinia?

✓ To what extent the model works in identifying the credit risk?

1.4 OBJECTIVE OF THE STUDY

1.4.1 GENERAL OBJECTIVE

The general objectives of the study is to design a predictive model for the credit risk assessment of Bank of Abyssinia in order to get a competitive edge and better customer satisfaction.

1.4.2SPECIFIC OBJECTIVES

The Specific objectives of this research was concentrated on the following key points:

✓ To identify suitable attribute for credit risk assessment of the bank;

✓ To prepare data set for training and testing purpose

✓ To select the best prediction model and find out interesting patterns from the output of the selected Model;

✓ To design a user interface that enables to assess the applicability of the pattern(knowledge) generated by the data mining approach;

1.5 SCOPE OF THE STUDY

The main target of this study was to apply data mining techniques in the credit risk assessment of bank of Abyssinia. This is an area of study that would be more fruitful if it will conducted widely by including all private and governmental banks of Ethiopia.

To conduct the study, there were some limiting factors such as schedule and budget in accordance with the objective and aim set by the researcher, the coverage of this research would be on bank of Abyssinia S.C. only. Bank of Abyssinia has been using "Temenos T24 (Temenos financials) core banking solution" with underlying oracle database for the last 9 years. It has also been using a legacy system before the introduction of Temenos 24. Due to resource and time limitations to merge the two system information into a data warehouse, this research will be conducted on the last 9 years Temenos T24 core banking data. In this study both descriptive and predictive modelling data mining tasks are applied.

In descriptive modeling, Creditor groups are clustered according to demographics, Credit behavior, expressed interests and other descriptive factors. Statistics can identify where the Creditor groups share similarities and where they differ. The most active Creditor get special attention because they offer the greatest (return on loan).

The aim of predictive data mining modelling is to find a description of how certain attributes within the data will behave in the future. For example, in credit risk applications, the analysis of creditor history to predict the probability of the return rate of the loan with in a given period of time.

1.6 SIGNIFICANCE OF THE STUDY

Primarily, the researcher gained an experience of conducting a research as this study was conducted for academic purpose; hence, the finding of this study, could motivate other researchers to conduct further researches in the area

Secondly, the study contributes a lot for the Bank of Abyssinia as a starting point if the bank needs to implement data mining on its database to get extra competitive edge. Hence this study will be conducted to provide sufficient information on credit risk assessment using data mining to uncover hidden knowledge based on the bank's historic data.

Thirdly, as the services and style of product development is nearly similar for Ethiopian banks, the research can also be used by other banks in the country to facilitate decision making and earn better profit margins by reducing credit risks. It also serves as a starting point for those individuals who would like to undertake broader research on the topic and also this research helps for those who make polices and different directive for the governors.

Fourthly, the finding of this research used by bank to increase the quality of service given to its creditors in order to maintain the standard or the quality of services. In other words, the customers can be beneficiaries of the quality service provision.

Fifthly, the national bank or the governing body used as an input to develop rules and regulations for credit facilities and services.

1.7 METHODOLOGY

This research is designed to apply data mining technology for the credit Risk assessment of Bank of Abyssinia. Methodology is the process used to collect information and data for the purpose of making business decisions. The methodology may include publication research, interviews, surveys and other research techniques, and could include both present and historical information.

1.7.1 RESEARCH DESIGN

This research follows experimental research. Experimental research is a study that strictly adheres to a scientific research design [30]. It includes a hypothesis, a variable that can be manipulated by the researcher, and variables that can be measured, calculated and compared. Most importantly, experimental research is completed in a controlled environment. The researcher collects data and results which is either support or reject the hypothesis. This method of research is referred to a hypothesis testing or a deductive research method [30].

To undertake the experiment in this research, the researcher has followed Hybrid Data Mining Process Models. According to, Swiniarski and Kurgan [21], the hybrid model is enhanced the knowledge discovery process by combining the academic and industrial models in data mining projects. The development of hybrid model was adopted from the CRISP-DM model as its can be used for academic research. Thus, these models are research-oriented, which present data mining step than the modeling step. The six steps of hybrid models allow a number of feedback mechanisms. Moreover, the knowledge discovered in the final step for a specific domain may be applied in other domains.

THE SIX STEPS OF HYBRID DATA MINING PROCESS MODEL INCLUDES

1.7.2 UNDERSTANDING OF THE PROBLEM DOMAIN

The initial step involves task such as the problem definition and project goal determination, identification of key people and grasping the current solution to the problem through close consultation with the domain experts. Then the project goals are transformed to DM goal and the preliminary selection of DM tools to be used in the study is conducted.

In this step of hybrid model the main task of the research which accomplished by using different mechanisms. Review of documents including different manuals; policy documents, procedure documents, Different National Bank Credit directives , different credit policy has an impact in understanding the problem domain and different reports; the reports which reviewed include annual financial report of the Bank . In this case the researcher became in a good position in order to define the problem, determine the domain objectives, to assess the problem and to determine the data mining goals, which are helpful for the next steps.

1.7.3 UNDERSTANDING OF THE DATA

This step includes collecting the initial Credit data from Bank of Abyssinia Portfolio Department. The data can be visualize using excel format. The visualization focuses on insuring the data completeness since there are creditor's records which has incomplete data. Similarly, the visualization can be focused on cross checking the suitability of the data concerning the data mining goals. Besides exploring the data and describe the data to identify the possible attributes is another task of understanding the data. Additionally, this step enables to determine the data mining methods and algorithms. Moreover, WEKA is another data visualization tool, which uses in describing the attributes.

1.7.4 PREPARATION OF THE DATA

In this step, the data used are prepared to apply the DM methods. It consist of tasks such as sampling, testing the correlation and significance of the data, cleaning the data, checking the completeness of the tuples, handling noisy and missing values. Then, the dimensionality of the data is reduced by feature selection and extraction algorithms. This step also comprises, the derivation new attributes, summarization of the data. Finally, the datasets that meet the input requirements of DM tools stated in the first step are selected for modeling purpose.

It is also include all activates that are needed to construct the final data set. The data which is taken from Bank of Abyssinia Portfolio Department preprocessed and cleaned for the application of different data mining techniques. Bank of Abyssinia Creditors data contain many missing values while felling creditor's details in to the system.

1.7.5 DATA MINING

This step involves DM methods that are applied on the preprocessed data to discover knowledge.

This step is application of the selected data mining methods to the prepared Creditors data and testing the generating rules whether they achieve the required minimum threshold. Beside it is a step of finding hidden, non - trivial and previously unknown information from the data. In this study, association rule mining approach is applicable on the processed data. WEKA is used for mining the data. WEKA is one of the open source DM tools having several functionalities. A study done by Meneses and Grinest [53], stated that, as compared to other DM tools WEKA toolkit:

✓ It's achieved the highest applicability.
✓ Achieved the highest improvements (when moving from the Percentage Split test mode to the Cross Validation test mode)
✓ Is the best tool in terms of the ability to run the selected classifiers
✓ Can better handle the problem of multiclass data sets For these reasons,

WEKA is selected as a modeling tool in this research. WEKA has different available classification algorithms among them J48, Randomforest and Randomtree used for experimentation of this study. In this study the above algorithms selected because of, it's easy of understanding and interpretation of the Result of the model. In other case the J48 algorithms of decision tree generate a model by constructing decision tree where each internal node is a feature or attribute.

1.7.6EVALUATION OF THE DISCOVERED KNOWLEDGE

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared finally user interface testing and Evaluation testing performed.

1.7.7 USE OF THE DISCOVERED KNOWLEDGE

The final step comprises planning the regarding the usage of the discovered knowledge. The knowledge discovered in the current domain may be applied in other domains. And also, a plan is created concerning the implementation of the knowledge discovered and the documentation of the whole project. Lastly, the deployment of the model takes place. To develop the proto type the researcher uses Microsoft Visual studio.2010 and to test the user acceptance questioner was developed based on ISO software Acceptance testing standard and asks the experts to give the feedback of the Prototype.

CHAPTER TWO

LITERATURE REVIEW

A review of both conceptual and empirical literature was conducted in order to have conceptual understanding about data mining and the area in which data mining can be applied. Also previous research work on the problem area was also reviewed in order to have a grasp of the potential applicability of data mining in credit risk assessment. Diverse books, journal, articles, Conceptual Papers and the internet pertaining to the subject matter of data mining and knowledge discover in the data base were reviewed. After extensive literature survey the research problem was formulated and specific problem area was identified.

2.1 OVERVIEW OF DATA MINING

Living in the age of digital information, the problem of data overload is an eminent phenomenon. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining [1].

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database.

Conventionally the idea of searching pertinent patterns in data has been referred using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely [10].

Data mining is a process concerned with uncovering patterns, associations, anomalies and statistically significant structures in data [1]. It typically refers to the case where the data is too large or too complex to allow either a manual analysis or analysis by means of simple queries. Data mining consists of two main steps, data pre-processing, during which relevant high-level features or attributes are extracted from the low level data, and pattern recognition, in which a pattern in the data is recognized using these features (see Figure 2.1). Pre-processing the data is often a time-consuming, yet critical, first step. To ensure the success of the data-mining process, it is important that the features extracted from the data are.

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

As Fayyad et al, [1] discussion in their articles organization keep record of data generated even long after their life time has expired because they believe that there is valuable information implicitly coded within it. In earlier days the data that was being collected was relatively easy to analyze manually but as data collection continues to grow in size and complexity there is growing need for more techniques of analysis and one such technique is data mining.

There are some misunderstanding among the researcher and the intellectual community in the field of data mining about the term data mining and knowledge discovery in the data base (KDD). According to Fayyad et al, [1] , KDD refers to the overall process of discovering useful knowledge from data and data mining refers to particular step in the process also it is the application of specific algorithm for extracting pattern from data. As Sidhant et.al, [2] describe in their text book Knowledge discovery is consists of an iterative sequence of the following step:-

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, although an essential one because it uncovers hidden patterns for evaluation.

From the discussion of Fayyad et.al, [1] Data mining involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, high- performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis. Furthermore that the term data mining has become more popular in industries, in media and data base research as synonym for knowledge discovery and hence the two terms are used interchangeably.

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. From the discussion of Fayyad et al, [1] the term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities.

It has also gained popularity in the database field. The phrase knowledge discovery in data base was coined at first KDD work shop in 1989 to emphasize that knowledge is end product of data driven discovery .it has also been popularized in artificial intelligence and machine learning field.

According to Gartner, Inc. [3] discussion in there company articles data mining is process of discovering meaningful pattern and relation by analyzing large data stored in the data base or data ware house and this definition is supported by Han and Kamber, [4] as they said that data mining is an analysis of large seen data base to find new relationship and summarize the data in novel way so that it can be easily understandable and it became useful to the data owner and the summarization derived from data mining is referred us pattern or model From the above definition we can see that data mining typically deals with data that have been collected for other purpose and doesn’t play a role in data collection strategy. For this reason it’s called secondary data analysis and the data set examined in data mining is often large.

From another author Zaki and Wong [5] also noted that, data mining is generally an iterative, interactive discovery process. The goal of the process is to mine pattern, association, changes, anomalies, and statistically significant structures from large amount of data. Furthermore the discover knowledge should be valid, novel, useful, and understandable such that Valid and useful qualities placed on the process while novel and understandable qualities are placed on the outcome of data mining.

2.2 WHAT IS DATA MINING?

It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term *data mining* does not really present all the major components in the picture. To refer to the mining of gold from rocks or sand, we say *gold mining* instead of rock or sand mining.

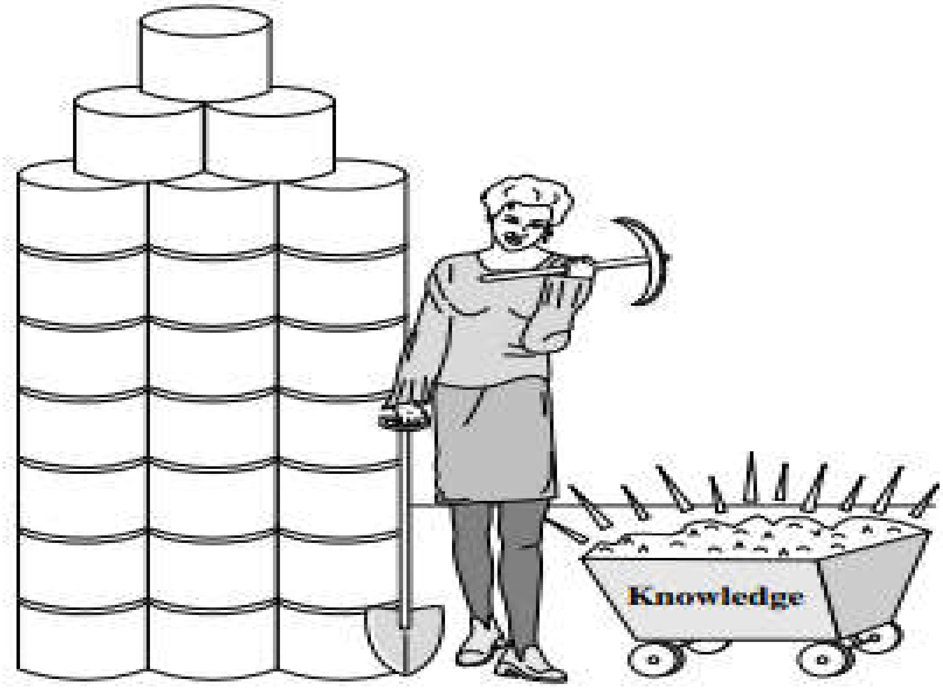


Figure 2.1: Data mining—searching for knowledge (interesting patterns) in data [5].

Analogously, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. However, the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer carrying both "data" and "mining" became a popular choice. In addition, many other terms have a similar meaning to data mining—for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [5].

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is shown in Figure 2.1 as an iterative sequence of the following steps:

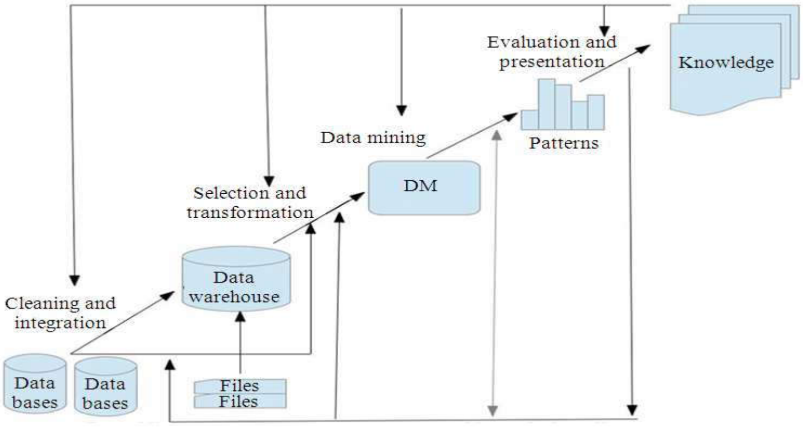


Figure 2.2: Data mining as a step in process of knowledge discovery [5]

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on Interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data). Therefore, we adopt a broad view of data mining functionality: Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

2.3 DATA MINING TASKS

The data mining tasks can be classified generally into two types based on what a specific task tries to achieve. Those two categories are descriptive tasks and predictive tasks. The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave [28].

The tasks of data mining are very diverse and distinct because there are many patterns in a large database. Different kinds of methods and techniques are needed to find different kinds of patterns. Based on the kinds of patterns we are looking for, tasks in data mining can be classified into Predictive and Descriptive modeling [23][24].

According to Kapil Sharma, et al., [7] the goal of any data mining can be one of the two tasks. One could use data mining to generate descriptive models to solve problems. Descriptive data mining tasks characterize the general properties of the data in the database, and focus on finding patterns describing the data that can be interpreted by humans, and produces new nontrivial information based on the available data set. In general the goal of descriptive model is to discover pattern and understand relationship between the attribute. The main activities in descriptive model are Clustering, summarization, association rules and sequence discovery.

One can also use data mining to generate predictive models to solve problems. Predictive data mining tasks perform inference of the current data in order to make prediction. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation. Therefore the main goal of predictive model is to predict the feature based on past record answer. Classification, is the main activity in predictive model to extract useful meaningful information from the data, which is also the concern of the study.

There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining [28].

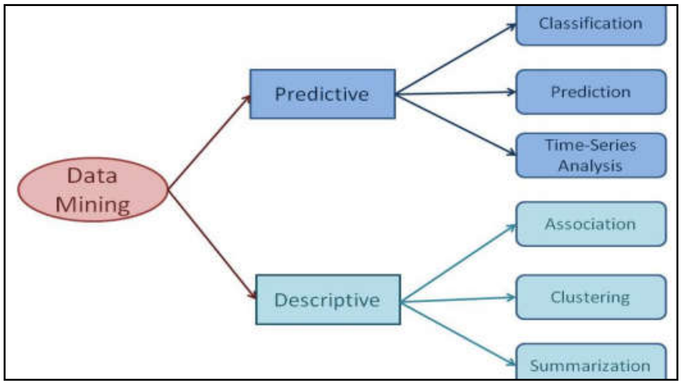


Figure 2.3: Tasks of data mining [28]

2.3.1 CLASSIFICATION ALGORITHMS

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised. It's also a modeling works by collecting data, creating a statistical model and applying probabilistic techniques to guess/predict the likely outcome. In IT, predictive modeling is used to evaluate and identify future trends related to a specific technology domain. For example, software usage statistics can be analyzed to predict future use trends. Moreover, predictive modeling is used on live systems to evaluate and make changes to the underlying system when fulfilling user and business demands [38][28].

Classification: - Classification is the derivation of a function or model which determines the class of an object based on its attributes [28]. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database.

For example, from a set of diagnosed patients, who serve as the training set, a classification model can be built, which concludes a patient's disease from his/her diagnostic data. The classification model can be used to diagnose a new patient's disease based on the patient's diagnostic data, such as age, sex, weight, temperature, blood pressure, etc.

The researcher chooses WEKA software constricting a predictive modeling using J48 Decision Trees, JRip rules induction and Naïve Bayes algorithm. These algorithms were proved to be important when applying them in credit data recommended because of easy to interpret result.

2.4 DATA MINING PROCESS MODELS

The process defines a sequence of steps (with eventual feedback loops) that should be followed to discover knowledge (e.g., patterns) in data. Each step is usually realized with the help of available commercial or open-source software tools. To formalize the knowledge discovery processes (KDPs) within a common framework, there is the concept of a process model [4]**.** The model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the project. This in turn results in cost and time savings, better understanding, and acceptance of the results of such projects. We need to understand that such processes are nontrivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners [4].

There are different DM process model standards that are used in different research and business data mining projects [4].

- ✓KDD process (Knowledge Discovery in Databases),
- ✓SEMMA (Sample Explore Modify Model Assess)
- ✓CRISP-DM (Cross Industry Standard Process for Data Mining), and
- ✓Hybrid model

2.4.1 THE KDD PROCESS

The KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database [1]. The general KDD process is depicted in Figure 2.4. It comprises the following steps [4].

✓ **Selection: -** This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

✓ **Preprocessing: -** This is the steps where the target data cleaning and preprocessing performed in order to obtain consistent, complete and better quality of data.

✓ **Transformation: -** This stage consists of the transformation and discretization of the data using dimensionality reduction and or sampling methods.

✓ **Data Mining:**- This is the knowledge extraction steps for extracting patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction).

✓ **Interpretation/Evaluation:** – in this stage the interpretation and evaluation of the mined patterns will be done using different effectiveness measures, such as accuracy, recall precession.

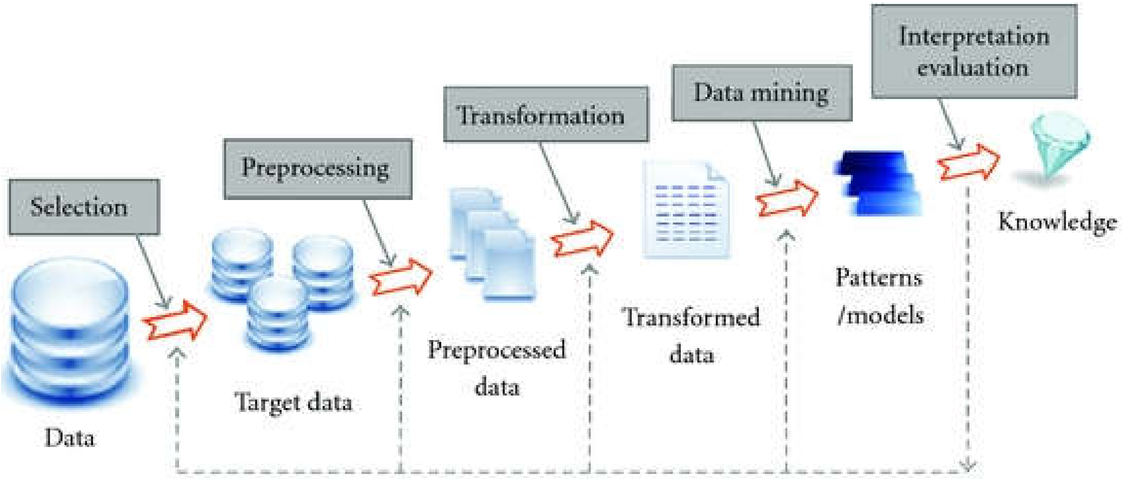


Figure 2.4. The six-steps of KDP model [6].

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user [31]. Additionally, the KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It also must be continued by the knowledge consolidation through incorporating this knowledge into the system [1].

2.4.2 THE SEMMA PROCESS

The SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. The SAS Institute considers a cycle with 5 stages for the process [32].

✓ **Sample:** - This stage consists of sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.

✓ **Explore:**-This stage consists of the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

✓ **Modify: -** This stage consists of the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

✓ **Model:** - This stage consists of modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

✓ **Assess:** - This stage consists of assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

Although the SEMMA process is independent from the chosen DM tool, it is linked to the SAS Enterprise Miner software and pretends to guide the user on the implementations of DM applications.

SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for the conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals.

2.4.3 THE CRISP-DM PROCESS

Analyzing the problems of DM & KD projects, a group of prominent enterprises (Teradata, SPSS - ISL, Daimler-Chrysler and OHRA) proposed a reference guide to develop DM & KD projects. This guide is called CRISP-DM (CRoss Industry Standard Process for Data Mining) [18]. CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem. The CRISP-DM methodology is described in terms of a hierarchical process model, comprising four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks, and process instances.

CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task.

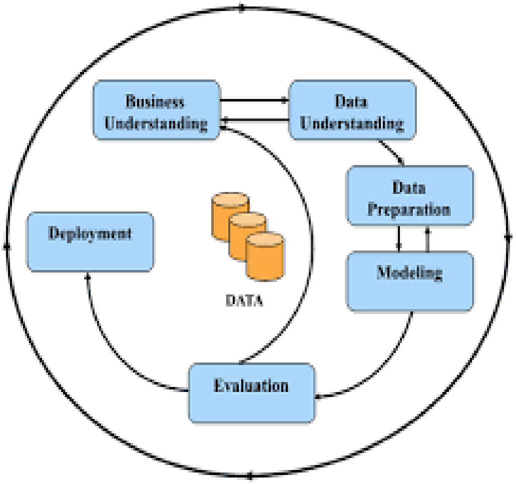


Figure 2.5: Phases of the CRISP-DM [16]

CRISP-DM divides the life cycle of a data mining project in to six phases which are shown in Figure 2.5 [16].The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase, or which particular task of a phase, has to be performed next.

The outer circle in Figure 2.5 symbolizes the cyclic nature of data mining itself. Data mining is not finished once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Each phase of CRISP-DM briefly describes as follows [16]:

- ✓ **Business Understanding**

This initial phase 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 project plan designed to achieve the objectives.

- ✓ **Data Understanding**

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. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

- ✓ **Data Preparation**

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, data cleaning, construction of new attributes, and transformation of data for modeling tools.

✓ **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 require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

✓ **Evaluation**

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

✓ **Deployment**

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

2.4.4HYBRID MODELS

The development of academic and industrial models has led to the development of hybrid models, Models that combine aspects of both KDD and CRISP-DM. One such model is a six-step KDP model and it was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include [21].

✓ Providing more general, research-oriented description of the steps,

✓ Introducing a data mining step instead of the modeling step,

✓ Introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and

✓ Modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

Figure 2.6 presents the six steps Hybrid Data mining Process model, which is designed for academic research

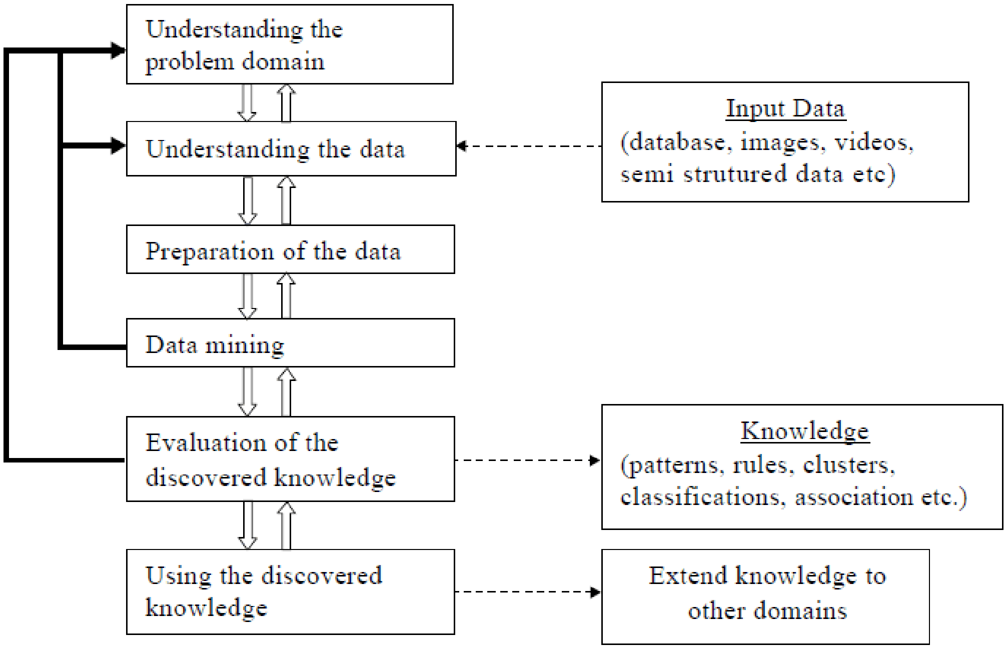

Understanding the problem domain
Understanding the data
Preparation of the data
Data mining
Evaluation of the discovered knowledge
Using the discovered knowledge
Input Data
(database, images, videos, semi strutured data etc)
Knowledge
(patterns, rules, clusters, classifications, association etc.)
Extend knowledge to other domains

Figure 2.6. The six-step Hybrid model. [21].

A description of the six steps follows

- ✓ **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 (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.

- ✓ **Data mining: -** Here the data miner uses various DM methods such as classification, prediction, time-series analysis, association, clustering, and summarization, to derive knowledge from preprocessed data.

✓ **Evaluation of the discovered knowledge:-** Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

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

SUMMARY OF DATA MINING PROCESS MODELS

KDD	**SEMMA**	**CRISP-DM**	**HYBRIDGE**
Problem Understanding	Sample	Business Understanding	Understanding of the problem domain
Selection	Explore		Understanding of the data
Pre processing	Modify	Data Understanding	Preparation of the data
Transformation	Model	Data Preparation	Data mining
Data Mining	Assessment	Modeling	Evaluation of the discovered
Interpretation/Evaluation		Evaluation	Use of the discovered knowledge
Post KDD		Deployment	

Table 2.2 Summary of phases KDD-SEMMA- CRISP-HYBRD

The data mining process is multi stage involving data selection, data cleaning and application of data mining algorithms. According to Shmuel el.al, [6] the following are the major process of data mining.

1. Develop an understanding of the purpose of data mining project or application.
2. Obtain data set to be used in the analysis. This involves random sampling from a large data base to capture records to be used in analysis.
3. Explore, clean and preprocess the data. This involve verifying that the data are in reasonable condition, that is make sure that the values are in reasonable range by giving what you will expect for each variable ,handling missing data, and also make sure that the consistency of the data.
4. Reduce the data if necessary and when supervised training is involved separate them into training, validation and test data set. This involves operation like eliminating unneeded variable; transform variable and creating new variable.
5. Determine the data mining task (classification, prediction, clustering, etc…).

6. Choose data mining technique to be used (regression, neural nets, hierarchical clustering etc...).
7. Use algorithm to perform the task and its iterative process by trying multiple variant of the same algorithm.
8. Interpret the result of algorithm by making a choice of best algorithm to deploy, test the final choice on the test data to know how well it will perform.
9. Deploy the model. This involves integrating the model into operational system and running it on real record to produce decision or action.

2.5 DATA MINING APPLICATION

Kpal [34], discussion data mining can used for predictive descriptive in a variety of applications in order to achieve organizational goal. Some of the common uses of data mining are as follows.

Fraud or noncompliance detection:- data mining isolate the factors that lead to fraud , waste and abuse, where the process of Compliance Monitoring for Anomaly Detection (CMAD) involve a primary monitoring system comparing some predetermined condition of acceptance with actual data or event like credit card fraud detection monitoring , privacy compliance monitoring can be done more effectively.

Intrusion detection:- it is a passive approach to security as it monitor information system and raise alarm when security violation are detected and the system can be either host based or network based according to the kind of input information they analyze.

Last few years research project like MADAM ID, clustering project have applied data mining approach to various problem like construct IDS, cluster audit record of intrusion detection.

Lie detection (SAS text miner):- using SAS tool manager can able to detect automatically when email or web information contains lies. Here the data mining can be applied successfully to identify uncertain in deal or angry customer and also have many other potential applications such as.

Customer segmentation and target marketing: - data mining can be used in grouping or clustering customer based on behavior which in turn helps in customer relationship management (CRM).

Financial banking and credit or risk scoring: - data mining can assist financial institutions in various ways such as credit reporting, credit rating, loan or credit card approval by predicting good customer, risk on sanctioning loan, mode of service delivery and customer retention.

Medical and health care: - applying data mining it’s possible to find relationship between diseases, effectiveness of treatment to identify new drug market activity in drug delivery service and so on.

2.6 RELATED WORKS REVIEW

Because of the sensitivity of the banking sector, there are various related works available.

Koyuncugil and Ozgulbas [8], from Turkey, have explored the possibility of designing data mining techniques for financial institutions using data obtained by means of financial analyses of balance sheets and income statements of companies under Turkish Central Bank. They came up with a model for detecting financial and operational risk indicators of Small and Medium Enterprises (SMEs).The study focused on creating segmentation model using decision tree algorithm for customer profiling where the method of CHAID (Chi-Square Automatic Interaction) is applied as defined in the scope of their study. They reported that the decision tree approach using the CHAID method helped to construct best model with acceptable accuracy that could be used to detect financial and operational risk indicators of SMEs. And their recommendation states that further data mining researches can be conducted utilizing the potentially useful customers' data through application of different classification techniques such as decision tree and neural network for different mining objectives.

Sirikulvadhana [9], from Swedish School of Economics and Business Administration, has conducted research on data mining as financial auditing tool with the objective to determine if data mining tools can directly improve audit performance. The data mining employed clustering analysis to develop cluster model. The model is aimed to assist auditors to select the samples from some representatives of the groups categorized in the way that they have not distinguished before and the obviously different transactions from normal ones or outliers. An accounting transaction archive with vast amount of datasets was used for the data mining process. In his conclusion he states that there are some interesting patterns discovered automatically by data mining technique, clustering, that cannot be found by generalized audit software this proving the vast significant application of the technology in the area namely banking and finance and micro finances as well.

The research has shown an encouraging result of an application of data mining techniques to assist auditors. Further, the researcher recommends that other techniques such as classification mining for risk assessment and customer churning problems could be further research areas worth conducting to come up with models useful to the owner of the data.

There are researches that have been conducted in credit risk assessment using data mining technique for banks in Ethiopia,[22][44][43]. Their result show that the problem in credit risk assessment can be solved by the use of data mining technology.

Askale [22] has conducted a research on the possible application of data mining technology in supporting loan disbursement activity at Dashen bank S.C., Ethiopia. The objective of the research conducted by Askale was to develop a model that supports the loan decision making process that would contribute in alleviating the high default rate in the company. The research focused on developing a classification model for the customer's repayment behaviors (Regular, Loss or Default) that in turn could support the credit risk assessment. The research came up with a predictive model that could help predict whether a potential borrower would default or not, which further guides in assessing the credit decision processes. Hence it was mainly based on collection data (credit report) and loan approval data pertaining to the individual borrower of the bank for extracting predictor features for the dependent class called repayment behaviors. The features mainly considered include security types (building, vehicle, personal guarantee, etc) and collection related attributes such as term of repayment, month/date of loan disbursement, as inputs for the model building.

The researcher recommends that other data mining techniques such as decision tree classifications would also be applied in financial areas to effectively utilize the borrowers' data for supporting loan decisions with respect to different purposes such as customer relationship management in addition to credit risk assessments.

Sara Worku [44], works on applying data mining technology in supporting credit risk assessment in case of Addis credit and Saving Association, Her research was more concerned to find out the most important variable affecting the loan repayment of borrower and to predict the pattern which borrowers are likely to be bad or good borrower at Addis Credit And Saving Institution by developing a classification model using WEKA tool. In her research a total 4000 customer records were collected. To build the model J48, PART and NAIVE BAYES algorithm were used to build the classification model using their different parameter.

During the experiment attribute selection method was applied in each phases and the attributes were selected based on three categories the first one was based on automatic attributes selected by the system using information gain evaluator, the second one was best attribute selected by previous work and the third one was best attribute selected based on the opinion of domain expert of the institution.

She conducts a total of 9 experiments for the research, after conducting a total of nine experiments in each phase the data mining technique that results in better performance is PART rule induction algorism and has shown a highest classification accuracy which is 99.825%.

The finding of her research had generated various rules of risky and risk free contracts which do have an acceptance by the domain expert that help to make accurate decision for credit risk assessment so that the research has an answer that it is possible to extracted useful pattern through data mining algorithm that help to make accurate decision for credit risk assessment.

Her research work has covered the potential applicability of data mining technology to support the loan disbursement activity at Addis Credit and Saving Institution based on historical data accumulated on borrower thus based on the finding of her research work the following recommendation was forward. The research has been attempted to determine the credit risk assessment with limited data set which is 4000 data set and 8 attributes (7 independent and 1 dependent variable) collected from Addis Credit and Saving Institution manually. However further experimentation on the larger data set and other additional attribute that are not used in the research should be made in order to drive more additional interesting rules.

However the application was not developed to use the rules in practice so that recommendation is forwarded to build credit risk assessment application for the practical implementation of the output generated by the system.

From the experiment done in her research and previous work Data mining techniques could contribute a lot in identifying potential customers that could be bad creditors thus it could be more important to use the data mining technique as a tool for the decision making process in other word the institution could optimize its credit assessment effort by employing data mining technology.

Mengestu [43], applying credit risk assessment in case of united bank. And their result show that the problem in credit risk assessment can be solved by the use of data mining technology.

His research has been conducted according to the CRISP_DM Model approach. After many pre-processing effort a data set with 27,310 total credit records was used to develop a classification model.J48Decision Tree and Naive Bayes algorithms were employed to conduct various experiments on the prepared dataset. A Model built by 10-fold cross-validation test mode of unprunedJ48 Decision Tree which registered the highest accuracy (96.6%), was selected as best model for prediction purpose. The finding of his research has generated various rules of risky and risk free contracts which do have an acceptance by the domain experts. The researcher suggests the use of his model to assist in the non-structured decisions where only the credit committee or managers' intuition and experience are used in the granting process of loans.

And also he recommended Even though results from his study were encouraging, further classification techniques like neural network should be undertaken by including data before the implementation of the core banking system (2006) to have the full picture of the bank's credit history. There is a need to develop a credit risk assessment prototype or knowledge base system for the practical implementation of his academic research endeavor.

2.7 RESEARCH GAP

Though there are different works for local banks that apply data mining, analyzing data and extracting patterns is context dependent, hence this study aims to identify attribute that describe the context of Abyssinia bank and extract pattern to improve credit risk scoring there by designing a prototype using credit risk prediction model.

CHAPTER THREE

PROBLEM UNDERSTANDING AND DATA PREPARATION

3.1 UNDERSTANDING OF THE PROBLEM

Credit facilities and investments are the cornerstones of the growing economy of Ethiopia. Bank of Abyssinia being one of the former private banks has played its own role in the economy by rendering loan facilities to the individuals and companies which are running business in various sectors. The bank uses internal and NBE credit policies, procedures and strictly followed manuals in various levels of credit committees before disbursing loan to customers. However, there are total defaulters and inconsistent loan repaying customers which declines the profitability of the bank in particular and threatens the growing economy of the country in general. While fueling the sprinting economy in the country, minimizing the possible defaulters is the prime concern of the bank.

Data mining is an innovative way of gaining new and valuable business insights by analyzing the information held in a company database. Thus the overall goal of this data mining process is to extract information from Abyssinia Bank's Credit data set and transform it into new, valuable and an understandable structure by identifying risky credit scenarios and facilitate the making of well-informed business decisions.

This data mining will be applied on the bank's credit related information extracted from T24 Core banking system and is expected to identify rules that will assist to classify contracts as good and bad loans. By uncovering patterns of credit failure and success through data mining the bank will benefit in many ways:

- ✓ Will enable the loan officers to scrutinize control on the existing loans which do have a bad loan pattern.
- ✓ Will be able to avoid/screen expected bad loan applicants and target a marketing campaign for possibly good loan applicants.
- ✓ Can lead to an improvement in the quality and dependability of strategic business decision making with regard to banks credit activity.

3.1.1 OVERVIEW OF CREDIT AND CREDIT RISK ASSESMENT

According to Basel [35], Credit risk refers to the risk that a bank borrower or counterparty will default (fail to meet its obligations) on any type of debt by failing to make payments which it is obligated to doing accordance with agreed terms. The risk is primarily that of the lender and includes lost principal and interest, disruption to cash flows, and increased collection costs. The loss may be complete or partial and can arise in a number of circumstances.

To reduce the lender's credit risk, the lender may perform a credit check on the prospective borrower, may require the borrower to take out appropriate insurance, such as mortgage insurance or seek security or guarantees of third parties, besides other possible strategies. In general, the higher the risk, the higher the interest rate that the debtor will be asked to pay on the debt.

Credit risk assessment/analysis is a largely standardized process that attempts to evaluate the desirability of a particular account based on its estimated reliability and profitability as part and parcel of banks money lending activity. Banks and other lenders conduct credit investigations in order to minimize the probability that they will experience losses from late and delinquent payments [35].

The world of credit risk assessment is a wide one. This is why so many credit assessment institutions can be found all over the world and why each player focuses on a specific population. This specification can be caused by location (National Central Banks for example), portfolio limits (commercial banks), whether a company is publicly listed or not, has recourse to bond and other traded securities markets or not (international agencies) etc. Due to these different scopes, each player uses the information he/she has access to in order to design the most appropriate rating scale. As a result a variety of definitions of default are in use nowadays (European Committee of Central Balance Sheet Data Offices (ECCBSO).

The goal of credit risk Assessment as discussed by Basel [35], is to maximize a bank’s risk-adjusted rate of return by maintaining credit risk exposure within acceptable parameters. Banks need to manage the credit risk inherent in the entire portfolio as well as the risk in individual credits or transactions. Banks should also consider the relationships between credit risk and other risks. The effective management of credit risk is a critical component of a comprehensive approach to risk management and essential to the long-term success of any banking organization.

3.1.2 TYPES OF CREDIT RISK

Credit risk is the probable risk of loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Traditionally, it refers to the risk that a lender may not receive the owed principal and interest, which results in an interruption of cash flows and increased costs for collection. There are many types of risks that banks face. But as noted by Saul [51], the following are the major ones in risk.

1. Market risk

Market risk is the possibility of an investor experiencing losses due to factors that affect the overall performance of the financial markets in which he or she is involved. Market risk, also called "systematic risk," cannot be eliminated through diversification, though it can be hedged against.

2. Operational risk

Operational risk is the prospect of loss resulting from inadequate or failed procedures, systems or policies. Employee errors. Systems failures. Fraud or other criminal activity. Any event that disrupts business processes.

3. Liquidity risk

Liquidity risk is the risk that a company or bank may be unable to meet short term financial demands. This usually occurs due to the inability to convert a security or hard asset to cash without a loss of capital and/or income in the process.

4. Business risk

Business risk is the possibility a company will have lower than anticipated profits or experience a loss rather than taking a profit. Business risk is influenced by numerous factors, including sales volume, per-unit price, input costs, competition, and the overall economic climate and government regulations.

5. Reputational risk

Reputational risk, often called reputation risk, is a risk of loss resulting from damages to a firm's reputation, in lost revenue; increased operating, capital or regulatory costs; or destruction of shareholder value, consequent to an adverse or potentially criminal event even if the company is not found guilty.

6. Systemic risk

Systemic risk is the possibility that an event at the company level could trigger severe instability or collapse an entire industry or economy. Systemic risk was a major contributor to the financial crisis of 2008. Companies considered to be a systemic risk are called "too big to fail."

7. Moral hazard

In economics, moral hazard occurs when someone increases their exposure to risk when insured, especially when a person takes more risks because someone else bears the cost of those risks.

According to Bank of Abyssinia [29], Credit risk can be classified in to three; Credit default risk, concentration risk and country risk.

Credit default risk:- The risk of loss arising from a debtor being unlikely to pay its loan obligations in full or the debtor is more than 90 days past due on any material credit obligation; default risk may impact all credit-sensitive transactions, including loans, securities and derivatives.

Concentration risk: - The risk associated with any single exposure or group of exposures with the potential to produce large enough losses to threaten a bank's core operations. It may arise in the form of single name concentration or industry concentration.

Country risk: - The risk of loss arising from a sovereign state freezing foreign currency payment (transfer/conversion risk) or when it defaults on its obligations (sovereign risk).

3.1.3 FACTORS CONSIDERD IN CREDIT RISK ASSESMENT

The factors that affect credit risk range from borrower-specific criteria, such as debit rations, to market-wide considerations such as economic growth [51]. The idea is that liabilities can be objectively valued and predicted to help protect against financial loss.

Factors that should be considered in credit risk assessment are related to customer of the Bank and loan given by the bank.

CUSTOMERS:-

The term Customer has not been defined by any act. It is generally believed that any individual or an organization, which conducts banking transactions with a bank, is the customer of bank. However, there are many persons who do utilize services of banks, but do not maintain any account with the bank [39].

A customer is someone who has an account with a bank or who is in such a relationship with the bank that the relationship of a banker and customer exists. The legal position implies that opening an account is the crucial element in establishing the banker-customer relationship.

Attributes that describes a Customer record in bank credit are , Customer type, ,Customer Address , Collateral, Granted to Borrow, Ready to loan, Project Plan, Business Sector, Experience in Business sector, Asset, , Trade license ,and Tin number. Table 3.1 provides customer attributes and description.

No	FILDS	DESCRIPTION
1.	Sector type	Describes the customer type whether private limited or share company
2.	Location District	Describes the application address of the customer /district
3.	Collateral Value	Describes the Customer shows collateral
4.	Income of the customer	Describes monthly income of the customer
5.	Loan type	Describes the loan customer/for which purpose/
6.	Business Proposal ,	Describes the project plan that the customer interested to implement
7.	Payment frequency	Describes the payment type of the customer/loaner
8.	Business sector	Describes the trade sector that the customer engaged.
9.	Experience in Business sector	Describes how many years past since the organization is established and experiences in the business
10.	Asset	Describes the asset that the borrowers have
11.	Renewal Trade license	Describes the Trade license of the Customer

Table 3.1: list of customer attributes with description

LOAN:-

A loan is the lending of money by one or more individuals, organizations, or other entities to other individuals, organizations. The recipient (i.e. the borrower) incurs a debt, and is usually liable to pay interest on that debt until it is repaid, and also to repay the principal amount borrowed.

In finance, a loan is the lending of money by one or more individuals, organizations, or other entities to other individuals, organizations. The recipient (i.e. the borrower) incurs a debt, and is usually liable to pay interest on that debt until it is repaid, and also to repay the principal amount borrowed [52].

The document evidencing the debt, e.g. a promissory note, will normally specify, among other things, the principal amount of money borrowed, the interest rate the lender is charging, and date of repayment. A loan entails the reallocation of the subject asset(s) for a period of time, between the lender and the borrower.

The interest provides an incentive for the lender to engage in the loan. In a legal loan, each of these obligations and restrictions is enforced by contract, which can also place the borrower under additional restrictions known as loan covenants. Although this article focuses on monetary loans, in practice any material object might be lent.

Attribute that describes loan are Application amount, Deposit amount, other bank loan and Expire date of the loan. Table 3.2 presents loan attributes and there description.

No	FILDS	DESCRIPTION
1.	Application amount	Describes the requested amount of the loan.
2.	Deposit amount	Describes the deposit amount of the customer.
3.	Loan Period	Describes the length of the loan.
4.	Other bank loan	Describes the customer have other bank loan.
5.	Expire date	Describes the end of the loan date.

Table 3.2: list of Loan attributes with their description

3.1.4 SUMMARY OF IDENTIFIED VARIABLES

for this study attributes listed in table 3.1 and table 3.2 are used for credit risk assesment.In the next section using this attributes the extracted data passes through data understanding and data preparation.

3.2 DATA UNDERSTANDING

Having an insight about the need of data mining the next basic thing for the process is getting the credit data and creating an understanding for it. Data understanding step of hybrid-DM has different components of learning before the actual application of data mining techniques.

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.

3.2.1 INITIAL DATA COLLECTION

The data for this research has been compiled from Bank of Abyssinia Core banking system. The core banking system contains numerous tables regarding credit activities and many other modules of the bank. Among these tables the following are considered to be relevant for this study:

✓ Customer and Loan contract record table,

3.2.2 DESCRIPTION OF DATA

Only selected fields and records as shown below in Table 3.3, are to be used for the task of this thesis. Having an explanation of each table and relevant fields will give a glimpse of the dataset to be constructed from these tables.

1. Customer Base table: This table contains basic customer's information without attaching to the account type they are opening. This table contains a unique ID called CIF (Customer Identification File) which is used as a basis to open any type of account for the customer. the following are considered important for this thesis:

2. Loan contract record master table:-: This table contains most of the loan contract information based on the paper signed as a binding agreement between the bank and the customer while granting the credit facility. The following fields will be used for this thesis**.**

FIELD NAME	DATA TYPE	DESCRIPTION
Application amount	numeric	Amount granted in local currency
Payment_freq	String	Payment Methods
Depost amout	Numeric	the amount of the deposit that the company have
Income of the customer	numeric	income of the customer monthly base
Location_District	String	The location that Thelon requested
Requsted_date	date	loan requested date
Collateral_ value	numeric	collateral amount in cash value
Renwal_trade_lic	String	renewed trade license

Experiance_in_business	String	year of business experiences
Other_bank_loan	String	either the applicant have other bank loan
Business proposal	String	have business proposal/project plan
Loan type	String	the purpose of the requested loan
Expiry_date	date	the end of requested loan
Tin_no	String	Have tax identification no
Sector	String	business sector
Loan Period	numeric	The duration of the loan in year(derivative)
Risk Type	nominal	The type of Risks that faces (Class)

Table 3.3: Loan contract record master table and the selected attributes

3.2.3 DATA QUALITY VERIFICATION

From the above description it can be seen that extracting data needs joining different tables and dropping redundant fields in order to come-up with a reliable set of columns and content for the data-mining task. Moreover certain fields have null values and others need categorization to meaningful elements.

Missing values

The data collected on the institution borrowers has few records which has no value for the attributes such as liability capital and asset. The cause for the missing information was that some of the borrowers are not willing to write their proper asset, capital and their liability. When these types of data occur, value was given by doing some calculation; that is, when either of two values was written in the document the rest missing attribute value can be obtained using calculation that means for instance when the TOTAL ASSET is subtracted from TOTAL LIABILITY we can find the CAPITAL. However one of the challenges the researcher faced during the collection of the data is that to find fully filled business plan form which doesn't have a missing value. But when missing values occur in the form for the other attribute some domain experts helps to enter the values if they remember the customer. Missing value of attributes such as TRADE SECTOR 48 (0.27%), YEARS IN BUSINESS 12(0.03) were handled by this technique.

3.3 DATA PREPROCESSING

This is one of the crucial steps to construct dataset used for modeling by Waikato Environment for Knowledge Analysis (hence forth WEKA) software. At this stage, all necessary tasks needed to perform Data mining will be finalized. Data mining techniques, tools and algorithms were decided. The data sets are pre-processed for specific Data mining tasks. Pre- processing includes selecting, cleaning, deriving, integrating, and formatting data in order to apply specific DM tasks. There are a number of possible DM techniques such as classification, clustering, association rule, regression, and others. This phase is concerned on deciding to make data ready which is used as input for data mining process. To this end, data cleaning (such as filling missing values, detecting outliers) and data transformation are performed using statistical techniques with the help of SPSS tool.

As per Han and Kamber [4] preprocessing assists to fill some missing values; to detect some outliers that may jeopardize the result of data mining; and to detect and remove/correct some noisy data. In relation to this, data normalization, discretization and related activities need to be performed. Moreover, to conduct the experimentation, the dataset must be prepared in the appropriate format.

3.3.1 Data Selection

On this phase, the relevant data for the data mining process is selected. The cornerstone of selection of data was based on relevance, availability and quality of variables.

The most challenging part of this step was getting the relevant information as needed. Most of the core banking reports is designed for daily routine and periodic administrative decisions per branch and specific products. Getting credit information bank wide for the whole period and fiscal year needs a tailored approach and exposure to the core banking system.

The core banking system, stacked with various tables and views is not an easy prey to get the desired information without losing focus. There were more than 3,500 tables and 600 views in the Core banking system. Identifying the important ones and further reducing the number of tables to minimize the number of joins has taken extended time. Finally the tables and views mentioned in the data description part were selected with the relevant fields.

While selecting the data the following criteria were used:

✓ Customer and Loan contract record table were used in one query to generate the data

✓ The final modified versions of each loan contract were used

✓ Only contracts of Loan Module were used.

✓ NPL (Non-Performing Loan) GL entries were counted per contract

3.3.2 Data Cleaning

The data cleaning task was partially managed in the data extraction step. While making an inner join and setting certain filtering criteria, the invalid or incomplete data was suppressed before extraction.

The data cleaning task was partially managed in the data extraction step. While making an inner join and setting certain filtering criteria, the invalid or incomplete data was suppressed before extraction. The researcher has used extracted data from oracle database of Bank of Abyssinia core banking system. The following data mining activities were performed:

✓ 1,418 records of reversed loans were avoided as they create duplication. These Loans are recaptured to the system with modifications after reversal. Since the reversed contract doesn't stay to the full life time of the contract (as it is replaced with the new one), this competes 7.05% of the total data removing it has a positive impact on the final dataset analysis.

3.3.3 DATA TRANSFORMATION AND AGGREGATION

Data transformation and aggregation assists in reducing the variations of field values and changes to a meaningful and understandable form.

In this research all the data was collected from one data base and there is no need of data aggregation.

- ✓ The Loan Period attribute is a derivative attribute that shows the duration of the loan in years.

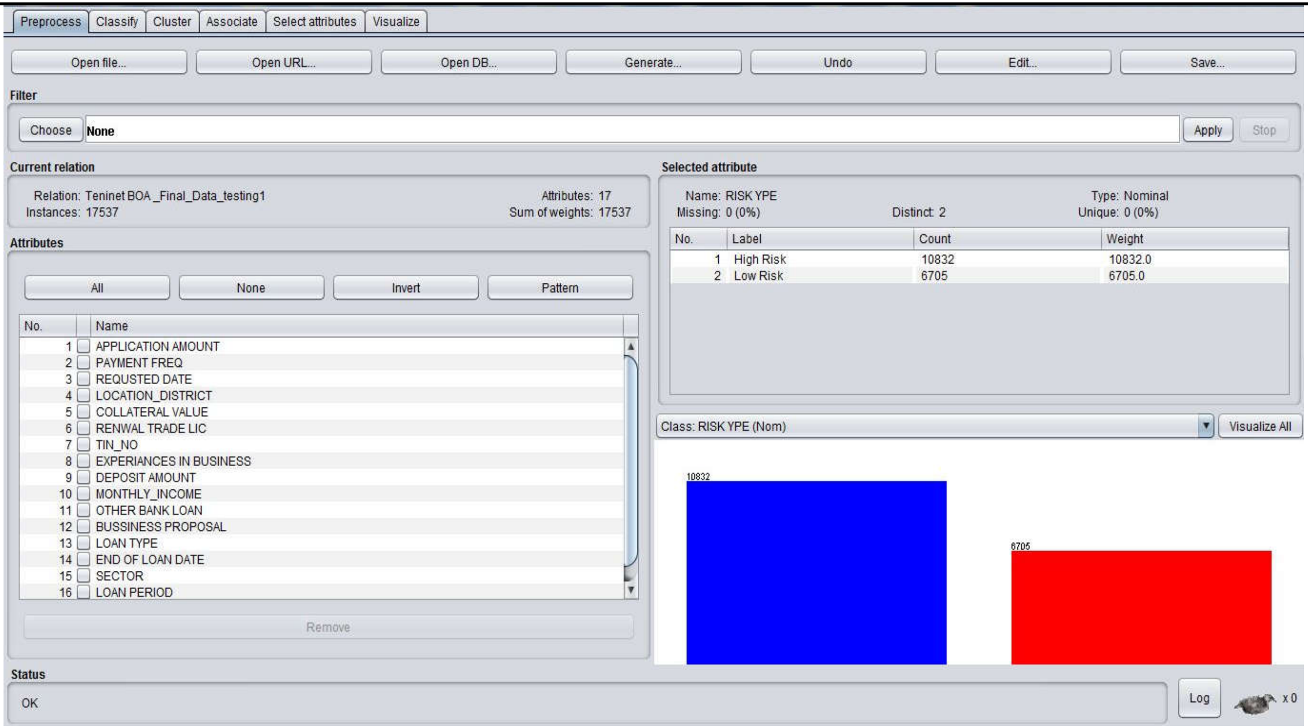
  Calculating the loan period by subtracting Requsted_Date from Expiry_Date.

- ✓ The Risk type attribute is a class attribute that shows whether the customer is predicted as a high risk or a low risk.

3.3.4 BALANCING DATA

According to Chawla, [69], a dataset is imbalanced if the classification categories are not approximately equally represented. Performance of DM algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large.

Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. However, Chawla[52], showed a method of over-sampling the minority (abnormal)class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) receiving operation characteristics, than only under sampling the majority class. The Credit Data of BOA has a higher imbalance for the class variable (Credit Risk). Therefore, the researcher used re-sampling of Weka (weka.filters.supervised.instance.Resample) to over sample the minority classes (High Risk instances) and under sample the majority (Low Risk instances) of loans. As a result, the class distribution in the dataset changes and probability of correctly classifying the instances of the class increases.



Filter

Choose None

Current relation

Relation: Teninet BOA _Final_Data_testing1 Attributes: 17
Instances: 17537 Sum of weights: 17537

Attributes

No.	Name
1	APPLICATION AMOUNT
2	PAYMENT FREQ
3	REQUSTED DATE
4	LOCATION_DISTRICT
5	COLLATERAL VALUE
6	RENWAL TRADE LIC
7	TIN_NO
8	EXPERIANCES IN BUSINESS
9	DEPOSIT AMOUNT
10	MONTHLY_INCOME
11	OTHER BANK LOAN
12	BUSSINESS PROPOSAL
13	LOAN TYPE
14	END OF LOAN DATE
15	SECTOR
16	LOAN PERIOD

Selected attribute

Name: RISK YPE Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

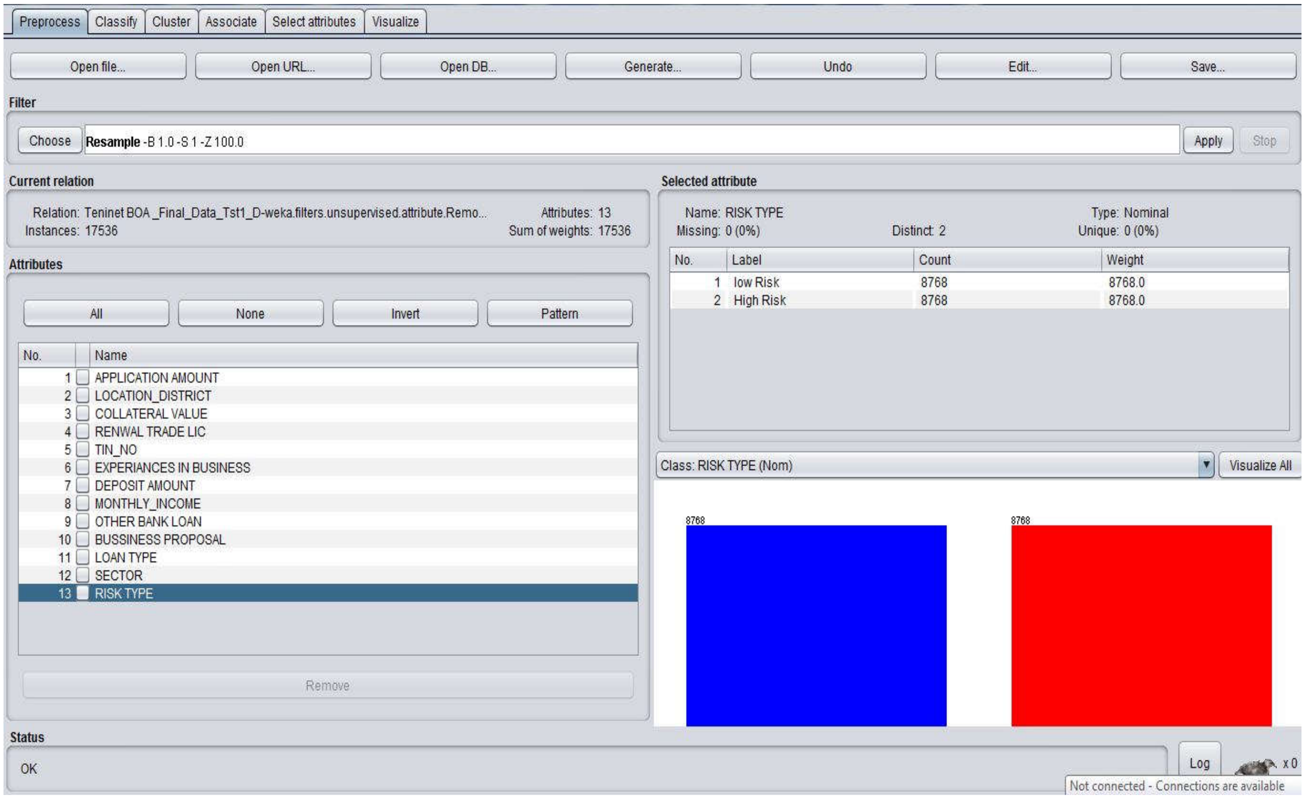
No.	Label	Count	Weight
1	High Risk	10832	10832.0
2	Low Risk	6705	6705.0

Class: RISK YPE (Nom)

Status

OK

Fig **A**



Filter

Choose Resample -B 1.0 -S 1 -Z 100.0

Current relation

Relation: Teninet BOA _Final_Data_Tst1_D-weka.filters.unsupervised.attribute.Remo... Attributes: 13
Instances: 17536 Sum of weights: 17536

Attributes

No.	Name
1	APPLICATION AMOUNT
2	LOCATION_DISTRICT
3	COLLATERAL VALUE
4	RENWAL TRADE LIC
5	TIN_NO
6	EXPERIANCES IN BUSINESS
7	DEPOSIT AMOUNT
8	MONTHLY_INCOME
9	OTHER BANK LOAN
10	BUSSINESS PROPOSAL
11	LOAN TYPE
12	SECTOR
13	RISK TYPE

Selected attribute

Name: RISK TYPE Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	low Risk	8768	8768.0
2	High Risk	8768	8768.0

Class: RISK TYPE (Nom)

Status

OK

Not connected - Connections are available

Fig B

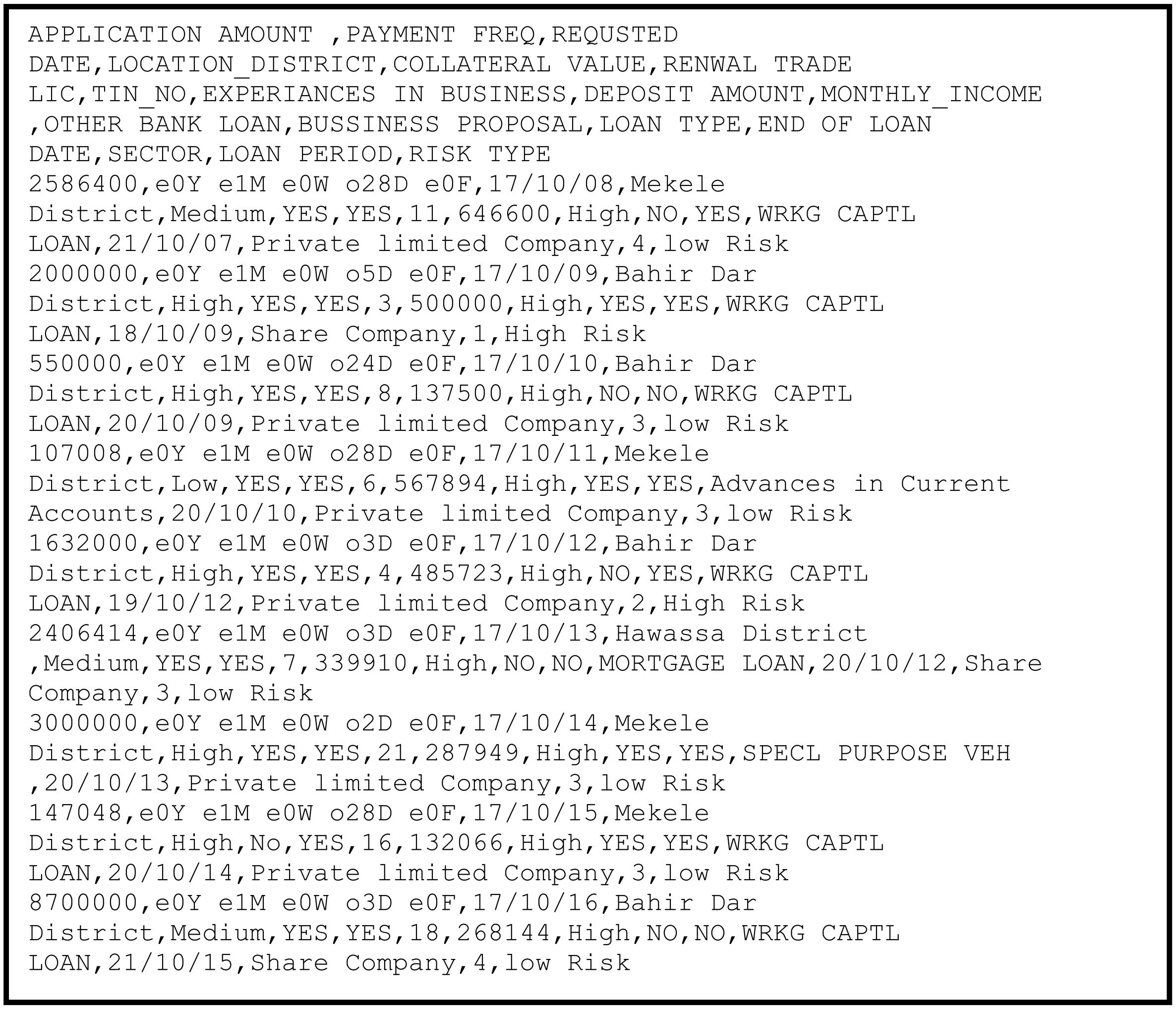
Figure 3.1 Side by side review of the class attribute creditors data (Fig.A) Original data; (Fig.B) balanced data.

A dataset is imbalanced if the classification categories are not approximately equally represented the creditors. Performance of imbalance data is deviates to doming value and bias the predictive accuracy. In the case of prescreening data the class variable has a higher imbalance with ratio 1:3 Therefore; the investigator decide to balance reduce the class attribute 1:1 ratio with minority classes to prevent false positive result

Figure 3.1 shows a side by side review of the class attribute is status of Creditors the majority class. Originally there were 10862 records in the majority class and only 6705 records the minority class but after applying balance by reduction the class is equal with the minority class.

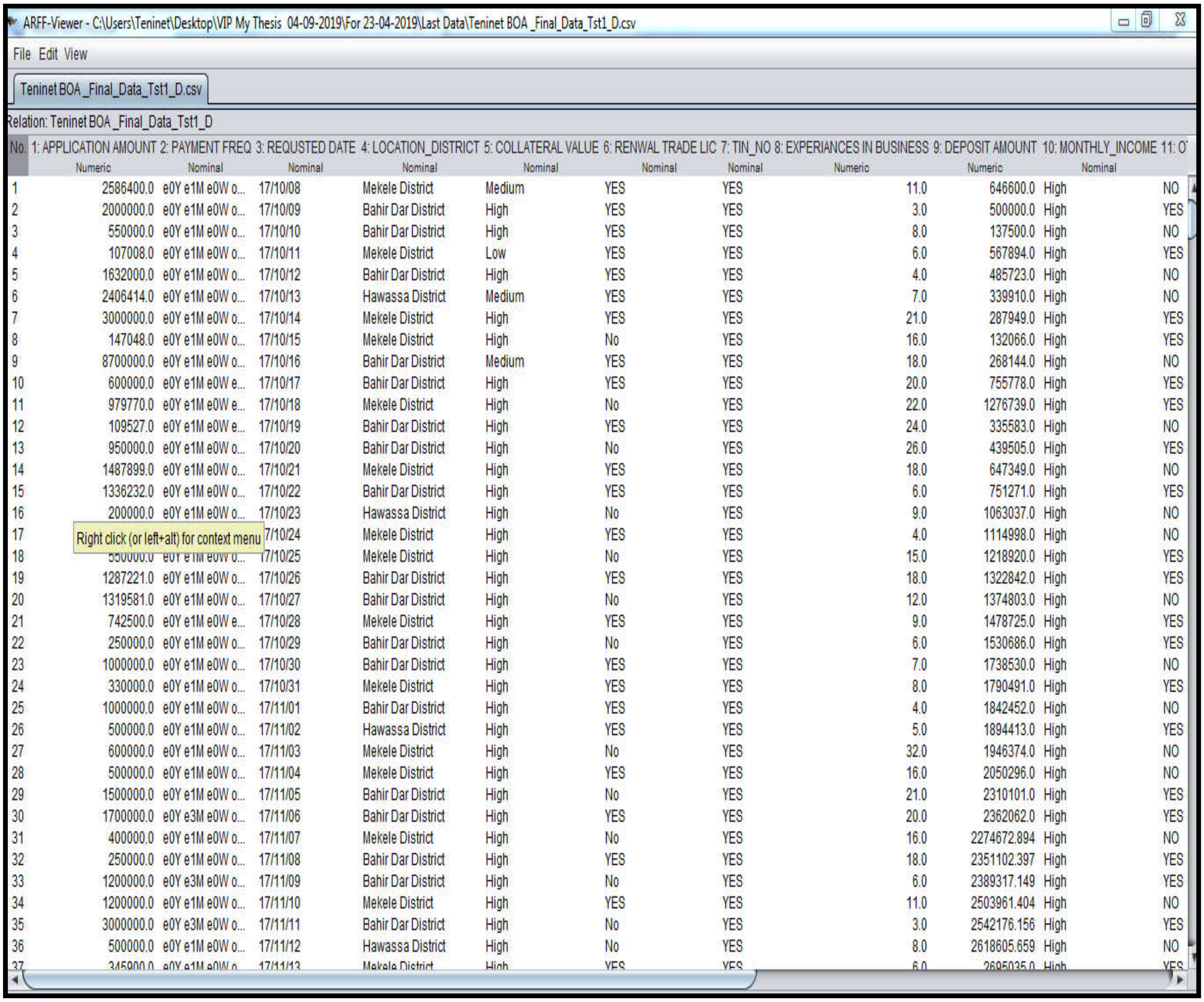
3.3.5 Final Dataset Preparation

This task was done after data cleaning. After passing the above preprocessing stage the researcher has got 14 attributes and 1 class attributes and 1 attribute was derived (a total of 16 attributes), 17,537 records. Then the preprocessed dataset in excel is converted to Comma Separated Values (.csv) and Attribute Relation File Format (.raff) to make it compatible with WEKA software.



```
APPLICATION AMOUNT ,PAYMENT FREQ,REQUSTED
DATE,LOCATION_DISTRICT,COLLATERAL VALUE,RENWAL TRADE
LIC,TIN_NO,EXPERIANCES IN BUSINESS,DEPOSIT AMOUNT,MONTHLY_INCOME
,OTHER BANK LOAN,BUSSINESS PROPOSAL,LOAN TYPE,END OF LOAN
DATE,SECTOR,LOAN PERIOD,RISK TYPE
2586400,e0Y e1M e0W o28D e0F,17/10/08,Mekele
District,Medium,YES,YES,11,646600,High,NO,YES,WRKG CAPTL
LOAN,21/10/07,Private limited Company,4,low Risk
2000000,e0Y e1M e0W o5D e0F,17/10/09,Bahir Dar
District,High,YES,YES,3,500000,High,YES,YES,WRKG CAPTL
LOAN,18/10/09,Share Company,1,High Risk
550000,e0Y e1M e0W o24D e0F,17/10/10,Bahir Dar
District,High,YES,YES,8,137500,High,NO,NO,WRKG CAPTL
LOAN,20/10/09,Private limited Company,3,low Risk
107008,e0Y e1M e0W o28D e0F,17/10/11,Mekele
District,Low,YES,YES,6,567894,High,YES,YES,Advances in Current
Accounts,20/10/10,Private limited Company,3,low Risk
1632000,e0Y e1M e0W o3D e0F,17/10/12,Bahir Dar
District,High,YES,YES,4,485723,High,NO,YES,WRKG CAPTL
LOAN,19/10/12,Private limited Company,2,High Risk
2406414,e0Y e1M e0W o3D e0F,17/10/13,Hawassa District
,Medium,YES,YES,7,339910,High,NO,NO,MORTGAGE LOAN,20/10/12,Share
Company,3,low Risk
3000000,e0Y e1M e0W o2D e0F,17/10/14,Mekele
District,High,YES,YES,21,287949,High,YES,YES,SPECL PURPOSE VEH
,20/10/13,Private limited Company,3,low Risk
147048,e0Y e1M e0W o28D e0F,17/10/15,Mekele
District,High,No,YES,16,132066,High,YES,YES,WRKG CAPTL
LOAN,20/10/14,Private limited Company,3,low Risk
8700000,e0Y e1M e0W o3D e0F,17/10/16,Bahir Dar
District,Medium,YES,YES,18,268144,High,NO,NO,WRKG CAPTL
LOAN,21/10/15,Share Company,4,low Risk
```

Figure 3.2 Sample Comma Separated Values (.csv)



ARFF-Viewer - C:\Users\Teninet\Desktop\VIP My Thesis 04-09-2019\For 23-04-2019\Last Data\Teninet BOA _Final_Data_Tst1_D.csv

Relation: Teninet BOA _Final_Data_Tst1_D

No.	1: APPLICATION AMOUNT	2: PAYMENT FREQ	3: REQUSTED DATE	4: LOCATION_DISTRICT	5: COLLATERAL VALUE	6: RENWAL TRADE LIC	7: TIN_NO	8: EXPERIANCES IN BUSINESS	9: DEPOSIT AMOUNT	10: MONTHLY_INCOME	11: O
	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Numeric	Nominal	
1	2586400.0	e0Y e1M e0W o...	17/10/08	Mekele District	Medium	YES	YES	11.0	646600.0	High	NO
2	2000000.0	e0Y e1M e0W o...	17/10/09	Bahir Dar District	High	YES	YES	3.0	500000.0	High	YES
3	550000.0	e0Y e1M e0W o...	17/10/10	Bahir Dar District	High	YES	YES	8.0	137500.0	High	NO
4	107008.0	e0Y e1M e0W o...	17/10/11	Mekele District	Low	YES	YES	6.0	567894.0	High	YES
5	1632000.0	e0Y e1M e0W o...	17/10/12	Bahir Dar District	High	YES	YES	4.0	485723.0	High	NO
6	2406414.0	e0Y e1M e0W o...	17/10/13	Hawassa District	Medium	YES	YES	7.0	339910.0	High	NO
7	3000000.0	e0Y e1M e0W o...	17/10/14	Mekele District	High	YES	YES	21.0	287949.0	High	YES
8	147048.0	e0Y e1M e0W o...	17/10/15	Mekele District	High	No	YES	16.0	132066.0	High	YES
9	8700000.0	e0Y e1M e0W o...	17/10/16	Bahir Dar District	Medium	YES	YES	18.0	268144.0	High	NO
10	600000.0	e0Y e1M e0W e...	17/10/17	Bahir Dar District	High	YES	YES	20.0	755778.0	High	YES
11	979770.0	e0Y e1M e0W e...	17/10/18	Mekele District	High	No	YES	22.0	1276739.0	High	YES
12	109527.0	e0Y e1M e0W e...	17/10/19	Bahir Dar District	High	YES	YES	24.0	335583.0	High	NO
13	950000.0	e0Y e1M e0W o...	17/10/20	Bahir Dar District	High	No	YES	26.0	439505.0	High	YES
14	1487899.0	e0Y e1M e0W o...	17/10/21	Mekele District	High	YES	YES	18.0	647349.0	High	NO
15	1336232.0	e0Y e1M e0W o...	17/10/22	Bahir Dar District	High	YES	YES	6.0	751271.0	High	YES
16	200000.0	e0Y e1M e0W o...	17/10/23	Hawassa District	High	No	YES	9.0	1063037.0	High	NO
17			7/10/24	Mekele District	High	YES	YES	4.0	1114998.0	High	NO
18	550000.0	e0Y e1M e0W o...	17/10/25	Mekele District	High	No	YES	15.0	1218920.0	High	YES
19	1287221.0	e0Y e1M e0W o...	17/10/26	Bahir Dar District	High	YES	YES	18.0	1322842.0	High	YES
20	1319581.0	e0Y e1M e0W o...	17/10/27	Bahir Dar District	High	No	YES	12.0	1374803.0	High	NO
21	742500.0	e0Y e1M e0W e...	17/10/28	Mekele District	High	YES	YES	9.0	1478725.0	High	YES
22	250000.0	e0Y e1M e0W o...	17/10/29	Bahir Dar District	High	No	YES	6.0	1530686.0	High	YES
23	1000000.0	e0Y e1M e0W o...	17/10/30	Bahir Dar District	High	YES	YES	7.0	1738530.0	High	NO
24	330000.0	e0Y e1M e0W o...	17/10/31	Mekele District	High	YES	YES	8.0	1790491.0	High	YES
25	1000000.0	e0Y e1M e0W o...	17/11/01	Bahir Dar District	High	YES	YES	4.0	1842452.0	High	NO
26	500000.0	e0Y e1M e0W o...	17/11/02	Hawassa District	High	YES	YES	5.0	1894413.0	High	YES
27	600000.0	e0Y e1M e0W o...	17/11/03	Mekele District	High	No	YES	32.0	1946374.0	High	NO
28	500000.0	e0Y e1M e0W o...	17/11/04	Mekele District	High	YES	YES	16.0	2050296.0	High	NO
29	1500000.0	e0Y e1M e0W o...	17/11/05	Bahir Dar District	High	No	YES	21.0	2310101.0	High	YES
30	1700000.0	e0Y e3M e0W o...	17/11/06	Bahir Dar District	High	YES	YES	20.0	2362062.0	High	YES
31	400000.0	e0Y e1M e0W o...	17/11/07	Mekele District	High	No	YES	16.0	2274672.894	High	NO
32	250000.0	e0Y e1M e0W o...	17/11/08	Bahir Dar District	High	YES	YES	18.0	2351102.397	High	YES
33	1200000.0	e0Y e3M e0W o...	17/11/09	Bahir Dar District	High	No	YES	6.0	2389317.149	High	YES
34	1200000.0	e0Y e1M e0W o...	17/11/10	Mekele District	High	YES	YES	11.0	2503961.404	High	NO
35	3000000.0	e0Y e3M e0W o...	17/11/11	Bahir Dar District	High	No	YES	3.0	2542176.156	High	YES
36	500000.0	e0Y e1M e0W o...	17/11/12	Hawassa District	High	No	YES	8.0	2618605.659	High	NO
37	345900.0	e0Y e1M e0W o...	17/11/13	Mekele District	High	YES	YES	6.0	2695035.0	High	YES

Table 3.4 Sample Attribute-Relation File Format (.arff)

The Risk Type field which was chosen for the class attribute has an imbalanced count of records. The count of Low Risk's is 838 and weighted (4.78%) while the highly risk's ones are 16699 weighted (95.22 %).

3.4 ARCHITECTURE

The major components of any data mining system architecture are the guideline of the data modeling for prediction data source, data mining engine, pattern evaluation module, and graphical user interface. The data mining process is identifying the most effective model of credit Risk of the bank. It is divided into six steps. The processing blocks are shown in Figure on Data Mining Architecture.

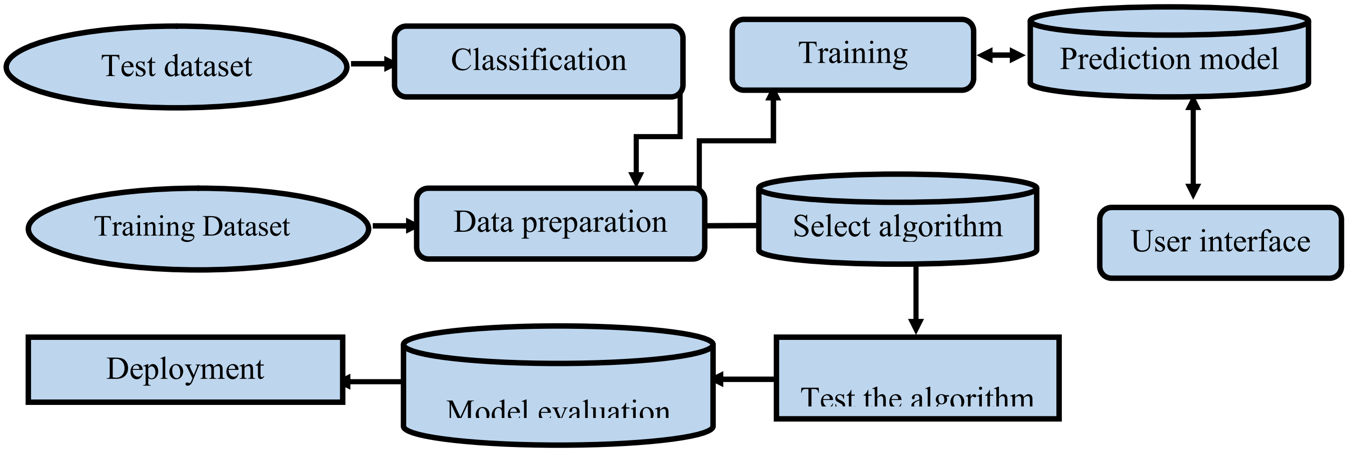


Figure 3.3 Credit risk Prediction Model architecture

Data collection: The first stage of the mining process is data collection from Bank of Abyssinia Portfolio and Credit department. Data collection period from March to April 2019.

Data preparation: The data preparation stage is crucial for data analysis. Dataset stored in proper format were found to be insufficient. The WEKA Data Miner software requires input to be provided in a particular format. Consequently, it was deemed necessary to convert the data **csv** format.

Data analysis: In the data analysis stage, data are analyzed to achieve the desired research objectives. In the data mining techniques comprise a suite of algorithms such as J48 Decision Trees, JRip classifiers rules and NaiveBayes. In this study, J48 Decision Trees, JRip classifiers rules and NaiveBayes algorithm.

Test the algorithm: At this stage, the desired algorithm and associated parameters have been chosen based on the parameter and tested.

Model evaluation and pattern prediction: This stage extracts new knowledge or patterns from the result dataset.

Deployment: The final stage of this process applies a previously selected model to new data to generate predictions.

CHAPTER FOUR

EXPERIMENTATION AND RESULT ANALYSIS

4.1 OVERVIEW

In this chapter, the researcher depicts the actual application of data mining process in a stepwise fashion on the credit data of Bank of Abyssinia customers. The classification techniques that have been used in developing a predictive model that Predict the credit risk of Bank of Abyssinia. Include J48Pruned and Unpruned decision tree, JRip classification algorithm and NaiveBayes classifiers were extract the dataset required for training and testing the models created by the classifiers. For creating; predictive model, a total size of 17,536 records was used for training and testing. The validations were done using 10, and 20 -fold cross validation, 75, and 25% and 66% 34 %(the default) split test options.

4.2 MODEL BUILDING

The researcher took 16 attributes: 14 features, 1 derived and 1 class attribute for building predictive model. The selection of attribute was made using subjective judgment of the investigator, reviewing literature and discussion with Portfolio and credit department professionals .To build the predictive model, the. arff and/or .csv format of the selected dataset was given to WEKA. The researcher used J48Prouned and Unpruned decision tree, JRip classification algorithm and .NaiveBayes classifiers**.**

4.2.1SELECTION OF MODELING TECHNIQUE

Modeling is one of the major tasks which are undertaken under the phase of data mining in hybrid methodology. Modeling is one of hybrid data mining technique to select an appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification techniques has been selected for model building. The analysis was performed using WEKA environment. WEKA is one of the open source DM tools having several functionalities. In their study done by Meneses and Grinste [53] stated that as compared to other DM tools Weka:

- ✓ achieved the highest applicability.
- ✓ achieved the highest improvements (when moving from the Percentage Split test mode to the Cross Validation test mode)
- ✓ is the best tool in terms of the ability to run the selected classifiers
- ✓ can better handle the problem of multiclass data sets For these reasons,

WEKA is selected as a modeling tool in this research. WEKA has different available classification algorithms among them J48, Randomforest and Randomtree used for experimentation of this study. In this study the above algorithms selected because of, it's easy of understanding and interpretation of the Result of the model. In other case the J48 algorithms of decision tree generate a model by constructing decision tree where each internal node is a feature or attribute.

Model building is one of the major tasks that are undertaken under the data mining phase in the hybrid methodology of conducting Data Mining Researches. In this study Weka 3.9.3 is used, WEKA 3-9-3 supports many types of classification algorithms. Among the classification algorithms that WEKA 3-9-3 supports the J48, JRip, Naive Bayes algorithm was used with different input parameters as well as different types of related classifiers.

4.2.2 TEST DESIGN

In this section, it is discussed how the samples are prepared for modeling, how the predicting accuracy of each modeling is evaluated and the major tasks to be conducted at each experiment for the selected algorithms. It is necessary to define a procedure to test the model's quality and validity, which means the model prepares for experimentation plan for training, testing and evaluating is required. Thus, in this experiment 16 attributes were selected during the data preparation phase to train and test the selected classifiers. A total of 17,537 records, extracted from the original dataset. In order to check whether the prediction to Credit risk the models was affected by the percentages and fold cross validation of partitioning the dataset into training and test sets.

The dataset is systematically partitioned in to three pairs of disjoint sets of training sets and test sets (for validating the reliability of the models): a

- 75% Training set and 25% Test set, and 66% Training set and 34% Test set, and
- 10 and 20-fold cross validation

Besides, other standard measure including precision, recall, sensitivity and specificity are available. Therefore, the test design specifies that the dataset should be separated into training and test set, and builds the model on the training set and estimate its quality on the separate test set. Process of building predictive models requires a well-defined training and validation protocol in order to insure that most accurate and robust prediction. As above suggestion as in this Research use the data set as training and testing. In WEKA Environment has used to set up and measure the quality, validity and test of the selected model. For purpose of this study k-fold (10-folds) cross validation percentage than 75-25 and 66-34 split test options are used because of its relatively low bias and variations.

A total of 16 experiments were carried out, where 8 of the experiments were constructing for J48 algorithm with 10 and 20-fold cross validation, and 75 and 66 % split test with pruned and unpruned .

Secondly, where 4 of the experiment were constructing for JRip algorithm with 10 and 20-fold cross validation, and 75 and 66% split test, and the remaining 4 of the experiment were constructing for Naive Bayes algorithm with 10 and 20-fold cross validation, and 75 and 66% split test. In relation to this, J48 was the algorithm used to construct and extraction of the corresponding rules.

There are different experiments carried out by using all the 16 attributes of the records with different schemes were applied in the experiment. Three data mining algorithms. Analysis of the J48 decision tree JRip rule induction and Naive Bayes are compered to Predict Banks Credit Risk were made in terms of detailed accuracy, precision, recall, F-measure and ROC curve of the classifier based on a confusion matrix of each predictive model resulted of different classes (Low Risk and High Risk in this research thesis).

Algorithm	Name of experiment	Description of experiment
J48 pruned decision tree	Experiment #1	J48 pruned decision tree with75% split test mode
	Experiment #2	J48 pruned decision tree with 64% split test mode
	Experiment #3	J48 pruned decision tree with 10-fold cross validation test mode.
	Experiment #4	J48 pruned decision tree with 20-fold cross validation test mode.
J48 unpruned decision tree	Experiment #5	J48 unpruned decision tree with75% split test mode
	Experiment #6	J48 unpruned decision tree with 64% split test mode
	Experiment #7	J48 unpruned decision tree with 10-fold cross validation test mode.
	Experiment #8	J48 unpruned decision tree with 20-fold cross validation test mode.
JRip pruned Induction	Experiment #9	JRip pruned Induction with75% split test mode
	Experiment #10	JRip pruned Induction with 64% split test mode
	Experiment #11	JRip pruned Induction with 10-fold cross validation test mode.
	Experiment #12	JRip pruned Induction with 20-fold cross validation test mode.
JRip unpruned Induction	Experiment #13	JRip unpruned Induction with75% split test mode
	Experiment #14	JRip unpruned Induction with 64% split test mode
	Experiment #15	JRip unpruned Induction with 10-fold cross validation test mode.
	Experiment #16	JRip unpruned Induction with 20-fold cross validation test mode.
Naive Bayes pruned	Experiment #17	Naive Bayes pruned with 75% split test mode
	Experiment #18	Naive Bayes pruned with 64% split test mode
	Experiment #19	Naive Bayes pruned with 10-fold cross validation test mode.
	Experiment #20	Naive Bayes pruned with 20-fold cross validation test mode.
Naive Bayes unpruned	Experiment #21	Naive Bayes unpruned with 75% split test mode
	Experiment #22	Naive Bayes unpruned with 64% split test mode
	Experiment #23	Naive Bayes unpruned with 10-fold cross validation test mode.
	Experiment #24	Naive Bayes unpruned with 20-fold cross validation test mode.

Table 4.1 List of experiments conducted using the three algorithm

These experiments were analyzed to compare them in terms of different performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC using three algorithms.

4.2.3 Model Building using J48 Decision Tree

J48 decision tree is one model building classification algorithm in data mining technique to conduct data mining research. In relation to this, J48 was the algorithm used to construct the decision trees using 16 attributes with experiment.

4.2.4 Model Building using J48 pruned Decision Tree

J48 decision tree is one model building classification algorithm to conduct data mining research. Summary of performance result is presented in table 4.2 using J48 pruned algorithm by changing test modes.

S. No	**Comparing parameters**	**Experiments' No**			
		1	**2**	**3**	**4**
1	Testing Mode	75% train	66% train	10-fold cross-validation	20-fold cross-validation
3	Confidence Factor	0.25	0.25	0.25	0.25
4	No. of Leaves	1307	1307	1307	1307
5	Size of Tree	2036	2036	2036	2036
6	TP Rate	0.944	0.833	0.957	0.960
7	FP Rate	0.054	0.168	0.043	0.040
8	Time Taken (sec.)	0.77	0.78	0.84	0.76
9	Precision	0.944	0.834	0.957	0.960
10	Recall	0.944	0.833	0.957	0.960
11	F-Measure	0.944	0.832	0.957	0.960
12	MCC	0.889	0.666	0.914	0.921
13	ROC area	0.970	0.868	0.979	0.984
14	PRC Area	0.961	0.839	0.973	0.977
15	Accuracy (%)	94.4 %	83.2 %	95.7 %	**96.0 %**

Table 4.2: Summary of experimental result of J48 pruned Decision Trees' algorithm

The above J48 pruned decision tree experiments were basically done by using 75% and 66% split test modes and Cross-validation fold of 10 and 20. Classification using the 20-fold cross-validation validation has been tested by the default value of confidence factor (0.25) and achieved highest accuracy of 96.0**%**. 10-fold cross validation also registers result with an accuracy 95.7 %

4.2.5 Model Building Using J48 unpruned Decision Tree

Summary of experimental results of J48 unpruned Decision Tree is summarized in table 4.3.

S. No	Comparing parameters	Experiments' No			
		5	**6**	**7**	**8**
1	Testing Mode	75% train	66% train	10-fold cross-validation	20-fold cross-validation
3	Confidence Factor	0.25	0.25	0.25	0.25
4	No. of Leaves	1717	1717	1717	1717
5	Size of Tree	2573	2573	2573	2573
6	TP Rate	0.958	0.855	0.967	0.970
7	FP Rate	0.043	0.145	0.033	0.030
8	Time Taken (sec.)	0.42	0.56	0.5	0.46
9	Precision	0.958	0.855	0.967	0.970
10	Recall	0.958	0.855	0.967	0.970
11	F-Measure	0.958	0.855	0.967	0.970
12	MCC	0.915	0.710	0.935	0.940
13	ROC area	0.977	0.886	0.984	0.987
14	PRC Area	0.968	0.853	0.978	0.982
15	Accuracy (%)	95.76 %	85.50 %	96.73 %	**97.02 %**

Table 4.3: Summary of experimental result of J48 unpruned Decision Trees' algorithm

The above J48 unpruned decision tree experiments were basically done using 75% and 66% split test modes and Cross-validation fold of 10 and 20. Classification using the 20-fold cross-validation validation folds has been tested by the default value of confidence factor (0.25) and got 97.2 **%** highest accuracy.

4.2.6 Model Building Using JRip pruned Rules algorithm

This experiment was performed on the JRip pruned Rules algorithm; it is an alternative representative of a classification Rules follows the same procedure applied on the previous experiment which are presented above. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. The result for four experiments conducted with 10-fold and 20-fold cross validation, and 75 and 66% split test is presented in table 4.4.

S. No	Comparing parameters	Experiments' No			
		9	**10**	**11**	**12**
1	Testing Mode	75% train	66% train	10-fold cross-validation	20-fold cross-validation
3	Confidence Factor	-	-	-	-
4	No. of Leaves	-	-	-	-
5	Size of Tree	-	-	-	-
6	TP Rate	0.890	0.763	0.920	0.916
7	FP Rate	0.111	0.241	0.082	0.085

8	Time Taken (sec.)	137.91	126.91	95.2	117.97
9	Precision	0.891	0.782	0.921	0.918
10	Recall	0.890	0.763	0.920	0.916
11	F-Measure	0.889	0.758	0.920	0.916
12	MCC	0.781	0.544	0.841	0.834
13	ROC area	0.905	0.771	0.935	0.934
14	PRC Area	0.875	0.728	0.921	0.922
15	Accuracy (%)	88.96 %	76.26 %	**91.99 %**	91.64 %

Table 4.4: Summary of experimental result of JRip pruned Rules algorithm

When we compare the performance of the models produced by JRip Rules algorithm the use of 10-fold validation registered the highest accuracy of 91.99 %. This is followed by 20 fold cross validation mode with the score of 91.64% accuracy**.**

4.2.7 MODEL BUILDING USING JRIP UNPRUNED RULES ALGORITHM

This experiment was performed on the JRip unpruned Rules algorithm it is an alternative representative of a classification Rules follows the same procedure applied on the previous experiment which are presented above. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. The result for four experiments conducted using with 10 and 20-fold cross validation, and 75 and 66% split testis presented in table 4.5

S. No	**Comparing parameters**	**Experiments' No**			
		13	**14**	**15**	**16**
1	Testing Mode	75% train	66% train	10-fold cross-validation	20-fold cross-validation
3	Confidence Factor	-	-	-	-
4	No. of Leaves	-	-	-	-
5	Size of Tree	-	-	-	-
6	TP Rate	0.848	0.882	0.895	0.909
7	FP Rate	0.150	0.117	0.105	0.091
8	Time Taken (sec.)	5.72	4.5	4.34	4.18
9	Precision	0.879	0.898	0.895	0.909
10	Recall	0.848	0.882	0.895	0.909
11	F-Measure	0.845	0.881	0.895	0.909
12	MCC	0.727	0.780	0.790	0.818
13	ROC area	0.849	0.882	0.941	0.948
14	PRC Area	0.807	0.843	0.932	0.943
15	Accuracy (%)	84.9 %	88.2 %	89.48 %	**91.0 %**

Table 4.5: Summary of experimental result of JRip unpruned Rules algorithm

When we compare the performance of the models produced by JRip Rules algorithm it registered the highest accuracy of 91.0 %.

4.2.8 Model Building Using Naive Bayes pruned Classifier Algorithm

In the study, Nave Bayes classifier is also tested to see its performance in predicting credit risk of customers. Experimental result of Nave Bayes classifiers is presented in table 4.6 below

S. No	Comparing parameters	Experiments' No			
		17	**18**	**19**	**20**
1	Testing Mode	75% train	66% train	10-fold cross-validation	20-fold cross-validation
3	Confidence Factor	-	-	-	-
4	No. of Leaves	-	-	-	-
5	Size of Tree	-	-	-	-
6	TP Rate	0.527	0.528	0.534	0.533
7	FP Rate	0.478	0.485	0.488	488
8	Time Taken (sec.)	0.05	0.03	0.04	0.19
9	Precision	0.582	0.577	0.579	0.578
10	Recall	0.527	0.528	0.534	0.533
11	F-Measure	0.427	0.423	0.432	0.431
12	MCC	0.090	0.082	0.086	0.085
13	ROC area	0.628	0.634	0.633	0.634
14	PRC Area	0.607	0.611	0.613	0.613
15	Accuracy (%)	52.74 %	52.78 %	**53.39 %**	53.34 %

Table 4.6: Summary of experimental result of Naive Bayes pruned Classifier Algorithm

When we compare the performance of the models produced by Naive Bayes Classifier Algorithm, is registered the lowest accuracy as compared to J48 decision tree and JRip rule induction. Naive Bayes achieved the highest accuracy of 53.39 % with 10-fold cross validation.

4.2.9 Model Building Using Naive Bayes unpruned Classifier Algorithm

In the study, Nave Bayes classifier is also tested to see its performance in predicting credit risk of customers. Experimental result of Nave Bayes classifiers is presented in table 4.5 below

S. No	Comparing parameters	Experiments' No			
		21	**22**	**23**	**44**
1	Testing Mode	75% train	66% train	10-fold cross-validation	20-fold cross-validation
3	Confidence Factor	-	-	-	-
4	No. of Leaves	-	-	-	-
5	Size of Tree	-	-	-	-
6	TP Rate	0.779	0.771	0.799	0.805
7	FP Rate	0.221	0.229	0.201	0.195
8	Time Taken (sec.)	0.22	0.05	0.03	0.01
9	Precision	0.781	0.774	0.801	0.807
10	Recall	0.779	0.771	0.799	0.805
11	F-Measure	0.779	0.770	0.799	0.805
12	MCC	0.560	0.545	0.600	0.612
13	ROC area	0.851	0.843	0.868	0.873
14	PRC Area	0.846	0.841	0.865	0.871
15	Accuracy (%)	77.91 %	77.06 %	**79.94 %**	80.52 %

Table 4.7: Summary of experimental result of Naive Bayes unpruned Classifier Algorithm

When we compare the performance of the models produced by Naive Bayes Classifier Algorithm, is registered the lowest accuracy as compared to J48 decision tree and JRip rule induction. Naive Bayes achieved the highest accuracy of 79.94.39 % with 10-fold cross validation.

4.2.10 COMPARISON OF THE EXPERIMENTED CLASSIFICATION MODELS

A comparison of J48 pruned, J48 unpruned, JRip pruned, and JRip unpruned Rules and Naive Bayes pruned, Naive Bayes pruned Classifier models is predicted in table 4.8

Comparison of J48 pruned, J48 unpruned, JRip Rules and Naive Bayes Classifier models

S. No	Comparing parameters	Algorithm's					
		J48 pruned	**J48 Unpruned**	**JRip Unpruned Rules**	**JRip Unpruned Rules**	**Naïve Bayes Unpruned**	**Naive Bayes pruned**
1	TP Rate	0.960	0.970	0.920	0.909	0.534	0.799
2	FP Rate	0.040	0.030	0.082	0.091	0.488	0.201
3	Time Taken (sec.)	0.76	0.46	95.2	4.18	0.04	0.03
4	Precision	0.960	0.970	0.921	0.909	0.579	0.801
5	Recall	0.960	0.970	0.920	0.909	0.534	0.799
6	F-Measure	0.960	0.970	0.920	0.909	0.432	0.799
7	MCC	0.921	0.940	0.841	0.818	0.086	0.600
8	ROC area	0.984	0.987	0.935	0.948	0.633	0.868
9	PRC Area	0.977	0.982	0.921	0.943	0.613	0.865
10	Accuracy (%)	96.0%	**97.02 %**	91.99 %	90.90 %	79.94 %	**53.39** %

Table 4.8: Comparison of J48 pruned, J48 unpruned, and JRip pruned Rules, JRip unpruned Rules, pruned Naïve Bayes and unpruned Naïve Bayes Classifier models

Among the experimented classification algorithms J48 unpruned decision tree registered the highest accuracy of 97.02 %.Accordingly, this algorithm is selected for classifications of banks credit risk.

The Confusion matrix of the selected algorithm is presented as follows

```
=== Confusion Matrix ===
  a      b   <-- classified as
 8723  266 |    a = low Risk
  257 8285 |    b = High Risk
```

The entries in the confusion matrix have the following meaning:

✓ 8723 is the number of correct predictions that an instance is **Low Risk**

✓ 266 is the number of **incorrect** predictions that an instance is **High Risk**

- ✓ 257 is the number of **incorrect** predictions that an instance is **Low Risk**
- ✓ 8285 is the number of **incorrect** predictions that an instance is **High Risk**,

The above confusion matrix of J48 Unpruned decision tree depicts that of 8989 Low Risks contracts 8723 are classified as Low risk (97.04%) and the actually good 266 were classified as High Risk (2.96%). On the other hand out of the resampled 8542 High Risk 8285 were classified as High Risk (96.99) and 257 of them were wrongly classified as Low Risk (3.01%). This entails that the records with the low risk class are classified with better accuracy than high risk loans.

4.3 EXTRACTING RULES

The experiments of credit information using the J48 unpruned decision tree technique showed better performance as compared to JRip and NaiveBayes. Hence this algorithm is selected for generating rules. The set of rules are extracted simply by traversing through the output of the decision tree

The researcher has extracted rules that are believed to be unambiguous, relevant and novel to the domain experiments and shared and discussed the result with the loan portfolio officers and credit department domain experts.

The following are the selected set of rules which are in line with the survey of credit risk assessment and primarily gain the attention of domain experts.

1. IF APPLICATION AMOUNT <= 999248 and EXPERIANCES IN BUSINESS > 3 and SECTOR = Private limited Company and LOCATION_DISTRICT = West Addis District Risk type: low Risk (51.0)
2. IF APPLICATION AMOUNT <= 999248 and SECTOR = Share Company and EXPERIANCES IN BUSINESS > 3 and LOCATION_DISTRICT = Central Addis District: low Risk (45.0/1.0)
3. IF APPLICATION AMOUNT <= 999248 and SECTOR = Share Company and EXPERIANCES IN BUSINESS > 15 and LOAN TYPE = WRKG CAPTL LOAN and OTHER BANK LOAN = NO and LOCATION_DISTRICT = Central Addis District: low Risk (45.0/1.0)
4. IF APPLICATION AMOUNT <= 999248 and SECTOR = Share Company and EXPERIANCES IN BUSINESS > 15 and LOAN TYPE = WRKG CAPTL LOAN and OTHER BANK LOAN = yes and LOCATION_DISTRICT = Mekele District: High Risk (2.0)
5. IF APPLICATION AMOUNT <= 999248 and EXPERIANCES IN BUSINESS > 3 and SECTOR = Private limited Company and BUSSINESS PROPOSAL = YES and COLLATERAL VALUE = Medium and LOAN TYPE = WRKG CAPTL LOAN and LOCATION_DISTRICT = Mekele District and DEPOSIT AMOUNT <= 2274672.894 Risk type: low Risk (19.0)
6. IF APPLICATION AMOUNT <= 999248 and EXPERIANCES IN BUSINESS > 3 and SECTOR = Private limited Company and BUSSINESS PROPOSAL = YES and COLLATERAL VALUE =

Medium and LOAN TYPE = WRKG CAPTL LOAN and LOCATION_DISTRICT = Hawassa District and DEPOSIT AMOUNT <= 2274672.894 Risk type : low Risk (33.0/1.0)

7. IF APPLICATION AMOUNT <= 999248 and EXPERIANCES IN BUSINESS > 3 and SECTOR = Private limited Company and OTHER BANK LOAN = YES and COLLATERAL VALUE = Medium and LOAN TYPE = WRKG CAPTL LOAN and LOCATION_DISTRICT = Bahir Dar District and DEPOSIT AMOUNT <= 2274672.894 Risk type: low Risk (18.0

8. IF APPLICATION AMOUNT <= 999248 and EXPERIANCES IN BUSINESS > 3 and SECTOR = Private limited Company and BUSSINESS PROPOSAL = YES and COLLATERAL VALUE = Medium and LOAN TYPE = WRKG CAPTL LOAN and LOCATION_DISTRICT = Hawassa District and DEPOSIT AMOUNT <= 2274672.894 Risk type : low Risk (33.0/1.0)

9. IF SECTOR = Private limited Company and COLLATERAL VALUE = Medium and LOAN TYPE = WRKG CAPTL LOAN and LOCATION_DISTRICT = East Addis District and APPLICATION AMOUNT > 106416 and TIN_NO = YES and EXPERIANCES IN BUSINESS <= 6 Risk type : low Risk (21.0)

10. IF APPLICATION AMOUNT <= 999248 and IF COLLATERAL VALUE = Medium and DEPOSIT AMOUNT <= 47248 and SECTOR = Share Company and LOAN TYPE = WRKG CAPTL LOAN and APPLICATION AMOUNT <= 125820And EXPERIANCES IN BUSINESS <= 6 and OTHER BANK LOAN = YES and LOCATION_DISTRICT = East Addis District Risk Type: low Risk (27.0)

11. IF APPLICATION AMOUNT >= 999248 andIF APPLICATION AMOUNT > 110000RENWAL TRADE LIC = YES and OTHER BANK LOAN = NO and LOAN TYPE = WRKG CAPTL LOAN and EXPERIANCES IN BUSINESS <= 16 and DEPOSIT AMOUNT <= 5140779 Risk type : low Risk (33.0)

12. IF APPLICATION AMOUNT <= 197703 and APPLICATION AMOUNT > 740000 and LOAN TYPE = MORTGAGE LOAN and EXPERIANCES IN BUSINESS <= 15 Risk Type: High Risk (20.0)

13. IF APPLICATION AMOUNT <= 887832 and SECTOR = Private limited Company and

DEPOSIT AMOUNT > 172700 and TIN_NO = YES and LOAN TYPE = WRKG CAPTL LOAN Risk type : low Risk (21.0)

14.IF APPLICATION AMOUNT <= 785963 And IF COLLATERAL VALUE = Medium and DEPOSIT AMOUNT > 165000 and SECTOR = Private limited Company and LOCATION_DISTRICT = Mekele District and EXPERIANCES IN BUSINESS > 15 and LOAN TYPE = WRKG CAPTL LOAN Risk Type: High Risk (54.0)

15. IF APPLICATION AMOUNT > 999248 IF LOAN TYPE = WRKG CAPTL LOAN and COLLATERAL VALUE = Medium

And EXPERIANCES IN BUSINESS <= 16 and LOCATION_DISTRICT = Hawassa District and OTHER BANK LOAN = YES and SECTOR = Private limited Company and APPLICATION AMOUNT > 654720 Risk Type: High Risk (44.0)

4.4 USE OF KNOWLEDGE DISCOVERED

The results of data mining models are evaluated whether the discovered knowledge is novel and interesting and the results of the models are interpreted with respect to domain experts' knowledge. This step includes interpretation of the customers, cross checking the customers and observing the interestingness and relationship of the discovered knowledge, review the process and another means of discovering knowledge can be assessed [24]. Based on the review another hybrid data mining step can be determined for evaluating the discovered results. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

4.4.1PROTOTYPE DEVELOPMENT

A typical decision support system consists of the following components: The screening model consists, the data base which store user information user name and pass word, rules, the user interface, and the users. One of the major differences between decision support systems employing data mining tools and those that employ rule-based expert systems rests in the knowledge engine.

In the decision support systems that utilize rule-based expert systems, the inference engine must be supplied with the facts and the rules associated with them that are often expressed in sets of "if–then" rules. In this sense, the decision support system requires an extracted knowledge on the part of the decision maker in order to provide the right answers to well-formed questions. On the contrary, the decision support systems employing data mining tools on the part of the decision maker. Instead of the system is designed to find new and unsuspected patterns and relationships in a given set of data. This is assisting credit and portfolio department professional, reduce time of decision making simple and easy to implement and integrated to other service delivery system and the management of the credit planning.

4.4.2 BANK CREDIT RISK PREDICTION USER INTERFACE (BCRP)

The user interface is a channel for communication between the system and the end-user. Therefore, in order to design the (BCRP) to be an interactive tool, decision was made by referring the set of rules or command in a simple manner. Examples of information to be shown are the consequences made by against with the set of rules. The events made back of screen and an explanation for the actions made by the system. The reason for the significance of the user interface component is the end-users usually evaluate BCRP based on the quality of the user interface instead of the system itself. The user insert the risk assessment data and triangulated with set of rule, such rules are mandatory attributes and optional.

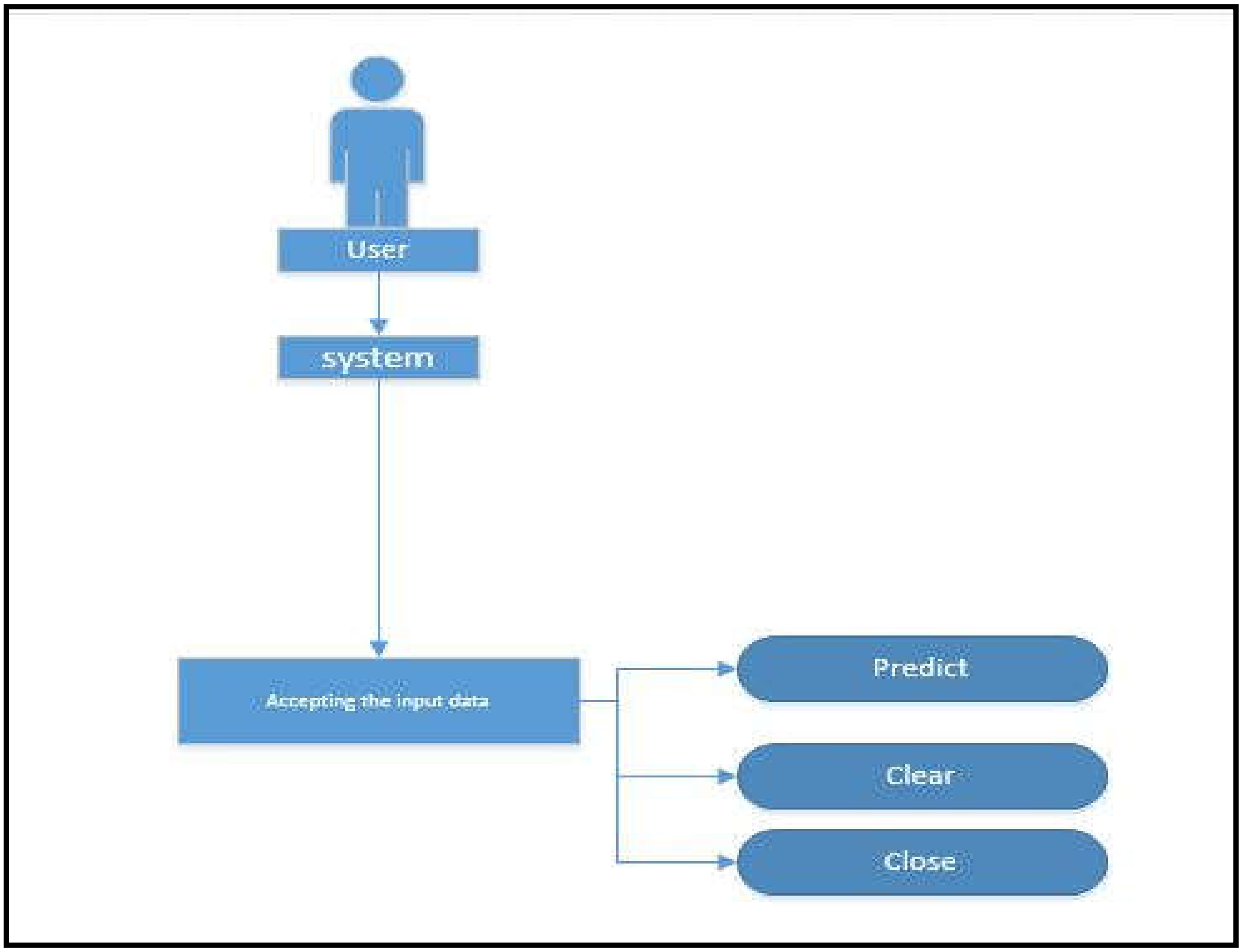


Figure 4.1 User Interface Flow Diagrams

If the user not enters the required data in to system the interface display error massage, the system satisfy mandatory attribute there is two options either High Risk or Low Risk. Home of interface screen display all attributes with drop box with alternative choice, application amount, Deposit amount and experience attribute is take integer and three decision option such as prediction, reset and exit button

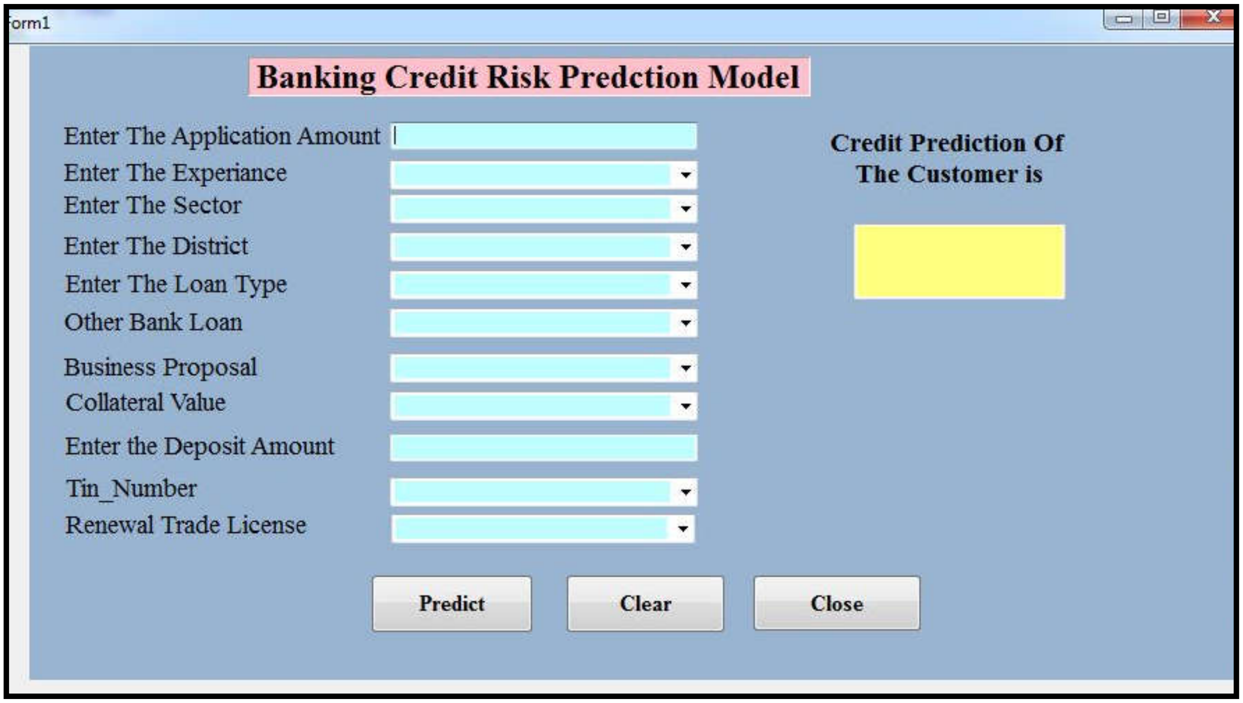
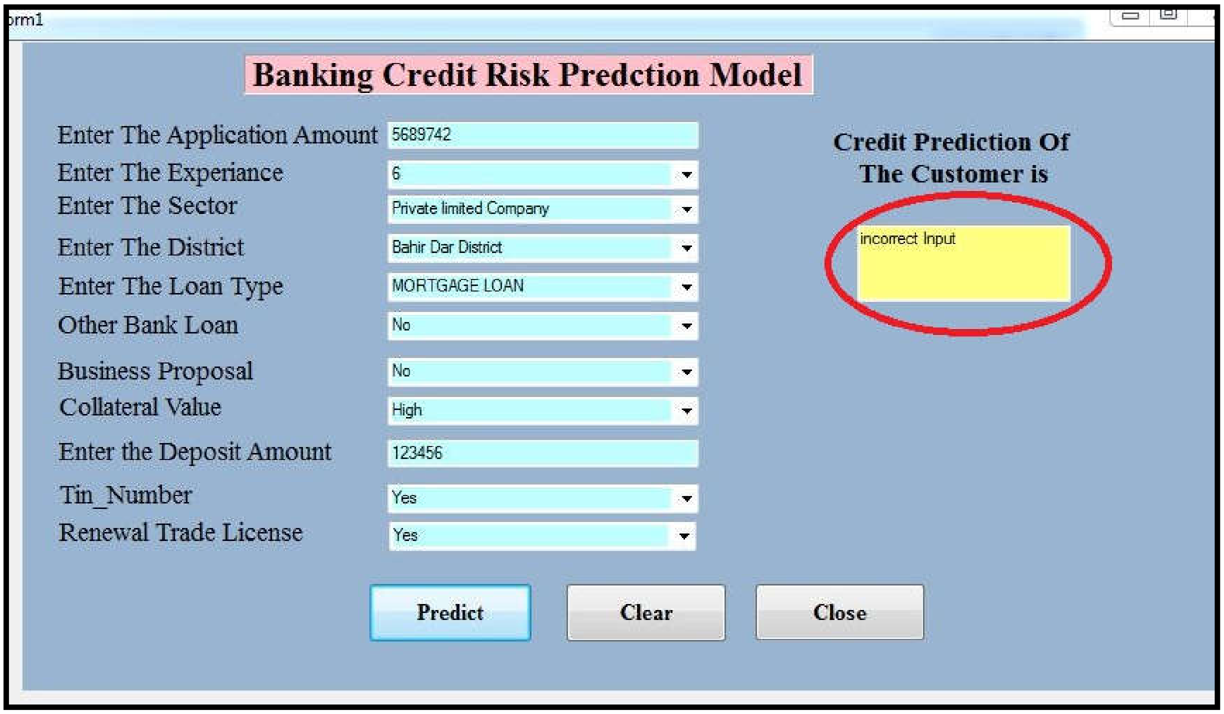


Figure 4.2 Graphical User Interface of the BCRPM (Banking Credit Risk Prediction Model Prototype)

The second option error massage displayer attribute. This system is displaying error massage when the user dose not input and select the appropriate data from the drop list.

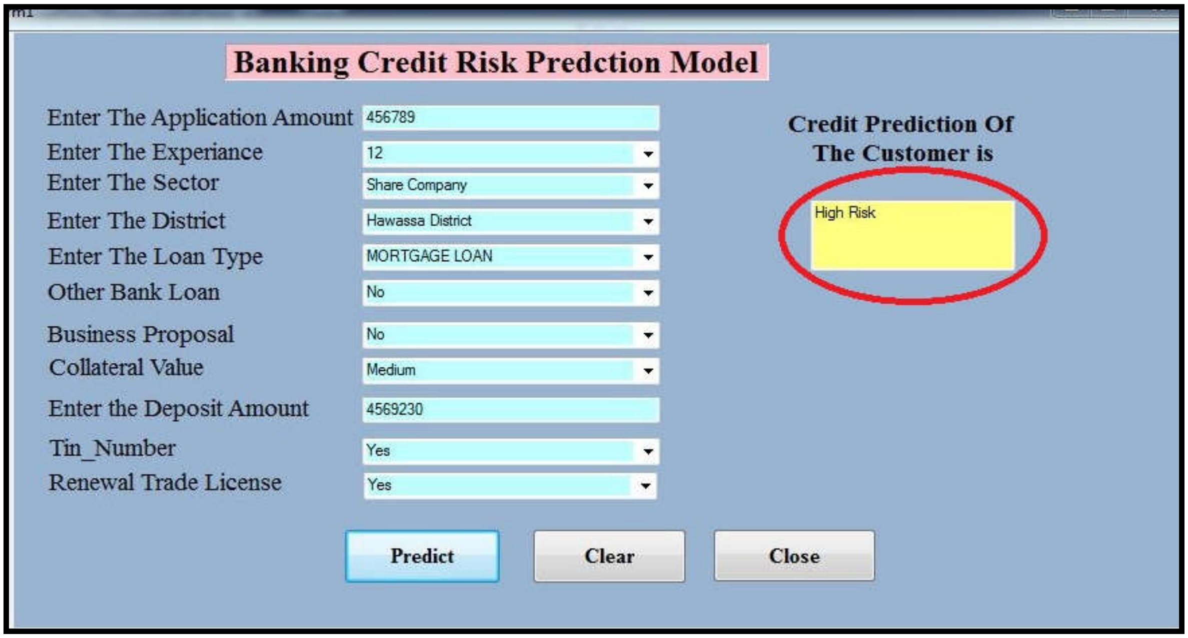


Enter The Application Amount 5689742
Enter The Experiance 6
Enter The Sector Private limited Company
Enter The District Bahir Dar District
Enter The Loan Type MORTGAGE LOAN
Other Bank Loan No
Business Proposal No
Collateral Value High
Enter the Deposit Amount 123456
Tin_Number Yes
Renewal Trade License Yes

Credit Prediction Of The Customer is
incorrect Input

Figure 4.3 Graphical User Interface with error massage unsatisfied mandatory attribute

The other dimension of the interface output displaying interface. The output displaying screen contain results of the prediction either “High Risk” when the credit predicted as a Highly Risky. The result is displaying in following figure.



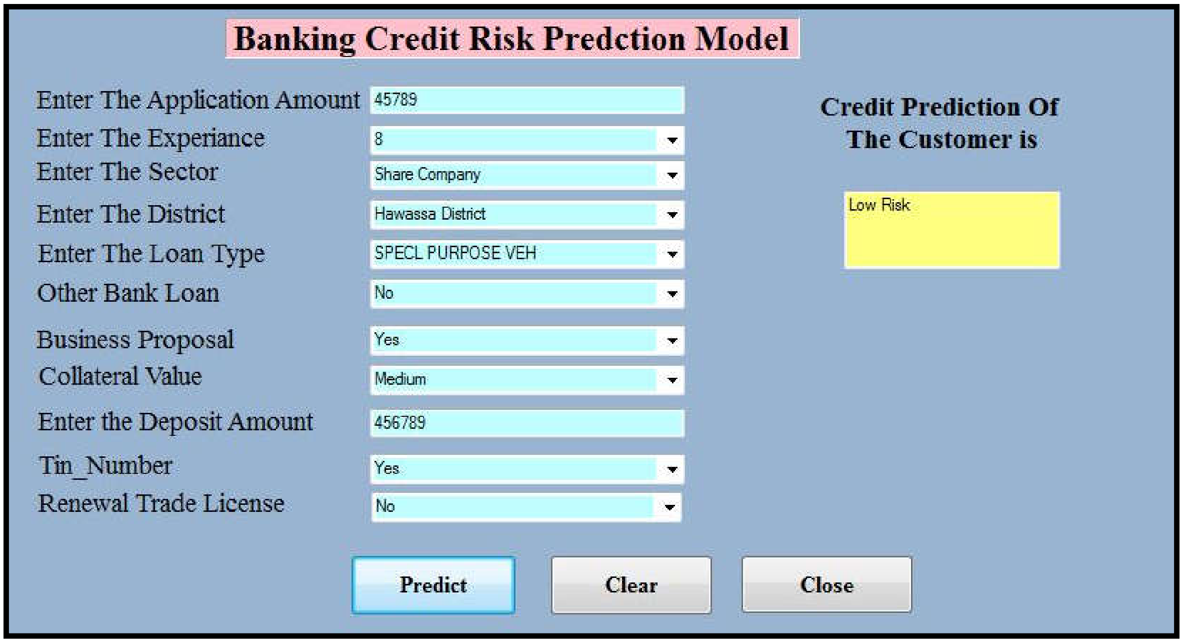
Banking Credit Risk Predction Model

Enter The Application Amount 456789
Enter The Experiance 12
Enter The Sector Share Company
Enter The District Hawassa District
Enter The Loan Type MORTGAGE LOAN
Other Bank Loan No
Business Proposal No
Collateral Value Medium
Enter the Deposit Amount 4569230
Tin_Number Yes
Renewal Trade License Yes

Credit Prediction Of The Customer is
High Risk

Figure 4.4 display result User Interface with displaying massage High Risk

The other dimension of the interface output displaying interface. The output displaying screen contain results of the prediction either “Low Risk” when the credit predicted as a Low Risky. The result is displaying in following figure.



Banking Credit Risk Predction Model

Enter The Application Amount 45789
Enter The Experiance 8
Enter The Sector Share Company
Enter The District Hawassa District
Enter The Loan Type SPECL PURPOSE VEH
Other Bank Loan No
Business Proposal Yes
Collateral Value Medium
Enter the Deposit Amount 456789
Tin_Number Yes
Renewal Trade License No

Credit Prediction Of The Customer is
Low Risk

Figure 4.5 display result User Interface with displaying massage Low Risk

4.4.3USER ACCEPTANCE TESTING

After the system was development process complete, the next step was testing and evaluating the system whether the system satisfies the users need and asses‟ performance of the system. The scope of testing and evaluation that is accomplished and the significance involved to depend on the complexity, and other core features of the system. As the aim of testing and evaluation of the system is to assure that the system expected what it is required to do. The Loan officers, Credit professional and Portfolio managers selected from Bank. However, the participants were oriented about the system's flow and what are the system features.

The goal of the system usability test was to determine the usability of Banking credit Risk prediction model Usability often refers as the question of how well users can use system functionality. This is also

not one-dimensional property of user interface. It's associated with five attributes learnability, efficiency, memorability, errors (error committed rate) and satisfaction. Ten participant tested by 5 question on each participant and the summary presented by no of question time 10 participant divided by 50 mathematically presented as follows.

The result of usability testing is demonstrated as follows. The values for all measurement tools Likert scale in table are fixed as: Strongly agree = 5, Agree = 4, Neutral = 3, Disagree = 2 and strongly Disagree = 1.based on evaluation demonstrated five question for the ten selected people 76% responds strongly agree , 16% agree only 8% users are undecided to use the software.

USER ACCEPTANCE TESTING							
No	**Criteria of evaluation**	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	**Average**
1	I would like to use this system Frequently.				20%	80%	100%
2	I found the system not complex.					100%	100%
3	I thought that the system was easy to use			10%	10%	80%	90%
4	I think that I wouldn't need the support of a Technical Person to be able to use this system.			10%	20%	70%	90%
5	I thought the system doesn't have inconsistency.			20%	30%	50%	80%
6	**Total**			**8%**	**16%**	**76%**	**92%**

Table 4.7 usability testing of CRPM

Finally, the average usability of the Credit Risk Prediction Model prototype according to the evaluation results filled by the participants (domain experts) majority of people 92% agreed that the system prototype has a good and clear informational and functional explanation regarding the objective of research.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATIONS

5.1 CONCLUSION

Most of the Ethiopian banks have automated their operations by branch computerization and local and wide area network connectivity. This automation and implementation of core banking solutions have created centralized terabytes of data. There are Valuable bits of information which are embedded in these databases of each bank. The historic data is used commonly for customer statement, auditors' verification and sometimes for functional level report consumption purpose only.

The bulky nature of the banking data is inconvenient to harness interesting information by a human analyst as it was used to be in the old manual days. As a result we need a means of extracting valuable information which is hidden in the terabytes of data. Data mining techniques become important to uncover useful but hidden knowledge through an efficient use of information stored in the databases to support the decision-making process of the business owners and other interested parties.

Bank of Abyssinia being one of the former private banks in Ethiopia, has played its own role in the economy by rendering credit facilities to the individuals and companies which are running business in various sectors. The bank uses internal and BOA credit policies, procedures and National bank of Ethiopia (NBE) credit policy and directives and strictly followed manuals in various levels of credit committees before disbursing loan to customers. However, there are total defaulters and inconsistent loan repaying customers which declines the profitability of the bank in particular and threatens the growing economy of the country in general. While fueling the sprinting economy in the country, minimizing the possible defaulters is the prime concern of the bank.

The presence of political, economic, social and technological correlations in the financial market forces the creditors to use substantial amount of subjective elements in the identification of risk free customers as it becomes hard to express through deterministic rules.

This research has assessed the application of DM technology on the credit information of Bank of Abyssinia to predict the pattern of high risky and low risky contracts by developing a classification model using Weka tool.

This research has been conducted according to the HYBRID -DM Model approach. After many pre-processing effort a data set with 17,536 total credit records was used to develop a classification model. J48 Pruned Decision Tree, J48 Unpruned Decision Tree, JRip Classification Rule algorithm and Naive Bayes algorithms were employed to conduct various experiments on the prepared dataset. A Model built by 20-fold cross-validation test mode of unprunedJ48 Decision Tree which registered the highest accuracy (97.0167 %), was selected as best model for prediction purpose.

The finding of this research has generated various rules of high risky and low risky contracts which do have an acceptance by the domain experts. The researcher suggests the use of this model to assist in the non-structured decisions where only the credit committee or managers' intuition and experience are used in the granting process of loans.

5.2 RECOMMENDATIONS

Banks do have the most liquid asset (cash) in their control. This cash comes through various marketing and deposit mobilization techniques and an interest is paid for it. As a result it should not be granted for customers who are not to pay it consistently. So banks need Information on creditworthiness of customers which can be converted to knowledge, which is the most valuable asset in this generation. The researcher believes that findings of this study will give an insight on the application of data mining techniques to make an informed decision by the bank officials, policy makers and governmental bodies.

Based on the findings discussed above, the following recommendations are forwarded:

✓ Even though results from this study were encouraging, further classification techniques like Neural network and Bayesian networks(or combinations of any of the techniques) should be undertaken by including data before the implementation of the core banking system (2012) to have the full picture of the bank's credit history.

✓ Currently the bank performs credit scoring activities which includes relationship with the bank, management quality to grade the risk level of the customer and business. However, these grading is filled on papers and not encoded into the core banking system. A data mining research which includes this data source will have a better chance of predicting the future status of any contract before disbursement. So capturing the workflow of loan processing will be a great input for any data mining or decision support system to be carried out for the bank.

✓ From the experiment done in this research and previous work Data mining techniques could contribute a lot in identifying potential customers that could be high risky thus it could be more important to use the data mining technique as a tool for the decision making process in other word the Bank could optimize its credit assessment effort by employing data mining technology

✓ There is a need to develop a credit risk assessment prototype or knowledge base system for the practical implementation of this academic research endeavor.

REFERENCE

[1] Fayyad, U., Piatesky -Shapiro, G., and Smyth, P. (1996). From Data Mining To Knowledge Discovery in Databases, AAAI Press / the MIT Press, Massachusetts Institute of Technology. ISBN 0-26256097-6 MIT.

[2] Sidhant Sethi , Dheeraj Malhotra and Neha Verma (2016) :Data Mining: Current Applications &Trends International Journal of Innovations in Engineering and Technology (IJIET)

[3] Gartner, Inc (2012),Data Mining, Retrieved from: http://www.gartner.com/it-glossary/data-Mining taken December 2018

[4] Jiawei Han, Micheline Kamber and Jian Pei. (2012): Data Mining. Concepts and Techniques, 3rd Edition.

[5] Zaki,M.J., & Wong, L.(2003), Data Mining Techniques, John Wiley & Sons Ltd, Chichester , England.

[6] Shmueli G , Patel N , Bruce P (2010) data mining for business intelligence: concept, technique and application in Microsoft office excel with XL Miner (2nded). John wily and Sons Inc,Hoboken, New Jersey

[7] Kapil Sharma , Ashok Mani, and Harish Rohil(2014): A Study of Sequential Pattern Mining Techniques, International Journal of Engineering and Management Research, Vol 4, PP 48-55

[8] Ali Serhan Koyuncugil and Nermin Ozgulbas Baskent University, Turkey,2011,Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection, Published In the United States of America by Information Science Reference (an imprint of IGI Global)

[9] Supatcharee Sirikulvadhana ,2002 Data Mining As A Financial Auditing Tool, M.Sc. Thesis in Accounting Swedish School of Economics and Business Administration

[10] KDnuggets (2012), CRISP-DM (Cross Industry Standard Process for Data Mining) phases https://www.kdnuggets.com/data_mining_course/index.html#modules

[11] French Polls and the Aftermath of 2002, 2004, and 2007 by Claire Durand, professor, Department of Sociology, Universities de Montreal.

[12] Hand, D.,Manila., & Smyth,P. (2001). Principles of Data Mining. The MIT Press. ISBN: 026208290x

[13] Shmueli G , Patel N , Bruce P (2010) data mining for business intelligence: concept, technique and Application in Microsoft office excel with XL Miner (2nded). John wily and Sons Inc, Hoboken, New Jersey

[14] Thearling, K. (2012), An Introduction to Data Mining, Retrieved on 25- Feb-2013,from: http://www.thearling.com/text/dmwhite/dmwhite.htm

[15] Anne Marie Donovan,2003: Knowledge Management Systems, LIS 385T The University of Texas at Austin

[16] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. & Wirth, R. (2000).CRISPDM 1.0 step-by-step data mining guide. Technical report, CRISP-DM

[17] Globalization and the Poor: Waiting for Nike in Ethiopia, by S.DERCON, Tijdschrift voor Economy Management, Vol. XLVIII, 4, 2003

[18] Ms. A J. Chamatkar et al Int. Importance of Data Mining with Different Types of Data Applications and

Challenging Areas, Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 4, Issue 5(Version 3), May 2014, pp.38-41

[19] Rajanish, D. (n.d), Data Mining in Banking and Finance: A Note for Bankers. Indian Institute of Management Ahmedabad, 2006

[20] Collaborative Research Center (CRC) 649. (2013). Economic Risk.Humboldt-Universitatzu Berlin. School of Business and Economics. Retrieved on 20-Mar-2013 from: http://sfb649.wiwi.hu-berlin.de/fedchomepage/xplore/ebooks/html/csa/node204.html

[21] K. J. Cios, W. Pedrycz, W. Swiniarski and L. A. Kurgan Data Mining:A Knowledge Discovery Approach New York : Springer Science+Business Media, LLC, 2007

[22] Askale Worku (2001).Possible application of data mining technology in support of Loan Disbursements Activities at Dashen Bank S.C. Unpublished Thesis, Addis Ababa University.

[23] Al. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering*, 8:866-883, 1996.*

[24] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.A1. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 1-35. ***AAAI/Mil'*** Press, 1996.

[25] Arun Yadav and Richa Jindal Article: Security Information Hiding Data Mining Privacy Preserving Technique. . International Journal of Computer Applications 1(15):46–49, February 2010. Published By Foundation of Computer Science

[26] Bank of Abyssinia financial report of 2011/12 taken from :https://www.Bank of Abyssinia .com/annual financial report 2011/12 taken December 2018

[27] Bank of Abyssinia , taken from :https://www.Bank of Abyssinia .com/Home taken 20,December 2018

[28] Data Mining Tasks taken from http://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks taken 20, December 2018

[29] Bank of Abyssinia S.C. (2017). Credit polic

[30] Babbie, Earl. The Practice of Social Research. 8th ed. Detroit: Wadsworth Publishing Company, 1998.

[31] Brachman, R. J. &Anand, T., 1996. The process of knowledge discovery in databases.

[32] SAS Institute Inc., (1998), A SAS Institute Best Practices Paper, Data Mining and the Case for Sampling: Solving Business Problems Using SAS® Enterprise Miner™ Software, Cary, NC: SAS Institute Inc.

[33] Kazi Imran Moin* and Dr. Qazi Baseer Ahmed (2012) Use of Data Mining in Banking International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, pp.738-742

[34] Jiban Kpal (2011) Usefulness application of data mining in extracting information from different perspective Annals of Library and Information Studies Vol. 58.Pp7-16

[35] Basel (1999). Principles for the Management of Credit Risk. Consultative paper issued by the Basel Committee on Banking Supervision

[36] Wills, B. (2012).How Banks Conduct Credit Risk Analysis-and How It Can Affect Your Business. Retrieved on 01-Apr-2013. From: http://creditbuilding.dnb.com/corporate-credit/how-banks-conduct-credit-risk-analysis-and-how-it-can-affect-your-business/

[37] D. L. Olson and D. Delen , Advanced Data Mining Techniques, ,Berlin Heidelberg: Springer-Verlag , 2008

[38] N. Sharma, A. Bajpai, R. Litoriya ―Comparison the various clustering algorithms of Weka tools. International Journal of Emerging Technology and Advanced Engineering, Vol. 2, pp.73- 80, May 2012

[39] Wills, B. (2012).How Banks Conduct Credit Risk Analysis-and How It Can Affect Your Business. Retrieved on 01-Apr-2013. From: http://creditbuilding.dnb.com/corporate-credit/how-banks-conduct-credit-risk-analysis-and-how-it-can-affect-your-business/

[40] Whatis.Com Digital library (2012), What is Descriptive and Predictive Modelling Retrieved on 10-feb-2019 https://whatis.techtarget.com/definition/descriptive-modeling

[41] K. Umamaheswari and S. Janakiraman ―Role of Data mining in Insurance Industry, international journal of advanced computer technology, Vol.3 Issue-6, June-2014

[42] Meretework Shawul (2004) possible application of data mining technology in supporting credit risk assessment: the case of NIB International bank sc. **A** thesis submitted to partial fulfillment of the requirement for the degree of Master of Science in information science Addis Ababa University

[43] Mengistu Tesefaye (2013) The application of Data mining in Credit Risk Assessment: in the case of United Bank S.C, A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science in information science Addis Ababa University

[44] Sara Worku (2016) The application of Data mining technology foe Credit Risk Assessment: in the case of Addis Credit and Saving Institution, A thesis submitted in partial fulfillment of the requirement for the Degree of Master of Science in information science Addis Ababa University.

[45] Henock, W. (2002). Application of Data Mining Techniques to support Customer Relationship Management at Ethiopian Airlines Master's Thesis Addis Ababa University

[46] Denekew, A. (2003). The Application of Application of Data Mining to Support Customer Relationship Management at Ethiopian Airline, Master's Thesis Addis Ababa University

[47] Kumneger, F. (2006). Application of Data Mining Techniques to support Customer Relationship Management (CRM) for Ethiopia Shipping Lines (ESL). Master's Thesis Addis Ababa University

[48] Tariku, A. (2011). Mining Insurance Data For Froude Detection: The Case of Africa Insurance S.C Master's Thesis Addis Ababa University

[49] Luel, B. (2011). The Role of Data Mining Technology in Electronic Transaction Expansion at Dashen Bank S.C Master's Thesis Addis Ababa University

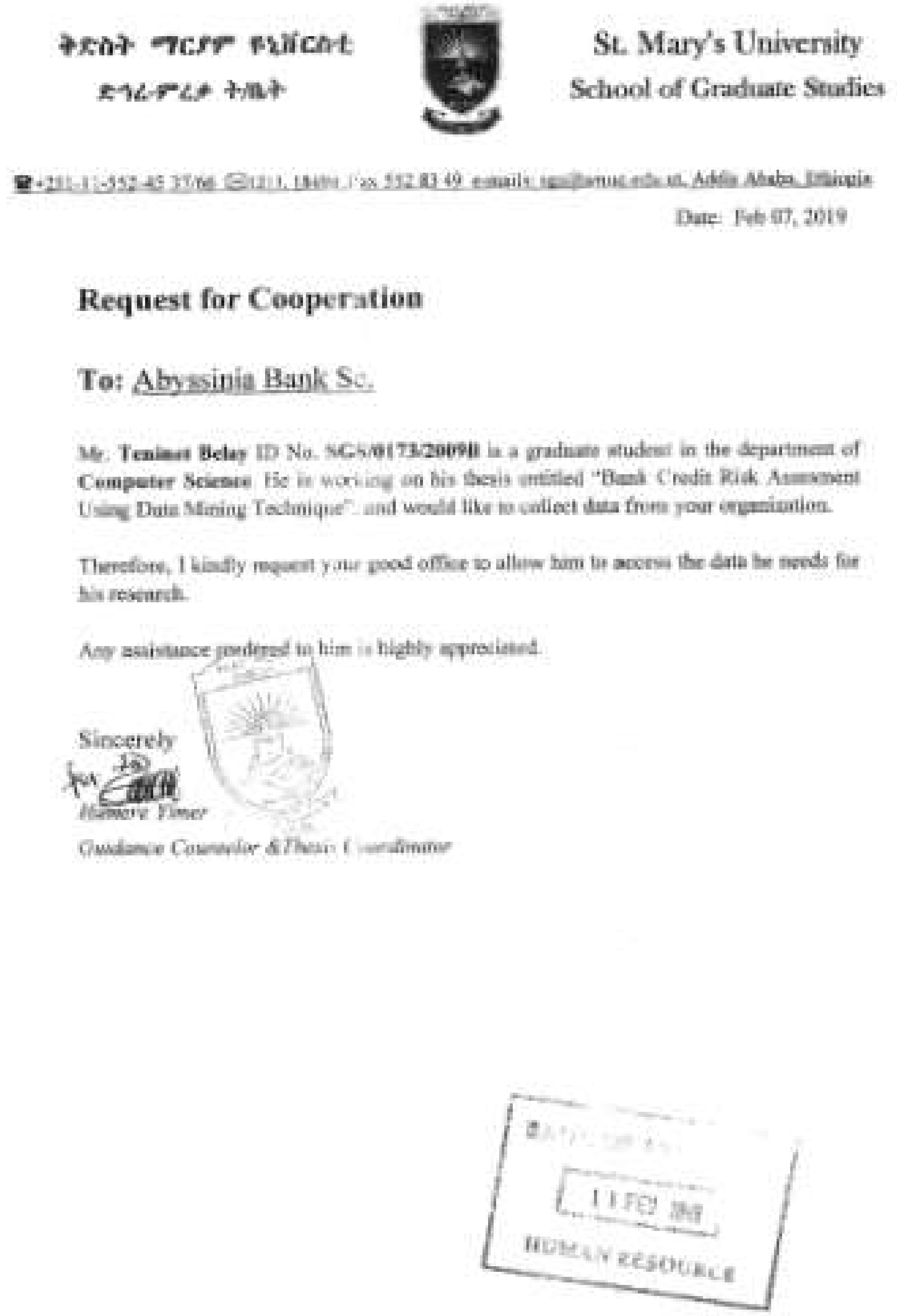
[50] Tesfaye, H. (2002). Predictive Modeling Using Data Mining Technique In Support of Insurance Risk assessment, Master's Thesis Addis Ababa University

[51] Saul Perez, Market Realist, Retrieved on 21-Feb-2019. From: https://articles.marketrealist.com/2014/09/must-know-8-types-bank-risks/

[52] Chawla, N.,Bowyer,K., Hall, L.,Kegelmeyer, P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16 (2002), 321-357

[53] Claudio J. Meneses, Georges G. Grinste (1998). Categorization and Evaluation of Data Mining Techniques Transactions on Information and Communications Technologies vol 19 1998 WIT Press.

APPENDICES



ቅድስት ማርያም ዩኒቨርስቲ
ድኅረ-ምረቃ ት/ቤት

St. Mary's University
School of Graduate Studies

☎+251-11-552-45 37/66 ✉1211, 18490 Fax 552 83 49 e-mails sgs@smuc.edu.et, Addis Ababa, Ethiopia

Date: Feb 07, 2019

Request for Cooperation

To: Abyssinia Bank Sc.

Mr. **Teninet Belay** ID No. SGS/**0173/2009B** is a graduate student in the department of **Computer Science**. He is working on his thesis entitled "Bank Credit Risk Assessment Using Data Mining Technique", and would like to collect data from your organization.

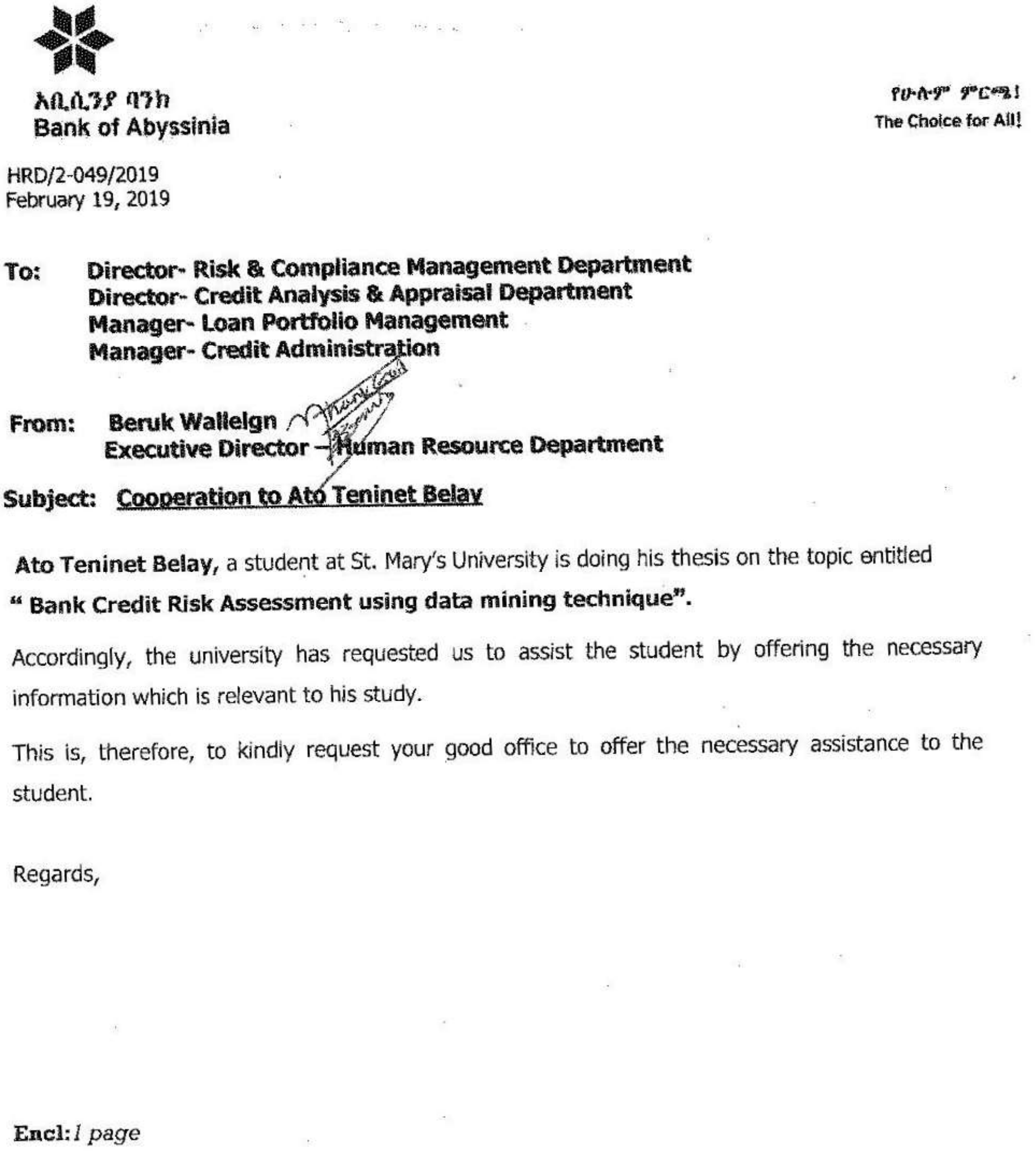
Therefore, I kindly request your good office to allow him to access the data he needs for his research.

Any assistance rendered to him is highly appreciated.

Sincerely

Hamere Yimer

Guidance Counselor &Thesis Coordinator



አቢሲንያ ባንክ
Bank of Abyssinia

የሁሉም ምርጫ!
The Choice for All!

HRD/2-049/2019
February 19, 2019

To: **Director- Risk & Compliance Management Department**
Director- Credit Analysis & Appraisal Department
Manager- Loan Portfolio Management
Manager- Credit Administration

From: **Beruk Wallelgn**
Executive Director – Human Resource Department

Subject: **Cooperation to Ato Teninet Belay**

Ato Teninet Belay, a student at St. Mary's University is doing his thesis on the topic entitled **" Bank Credit Risk Assessment using data mining technique".**

Accordingly, the university has requested us to assist the student by offering the necessary information which is relevant to his study.

This is, therefore, to kindly request your good office to offer the necessary assistance to the student.

Regards,

Encl: *1 page*

SYSTEM PERFORMANCE TEST

Its requesting you to fill the following Usability testing evaluation form for the performance of **Banks Credit Prediction Model Prototype**, put the numbers which most appropriately reflect the performance about using this system from the list. The values for all measurement tools likert scale in table are fixed as: Strongly agree = 5, Agree = 4, Neutral = 3, Disagree = 2 and strongly Disagree = 1.

USABILITY TESTING						
Criteria of evaluation	**Strongly Disagree**	**Disagree**	**Undecided**	**Agree**	**Strongly Agree**	**Average**
I would like to use this system Frequently.						
I found the system not complex.						
I thought that the system was easy to use						
I think that I wouldn’t need the support of a Technical Person to be able to use this system.						
I thought the system doesn’t have inconsistency.						