ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES FOR CUSTOMERS SEGMENTATION AND PREDICTION: THE CASE OF BUUSAA GONOFA MICROFINANCE INSTITUTION

BELACHEW REGANIE

January, 2013

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Application of Data Mining Techniques for Customers Segmentation and Prediction: The Case of Buusaa Gonofa Microfinance Institution

Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

Belachew Reganie

January, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Application of Data Mining Techniques for Customer Segmentation and Prediction: The Case of Buusaa Gonofa Microfinance Institution

By

Belachew Reganie

June, 2013

Name and signature of Member of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____

ACKNOWLEDGMENT

First and foremost I would like to thank my advisor Dr. Gashaw Kebede for his guidance, suggestion, and support throughout my study of this research work. His comments and suggestions have helped me in maintaining the right direction for my study and making it meaningful.

I would also want to thank Daniel Mamo (Head Librarian at IES), and Meseret Ayyana (Coordinator of our research work and Head of Bibliography at IS department), who have provided me important material and information for my success.

I am greatly indebted to my instructor Million Meshesha (PhD) and Solomon Teferra (PhD) for their support during my thesis work.

I would also like to thank sincerely all my friends those who helped me with their valuable support during the entire process of this thesis, specially, Solomon G/Mariam, Tagel Aboneh and Biazen Getnet.

I would also thank my family Ato Shiferaw Hagos and W/r Askale Atomsa, who have been always with me through this study by encouraging and asking me about the progress of my thesis work.

Finally, I would also like to thank BG MFI staffs, particularly, Getachew Mekonnen (Marketing and promotion officer), for providing relevant data and other necessary information for my study. And also my special thanks go to Mesfin Admasu (Branch Manager), for providing relevant data, consulting as domain expert, providing other necessary information and encouraging me by providing necessary material for my research work.

LIST OF ACRONYMS

ARFF - Attribute Relation File Format

BG MFI – Buusaa Gonofa Microfinance

CART - Classification and Regression Trees

CIF - Customer Information File

CRA - Customer Recommended Actions

CRISP-DM – **CR**oss-Industry **S**tandard **P**rocess for **D**ata **M**ining

CRM – Customer Relationship Management

CSV – Comma Separated Values

DM –Data mining

ERP – Enterprise Resource Planning

KDD - Knowledge Discovery and Data Mining

MFI – Microfinance Institution

SCRMS - Strategic Customer Relationship Management System

SFA - Sales Force Automation

SSE - Sum of Squared Error

TABLE OF CONTENTS

Contents **Pages**

LIST OF TABLES

LIST OF FIGURES

Abstract

Identifying customers which are more likely potential to a product and service offering is an important issue. In customers identification data mining has been used extensively to predict potential customers for a product and service.

The final goal of this thesis is to build a model that helps to classify customers for Buusaa Gonofa microfinance institution product and service. Since there are no predefined classes, that describe the customers of the institution, the researcher uses clustering techniques that resulted in the appropriate number of clusters. Then, a predictive model was developed to predict potential customers. This predictive model achieved an accuracy of 99.95%.

For modeling purpose, data was gathered from the institution head office. Since irrelevant features result in bad model performance, data preprocessing was performed in order to determine the inputs to the model.

Thus, various data mining techniques and algorithms were used to implement each step of the modeling process and alleviate related difficulties. K-means was used as a clustering algorithm to segment customers' record into clusters with similar characters. Different parameters were used to run the clustering algorithm before reaching at segment that made business sense. J48 decision tree algorithm was used for classification purpose. In addition to those attributes that are believed by the experts to have high impact on customer segmentation, attributes value of loan amount have a big influence.

Generally, the result of the study was encouraging, which reinforces the possible application of data mining solution to the microfinance industry, particularly, in customer segmentation and prediction in Buusaa Gonofa microfinance institution.

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

The work culture in organizations has changed due to the growth of computer networks and the World Wide Web. The growth of interest in knowledge management has been helping many organizations to digitize their own knowledge for effective use of knowledge in future prediction. Having information resources in the organization is no longer sufficient to access through intranets and/or extranets, but also to efficiently exploit what the system actually knows should be the main goal of knowledge management in the organizational work culture. The successful implementation of knowledge management is that people will share what they know and reuse the knowhow of others. Based on this fact, it is important to uncover knowledge from existing database using data mining techniques. It is also important to focuses on the vision, strategy and appropriate mind set required for successful implementation of knowledge to solve problems using data mining concept and tools (Zhixian Yi, 2008).

The term data mining has no precise definition (Hand, et al, 2001). In this research it is defined as, a process that uses various techniques to discover hidden relevant knowledge from heterogeneous and distributed data stored in large databases, data warehouses and other massive information repositories. Data mining refers to the process of finding interesting patterns in a large database that are not explicitly part of the data. The extracted interesting patterns can be used to tell something new and to make predictions and description (Lori, 2006).It is a young interdisciplinary field drawing from areas such as: database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, neural networks, pattern recognition, spatial data analysis, and many more. (Han and Kamber, 2006).

Data mining techniques can be applied to a wide variety of data repositories including databases, data warehouses, spatial data, multimedia data, Internet or Web-based data and complex objects (Lori, 2006). Advances in computer hardware and data mining software have made data mining systems accessible and affordable to many businesses (SIM, 2002). Its application areas can be educational, banking, medicine, financial, criminology, and many others (Deshpande and

Thakare, 2010). Hence, it is not surprising that data mining has gained widespread attention and increasing popularity among bankers in recent years.

Banking and finance involves different activities including: collecting money from customers with little or justifiable interest rate; lending the collected money to different customers based on some criteria; performing foreign exchange; funding development and supporting payment activities of other businesses.

In doing all these activities, the fundamental and the primary task of banks and finances are finding new customers and retaining the existing customers. Prior to any activities and any investment, banks should have an identification mechanism of their potential and profitable bank customers (OECD, 2009).

Currently, huge electronic data repositories are being maintained by banks and other financial institutions across the globe. Valuable bits of information are embedded in these data repositories. The only problem is that this storehouse of data has to be mined for useful information. Normally, these terabytes of transaction data are collected, generated, printed, stored, only to be filled and discarded after they have served their short-lived purposes as audit trails and paper trails (Rene, 2010).

However, Global competitions, dynamic markets, and rapidly decreasing cycles of technological innovation provide important challenges for the banking and finance industry. As banking competition becomes more and more global and intense, banks have to fight more creatively and proactively to generate knowledge from large databases. This is possible by introducing application of data mining system in banking and finance sector (Rene, 2010).

Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts (Dass, 2012).

Banking sector cannot reach millions of poor for whom small loans could make huge differences. According to Alemayehu (2008), most of the poor are living in rural, and they are

much dispersed and they have low education levels. Thus, the cost of supplying loans to the poor population is extremely high. They do not have any assets to use as collateral to be served through banking. So, to reach those poor, microfinance institutions started giving service.

Accordingly, today, there are millions of poor people around the world who turn out to be entrepreneurs through the micro-credit sector(Alemayehu, 2008).

In Ethiopia Microfinance Institutions (MFI), microfinance credit is considered as one of the methods of alleviating poverty since it is believed that the provision of micro credit to poor households would increase their assets and income. As a result, in the mid-1980s, many non-governmental organizations' in Ethiopia have started providing microcredit to poor households for income generating activities. The Development Bank of Ethiopia, in collaboration with the Ministry of Trade, has also launched a Micro Enterprise Lending Program. Recognizing the importance of Microfinance facility, the present government issued a proclamation that laid down the framework for Licensing and Supervision of the Business of MFIs in July 1996, through Proc. No. 40/1996**(**Micha'el, 2006**)**.

According to Micha'el (2006**)**, the proclamation allows MFIs to undertake both financial and non-financial activities. Therefore, licensing and supervision of MFIs enhanced the status of MFIs as it authorized them, among many other things, to legally accept deposits from General Public, draw and accept deposits, and manage funds for microfinance businesses.

Presently, there are around 27 MFIs operating throughout the country licensed under The National Bank of Ethiopia (Letenah, 2009). However, this Institutions have met only very small of the demand for financial services of the active poor. Even if it is not common in these institutions, they should have customer relationship management system to reach the poor and to be more profitable.

1.2. Background of Buusaa Gonofa

Buusaa Gonofa microfinance institution (BG MFI) is fathered by HUNDEE, a local NGO that implements various development projects including credit scheme.

Microcredit scheme was one of HUNDEE's core programs since its establishment. HUNDEE was providing microcredit services in four districts by targeting poor women and landless youth; this scheme was discontinued in June 1998 due to a new policy issued by the Ethiopian government. This regulation prohibited NGOs from directly engaging in micro credit and saving schemes unless they are register as Micro financing Institutions as per Proclamation number 40/1996 and licensed by the National Bank of Ethiopia to do so. This sudden change in government policy resulted in substantial disruption in the micro credit operations of HUNDEE; many beneficiaries have already received loans and they were actually making repayments to access new higher loans; many groups were organized and were waiting for disbursement of loans; other micro credit projects were in the pipeline for further implementation. These micro credit schemes had significant share in the project budget of HUNDEE at the time and it was not reasonable to abruptly discontinue a scheme that was benefiting the target communities at that stage. Accordingly, HUNDEE decided to establish Buusaa Gonofaa as a separate microfinance institution as per the requirements of the law; this way, the micro credit operations of HUNDEE will be continued in the existing districts and in other areas to be accessed in the future.

The Microfinance Proclamation number 40/1996 has stipulated certain specific requirements to be registered as MFI. Some of the licensing requirements include the following,

- MFIs must be established as Share Company as per the commercial code of Ethiopia;
- All of the owners of MFIs must be Ethiopian nationals (foreigners are prohibited);
- Minimum registered capital of birr 200,000 must be deposited in blocked account in order to process the application for license;
- A minimum of 5 shareholders are required to form MFI; and
- The maximum share of a single shareholder must be limited to 20% of the capital;

It was triggered by this new policy direction that Buusaa was established with a separate legal mandate and started to undertake the credit scheme HUNDEE has established. BG MFI is licensed under Proclamation No 40/1996 and is supervised by the National Bank of Ethiopia. Being a non-bank financial institution, BG has a legal mandate to take deposits from its specific clients as well as the public at large.

BG MFI started its operations not simply as legally separate from HUNDEE but also with fundamental re-orientation and change in some key areas. Business-like approach and financial viability of the service was one of the key priority areas and this involved a two-pronged approach: charging sufficiently high interest rate on its costly micro-loans and efficiency. Continuity of the service is very important even for the clients but many MFIs shied away from setting an interest rate that would ensure repeatable service. BG MFI is the first licensed financial institution that set the highest lending interest rate in Ethiopia – 24% p.a.

1.2.1. Vision/Mission Statement, Objectives and Strategies

The vision of BG MFI is to see the development of an inclusive, efficient and mature financial system that works for all people, rural and urban, the poor and the rich alike.

The mission of BG MFI is to provide flexible and efficient microfinance services on a sustainable basis to improve the livelihood of the resource poor in rural and per-urban areas, particularly, women, small holder farmers and landless youth.

BG MFI will realise its mission through the delivery of poor-friendly micro-loans that meet the needs and capacities of its target market, and through mobilisation of savings. BG will render its services by building an institution that is efficient, profitable, transparent and responsive to the needs of its clients.

BG MFI has the following strategies to achieve its overall objectives and business goals:

- Competitive pricing - setting competitive interest rate on its lending and saving services to cover full costs and ensure reasonable return;
- Promote the exchange of skills and experiences, mutual problem solving and group-based initiatives among the target clients;
- Offering flexible and responsive products that are well suited to the needs and livelihood priorities of the target clients;
- Excellence in customer services - speedy service delivery for repeat borrowers;
- Mobilizing local savings and accessing commercial loans for on lending;

- Building strong institutional base that enables expansion and growth in safe and sustainable way;
- Introducing innovative products and services that are appropriate for deepening outreach into remote rural areas and reaching other under-served market segments;
- Improving efficiency, productivity, and profitability;
- Maintain high quality of loan portfolio;
- Develop and maintain competent human resources committed to quality service for the poor and profitability;
- Establish and maintain effective and efficient policies, management information systems and procedures;
- Implement market studies and client satisfaction surveys on regular basis to respond to client needs;
- Implement social performance management (SPM) system that shows the profile of client at entry and measure progress over time.

This study focuses on mining and discovering the hidden knowledge related to customers' relationship management that helps the BG MFI to implement its strategies and maximize its profit.

1.3. Problem Statement and Justification of the Study

Microfinance Institutions in Ethiopia have registered positive outcome over the past few years. But the service in the area is still low and need higher effort to achieve better result. Some of the weaknesses that characterize the industry are: gap in serving more structured micro enterprises; low technical capacity; difficulties in accessing funds from donors; lack of product diversification and inadequate management of information system (MIS); lack of research to understand the needs of clients; lack of business development services to clients; weak internal control system, and weak marketing strategy (Befekadu 2007; Meklit et al, 2004; Wolday, 2004).

In Ethiopia, awareness and knowledge about microfinance services has progressed in last few years. However, Meklitet.al (2004) show that only 15% of the potential market was reached showing that there is wide demand for financial services in the country.

High competitive environment is one of major challenges of financial service providers in urban area. According to the report of Meklit et. al,(2004), depending on availability of services, customers would like to look for cheaper products and quality service. As there are a number of financial service providers, customers would like to know the details of services and products. This would make the financial service providers in urban areas more competitive as compared with financial service providers in the rural areas.

As reported in Meklit et. al (2004), with the cumulative effect of poor marketing strategy most MFIs are following, they have high retention rate than increased in number of active clientele. Furthermore, the report shows that the service MFIs deliver is the minimum outreach as explained in the following paragraph.

> *"Depending on word-of-mouth marketing as best means of promoting services and products, MFIs would last for long period of time with the same client, boasting high retention rate. The market is in shortage of financial services both for start-up and working capital, whereas MFIs are concentrating on retention or keeping customers instead of looking for additional clients and opportunities. Struggling to keep customers for repeat loan might be one of the cases for un-healthy competition."*

As the report of Meklit et. al (2004) shows, retention of customers without identifying best customer segments from bad customers segments creates un-healthy competition. To identify and predict such segments, the marketing promotion, appraisals and screening department of MFIs are not using modern technology while providing financial service to their customers.

BG MFI shares the same problems as discussed above in microfinance industry. It is not using modern technology while providing financial service to its customers. It is a wholly private owned institution and is responsible for providing loan and saving services all over the region of Oromia nowadays. The institution is concentrating on retention or keeping customers without identifying best customers segments from bad customers segments. Furthermore, the institution

is looking for additional clients and opportunities without identifying and knowing about the existing customers and markets.

The institution has the same business policy and procedure all over its market area. But, individuals' interest in each market area is not the same. Services individual customers demand from the institution cannot be handled accordingly due to institution's business policy. To solve such problems, segmenting customers to identify profitable and potential customers from the others will be the best solution. But, the institution has not implemented modern system and technology to identify and predict customers' need before providing loan.

BG MFI seems to forget the importance of keeping and analyzing of customers data with the appropriate tools and techniques, which help to come up with a better view of its customers and design the appropriate business strategy. Currently, there is a very poor traditional means of knowing the unique features and need of their customers and how to use this knowledge to make better future decisions.

According to the interviews and discussion made with senior managers of the organization on the way they are retaining customers' and looking for additional clients and opportunities, the institution is attempting to attract and retain more customers while expanding the market to outreach potential customers. However, the institution is not using appropriate tools and techniques that help it to segment and identify existing customers in the market before looking for further potential customers. If so the institution cannot pass proper decisions to improve the service according to customers need in each customer segments. As the officials say, the current system does not help them to identify and predict high value and high profitable customers (segments). Thus, some market areas (branches) are not generating the expected profit while few are closed due to lost revenue.

Based on the above facts and problems, existing system requires significant improvement. If a system or methods available that would segment the customers and predict customers' type with respect to the loan usage, it would be highly beneficial for the institution in generating high profit as well as for better customers' relationship management.

Thus, the underlying problem that necessitates this research is the inability of the institution to

identify a group of similar customers in the market to conduct customers' segmentation and predict. This problem interns has its long lasting penalties such as unable to have a clearer and bigger picture of their customers, lost revenue, lessening of customers or a market.

Therefore, this research work is initiated to come up with a data mining techniques that helps to segment and predict profitable customers, so that the institution can make proper decisions in designing strategies in looking for additional clients and opportunities for provision of loan services. This has a significant impact in improving customer relationship management of the institution.

1.4. Objective of the Study

1.4.1. General objective

The general objective of this study is to apply data mining methods and techniques to segment and predict profitable customers of BG MFI for better customer relationship management.

1.4.2. Specific objectives

To achieve the general objective, the research has the following specific objectives:-

- To review related literatures and previous work in the area to have better understanding about the work and to know others contribution in the area if any.
- Collect data on which the mining process will be conducted.
- Preparing the data that will be used for model building by selecting important attributes, and cleaning them.
- Exploring the relationship between different variables (attributes).
- To come up with appropriate number of clusters based on the similarity of instances.
- Building and then, testing a predictive model that will help in the classifying customers in the market to be in one of the identified clusters.
- Finally, interpreting, and then coming up with conclusion and recommendations.

1.5. Methodology

Methodology is the steps or procedure that the researcher follows to achieve the stated objectives. It is a road map that shows the direction how the researcher is going to conduct the research to reach the end. Accordingly, this study followed different methods in order to develop good customers' segmentation and classification models for designing and implementing successful customer relationship management. Specifically, descriptive and predictive methods have been implemented. Through descriptive, the behaviors of individual customers have been studied and categorized into similar clusters. Then, through predictive method, customers assigned and classified into their class.

1.5.1. Data mining model

Modeling is an iterative process - different for supervised and unsupervised learning and it may be either description or prediction. Selection of the data mining modeling techniques is based upon the objective/task of data mining, problem type and type of data found in the real world. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. In the business environment, complex data mining projects may require the coordinate efforts of various experts, stakeholders, or departments throughout an entire organization. In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements (Berry and Linoff, 2000).

The current process model for data mining provides an overview of the life cycle of a data mining project model that used to find patterns from data. It contains the corresponding phases of a project model, their respective tasks, and relationships between these tasks. There exists a relationship between all data mining tasks depending on goals, background and interest of the user, and most importantly depending on the data (Hegland, 2003).

To achieve the objective of this study **CR**oss-Industry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) has been used. This is because CRISP-DM has been widely applied in data mining

studies and also it is flexible for different data. CRISP-DM model is divided into six phases. These are Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment.

1.5.2. Review of related literature

To see what others have done in the area, related literatures have been reviewed. The literature review included data mining algorithms, application of data mining in banking finance and customer relationship management in general and in MFIs in particular.

1.5.3. Exploration of the domain problem

In order to understand the problem, experts have been interviewed in the area. Observation at institution while serving customers was conducted. Finally, related work reviewed to have better understanding, to see the gabs in the institution and others microfinance how they are working now in world wide.

1.5.4. Identification and selection of target dataset

Dataset identification for the data mining task has been done based on the objective of the study. Beside this, advice of domain expert also considered during selection of dataset (explained in detail under Section 4.3).

Hence major task performed at this stage identifying and collecting relevant data for the purpose of research work at hand. Since final result of the research work determined by this task so much effort was done in creating the right dataset.

1.5.5. Data preprocessing

Here after data had been gathered, refining and cleaning of data had been done as it has been convenient to the data mining process. The collected data was stored in Microsoft excel. Then the researcher conducted the preprocessing steps on the data in order to improve the accuracy, efficiency and scalability of clustering and classification process. The preprocessing steps deal with missing values, exclusion of some attributes that are believed to have no use, like Id No, name, etc of the customers; inclusion of attributes that are not explicitly stated in the table, which

are important in the decision making process.

1.5.6. Building, training and evaluating model

Data mining tasks chosen to apply to the problem here are clustering and classification. These techniques are capable of processing a wider variety of data and easier to interpret in giving meaning to the problem. K-means clustering algorithm was used for customers' segmentation while J48 decision tree algorithm was used for prediction purpose.

Based on the output of clustering, class assigned and then a predictive model built to predict individual behavior for the institution services. In building models the algorithm is expected to learn different patterns in the dataset and this learned knowledge by the algorithm could be applied on new dataset. Different experiments have been done with different parameters; models evaluated and the best performing once selected.

The obtained K-means clustering models are evaluated using the classes to clusters evaluation approach. Objects were clustered in classes based on their centroids. In the case of classification the obtained decision tree models are evaluated using the 10 fold cross validation approach. The dataset is divided into 10 subsets, insuring that each class is represented with approximately equal opportunities subsets. Then each subset is used for testing and the remaining 9 for training purposes. The error estimates are averaged to yield an overall estimate.

1.6. Scope of the Study

The scope of this research was restricted to building data mining models for segmenting the customers into similar groups, interpreting the result of segmentation and generating classification rules that would help BG MFI to come up with better CRM. To do these tasks, clustering and classification techniques with K-means and Decision Tree J48 algorithms were employed respectively. However, nowadays, business problems become complex and diversified. It needs testing models with combination of different algorithms in each techniques. But this fact was not employed in the study due to time constraints.

1.7. Significance of the Study

The study resulted in how to handle customers by identifying individual behavior. Beside this, appropriate infrastructures and customer relationship tools/software have been recommended. This could lead the institution to be more competitive in the industry. Therefore, the study can support the routine and strategic decision made by the institution in the industry. By implementing appropriate customer relationship management strategy and by providing an attractive service through the right channel, at the right time and to the right customer, with each customer contact, the institution services and goal can be more likely to be achieved.

The research is also believed to initiate further research in the area, as it is an initial attempt for exploiting the potentials of data mining techniques in the institution for the purpose of achieving better customer relationship management.

1.8. Organization of the Paper

This thesis consists of five chapters. The first chapter deals with the general overview of the study including background of the study, background of the institution, statement of the problem and its justification, objective, scope, application and methodology of the research. The second chapter is devoted to literature review of data mining technology and about customer relationship management including CRM and data mining with its application in Microfinance Industry. The third chapter is review of application of data mining techniques. The fourth chapter is experimentation. It compromises the CRISP-DM process model steps such as business understanding, data understanding, data preparation, modeling, evaluation and deployment of the result. Result of the experiments are also analyzed and interpreted in this chapter. Chapter five presents conclusion that summarizes major points of the research and recommendations have been forwarded for further research.

CHAPTER TWO

LITERATURE REVIEW

DATA MINING CONCEPTS

2.1. Introduction

It is obvious that more than ever huge amount of data is produced in different fields of study. Studying the relationship, locating specific data group, and retrieving information from this bulk of data is challenging task. As a result of the aforementioned reasons and the wide interest on data mining, a way of grouping similar data into the same cluster and classifying into the same label based on their predetermined class is becoming important more than ever. To address this issue, many techniques have been developed over the years. But, there is long way ahead to get the most out of what clustering and classification can do in data mining (Xu, et al,2005).

Thus, data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data. It also includes analysis and prediction.

One of the main points in data mining is to analyze all the data in hand and understand the business very well. These two key points are among the basic facts that affect the success of the data mining application. Choosing the right data mining technique is another important fact to overcome the business problem. Another important process in data mining which makes the process worth conducting is the interpretation given for patterns discovered. In addition to this, the last step of a data mining application which is the representation of the information for the users is also an important issue that should be considered carefully.

2.2. Data Mining

Data mining refers to extracting or mining hidden knowledge from large amounts of data. Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns (Koh and Tan, 2005).

Data mining is the work of finding the information that enables us to find the relationship among the huge data collection which will be very helpful for making guesses about the future by using a computer program. Data mining is used in many fields such as biomedicine, gene functions, data analysis of DNA arrangement pattern, diagnosis of illnesses, retail data, telecommunication industry, selling, financial analysis and astronomy. All these works can be supported using data mining algorithms.

Data mining algorithm is the mechanism that creates a data mining model. To create a model, an algorithm first analyzes a set of data and looks for specific patterns and trends. The algorithm uses the results of this analysis to define the parameters of the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics (Brefelean, 2007).

The mining model that an algorithm creates can take various forms, including:

- A set of rules that describe how products are grouped together in a transaction.
- A set of classifiers that predicts whether a particular customer will buy a product.
- A mathematical model that forecasts sales.
- A set of clusters that describe how the cases in a dataset are related.

Choosing the best algorithm to use for a specific business task can be a challenge. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result. Different types of Data Mining Algorithms have different functions (SQL Server, 2012). For instance

- Classification algorithms predict one or more discrete variables, based on the

other attributes in the dataset.

- Regression algorithms predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset.
- Segmentation algorithms divide data into groups, or clusters, of items that have similar properties.
- Association algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis
- Sequence analysis algorithms summarize frequent sequences or episodes in data, such as a Web path flow.

2.3. Data Mining and Knowledge Discovery

Data mining and Knowledge discovery in database (KDD) is the rapidly growing interdisciplinary field which merges together database management, statistics, machine learning and related fields in extracting useful knowledge from large collections of data. Knowledge Discovery is a process that seeks new knowledge about an application domain. It consists of many steps, one of them being data mining each aimed at completion of a particular discovery task, and accomplished by the application of a discovery method (Klosgen, 2002).

KDD is the process of extracting novel information and knowledge from large databases. This process consists of many interacting stages performing specific data manipulation and transformation operations with an information flow from one stage onto the next. The process can be very complex and may exhibit much variety in the context of the variety of tasks undertaken within KDD. Generally, KDD can be simply defined as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” ((Matheus et al, 1993; Fayyad, et al, 1996).

Data mining is the search for relationships and global patterns that exist in large databases but are hidden among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database, if the database is a faithful mirror of the real world registered by the database (Fayyad, 1996).

As time passed, the amount of data in many systems, organizations and business agents grew to larger than terabyte size, and could no longer be maintained manually. KDD used to overcome such problems to manage data properly and to become successful in business activities by leading focuses to discovering underlying patterns in data as essential. As a result, several software /tools and techniques have been developed to discover hidden pattern. Thus, the term KDD process refers to the whole process of changing low level data into high level knowledge which is automated discovery of patterns and relationships in large databases while data mining is one of the core steps in the KDD process as illustrated in figure 2.1.

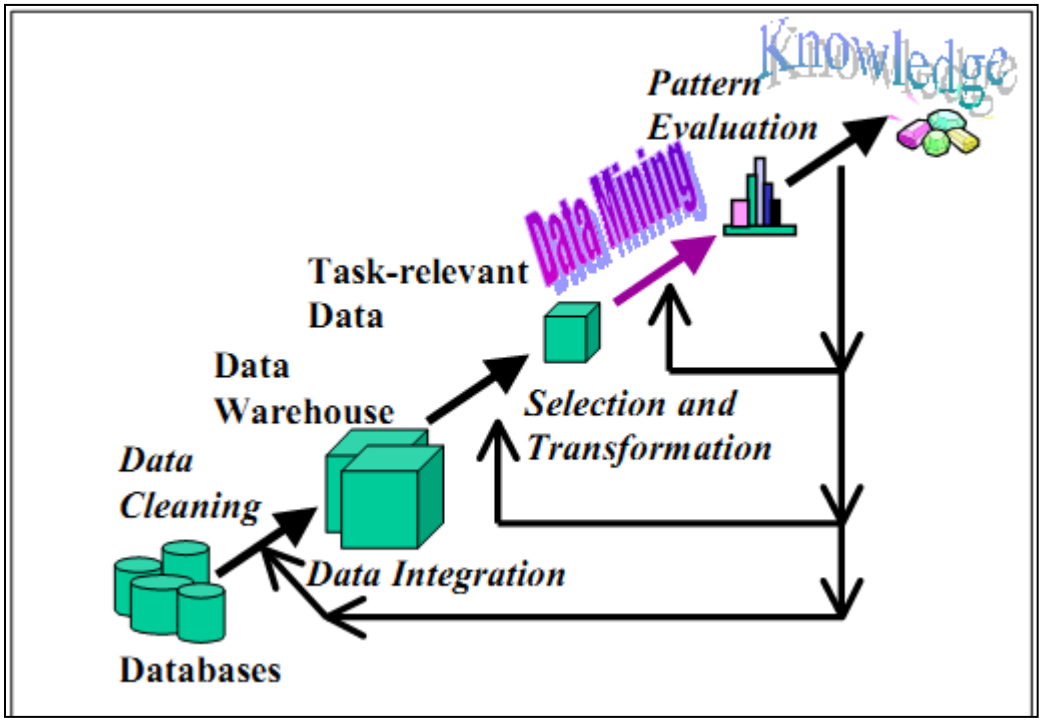

Fig.2.1 Illustration of data mining as core of knowledge discovery process

The KDD process is interactive and highly iterative, user involved, multistep process; comprising a number of phases requiring the user to make several decisions. KDD employs methods from various fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization (Fayyad, et al, 1996). It is said to employ a broader model view than statistics and strives to automate the process of data analysis, including the art of hypothesis generation.

The goal of KDD and DM is to find interesting patterns and/or models that exist in databases but are hidden among the volumes of data. Hence, a new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. This technique and tool is the emerging field of KDD.

The phrase knowledge discovery in databases was introduced during the first KDD workshop in 1989 to stress that knowledge is the ultimate output of a data-driven discovery. KDD was then widely accepted in artificial intelligence and machine-learning fields by modeling of real-world phenomena (Fayyad, et al, 1996).

According to Luo (2008), large databases of digital information are everywhere. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge databases. Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer goods company may yield knowledge of correlations between sales of certain items and certain demographic groupings. This knowledge can be used to introduce new targeted marketing campaigns with predictable financial return relative to unfocused campaigns. Databases are often a dormant potential resource that tapped, can yield substantial benefits (Luo, 2008).

The value of storing volumes of data depends on our ability to extract useful reports, spot interesting events and trends, support decisions and policies based on statistical analysis and inference and exploit the data to achieve business, operational, or scientific goals (Frawley, 1991).

KDD is the process of discovering useful knowledge from a collection of data. It concerns the knowledge discovery process applied to databases (Klosgen, et al, 2002). The KDD process as described by Fayyad, et al, (1996) consists of five major phases, involved in the entire iterative KDD process. Namely the steps are the following and illustrated in figure 2.2:

1. Understanding the domain and Defining problems
2. Collecting and Preprocessing Data
3. Extracting Patterns/Models
4. Interpretation and evaluation of patterns
5. Putting the results in practical use

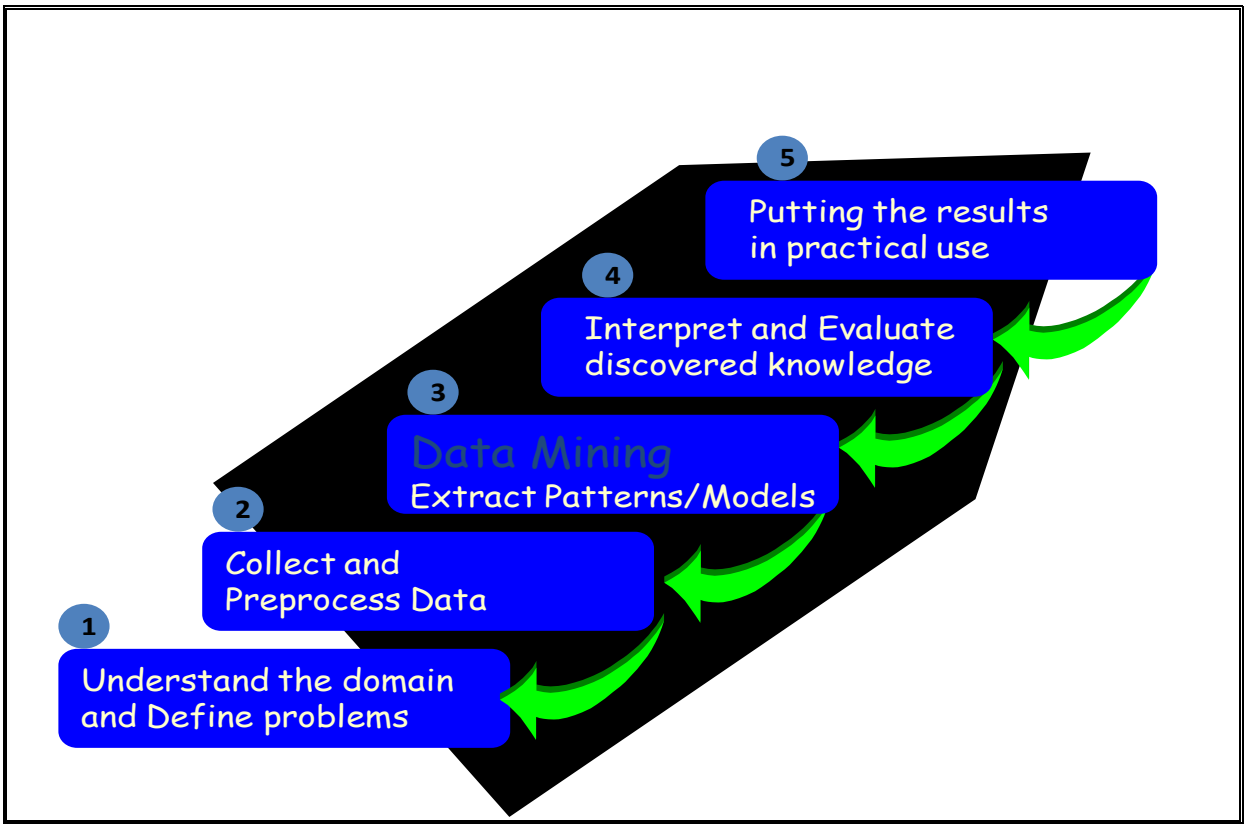

Fig.2.2Thesteps in knowledge discovery process

The KDD process can be classified into sub major phases. i.e. per and post processing to control the activities and task parallel for finding effective and efficient knowledge from unseen patterns using different data mining techniques and tools iteratively. Basic tasks of Pre-processing of KDD process are: learning the domain, creating a datasets, data cleaning, integration and transformation, data reduction and projection and choosing the data mining task. After completing data mining, the visualized results are called post-processing. A possible post-processing methodology includes: finding all potentially interesting patterns according to some criteria; providing flexible methods for iteratively and interactively creating different views of

the discovered patterns, and use of discovered Knowledge to find required patterns (Fayyad, et al, 1996).

2.4. Data Mining Task

The tasks of data mining can be either Predictive or Descriptive in nature (Dunham and Stamatis, 2003). Predictive model data mining tasks include classification, prediction, regression and time series analysis. The Descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis Selection of the modeling techniques (fig.2.3). Modeling is an iterative process - different for supervised and unsupervised learning(Siraj and Abdoulha, 2012).

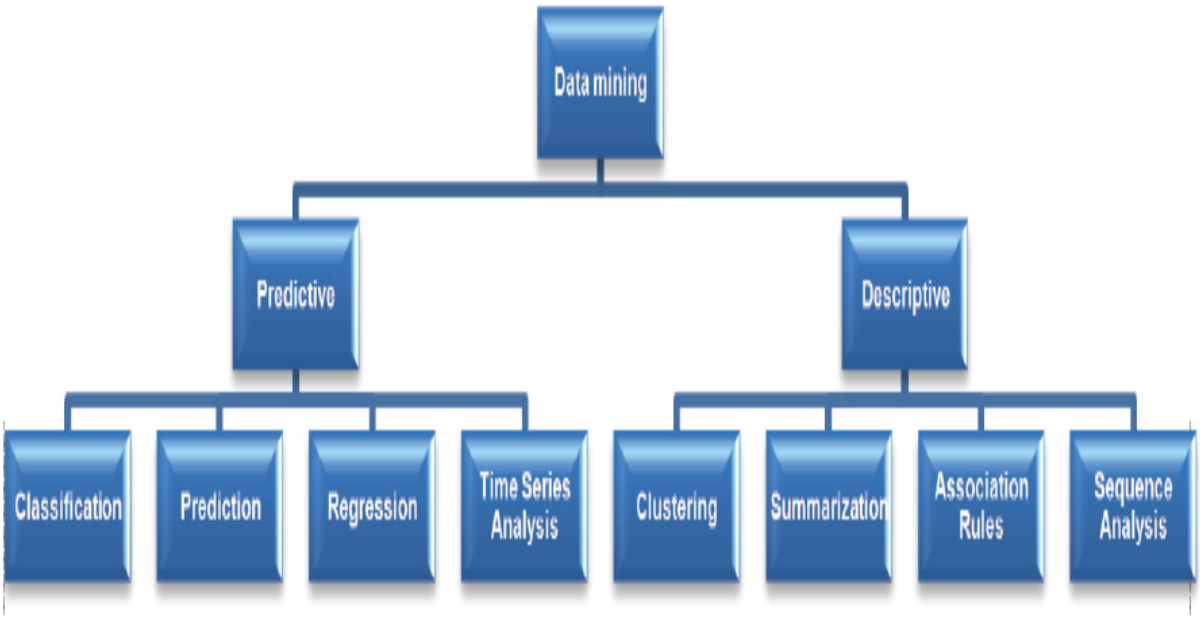

Fig.2.3 Data mining tasks and models

According to Two Crows Corporation, (1999) and Larose, (2005) the tasks can also be classified as Predictive model, Descriptive model, Exploratory Data Analysis and Discovering Patterns and Rules.

2.4.1. Predictive model

This model permits the value of one variable to be predicted from the known values of other variables. Predictive modeling is the event in which a model is made or chosen to accurately predict an outcome. Businesses have grown and science is a lot more complex, therefore data mining requires more computerized method of doing this. Software and computer applications have made data mining seem an effortless job, because it does not require a direct hands-on approach (Dunham, 2006).A Predictive model makes a prediction about values of data using known results found from different data.

Classification and regression represent the largest part of problems to which data mining is applied today, creating model to predict class membership (classification) or value (regression). Classification is used to predict what group a case belongs to. Regression is used to predict a value of a given continuous valued variables based on the values of other variables, assuming a linear or non-linear model of dependency. Logistic regression is used for predicting a binary variable. It is a generation of linear regression; the binary dependent variable cannot be modeled directly by linear regression. Logistic regression is a classification tool when used to predict categorical variables such as whether an individual is likely to purchase or not, and a regression tool when used to predict continuous variables such as the probability that an individual will make a purchase. There are several classification and regression techniques including decision trees, neural networks etc(Dunham, 2006).

2.4.2. Descriptive model

Descriptive data mining models is unsupervised learning functions. These functions do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness, etc. of the data. It presents the main features of the data or a summary of the data. Data randomly generated from a "good" descriptive model will have the same characteristics as the real data(Siraj and Abdoulha, 2012).

Descriptive modeling techniques, such as Summarization, Association Rule, Sequence Analysis and clustering which produces classes (or categories), are not known in advance. Summarization maps data into subsets with associated simple descriptions. Basic statistics such

as Mean, Standard Deviation, Variance, Mode and Median can be used as Summarization approach (Dunham, 2003).Association Rule is a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. The investigation of relationships between items over a period of time is also often referred to as Sequence Analysis. Sequence Analysis is used to determine sequential patterns in data .The patterns in the dataset are based on time sequence of actions, and they are similar to association data, however the relationship is based on time. In Market Basket analysis, the items are to be purchased at the same time, on the other hand, for Sequence Analysis the items are purchased over time in some order (Dunham, 2003).

Clustering is a technique useful for exploring data. It is particularly useful where there are many cases and no obvious natural groupings. Here, clustering data mining algorithms can be used to find whatever natural groupings may exist. Clustering analysis identifies clusters embedded in the data. A cluster is a collection of data objects that are similar in some sense to one another. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words, members of a cluster are more like each other than they are like members of a different cluster. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build predictive models(Berkhin (2012)

Clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are formed. The term “class” is in fact frequently used as synonym to the term “cluster” (Dunham, 2003).

2.5. Data Mining Techniques

Data mining techniques mostly used nowadays are Classification, Clustering, Association and Sequence/Temporal, Regression, Summarization, Dependency modeling, and change and Deviation detection (In-Tech, 2009). Among the techniques used in data mining clustering and classification analysis are two common data mining techniques used for finding hidden patterns

in data (E. Colet, 2000). However they have basic difference. Let's discuss about some of the techniques in next section.

2.5.1. Clustering

There is a tremendous amount of information on the largest shared information source -the web. The task of searching relevant information from the web is very difficult. Search engines index the web and rank documents, however, it is still difficult to find relevant information since they returned large amount of document for a particular query. This fact leads to the need to organize a large set of documents into category through clustering. Grouping similar documents together into clusters will help the users locate relevant information quicker and enhance searching efficiency (Hammouda, 2001). As it is presented in Hammouda, (2001) and Han and Kamber, (2006), the motivation behind clustering is not only the need in information retrieval, it is also useful in exposing inherent structure within group of data, and meaningful relationship among data in various domains.

There is no agreed upon definition of clustering (Steinbach, et al, 2003). But, in this study a cluster is a collection (group) of data objects that are similar whereas the process of grouping a set of physical or abstract objects into classes of similar objects is called clustering (Han and Kamber, 2006). Ali et al. (2000) defined a cluster as it is an ordered list of objects, which have some common characteristics. Generally, the main goal of clustering is as explained in Xu, et al (2005) is to increase the similarity or homogeneity within a group, and reduce the similarity among groups so that new knowledge about a given set of data is acquired. Different literatures use various names like unsupervised learning, numerical taxonomy, vector quantization, and learning by observation which are all equivalent with clustering.

As Jain et al(1999) discussed it; history of clustering is as old as the history of mankind. It has long history in disciplines like biology, psychiatry, psychology, archaeology, geology, geography, and marketing. Clustering techniques initially started to be used in database systems. Since then a lot of models have been developed and there is an increasing interest in the use of clustering techniques in pattern recognition, image processing and information retrieval (Hammouda, 2001). Especially in information retrieval, Cui et al(2005) indicated that fast and

high-quality document clustering algorithms play an important role in effectively navigating, summarizing, and organizing information.

Even if there are several algorithms available in cluster, it is still a challenging task in data mining (Chang and Bai, 2010). According to Jain and Dubes (2005) in (Hammouda, 2001) most of the methods, have the following some common features:

- Many of the clustering algorithms were motivated by a certain problem domain.
- There is no explicit supervision effect and patterns are organized with respect to an optimization criterion.
- They all adopt the notion of similarity or distance.

Generally, as it is shown in Fig.2.4 the clustering process consists of four basic processes: Feature Selection, Clustering Algorithm Design, Clustering Validation and Result Interpretation. These four clustering procedures are aligned with a feedback pathway. This is because clustering is not a one shot process rather repetitive task.

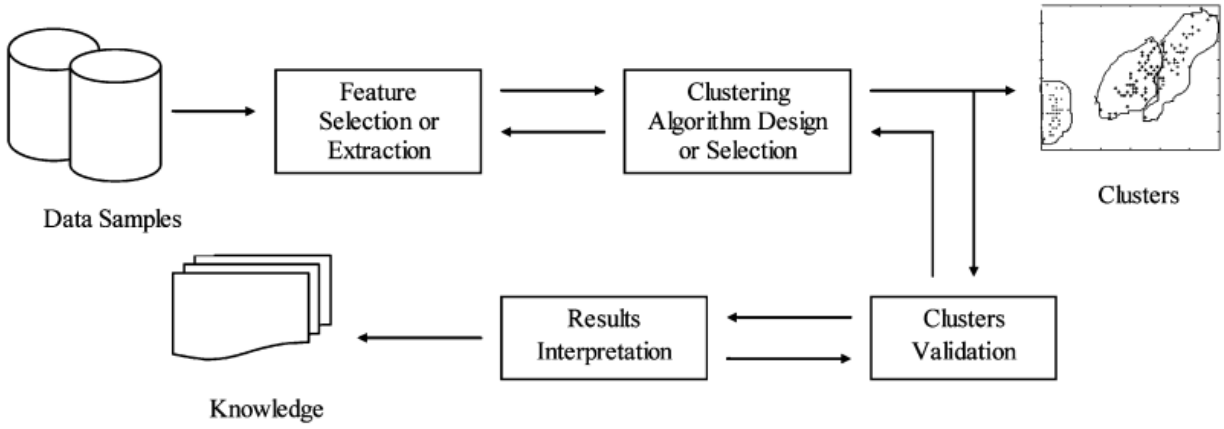

Fig.2.4 The clustering process (Xu, et al, 2005)

Feature selection is the process of selecting useful and novel features from the original dataset which can discriminate among patterns. Feature selection is a critical task in clustering which significantly improve or deteriorate the quality of the clusters generated (Xu, et al, 2005; Jain, et al, 1999).

Clustering algorithm Design or selection is step that incorporates proximity measure selection and appropriate clustering algorithm identification. Since there is no algorithm that can be used for every problem domain, the selection of the algorithm is dependent on many factors. Therefore, the problem at hand, and the number and type of attributes should be taken into consideration while deciding on the algorithm and proximity measure to be used.

Cluster validation-there is no universal clustering algorithm that can be used for every application area. As a result of this, different algorithms give varying number of clusters and cluster structures. Therefore, it is important to asses issues like how many cluster are hidden in the data, whether the final clusters are meaningful or artifact of the algorithms and how to choose among various clustering algorithms available (Xu, et al, 2005). These are the main tasks of this stage of clustering. There are different cluster validation approaches which will be discussed briefly later in this thesis.

Results interpretation- the final goal of clustering is to expose meaningful information about data. Such kind of interpretation of clusters is most of the time conducted by experts of the domain.

In clustering the following are some challenges that should be considered while dealing with it (Xu, et al, 2005). These are:

- There is no general purpose clustering algorithm. As a result of this, the selection among the available clustering techniques has great impact on the final output. There is no universally accepted guideline on how to make this selection.
- New developments in various fields of study are demanding more powerful clustering algorithms which satisfy the aforementioned properties.
- One of the challenging tasks in clustering is feature identification and evaluation of the resulting cluster but, there is no standard approach for these activities.
- In application areas like information retrieval number of documents grow from time to time and it is difficult to rebuild the cluster every time upon arrival of new documents. Therefore, how to incrementally update an existing cluster is another challenge.

2.5.2. Classification

People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines.

Classification is a data mining (machine learning) technique used to predict group membership for data instances (Thair, 2009). It aims at building a model to predict through classifying database records into a number of predefined classes based on certain criteria. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.

Classification method assigns new knowledge to the classes that are known a priori. In classification there is a set of predefined classes and the task is to know which class a new object belongs to. In the context of machine learning, classification is supervised learning*(*Thangavel, 2006).

According to Colet (2000) classification is much less exploratory than clustering. The reason is that the objective of a classifier is not to explore the data to discover interesting segments, rather to decide how new records should be classified. Generally, clustering and classification are often used for purposes of segmenting data records, but they have different objectives and also achieve their segmentations through different ways.

2.5.3. Summarization

Summarization technique is task of discovering common and/or essential characteristics of groups of data in a compact form. Typically, the compact form is summarization rules, and more sophisticated techniques make possible to generate functional dependencies among data items.

Summarization maps data into subsets with associated simple descriptions. Basic statistics such as Mean, Standard Deviation, Variance, Mode and Median can be used as Summarization approach (Dunham, 2003).

2.6. Data Mining Model

Modeling is an iterative process - different for supervised and unsupervised learning and it may be either description or prediction. Selection of the data mining modeling techniques is based upon the objective/task of data mining, problem type of data mining and type of data found in the real world. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. In the business environment, complex data mining projects may require the coordinate efforts of various experts, stakeholders, or departments throughout an entire organization. In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements (BerryandLinoff,2000).

The current process model for data mining provides an overview of the life cycle of a data mining project model that used to find patterns from data. It contains the corresponding phases of a project model, their respective tasks, and relationships between these tasks. At this description level, it is not possible to identify all relationships. There exists a relationship between all data mining tasks depending on goals, background and interest of the user, and most importantly depending on the data (Hegland, 2003). Few of data mining process model will be discussed in the following section.

2.6.1. The CRISP-DM process model

Cross-Industry Standard Process for Data Mining (CRISP-DM) is a standard process model in industries which consisting of a sequence of steps that are usually involved in a data mining study (SPSS,2000). CRISP-DM stands for CRoss Industry Standard Process for Data Mining developed by two vendors; ISL (now part of SPSS) and NCR Corporation which are the world's leading supplier of data warehouse solutions.

CRISP-DM was developed by the means of the efforts of an association initially composed of Daimler Chrysler, SPSS and NCR. The life cycle of a data mining project consists of six

phases. These are business understanding, data understanding, data preparation, modeling, evaluation and deployment (SPSS 2000).

CRISP-DM is the standard model which borrowed ideas from the most important pre-2000 models and is the groundwork for many later proposals.

CRISP-DM is vendor independent so it can be used with any DM tool and it can be applied to solve any DM problems. It defines the phases to be carried out in a DM project, with related tasks and the deliverables for each phase (fig.2.5).

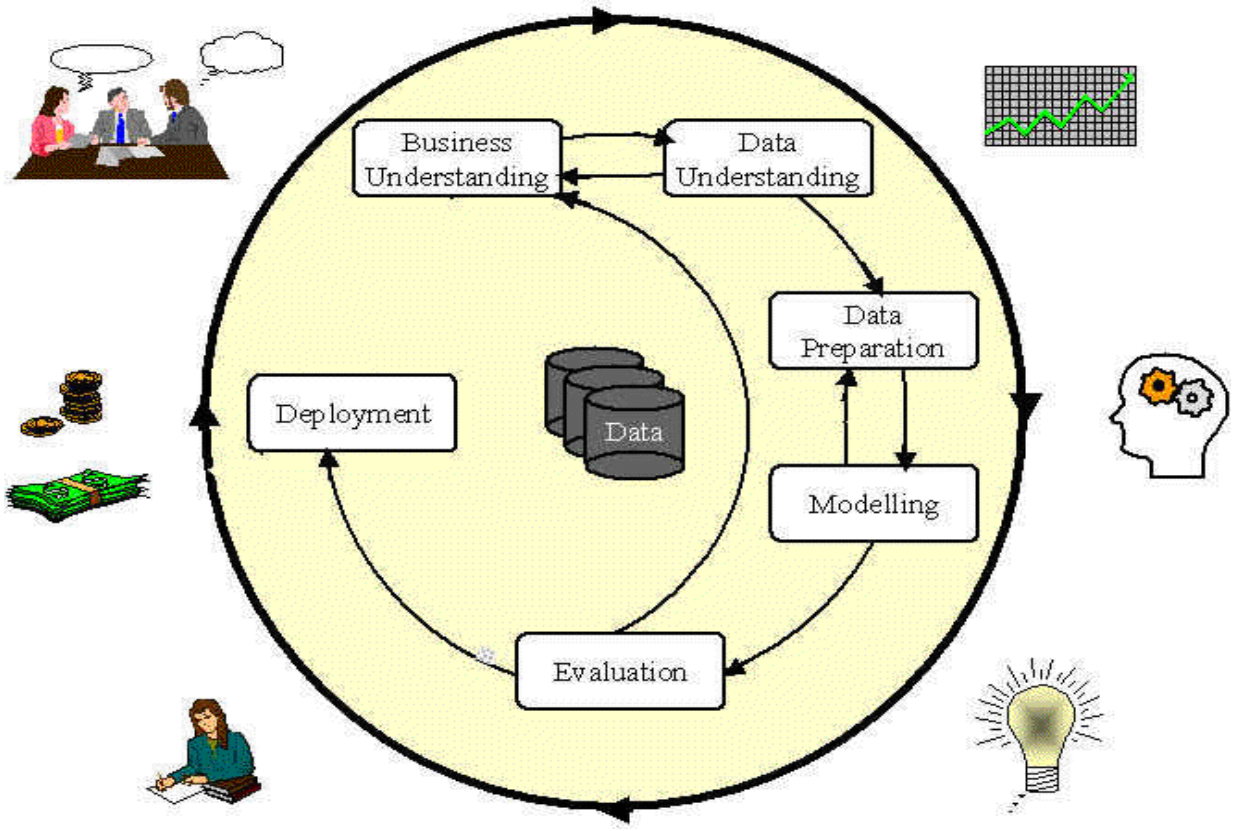

Fig.2.5The**CR**oss-Industry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM)

CRISP-DM model is divided into six phases and related tasks as described in the (Peter et al 1998).

1. **Business understanding**: This phase focuses on understanding the project objectives and requirements from a business perspective. Then, converting this knowledge into a DM

problem definition and a preliminary plan designed to achieve the objectives (SPSS, 2004). Main tasks of this phase are: first, determining business objectives by understanding the client's needs from the business perspective; secondly, assessing situations through investigation of facts about the factors influencing the project; thirdly, determining data mining goals based on the project objectives in technical terms and lastly, producing project plan by preparing a detailed plan to reach the project objectives.

2. **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. Major activities are collecting initial data (initial data collection report), describing data (data description report), exploring data (data exploration report) and verifying data quality (data quality report).
3. **Data preparation:** The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order. Data preparation and cleaning is an often neglected but extremely important step in the data mining process models. Often, the method by which the data gathered was not tightly controlled. Thus, the data may contain outlier values and impossible data combinations. Analyzing data that has not been carefully screened for such problems can produce highly misleading results, particularly in predictive data mining (SPSS, 2004). The main tasks in data preparation are; data selection which is rationale for industry, data cleaning and documenting, constructing data **,** integrating data, data reformatting and have dataset description.
4. **Modeling:** In this phase, various modeling techniques are selected and applied. Their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Here, modeling technique selection, test design generation to validate the model and test it's quality, building model and model assessment through interpretation, evaluation, comparison and ranking of models according to the evaluation criteria from a data mining perspective are main tasks of this phase.
5. **Evaluation:** Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certainto achieve business

objectives. At the end of this phase, a decision should be reached on how to use the DM results through assessment of data mining results with respect to business success criteria to approve models, review of process, and determine next steps such as possible actions and decisions making based on patterns and relationships.

6. **Deployment**: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The main activities in this phase are planning deployment, planning monitoring and maintenance, producing final plan and presenting, reviewing project and producing documentation.

2.7. CUSTOMER RELATIONSHIP MANAGEMENT

Customer Relationship Management (CRM) is an important business approach that focuses on marketing to each customer individually rather than marketing to a mass of people or firms. It is a customer centric – doing business rather than a simple marketing strategy. CRM involves all of the corporate functions like marketing, manufacturing, customer services, field sales, and field service who are required to contact customers directly or indirectly. Its objective is to return to the world of personal marketing to increase profitability, revenue, and customer satisfaction. The approach is made possible by advancing information technology (Gray and Byun, 2001).

In traditional marketing strategies, the main concern was to increase the volume of transactions between seller and buyer, based on the four Ps (price, product, promotion, and place) to increase market share. Performance of marketing strategies was measured through volume of transactions.

Nowadays's, CRM is a business strategy that goes beyond increasing transaction volume. It is primarily a strategic business and process issue rather than a technical issue (Gray and Byun, 2001).

It is an era of company loyalty to the customer in order to obtain customer loyalty to the company. The reason is that consumers are more knowledgeable than ever before and, because the customer is more knowledgeable, companies must be faster, more agile, and more creative than a few years ago(Gray and Byun, 2001).

In recent years, CRM has become widely accepted as an important management discipline. Thus, it is one of the top ten tools used by manager (Iriana and Buttle, 2006). CRM is defined as the business strategy, process, culture and technology that enable organizations to optimize revenue and increase value through a more complete understanding and fulfillment of customer needs.

Buttle (2004) in (Iriana and Buttle, 2006) also defined CRM as "the core business strategy that integrates internal processes and functions, and external networks, to create and deliver value to targeted customers at a profit. It is grounded on high quality customer-related data and enabled by information technology".

Others have also defined CRM emphasizing different aspect of it. Schoder and Madeja (2004) defined CRM as it is a concept for increasing companies' profitability by enabling them to identify and concentrate on their profitable customers.

Griffin, et al, (2000), defined CRM as it is a comprehensive approach which provides seamless integration of every area of business that touches the customer – namely marketing, sales, customer service and field support – through the integration of people, process and technology, taking advantage of the revolutionary impact of the internet.

Organizational culture has an impact on CRM. The reason is that successful CRM performance has been linked to an organization's ability to identify and respond to potential barriers within organizational culture. That means, people's resistance to working with newly created processes and to using the CRM software may lead to implementation failures. Thus, employee behavior and attitude have to be reviewed, and potentially changed, to create a culture that is conducive for the successful implementation of a CRM system (Iriana and Buttle, 2006).

Communicating and valuing changes in organization culture through rewards, ensuring that all employees understand the importance of adopting customer-centric behaviors as they strive to develop stronger customer relationships is important.

CRM is characterized by the application of information technology to the customer-facing functions of an organization; that means it includes selling, marketing and service functions.

The advent of Internet permits firms to establish a personalized customer experience through online help, purchase referrals, quicker turn-around on customer problems, and quicker feedback about customer suggestions, question and so on. Accordingly, the IT department needs extensive infrastructure and resources to implement CRM databases successfully. Executives must be willing to support the CRM implementation process forever because CRM never ends (Gray and Byun, 2001).

Microsoft Dynamics GP (2007), explained the following benefit of CRM:

- Share information - With easy and accurate access to a vast array of information, organizations can provide customers with first-class service, ensuring that the right technician is dispatched with an understanding of the complete history of the customer.
- Increase customer satisfaction - Improve customer service and reduce costs with Web-based tools that enable customers to resolve service issues themselves.
- Make quick, intelligent business decisions - Use standard reports and inquiries to track equipment service details, parts usage, and technician labor. Monitor customer call status, response times, and technician workload. Analyze customer and equipment call history, service contract profitability, and vital warranty issues.
- Give technicians fast access to maps and directions - Help ensure on-time arrival.
- Flexibility - Reverse a contract or credit a customer flat or prorated amount when cancelling a contract.
- Provide visibility across the organization to ensure that the right resource is assigned to the right work order.

2.7.1. Components of Customer Relationship Management

According to Almotairi (2009),there is a general acceptance among researchers of the categorization of CRM components. CRM consists of three major components: Technology, people, and business Process. The contribution to each component varies according to the level of CRM implementation(Almotairi, 2009).

Technology

Technology refers to computing capabilities that allow a company to collect, organize, save, and use data about its customer. Technology is the enabler for CRM systems to achieve their objectives of collecting, classifying, and saving valuable data on customers. Integration technology allows organizations to develop better relationship with customers by providing a wider view of the customer behavior. Thus, organizations are required to integrate IT to improve the capabilities of understanding customer behavior, develop predictive models, build effective communications with customers and respond to those customers with real time and accurate information. For an organization to integrate IT, concepts such as data warehouse, software customization, process automation, help desk and call centers, and internet influence should be addressed as Mendoza et al. (2007) in (Almotairi, 2009).

People

Employees and customers are a key factor for successful CRM projects. CRM is built around customers to manage beneficial relationships through acquiring information on different aspects of customers. The main objective of CRM is to translate the customer information into customized products and services that meet the changing needs of customers in order to gain their loyalty. Nevertheless, a full commitment of the organization's staff and management is essential for an effective CRM implementation to best serve customers and satisfy their needs.

Business process

CRM is a business strategy that has its philosophical basis in relationship marketing. CRM success requires a change of business processes towards customer – centric approach. As such, all business processes that involve both direct and indirect interaction with customers should be analyzed and assessed. Although CRM has an organization-wide impact, process that has direct interaction with customers should be dealt with as a priority when integrating and automating business processes. According to Mendoza et al (2007) in (Almotairi, 2009) the main business processes that should be addressed in CRM implementation are: marketing, sales, and services.

2.7.2. Customer relationship management architecture

CRM architecture addresses the requirements of enhancing/enriching and changing the customer experience by providing the functionality required to effectively interact with the customer, during the Sales and Marketing process (fig.2.6).

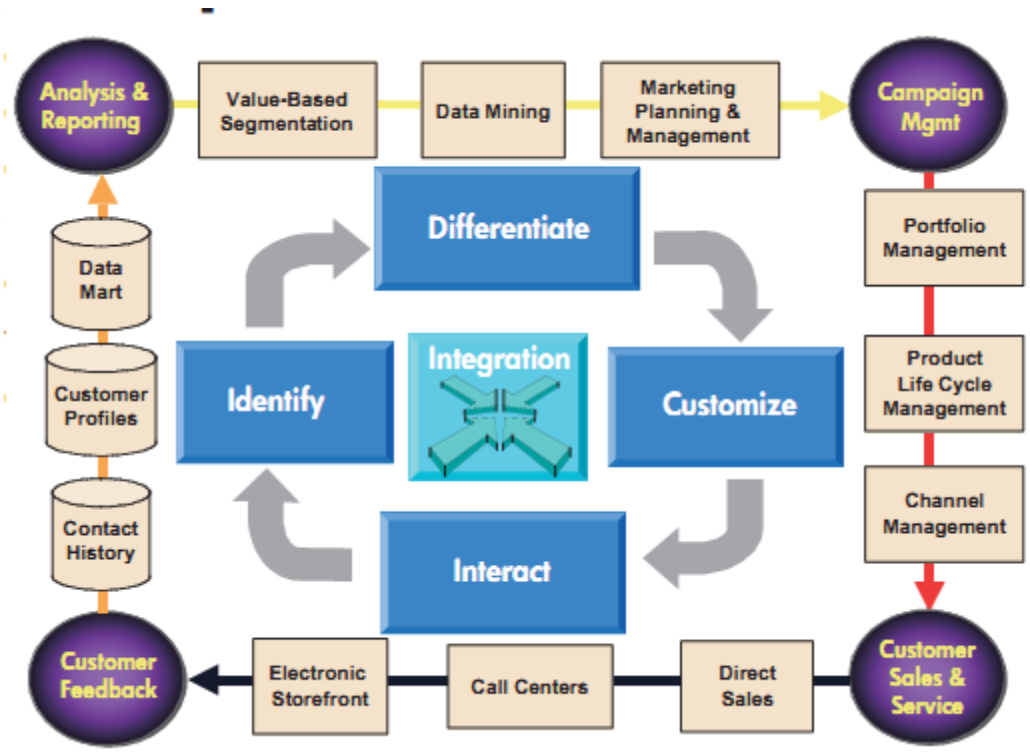

Fig.2.6 CRM Processes and Functions

Hp-invent, (2012),discussed that effective interaction with the Customer requires the following information:

1. Know your customer's needs and pro-actively engage your customer.
2. Know your customer and also his/her conversations/interactions with you (not only to build a better relationship with the customer, but also to serve the customer effectively).
3. Use the knowledge gained during customer interaction to improve the interaction and relationship with the customer.

Hp-invent, (2012),further explained about the core of this architecture is the Decision Engine component, which uses Business Rules (a set of Enterprise Rules that define the "actions" to be

taken and are termed as “Customer recommended actions or CRA”). The Decision engine takes as input, the customer profile, contact history and applies the stored business rules thereby creating a set of one or more recommended actions for the customers.

The business rules management component enables the creation, deletion, analysis and storage of the business rules in a repository. The customer profile component is generated based on the information in the Customer Information File (CIF), Customer Information Warehouse (segmentation, scoring). The relevant customer conversation/contact/dialogue information is captured and stored in contact management component.

The CRA effectiveness analysis component by using data mining technology provides a facility to explore rules, recommended actions and customer interaction effectiveness in general. The treatment provides unique treatment data for customer, products and services.

2.7.5. Principles and tasks of CRM

According to Gray and Byun (2001), the overall processes and applications of CRM are based on the following basic principles.

Treat customer individually

Remember customers and treat them individually. CRM is based on philosophy of personalization. Personalization means the ‘content and services to customer should be designed based on customer preferences and behavior’ according to Hagen, (1999) in (Gray and Byun 2001). Personalization creates convenience to the customer and increases the cost of changing vendors.

Acquire and retain customer loyalty

To a acquiring and retaining customer loyalty through personal relationship once personalization takes place, a company needs to sustain relationships with the customer. Continuous contacts with the customer – especially when designed to meet customer preferences – can create customer loyalty.

Select customer

Select “Good” Customer instead of “Bad” Customer based on Lifetime Value. Find and keep the right customers who generate the most profits. Through differentiation, a company can allocate its limited resources to obtain better returns. The best customers deserve the most customer care; the worst customers should be dropped.

In summary, personalization, loyalty, and lifetime value are the main principles of CRM implementation.

Hp-invent, (2012), discussed about the following principles for building strong customer relationships.

Principle 1: Knowing more about the customer value and anticipating relationship needs better than when the customer was involved in a high-touch relationship.

Principle 2: Consolidate and make available all customer interaction information from all channels/touch points

Principle 3: Develop a customer centric infrastructure that can consistently support the customized treatment of each customer.

Principle 4: Assign dedicated people, process and technology resources to achieve profitable results.

As Seiler and Gray (1999), discussed CRM differs from the previous method of database marketing in that the database marketing technique tried to sell more products to the customer for less cost. The database marketing approach is highly company centric. However, customers were not kept loyal by the discount programs and the one-time promotions that were used in the database-marketing programs. The CRM approach is customer-centric. This approach focuses on the long-term relationship with the customers by providing the customer benefits and values from the customer’s point of view rather than based on what the company wants to sell.

According to Peppers, et al., (2000) in (Gray and Byun, 2001), the four basic tasks that are required to achieve the basic goals of CRM are:

1. **Customer Identification**: To serve or provide value to the customer, the company must know or identify the customer through marketing channels, transactions, and interactions over time.
2. **Customer Differentiation**: Each customer has their own lifetime value from the company's point of view and each customer imposes unique demands and requirements for the company.
3. **Customer Interaction**: Customer demands change over time. From a CRM perspective, the customer's long-term profitability and relationship to the company is important. Therefore, the company needs to learn about the customer continually. Keeping track of customer behavior and needs is an important task of a CRM program.
4. **Customization / Personalization**: "Treat each customer uniquely" is the motto of the entire CRM process. Through the personalization process, the company can increase customer loyalty. The automation of personalization is being made feasible by information technologies.

Traditional (mass) marketing doesn't need to use information technologies extensively.Because there is no need to distinguish, differentiate, interact with, and customize for individual customer needs. Although some argue that IT has a small role in CRM, each of the four key CRM tasks depends heavily on information technologies and systems Peppers, et al., (2000) in (Gray and Byun, 2001).

2.7.6. TECHNOLOGIES USED IN CRM

CRM is a business strategy designed to help an enterprise understand and foresee the needs of its potential and current customers. Customer data is captured in several different areas of the enterprise, stored in a central database, analyzed, and distributed to key points (called touch points). Touch points can include a mobile sales force, inbound and outbound call centers, Web sites, point-of-sale, direct marketing channels, and any other parts of an enterprise that interact with the customer. The distributed data is intended to help foster effective, individual experiences between the company and the customer (Gartner Research, 2000).

In a sense, CRM is a natural and predictable extension of the evolution of marketing and sales. The first CRM-enabling technologies included basic contact management software linked to

individual PCs. This primitive form of Sales Force Automation (SFA) soon grew to include - addition to contact management--account management, opportunity management, mail merge, and forecasting. Client, product, marketing, and competitive information were eventually added to the mix. Other front-office applications, such as sales configuration engines, were added, as well as tight links to back-end ERP. Initially, CRM projects focused on unifying the spheres of sales and customer service, but in the last few years, a marketing function was added, as enterprises recognized both a need to tie marketing campaigns to sales and the significant impact service interactions have on sales lead generation. Now, CRM projects strive to provide data to every enterprise department that touches the customer (Gartner Research, 2000).

CRM is designed to empower the entire enterprise when managing customer relationships. Enterprises want their customers to see one, friendly, corporate face, as opposed to a collection of disconnected departments trying to work together. Ideally, an effective CRM strategy will enable the enterprise to utilize all of its resources when interfacing with a customer, including marketing, sales, finance, and manufacturing, as well as post-sales services. When carefully and strategically employed, econometric, demographic, lifestyle, and psychographic data; decision-support systems; the Internet; and customer access techniques and technologies can help promote effective CRM, despite the size of enterprise, the size of enterprise's customer base, or the size of relative market. The ability to gain value from CRM projects is contingent on the enterprise's capability to leverage and integrate all of these functions, technologies, and consolidated data in a way that promotes departmental synergy, as well as competitive advantage (Gartner Research, 2000).

2.7.7. DATA MINING TOOLS IN CRM

The process of data mining helps firms to analyze the customer data and extract the useful information, to gain competitive advantage over others. Hui and Jha, (2003) in (Jayanthi et al, 2008) described the database of customer service as a repository of valuable information and knowledge that can be utilized to improve customer service. Chen and Liu,(2002) in (Jayanthi et al, 2008) studied applications of data mining in bioinformatics, information retrieval, adaptive hypermedia and electronic commerce, which require interaction with the customer. Chang et al, (2002) discussed the importance of customer relationship management in enchasing the ability of

a firm to compete and retain key customers. CRM as described by Kwok et al, (2007) is Strategic Customer Relationship Management System (SCRMS) which collects, integrates and diagnoses various customer-related data from different operation systems in departments within an enterprise. According to Jayanthi et al, (2008) data mining tools helps CRM by providing the complete framework, which covers:

- To analyze the business problem.
- To prepare the data requirements.
- To build the suitable model with respect to business problem.
- To validate and evaluate the designed model.

The analytical engine of the data mining tool helps to discover the hidden patterns that help the firms in decision making. Earlier data mining tools used to focus on analytical problems that are on discovering the hidden patterns. Slowly the focus turned to other concern of the data mining like the data preparations, model building and evaluation of models. The data preparation is vital for the success of CRM as the data for the tools comes from various sources. So the missing data, outlier and other necessity work is carried out in data preparation phase. To standardize the process of data mining the CRISP-DM model is proposed which ensure that the standard required for the data mining is maintained.

Model building is the next phase of the Data mining tool, which builds the various models according to the data given in the data preparation phase. The last phase is the evaluation of the model, so that the proper results in the form of useful patterns can be drawn from the models built by the tools.

The tools of data mining for CRM should be able to detect the necessary information from the available data .To achieve this, Data mining tools should have some characteristic and it is discussed below by Jayanthi et al, (2008):

- User friendly environment
- Efficiency of the tool
- Basic task should be accomplished
- Low cost of implementation

Competitive firms with a future vision uses data mining to reduce fraud anticipate resource demand and curb customer attrition, according to CRM Today (2003). As pointed by Spangler et al(1999),availability of detailed customer data and advances in technology for warehousing and mining enable firms to better understand and serve their customers. This raises the concern over the issue of privacy of customer data, but the early adaptation of information technology in their business has gained them competitive advantage over others while the rest were resisting adapting it.

Organization should incorporate the best data mining tools to remove the short comings in their companies. An in-depth knowledge of your customer's information is essential for competitive advantage. Successful implementation of data mining tool in the organization can improve the relationship of the customer with the company, which is the demand of the present business environment. The companies will be able to analyze the customer data and understand their customer effectively and efficiently.

There are numerous data mining tools available in the market for effective CRM, but the right tools for the firms depend upon the methodology used in the firm and the goals that the firms need to achieve through the implementation of the tools. Non-effective implementation of tools of data mining has some limitations for the firms (Ranjan, 2008).

Generally, data mining tools can provide CRM with better understanding of customer relations and improved customer satisfaction, higher profitability for the company and higher probability of attaining competitive advantage. This creates an atmosphere in the companies which helps the executives to take better decision towards the success of the firms.

2.7.8. Application of CRM in microfinance industry

Microfinance industry as a whole is challenged by the need to reach out to the poor and sometimes being financially self-insufficient. Although the industry as a whole is growing at a faster pace, still the two critical questions of reaching the poor and building a financially sustainable microfinance industry that walk on their own leg freely are empirical questions(Letenah,2009).

Microfinance is the provision of financial serves to the poor people with very small business or business projects Marzys, (2005) in (Letenah, 2009). Only a small fraction of the world population has access to financial instruments, essentially because commercial banks consider the poor people as un-bankable due to their lack of collateral and information asymmetries.

There are a number of studies in the microfinance industry. The reason is that it has got the attention of academicians and practitioners, as an innovative method of fighting poverty. Thus, it is better to see some business drivers for CRM application in the industry.

According to Gray and Byun, (2001), competition for customers is intense. From a purely economic point of view, firms learned that it is less costly to retain a customer than to find a new one. The quoted statistics go something like this:

> "*…… An increase in retaining existing customers more increase in profitability. In the past, the prime approach to attracting new customers was through media and mail advertising about what the firm has to offer. This advertising approach is scattershot, reaching many people including current customers and people who would never become customers. For example, the typical response rate from a general mailing is about 2%. Thus, mailing a million copies of an advertisement on average yields only 20,000 responses.*
>
> *Another driver is the change introduced by electronic commerce. Rather than the customer dealing with a salesperson either in a brick and mortar location or on the phone, in electronic commerce the customer remains in front of their computer at home or in the office. Thus, firms do not have the luxury of someone with sales skills to convince the customer. Whereas normally it takes effort for the customer to move to a competitor's physical location or dial another 1-800 number, in electronic commerce firms face an environment in which competitors are only a few clicks away.* "

Here, the major cost goals of CRM are increasing revenue growth through customer satisfaction and minimizing customer support costs.

CHAPTER THREE

DATA MINING METHOD FOR MARKET CLUSTERING AND CLASSIFICATION

3.1. INTRODUCTION

Data mining techniques include clustering, classification, association rules, and regression. Most of the time clustering and classification techniques are used in customer segmentation and prediction.

A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. The clustering algorithm also finds the centroid of group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster (Dunham, 2003).

Classification is also a very important and frequently used technique in data mining. It is the process of finding a set of models that express and distinguish data classes or concept into pre-defined class.

Here, since there are no predefined classes that describe the customers of the institution, the researcher uses clustering techniques that resulted in the appropriate number of clusters. Techniques used in this study to perform data mining tasks are discussed below.

3.2. CLUSTERING TECHNIQUES

Clustering technique is unsupervised learning. It is grouping a given collection of unlabeled patterns into meaningful clusters. In this technique the modeling process is unsupervised that is no prior (pre-defined) knowledge is available to exactly guide the clustering process. The clustering algorithm clusters the database autonomously.

Clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, web analysis, CRM, marketing, medical diagnostics, computational biology and many others.

3.2.1. K-means algorithm

The K-means algorithm, one of the clustering algorithms proposed for this study, is based on a very simple idea. Given a set of initial clusters K (k-stands for numbers of clusters), assign each point to one of them and then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point (Rao, 2003). In K-Means, the centroids are computed as the arithmetic mean of the cluster all points of a cluster. The distances are computed according to a given distance measure, e.g. Euclidean distance.

Although K-means has the great advantage of being easy to implement, it has two big drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success (Rao, 2003).

3.2.2. Interpretation

After the clusters have been created the result should be interpreted. According to Berry and Linoff (2000), the three commonly used approaches to understand clusters or class are:

1. Examining the differences in the distributions of variables from cluster to cluster, one variable at a time.
2. Using visualization to see how the clusters are influenced by changes in the input variables.
3. Building a decision tree with the customer label as the target variable and using it to generate rules explaining how to assign new records to the correct cluster.

Using the above three approaches interpretation of the output was performed to have better understanding on customers clustering.

3.2.3. Cluster result validity

Cluster validity is a broad and a subject of endless argument since the notion of “good” clustering is strictly related to the domain applications and its specific requirements (Halkidi and Vazirgiannis, 2001). Different clusters are obtained with different parameter values from a given

database and clustering algorithm. So, there is a need to decide the best clustering that fits the dataset and the business under consideration.

Numerical measures that are applied to judge various aspects of cluster validity are classified into three types. The first type is External Index. It is used to measure the extent to which cluster labels match externally supplied class labels. Entropy is one of external index measures. The second type is Internal Index. It is used to measure the goodness of a clustering structure without respect to external information. Here Sum of Squared Error (SSE) is one of the measures. The third is Relative Index. It is used to compare two different clustering or clusters. Often an external or internal index is used for this function. Sometimes these are referred to as criteria instead of indices. However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion (Kumar, 2005).

Nevertheless, the two generally accepted measures of cluster result validity are separation among the clusters and cohesion with clusters. And also there are two aspects that should be considered in checking the validity of clustering result with regard to dataset. These are the choice of the appropriate input parameter values for clustering algorithm resulting in the optimal partitioning (Kumar, 2005).

Cohesion and Separation are internal measures. Cluster cohesion measures how objects closely related in a cluster like in case of SSE while cluster separation measure how distinct or well-separated a cluster is from other clusters like in case of Squared Error. For cohesion and separation a proximity graph based approach can also be used. Thus, cluster cohesion is the sum of the weight of all links within a cluster. And also cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster (Kumar, 2005).

In this research work from Internal Index measures the researcher used Sum of Squared Error measure to judge cluster validity.

3.3. Classification Techniques

Classification is a process of finding a set of models or pre-defined conditions that describes and distinguished data classes or concepts. It is supervised learning method. In supervised learning,

we are provided with a collection of labeled pattern and the problem is to label a newly encountered, yet unlabeled pattern. The given labeled patterns are used to learn the descriptions of classes which in turn are used to label (classify) a new coming pattern. Classification technique maps data into predefined groups. The derived model of classification may be represented in various forms such as "If-then" rule, decision tree, neural networking, Bayesian networks etc.

Here for the purpose of this study, the researcher used decision tree. This is due to:-

- ✓ Relatively faster learning speed than other classification methods
- ✓ Convertible to simple and easy to understand classification if-then-else rules
- ✓ Comparable classification accuracy with other methods.
- ✓ Does not require any prior knowledge of data distribution, works well on noisy data.

3.3.1. Decision tree classification techniques

Decision tree is a predictive modeling technique used in classification and prediction tasks. It uses a divide and conquers techniques to split the problem search space into subsets. Decision tree is a classifier expressed as a recursive partition of the instance space. It is used in data mining to classify objects into values of the dependent variable based on the values of independent variables. There are two main types of decision trees. These are classification trees and regression trees. Classification trees are decision trees used to predict categorical variables, because they place instances in categories or classes. And the second one is regression trees, which is a decision trees used to predict continuous variables (variables which are not nominal). Classification trees can provide the confidence to correctly classify the data. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class. On the other hand, regression trees estimate the value of a target variable that takes on numeric value (Loh, 2011).

The structure of decision tree is a tree like structure, where each internal node represents a test on an attribute, each branch characterizes an outcome of the test, and leaf nodes at the end represent classes in which the data is assigned. The top most nodes in a tree are the root node. Depending on the algorithm, each node may have two or more branches. For example, CART (Classification

and Regression Trees) generates trees with only two branches at each node. Such a tree is called a binary tree. If there are more than two branches at each node, the tree is called a multi-way tree. (Loh, 2011).

3.3.2. Decision tree construction

The basic algorithm for decision tree training is a greedy algorithm that constructs decision trees in a top-down recursive divide and conquer manner. The algorithm enables to select an attribute from the rest of attributes with a strategy of searching a local optimum solution (at each node) that leads to a global optimum solution (Loh, 2011; Liu, et. al, 2012).

There are different basic methods of attribute subset selection which includes the following techniques, where the stopping criteria for those different techniques may vary.

The first one is stepwise forward selection. In this technique the process starts with an empty set of attributes then the best attributes is determined and added to the set. At each succeeding iteration or step, the best of the remaining attribute is added to the set. The second is stepwise backward elimination. The process starts with the full set of attributes. At each step, it removes the most terrible attribute remaining in the set. The third is combination of forward selection and backward elimination. The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes poor attributes from among the remaining attributes (Tuv, et. al, 2009).

The “best”/””worst” attributes are typically determined using tests of statistical significance, which assume that the attributes are independent of one another. Many other attribute evaluation methods can also be used. Information gain measure is one of the methods used in building decision trees for classification (Liu, et. al, 2012).

In decision trees construction a greedy algorithm is used while the researcher used information gain measure for attribute selection and evaluation.

3.4. Criteria for selecting and evaluating of data mining software

Data mining has emerged as a technology for competitive advantage for business organizations and these business organizations are incorporating this technology into their business practices by using data mining software or tool. However, different data mining software are selected and implemented wrongly and this selection and evaluation of the wrong tool is expensive both in terms of money and time. So, there is a need to have a framework for evaluating data mining software/tools and to select the best data mining software.

Collier et al (1999) presented a framework consisting of four categories of criteria for evaluating and selecting data mining tools or software and these criteria have been employed in this research work.

- ✓ Performance – the ability to handle a variety of data sources in an efficient manner. This category of criteria is focuses on the qualitative aspects of a tool's ability to easily handle data under a variety of circumstances rather than on performance variables that are driven by hardware configurations and/or inherent algorithmic characteristics. The performance criterion includes platform variety; software architecture, heterogeneous data access, data size, efficiency, interoperability and robustness of software.
- ✓ Functionality – the inclusion of variety of capabilities, technique and methodologies for data mining. The software functionality helps to assess how well the tool will adapt to different data mining problem domains. The functionality criteria includes algorithmic variety agreed methodology, model validation and data type flexibility, algorithm modifiability, data sampling, reporting and model exporting.
- ✓ The usability and applicability by different levels and types of users without loss of functionality or usefulness. A tool should help guide the user toward proper data mining function since KDD is a highly iterative process. The usability criterion including user interface, learning curve, user types, data visualization error reporting, action history and domain variety.
- ✓ Ancillary task support – this criteria category tells the capability of the tool or software to allow the users to perform the variety of data cleansing, manipulation, transformation visualization and other tasks that support data mining. These tasks includes data

selection, cleansing, enrichment, value substitution, data filtering, binning of continuous data, generating derived variables, randomizing, deleting records, etc. since it is rare that a data set is truly clean and ready for mining, the software should be able to support data selection, cleansing, filtering the data for the model building in the KDD process.

CHAPTER FOUR

EXPERIMENTATION

4.1. INTRODUCTION

This chapter describes the data mining goals, sources of data and techniques used in preprocessing and model building phases. It also deals with the description of the data mining process undertaken based on the CRISP-DM approach. All the data mining process of this research has been done in line with the following CRISP-DM process models. The CRISP-DM method is described in terms of a hierarchical process model, consisting of sets of tasks described at six levels: business understanding, data understanding, data preparation, modeling, evaluation and deployment.

4.2. Business Understanding

To understand BG MFI business and coin data mining problems- researcher observe the Institutions at its different branches. Interviews and discussion was also made with senior managers of the organization. Domain expert consultation had made to have brief understanding on the problem area. Besides, the dataset is also thoroughly examined with domain experts.

From business understanding perspective, currently, there is a very poor traditional means of knowing what is unique features and need of their customers and how to use this knowledge to make better future decisions by facilitating the ordinary process of market segmentation that involves identification markets.

Through interviews and discussion made with senior managers of the organization, the institution is attempting to attract and retain more customers while expanding the market to outreach potential customers. However, the institution is not using appropriate tools and technique that helps them to identify their customers in the market. So, the institution could not pass proper decisions to satisfy their customers in providing loan services. As the officials say, the current system, do not helps them to identify as well as to predict high value and profitable customers (segments). Thus, some branches are not generating the expected profit while few are closed due to lost revenue.

Therefore, the researcher proposed a data mining technique that helps to segment and predicts profitable market, so that the institution can pass proper decisions on provision of loan services. This in turn has a significant impact in improving customer relationship management of the institution.

After understanding the business, the next step was selecting appropriate tool. The basic aim of using data mining tool is to discover hidden knowledge from a large database. So, for the smooth course of action, selecting appropriate data mining tool is indispensable (Kurgan and Musilek, 2006).As a result, to get useful knowledge, convenient data mining tool was selected. The researcher used an open source data mining tool, WEKA, which is developed by the University of Waikato in New Zealand (Shigeki, 2006). For this research *Weka 3.7.5* is selected for rules mining and MS Excel is employed for preprocessing the dataset. WEKA is selected since researchers are familiar with this tool.

4.3. Data Understanding

Data understanding and data preparation tasks should performed carefully to come up with good output in data mining process. The reason is that the models that will be built mainly depend on these tasks. Hence the next section discuss about the activities performed to understand data.

Initial dataset is collected from BG MFI. This data is organized in the format of Microsoft office Excel 2007 for further processing. The data has 19 attributes (Branch Name, Input Date, Group Name, Group Code, Certification Date, Group Address, Zone, Woreda, City, Kebele, Sex, Age, Marital Status, Location Category, Education Years, Education Level, Loan Amount, Loan Number, and Loan Cycle) and were total of 13390 records.

Describing the collected data is another activity that has been conducted to be familiar with the data that had been used for the research as described above. This has been achieved through careful study of the data. This means that the attributes have been analyzed individually to know what kind of value they accept (nominal, categorical, etc) and how missing data (not available data) are handled in the data. As a result of this, the description of the dataset has been prepared. In addition, attribute relevance analysis has been performed on data so as to remove some irrelevant and least important attributes based on domain experts' recommendation.

The description of the data features is shown in Table 4.1 below.

Attribute Names	Modified Attribute Names (if any)	Description	Selected
Branch Name	Branch	Institution's branches	
Input Date	Input Years	Date of new customer applied for loan	
Group Name	Group Name	Name given to the group	
Group Code	Group Code	Id No of the group	
Certification Date	Certification Date	Date of acceptance	
Group Address	Group Address	Specific place where group members live	
Zone	Zone	Zone	
Woreda	Woreda	Woreda	
City	City	City	
Kebele	Kebele	Kebele	
Sex	Sex	Sex	✓
Age	Age	Age	✓
Marital Status	Marital Status	Marital Status	✓
Location Category	Location Category	Where the customers categorized into Urban or Rural	✓
Education Years	Education Years	Number of years they attend formal education	
Education Level	Education Level	The level at customer categorized into primary, secondary education etc	✓
Loan Amount	Loan Amount	Loan size borrowed	✓
Loan Number	Loan Number	Loan code	
Loan Cycle	Loan Cycle	Loan repetition taken	

Table 4.1Attributes and their description

4.4. Data Preparation

Han and Kamber (2006) stated, preprocessing helps 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, etc need to be performed. Moreover, to conduct the experimentation, the dataset must be prepared in the appropriate format.

The first step toward accomplishing the research is gathering data, refine it and preparing for data mining process. The dataset which has been used for administrative and daily transaction purpose cannot directly be used in data mining process. It needs further preprocessing tasks. Here, data cleansing and data preparation are steps performed in this research work.

Based on the objective of the research, from total of 13390 records 13057 clean data are selected for modeling. The reason is that the data is full of outliers, noise and contain many missing values. From selected dataset, missing data are removed and some changes of attribute names are made. Some less important attributes are also removed from the selected dataset.

4.4.1. Data Selection

In every data mining task, one can get some attribute that has little or no impact on the overall mining output. As mentioned above there are many attributes in the data that have been used for this research work. After getting the description of the features, the domain experts were consulted for the selection of appropriate attributes that may help in discovering some patterns about the data as the objective of the research. In fact, one may know what pattern may be discovered. So, some intuitive decision is used to select some attributes. Hence, those attributes that have tick mark (✓) against them in Table 4.1, are selected for the research work. Then, by running small size sample data, other attributes are excluded automatically based on the recommendations of the domain experts and objective of the research.

Hence, Input Years, ID No, Group Name, Group Code, Certification Date, Group Address, City, Kebele, Education Years, and Loan Number have been removed from the dataset because they are not that much significant for the final result. In addition, attribute name changes and data

reshuffling had been performed to have the required dataset. A total of 6attribute(for clustering purpose) and 13057 records have taken for analysis.

All the selected attributes are the target attributes which are used for clustering and classification purpose. These are: Sex, Age, Marital Status, Location Category, Education Level, and Loan Amount.

4.4.2. Data Cleaning

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. In particular, data collected from different sources has to be managed into a form that will allow the data mining tools to be used to best advantage. This process of data cleaning and pre-processing is highly dependent on the technique to be employed. In this research since the collected dataset had some inconsistent values, such data were removed. Likely, some data contained missing value and they were also removed from the original dataset. Hence data cleaning had been required and performed on the selected dataset.

To do this the following activities were done:

- Even if it is possible to preprocess the data in Weka, the researcher has found easy to use MS Excel features for preprocessing; i.e. data cleaning is conducted while the data is in MS Excel format. To view some outlier value and /or very little values, the Data Filter feature of MS Excel is used.

  Then, attributes that have so many missing values can easily be detected and removed. Values that are not compatible are modified. For instance, Weka doesn't handle values with space unless they are in single quotation. Even if it is difficult to put all such values in single quotation, the researcher preferred to put them as one token by removing the spaces such as EducationalLevel, LoanAmount, etc.

- In order to make the dataset convenient for the data mining tool, the dataset is prepared in ARFF (Attribute Relation File Format) file format. In doing so, declarations, such as

@relation, @ attribute <attribute names> together with the data type and *@data* are added in the CSV file.

- Then, after going through the data cleaning, the data is saved as CSV file format in which the values are saved in comma delimited form in order to create an ARFF format file.
- The data that was converted into ARFF file format has been used for the experimentation.

4.4.3. Data transformation and aggregation

Data transformation and aggregation help to minimize the variation of the attribute values in some of the fields. And also help to make results more meaningful and easy.

The varying educational levels were transformed into more aggregated values by consultation with the domain expert. This is to format the data so that it can be analyzed easily by any users as well as for the software to easily interpret the output of the result. Customer with no formal education categorized as None, educational level with grade 1-4 as Primary, 5-8 as Junior, 9-10 as secondary, and above grade10 categorized as Higher.

4.5. MODELING

Modeling phase of data mining is the process of providing the processed data to the selected clustering and classification algorithm and selecting model that shows better performance. There are a number of tasks involved in the phase. Some of the tasks include selection of modeling technique, generating test design, building model and selection of the best model.

4.5.1. Selection of modeling techniques

A good segmentation model can divide customer into homogeneous group on the basis of shared common attribute. Clustering technique is a direct data mining technique (i.e. no predefined classes to be predicted)and means the instances are to be divided into natural groups. After clusters are identified, new customers should be classified to these cluster indexes. In this research, automatic cluster detection (K-means clustering algorithm), and decision tree classifier techniques are the selected data mining modeling techniques for customer segmentation and classification respectively. The selection of the techniques was made because of widely applicability of the techniques for segmentation / classification.

4.5.2. Testing design

A plan should be first set to guide the training, testing and evaluation process of the model. Mostly researchers split the dataset into training and test sets. Normally, training should be done on large proportion of the total data available, whereas testing is done on small percentage of the data that has been excluded during training of the model. In this research, all the sample size or available dataset has been used for training and testing. In the case of decision tree classification models, different experiments have been done by splitting the dataset into training and testing set and by adjusting the default parameter values. Finally, the classification model that shows better accuracy performance has been selected.

In this research, the analysis and interpretation of each and every cluster was made by the researcher and the domain experts. The domain experts in the institution involved in the process of segmenting and interpreting the segmented results.

4.5.3. Cluster Model building

The cluster model building process consists of three activities namely: attribute selection, clustering and selection of best clustering model. This is followed by building different decision tree classifier models and selecting the one that shows better accuracy are part of the process.

4.5.3.1. Attribute selection

The attribute selection activities involved selection and identification of best variables or attributes for segmenting and identifying customer from business perspective in which the institutions involved and develop clustering /classification model in microfinance. Clustering is formation of clusters based on the selected attributes. Moreover, interpretation and analysis of cluster results were also made in this selection (clustering). Next to this, there is selection of the clustering model that shows better clustering performance based on the evaluation criteria. The final selection is building and choosing the best classifier model using J48 decision tree algorithm.

4.5.3.2. Clustering of customers

This step involves clustering the dataset based on selected attributes and this clustering model should enable to develop the decision tree classifier models. The clustering result interpretation is also performed by researcher and domain expert.

The challenging task in the K-means clustering algorithm was determining the value of K, which finds out the optimal clustering model that creates dissimilar segments of customers according to organizations need. Most of the time, the number of customer segment (K) can range from 4 to 10. In addition, it depend on the organizations capacity to manage various clustering properly. However the optimal number of cluster size (the value of K) is obtained through trial of different experiments by adjusting default parameters.

In this research, after prototype experiments conducted and discussion with the domain experts in institution, the value of k was set to be three to five. Hence, different clustering models were built at three different values of K (two, three and four) as well as different seed size using WEKA version 3.7.5. Finally, the best clustering model was selected based on specified criteria that match with the business objective of institutions.

4.5.3.3. Assessment of clustering customer models

There is no actual defined good clustering output. Hence, assessing the clusters based on certain crucial attribute is reasonable. In addition, it was important to discuss with expert at institution. The discussion was focused on assessing the most influential attributes from selected ones for clustering customers in the institution. The expert discussed and stated the most important variables that are used to select or predict preferable customers for the Institutions.

According to the domain expert, loan amount was given a high weight from selected attributes based on the objective of the institutions. Here, the main objective of the institution is lending money/loan provision and retaining customer to maximize its profit. This objective has a directly relationship with loan amount and loan cycle. That means, as loan cycle increases, the retention rate also increases (i.e. retention rate has a direct relationship with CRM and indirect relationship with profit of the institution). And also as loan amount increases, the institution become more profitable (i.e. loan amount has direct relationship with profit of institution). As a result, the

analysis and interpretation of each and every cluster in the experimentation was highly dependent on these attributes. However, it doesn't mean that other attributes have no importance, rather to express the weight given to this attributes by the domain expert in the institutions to cluster customers.

To observe different changes in the distribution of the segments, the researcher compares the nine (where K= 3, 4and 5 with randomly selected seed size 10,100, 1000) clustering model.

The comparison of clustering model was done in a way that the attribute value of each cluster in model are compared to other clustering models with number of iterations, inter-class similarity error, the objective of institutions and the domain expert's judgment.

Experiment One

The first experiment was conducted using simple K-means to build cluster model. 13057 instances and 6 attributes were used in the experiment. Here to have natural segmentation (correctly clustered instances) of customers, classes to clusters evaluation mode was selected from four cluster mode of WEKA's tool. This was important to know measure of incorrectly clustered instances. In turn, this helps to choose better model from the others. Beside this, number of iterations, inter-class similarity error, and the objective of institution were examined through experiment. The output of the experiment one depicted in the following Table 4.2.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	3594 (28%)	Female	30.55	Married	Urban	Junior	1114.19
2	1703 (13%)	Male	40.38	Married	Rural	Primary	1137.71
3	7760 (59%)	Female	35.69	Married	Rural	None	974.96

Table 4.2 Cluster description based on values of attributes for K=3 and seed size 10

As indicated in the above Table 4.2, cluster one contains attribute values with frequency of 3594(28%) records, married female average aged 30, living in urban, completed junior school, capable of borrowing an average of 1114.19 loan amounts. In the second cluster, with frequency

of 1703(13%) records, male who are married, an average age of 40, living in rural, completed primary school, capable of borrowing an average of1137.71loan size are categorized. In the third cluster, with frequency of 7760 (59%) records, female who are married, an average age of 35, living in rural, with no formal education, capable of borrowing an average of 974.96 loan size are categorized.

In the experiment Sum of squared errors is 10770.78.This measurement was used to compare the goodness of this experiment with others. Beside this, based on the average loan size shown in Table 4.2, cluster one categorized into high preferred customer segmentation, cluster two is categorized into very high preferred segmentation and cluster three is categorized into less preferred segmentation after discussing with domain experiment.

According to the expectation of the researcher, the segmentation should contains proportional clusters segmentation like poorly, less, moderately, high and very high preferred customers category to have a good clustering model. But the result didn't identify as expected. This is because of the default K value and seed number used in the experiment as starting point. Since the main goal of this clustering experiment is to come up with a good clustering model, the researcher is forced to continue the experiment by increasing the seed number.

Experiment Two

In the second experiment, value of K set to be 2 and size of seed to 100 but everything the same as the previous experiment. The output of the experiment depicted in the following Table 4.3

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	2885 (22%)	Female	28.55	Married	Urban	Junior	1084.66
2	2091 (16%)	Female	42.49	Married	Urban	Primary	1193.32
3	8081 (62%)	Female	35.18	Married	Rural	None	975.52

Table 4.3 Cluster description based on values of attributes for K=3 and seed size 100

As indicated in the above Table 4.3, cluster one contain attribute values with frequency of 2885(22%) records, married female average aged 28, living in urban, completed junior school, and capable of borrowing an average of 1084.66 loan amount. In the second cluster, with frequency of 2091(16%)records, female who are married, an average age of 42, living in urban, completed primary school, and capable of borrowing an average of1193.32loan size are categorized. In the third cluster, with frequency of 8081(62%)records, female who are married, an average age of 35, living in urban, with no formal education, and capable of borrowing an average of 975.52loan size an average loan are categorized.

In the experiment Sum of squared errors is 11151. This measurement was used to compare the goodness of this experiment with the previous experiment. Beside this, based on the average loan size shown in Table 4.3, cluster one categorized into high preferred customer's segmentation, cluster two is categorized into very high preferred customer's segmentation and cluster three is categorized into less preferred customer's segmentation after discussing with domain experiment.

According to the expectation of the researcher, the segmentation should contain proportional clusters segmentation like as said so far in experiment one. But the result didn't identify as expected. Thus, the researcher is forced to continue the experiment by increasing the seed number.

Experiment Three

The third experiment was conducted by setting the value of K to 2 and seed size to 1000 and it produces the following output as depicted in Table 4.4.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	3290 (25%)	Female	30.52	Married	Urban	Primary	1097.35
2	8540 (65%)	Female	36.94	Married	Rural	None	998.05
3	1227 (9%)	Male	32.28	Married	Urban	Primary	1119.80

Table 4.4 Cluster description based on values of attributes for K=3 and seed size 1000

To describe the attributes value in Table 4.4, let's start with the first cluster. Cluster one contain attribute values with frequency 3290(25%) records, female with average age of 30, who had completed primary education, married and living in urban, capable of borrowing an average of 1097.35 loan sizes. Next to this, attribute values with frequency 8540(65%) records, female living in rural with average age 36 who are not completed any formal education, married and having capacity of borrowing an average of 998.05 amount of loan are categorized. Cluster three contain attribute values with frequency 1227(9%)records, male with average age of 32, who had completed primary education, married and living in urban, capable of borrowing an average of 1119.80 loan sizes.

In the third experiment Sum of squared errors is 11329.08. This measurements clearly showed that sum of squared error registered less result than the previous two experiments. Interpretation with domain experts based on the average loan size, as shown in Table 4.4segmentcluster one, into high preferred, cluster two into moderately preferred and cluster three into high preferred.

As shown in the above table and its description, Clusters segmentation did not contain uniquely identified proportional customers' segmentation. Therefore, to come up with a good clustering model, the researcher is forced to continue the experiment by increasing the value of K and seed number.

Experiment Four

The fourth experiment was conducted by setting the value of K to 3 and seed size to 10 and it produces the following output illustrated in Table 4.5.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	3271 (25%)	Female	32.86	Married	Urban	Junior	1170.71
2	1552 (12%)	Male	34.85	Married	Rural	Primary	1070.90
3	3388 (26%)	Female	48.82	Married	Rural	None	1056.16
4	4846 (37%)	Female	26.53	Married	Rural	None	915.79

Table 4.5 Cluster description based on values of attributes for K=4 and seed size 10

Table 4.5 shows that under category of cluster one attribute values with frequency 3271(25%)records, female who are married, with average age of 32, living in urban, completed junior school, and capable of borrowing an average amount of loan 1170.71 are categorized. In the second cluster, attribute values with frequency 1552 (12%) records, married male with average age of 34, living in rural, completed primary school formal education and capable of borrowing an average of 1070.90sizes of loan are categorized. In the third cluster, attribute values with frequency 3388 (26%) records, married female with average age of 48, living in rural with no formal education, capable of borrowing an average of 1056.16 loan size are categorized. In the fourth cluster, attribute values with frequency 4846 (37%) records, married female with average age of 26, living in rural with no formal education, capable of borrowing an average of 915.79loan size are categorized.

In this experiment Sum of squared errors is 10560.78. This measurement clearly showed that sum of squared error registered better result from all previously conducted experiments. Here also as interpreted with domain experts based on the average loan size as shown in Table 4.5, cluster one categorized into very high preferred customers' segmentation, cluster two and three categorized into high preferred segmentation, cluster four categorized into very less preferred customers' segmentation. Similarities exist in cluster two and three indicate that the mode is not a good model. Thus, to come up with a good clustering model, the researcher is forced to continue the experiment by increasing the seed number again.

Experiment Five

The second experiment was conducted by setting the value of K to 3 and seed size to 100 and it produces the following output illustrated in Table 4.6.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	2755 (21%)	Female	28.38	Married	Urban	Junior	1085.97
2	1849 (14%)	Female	40.18	Married	Urban	Primary	1202.21
3	6923 (53%)	Female	33.50	Married	Rural	None	979.71
4	1530 (12%)	Female	46.45	Widowed	Rural	None	987.18

Table 4.6 Cluster description based on values of attributes for K=4 and seed size 100

Table 4.6 shows that under category of cluster one attribute values with frequency 2755 (21%)records, married female with average age of38, living in urban, completed junior school, and capable of borrowing an average of 1085.97 loan size are categorized. In second cluster, attribute values with frequency 1849(14%)records, female that are married, with average age of40, living in urban, completed primary school, and capable of borrowing an average amount of loan1202.21 are categorized. In the third cluster, attribute values with frequency 6923(53%) records, married female with average age of 33, and living in rural, no formal education, capable of borrowing an average of 979.71 sizes of loan are categorized. In the fourth cluster, attribute values with frequency 1530 (12%) records, widowed female with average age of 46, living in rural, no formal education, capable of borrowing an average of 987.18sizes of loan are categorized.

In this experiment Sum of squared errors is 9967.598946169874. This measurement clearly showed that sum of squared error registered better result than the previous experiments. And also as interpreted with domain experts based on the average loan size in Table 4.6, cluster one categorized into high preferred customer segmentation, cluster three categorized into very high preferred, cluster three categorized into less preferred customer's segmentation and cluster three categorized into moderately preferred customer's segmentation. This experiment give better output than of all previously conducted.

Now, to come up with a good clustering model than this experiment's output, the researcher is continued the experiment by increasing the seed number.

Experiment Six

In the six experiment value of K is set to 3 and size of seed to 1000 but everything the same as the previous experiment. The output of the experiment depicted in the following Table 4.7.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	2840 (22%)	Female	29.61	Married	Urban	Primary	1092.20
2	7393 (57%)	Female	34.76	Married	Rural	None	976.97
3	1200 (9%)	Male	32.13	Married	Urban	Primary	1113.08
4	1624 (12%)	Female	46.72	Widowed	Urban	None	1137.55

Table 4.7 Cluster description based on values of attributes for K=4 and seed size 1000

As observed from Table 4.7, attribute values with frequency 2840 (22%) records, married female with an average age of 29, living in urban, completed primary school and capable for borrowing an average loan size of 1092.20 are categorized in the first cluster. Under cluster two, attribute values with frequency 7393(57%)records, female who are married, average age of34, living in rural, with no formal education, and capable of borrowing an average loan size of 976.97are categorized. The next cluster contain attribute values with frequency 1200 (9%)records, in married male with an average age of32, living in urban, completed primary school, and capable of borrowing an average loan size of 1113.08 as cluster three. In cluster four, attribute values with frequency 1624(12%) records, female who are married, average age of 46, living in urban, with no formal education, and capable of borrowing an average loan size of 1137.55 are categorized.

As observed from the experiment Sum of squared errors is 10292.50. In this measurement, result of sum squared error decrease than the previous experiment (exp. 5). Here also based on the average loan size in Table 4.7, as interpreted with domain experts, cluster one and three segmented into high preferred, cluster two segmented into less preferred and cluster four segmented into very high preferred customer's segmentation.

In case of this experiment, clusters segmentation did contain unique and proportional customers' segmentation. But measurement of sum of squared error did not register better result than the previous one. Again it needs another experimentation to come up with a good clustering model by increasing the value of k and seed number.

Experiment Seven

The fourth experiment was conducted by setting the value of K to 4 and seed size to 10 and it produces the following output as depicted in Table 4.8.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	968 (7%)	Male	28.20	Married	Urban	Secondary	1025.81
2	1851 (14%)	Male	36.52	Married	Rural	None	1103.51
3	2997 (23%)	Female	48.13	Married	Rural	None	971.70
4	4595 (35%)	Female	26.91	Married	Rural	None	817.11
5	2646 (20%)	Female	35.03	Married	Urban	Primary	1168.25

Table 4.8 Cluster description based on values of attributes for K=5 and seed size 10

To describe the attributes value in Table 4.8, let's start with the first cluster. Cluster one contain attribute values with frequency 968 (7%)records, male with average age of 28, who had completed secondary education, married and living in urban that are capable of borrowing an average of 1025.81loan size. Next to this, attribute values with frequency 1851(14%)records, male living in rural with average age 36 who learnt no formal education, married and having capacity of borrowing an average of 1103.51amount of loan are categorized under cluster two. Cluster three also contain attribute values with frequency 2997(23%) records, in married females with average age of48, living in rural, no formal education, and capable of borrowing an average of 917.70loan size. In fourth category, attribute values with frequency 4595 (35%) records, married females with age of26, their residence is rural, completed no formal school can borrow an average of 871.11 amount of loan are categorized. In five cluster, attribute values with

frequency 2646 (20%) records, married females with age of 35, their residence is rural, completed primary education, capable of borrowing an average of 1168.25 amount of loan are categorized.

In the seventh experiment Sum of squared errors is 9664.82. This measurement clearly showed that sum of squared error registered better result of all the previous experiments. Interpretation with domain experts based on the average loan size in Table 4.8segment cluster one into moderately preferred, cluster two into high preferred, three into very less preferred, cluster four into less preferred and cluster five into very high preferred customer's.

In this experiment it became possible to clearly distinguish between the clusters and also to classify each of the customers' to a different level according to their expected (very less preferred, less preferred, moderately preferred, high and very high preferred) customers' segmentation.

Experiment Eight

In the eighth experiment, value of K set to be 4 and size of seed to 100 but everything the same as the previous experiment. The output of the experiment depicted in the following Table 4.9.

Cluster number	Freq of records	Sex	Age (average)	Marital status	Residence	Educational level	LoanAmount (average)
1	2091 (16%)	Female	30.11	Married	Urban	Junior	1156.19
2	1454 (11%)	Female	36.87	Married	Urban	Primary	1187.20
3	1593 (12%)	Female	33.53	Married	Rural	Primary	977.49
4	3379 (26%)	Female	48.60	Married	Rural	None	1061.50
5	4540 (35%)	Female	26.71	Married	Rural	None	929.49

Table 4.9 Cluster description based on values of attributes for K=5 and seed size 100

In Table 4.9, cluster one contain attribute values with frequency 2091(16%) records, female with average age of 30, who had completed junior education, married and living in urban, and capable

of borrowing an average of 1156.19 loan size. Next to this, attribute values with frequency 1454 (11%) records, female living in urban with average age 36 who completed primary education, married and having capacity of borrowing an average of 1187.20amount of loan are categorized under cluster two. Cluster three contain attribute values with frequency 1593(12%)records, married female with average age of 33, living in rural, completed primary education, and capable of borrowing an average of 977.49loan sizes. In fourth cluster, an attribute values with frequency3379 (26%) records, married female with age of 48, their residence is rural, completed no formal school, capable of borrowing an average of 1061.50amount of loan are categorized. In fifth cluster, an attribute values with frequency 4540 (35%)records, married female with age of 26, their residence is rural, completed no formal school, capable of borrowing an average of 929.49amount of loan are categorized.

The eighth experiment Sum of squared errors is 9701.15. As clearly observed from the eighth experiment, other experiments conducted after the seventh experiment with increment of K values do no yield better result than the previous one. It registered results beyond minimum value registered in experiment seven. As explained previously, the main goal of clustering experiment is to come up with a good clustering model. After good clustering model resulted in seventh experiment, the researcher did not need to continue the experiment.

Generally, the overall result of this experiment (the seventh experiment) looks satisfactory because of the fact that it satisfies the criteria of a good segmentation model used in the research; it is the clarity of the segments to be explained by the domain experiments. The result shows different group of customers segments and most of the drawbacks indicated in the previous experiments are solved. As clearly indicated, some of the clusters in the previous experiments are suffering from having patterns which are difficult to interpret. In addition to this, the clustering algorithm put customers' segmentations with similar pattern in different clusters.

4.5.3.4. Choosing the best clustering model

In finding the best clustering model, eight experiments were conducted. This is to come up with the appropriate clustering model. Eight of them are presented and discussed and summary of all the experiments are illustrated in Table 4.10. Finally, based on Sum of squared errors, best clustering model that satisfy the criteria was selected.

To validate the obtained output of the experiments, Sum of squared errors is taken as measurement criteria. To measure the goodness of a clustering structure, Sum of Squared Error (SSE) is one of the measures as explained under Section 3.2.3. It measures how objects closely related in a cluster and how distinct or well-separated a cluster is from other clusters.

Beside the above criteria, the comparison of clustering result validity has been done in relation to the values of the key attributes (loan amount) in each cluster, business objectives of the institutions (maximizing profit) and finally, the domain expert's judgment based on business objective of the institutions.

The main objective of every business institutions is to maximize their profit. In BG MFI, this is done by increasing the amount of money lent to customers and by retaining customers for long period of time. That is why the values of loan amount play a significant role in validating the obtained clustering results.

Accordingly, the customers' segmentation which has high loan size has high probability to be the preferred customer of the institution. Whereas, customer segmentation which has low loan size, will have less probability to be preferred customer of the institution based on the objective of the institutions.

In the attempt to improve the distribution of instances in different segmentations, different seed values (10, 100, and 1000) with different values of K (3, 4and 5) have been tried and after a number of experiments conducted, the best cluster had been obtained (experiment 7). The seed value at 10 and value of K=5 gives the best distribution of instances in the segments. This is because of the minimum measurement values registered through the experiment compared to others. Comparison of clustering models with different values is shown in the following Table 4.10.

Experiment No.	K-values	Seed size	Number of iteration	Sum of squared errors
1	3	10	5	10770.78
2	3	100	5	11151.80
3	3	1000	5	11329.08
4	4	10	8	10560.79
5	4	100	5	9967.60
6	4	1000	7	10292.50
7	**5**	**10**	**9**	**9664.82**
8	5	100	7	9701.15

Table 4.10Comparison of clustering models

4.5.4. Classification modeling

In this research, as explained in methodology part, to build classification model, the output of clustering model that is built through different experiments and chosen as a best model, has been used as an input for the purpose of constricting rule to predict potential customers. The algorithm selected for classification purpose was J48 decision tree. The researcher tested the algorithm with different parameters and record numbers to improve the classification accuracy. Finally, models compared and the best model selected. The selected classification model generates rules that enable to identify profitable customers in the market segmentation.

4.5.4.1. Decision tree model building

A decision tree classifier is one of the most widely used supervised learning methods used for data exploration, approximating a function by piecewise constant regions, and does not necessitate previous information of the data distribution. Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. The true purpose of the decision tree is to classify the data into distinct groups or branches that generate the strongest separation in the values of the dependent variable, being superior to predict segments with a desired individual behavior such as response or activation, thus providing an easily interpretable solution (Brefelean, 2007).

Below, in the Table 4.11, input dataset and the result of decision tree output at different experiments are illustrated.

Experiments	No. of records	Number of attributes	No. of leaves	Size of trees	Test modes	Time taken to build model (Sec)	Accuracy (%)
1	13057	7	43	70	All training dataset	0.17	99.977%
2	13057	7	43	70	70/30%	0.19	99.8979%
3	**13057**	**7**	**43**	**70**	**10-fold cross-validation**	**0.2**	**99.9464%**

Table 4.11Output of decision tree with different test modes

As observed from Table 4.11, different experiments were conducted at different test modes in decision tree algorithm. Results showed that at different test modes there is no any impact on number of leaves and size of tree while there is difference on accuracy of the model built.

As explained under Section 4.5.2 in the decision tree classification models, different experiments have been done by splitting the dataset into training and testing set and by adjusting the parameter values into 70/30% to have better accuracy. In addition, all training dataset and 10 fold cross-validation test modes are used in experiment to compare models built with each other.

From all experiments conducted, using all training dataset test mode resulted in better classification accuracy (99.98%). However, using all dataset for training has its own limitation. There may be bias of classification. Therefore, the researcher decides to validate the developed model with the test modes that used some dataset for training and some dataset for testing. From the experiments carried out, 10-fold cross-validation test mode, with all record amounts (13057), number of attributes 7, numbers of leaves 43 and size of trees 70 scored better accuracy (99.95%) and selected as better decision tree model than others. The confusion matrix of selected model is shown in Table 4.13 below.

Actual	Predicted					Total	Accuracy
	very_high_preferred	high_preferred	less_preferred	very_less_preferred	moderately_preferred		
very_high_preferred	2642	3	0	0	1	2646	99.85
high_preferred	2	4593	0	0	0	4595	99.96
less_preferred	0	0	1851	0	0	1851	1.00
very_less_preferred	0	0	0	968	0	968	1.00
moderately_preferred	1	0	0	0	2996	2997	99.97
Total	2645	4596	1851	9092	2997	13057	99.95

Table 4.12Confusion matrix of 10 fold cross-validation model

Table 4.12 depict that out of total numbers records (13057) supplied, 13050 (99.95%) instances classified correctly, while the remaining 7 (0.05%) of instances classified incorrectly. In addition, as shown in confusion matrix table, 2642 (99.85%), 4593 (99.96%), 1851 (1.0%), 968 (1.0%) and 2996 (99.97) records are classified correctly as very high preferred customers' segmentation, high preferred customers' segmentation, less preferred customers' segmentation, very less preferred customers' segmentation and moderately preferred customers' segmentation respectively. The decision tree generated from this model is attached in Appendix 1.

4.5.4.2. Best rules generated

There are different rules generated from the selected decision tree model for class prediction. But for simplicity and manageability, only those rules with large number of instances correctly classified are taken as best rules and these are the following:

Rule #1

If Sex = Female, Age <= 37, Residence = Urban and Educationallevel = Junior: then customer will be classified as very_high_preferred customer (691.0)

Rule #2

If Sex = Female, Residence = Urban, Educationallevel = None, LoanAmount<= 2250, and Age <= 30: then customer will be classified as high_preferred customer (428.0)

Rule #3

If Sex = Female, Residence = Urban, Educationallevel = None, and Age > 31: then customer will be classified asvery_high_preferred customer (192.0)

Rule #4

If Sex = Female, Age <= 37, Residence = Urban and Educationallevel = Primary: then customer will be classified as very_high_preferred (355.0)

Rule #5

If Sex = Female, Residence = Urban and Educationallevel = Secondary, and Age <= 31: then customer will be classified as very_less_preferred (401.0)

Rule #6

If Sex = Female, Residence = Urban and Educationallevel = Secondary, and Age > 31: then customer will be classified as very_high_preferred (138.0)

Rule #7

If Sex = Female, Residence = Rural and Age <= 30: then customer will be classified as high_preferred (3302.0)

Rule #8

If Sex = Female, Residence = Rural, Age > 30 and Educationallevel = Junior: then customer will be classified as high_preferred (131.0)

Rule #9

If Sex = Female, Age <= 37, Residence = Rural and Educationallevel = None: high_preferred (678.0)

Rule #10

If Sex = Female, Age > 37, Educationallevel = Junior and Residence = Urban: then customer will be classified as very_high_preferred (237.0)

Rule #11

If Sex = Female, Educationallevel = None, Residence = Urban and Age <= 41: then customer will be classified as very_high_preferred (176.0)

Rule #12

If Sex = Female, Educationallevel = None, Residence = Urban and Age > 41: then customer will be classified as: moderately_preferred (576.0/1.0)

Rule #13

If Sex = Female, Age > 37, Educationallevel = None and Residence = Rural: then customer will be classified as moderately_preferred (2184.0)

Rule #14

If Sex = Female, Age > 37, Educationallevel = Primary and Residence = Urban: then customer will be classified as very_high_preferred (283.0)

Rule #15

If Sex = Female, Age > 37, Educationallevel = Primary and Residence = Rural and Age <= 41: then customer will be classified as very_high_preferred (151.0)

Rule #16

If Sex = Female, Age > 37, Educationallevel = Primary and Residence = Rural and Age > 41: then customer will be classified as moderately_preferred (130.0)

Rule #17

If Sex = Female, Age > 37, Educationallevel = Secondary, and Residence = Urban: then customer will be classified as very_high_preferred (105.0)

Rule #18

If Sex = Male, Residence = Urban and Educationallevel = Junior: then customer will be classified as very_less_preferred (218.0)

Rule #19

If Sex = Male, Residence = Urban and Educationallevel = Secondary: then customer will classified as very_less_preferred (141.0)

Rule #20

If Sex = Male, Residence = Rural and Educationallevel = Junior: then customer will be classified as less_preferred (279.0)

Rule #21

If Sex = Male, Residence = Rural and Educationallevel = None: then customer will be classified as less_preferred (1132.0)

Rule #22

If Sex = Male, Residence = Rural and Educationallevel = Primary: then customer will be classified as less_preferred (341.0)

As illustrated in the above rules generated from model built, the experimentation help to get feature which characterize customers with very high preferred, high preferred, moderately preferred, less preferred and very less preferred customers segmentation. These rules will help to predict customers to which class he/she may be grouped.

4.6. EVALUATION

In this phase the degree to which the model meets objectives of the study is assessed. The business objective of the institution is to come up with a model that could find the appropriate number of clusters of customers according to individual behavior and also to assign customers to the appropriate clusters index. Consequently, the business can have appropriate data mining techniques.

Data mining techniques applied to solve the problem of profitable customers' identification in the institution were clustering and classification. These techniques were chosen because of their capability of processing a wider variety of data and easier to interpret in giving meaning to the problem. Different experiments have been done; models have been evaluated and the best performed model selected. After different experiments have been conducted, optimal models built and the best models were selected, the next step is evaluating the output against the institutions business objective. Evaluation of the segmentation output is based on the dataset of the institution. These are different demographic and financial information of customers.

The obtained clustering models have been evaluated using the classes to clusters evaluation approach/cluster mode and SSE. In the case of classification, the obtained classification tree model has been evaluated using the 10 fold cross validation approach. The dataset is divided into 10 subsets, insuring that each class is represented with approximately equal opportunities subsets. Then each subset was used for testing and the remaining 9 for training purposes. A total of 12985 (99.95%) instances were correctly classified.

The analysis which was closely undertaken with domain expert revealed that the 7^{th} segmentation experiment indeed discovered patterns that are really interesting. As clearly indicated, the clustering model brought customers into different clusters according to their individual behaviors. In addition to this , the decision tree model provide a very good description

of the segments and it clearly shows a number of rules that have valuable help to assign potential customers to one of the clusters.

Generally, results identified individuals' customer behavior in each customer segment that can be used to improve the quality of the institution's products and services to have profitable and potential customers. Hence, appropriate customer relationship management strategies and programs can be designed and implemented based on market segmentation into meaningful groups according to the institutions need. These in turn lead institution to achieve their business objectives.

4.7. DEPLOYMENT OF THE RESULT

Deployment of the result means using the data mining results investigated through this research work i.e. result of clustering and classification. The new knowledge or pattern discovered should be organized and presented in a way that the organization can understand and use it for effective customer relationship management. For the application of the result, resources, technology and business process in the institution should be integrated.

At the very beginning, BG MFI had been established to maximize its profit by providing loan service. As discussed with the officials and domain expert, in the institution, based on the business objective, customers are identified as good and services delivered based on trust. Customers trust each other grouped together in lack of collateral (as like in banking) to get loan provision given priority to start the service. But, as business institution, to be more profitable and to handle customers according to their individual need in the market, it is better to look at and predict customers before providing loan. To predict customers, best rules generated like under Section 4.5.4.2. are solutions to such problem. Some sample rules generated from customers segmentation results based on institution important dataset are the following.

If sex of customer is female, her age is less than or equals to 37 years, her residence is Urban and her educational level is junior school completed then customer will be classified into very high preferred customer (**as defined in rule no. 1**). If sex of customer is female, her residence is Urban, her educational level is with none of formal education attended, loan amount she requested is less than or equals to 2250, and her age is less than or equals to 30 then customer

will be classified as high preferred customer (**as defined in rule no. 2**).If sex of customer is female, her educational level is none of formal education attended in school, her residence is Urban and her age is greater than 41years then customer will be classified as moderately preferred customer (**as defined in rule no. 12**).If sex of customer is male, his residence is rural and his educational level is primary school completed then customer will be classified into less preferred customer (**as defined in rule no. 22**).If sex of customer is female, her residence is Urban, her educational level is secondary school completed, and her age is less than or equals to 31years then customer will be classified into very less preferred customer (**as defined in rule no. 5**).The above rules are discussed here to show as sample. So, for detail information look at Section 4.5.4.2. or Appendix I.

As tried to show above, institution can simply identify customers as poorly, less, moderately, high and very high preferred customers' segmentation. Then, before decision passed to provide loan, best customers will be selected to have maximum profit. Else strategies will be planned to promote and motivate those customer like very less preferred customers segmentation to generate better profit.

Therefore, based on pattern mined from data, if the institution deploys it with little modification, it is possible to create a long lasting relationship with their customer, generating more profit through their optimal customer satisfaction.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The growth of interest in data, information and knowledge management has been helping many organizations to digitize and manage their information resource for effective use in future to prediction about their business processes, product and behaviour of their customers.

Implementation of data mining technologies in BG MFI help to discover pattern for customers segmentation and classification in order to enhance service delivery and to maximize profit in the institution. Data mining application could be used to cluster and classify customers' behaviour in providing loan service. Then, the data mining technology could help to predict potential customers' from the other in taking measures to improve the service delivery and profit generating in the future.

To uncover the hidden knowledge within the dataset of the Institutions, preprocessing, of the dataset were performed using Weka tool. The data was analysed and interpreted using the WEKA 3.7.5 version software. Clustering and then classification models were built to categorize and predict customers.

To cluster instances into similar groups, SimpleK-means algorithms were employed. Thus, using SimpleK-means algorithm, different experiments conducted with different K-values and seed sizes. Segmentation at K=5 and seed size 10 with 5clusters selected as best customers segment model in the institution.

To classify instances into same groups based on result obtained through clustering,Decision TreeJ48algorithm was employed. Using DecisionTreesJ48algorithm also different experiments conducted. Model built by 10-fold cross-validation test mode which registered high accuracy (99.95%), selected as best model for prediction purpose.

The finding's result demonstrates that there are patterns relationships among different attributes. Thus, based on discovered patterns, customers can be clustered in different groups according to

their similarities. Then, it will be easy to predict and classify customers accordingly. Experimental findings show that there are different options to attract and retain customers, as well as to gain competitive advantage in the industry and also to serve customers with optimal satisfaction based on the discovered patterns. These paternal relationships between the data indicate the hidden fact among the dataset. Therefore, the institution could plan and implement strategies for effective and efficient service to maximize profit gained through maximum customer satisfaction.

Generally, to predict the behaviour of customers, and business processes of the institution, data mining techniques offers great promise in helping institution to discover hidden patterns in their data. Thus, using combination of techniques likes clustering with classification (as done in this research work) will assist to overcome the complexity of problems in business processes.

5.2. Recommendation

The researcher believes that findings of the study will encourage the institution, to work on the application of data mining techniques for successful achievement of the institutional goal. Because the finding showed the importance of customers profile, and how new knowledge can be generated from those data, to improve their service in a new and modern methods than the traditional one. These in turn help them to identify their market.

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

- Performance measured through the study is promising. But this research was conducted for academic purpose. To deploy in the institution with little modification and to come up with more comprehensive models, it is recommended that experimental tests be conducted by the institution with inclusion of many dataset by using large training and testing datasets.
- Institution should develop an integrated data warehouse in order to apply data mining techniques they require. To facilitate this, selection of hardware, software and human resources should be considered as a key issue that will have direct influence on the success or failure of any data-mining application.

- Based on result of research, institution should consider and develop important customer relationship management strategies that could be applied, to expand their service by predicting their potential customers, to retain their customers for prolonged time by giving optimal satisfaction and to gain competitive advantage in the industry.
- Nowadays, business problems become complex and diverse. Thus, the researcher believes that, application of other data mining techniques (rather than clustering and classification) with different algorithms in new tool is potential research area to improve performance of customer relationship management in the institution.

References

1. A. Jain, M. Murty, and P. Flynn (1999). Data Clustering: A Review.ACM Computing Surveys, vol. 31.
2. Addison-Wesley (1977). Exploratory Data Analysis. Accessed on 29/02/2012, from http://www.press.princeton.edu/chapters/s02_8709.pdf
3. AlemayehuYirsaw (2008). The performance of Micro Finance Institutions in Ethiopia: A case of six microfinance institutions. Master's Thesis.
4. Befekadu B. Kereta (2007). Outreach and Financial Performance Analysis of Microfinance Institutions in Ethiopia. African Economic Conference United Nations Conference Center (UNCC), Addis Ababa, Ethiopia.
5. Berry M. J. A. and Linoff, G. S (2000). Mastering data mining. New York, Wiley
6. Berry, Michael J. A. and Gordon Lino (2000). Mastering Data Mining. The Art and Science of Customer Relationship Management, John Wiley & Sons, New York.
7. Bertels, Thomas. Rath and Strong's (2003). Six Sigma Leadership Handbooks. John Wiley& Sons Inc., Hoboken, NJ.
8. Bing Liu, Yiyuan Xia and Philip S. Yu (2012). Clustering Via Decision Tree Construction. Accessed on 10/03/2012, from http://www.cs.ucla.edu/~wwc/course/cs245a/CLTrees.pdf.
9. Breyfogle and Forrest W (1999). Implementing Six Sigma: Smarter Solutions Using Statistical Methods. The Nature of Six Sigma Quality, Motorola University Press.
10. Collier K, B. Carey, D. Sautter and C. Marjaniemiand (1999). A Methodology for Evaluating and Selecting Data Mining Software. Proceedings of the 2nd Hawaii International Conference on System Science.

11. D.H. Stamatis (2003). Six Sigma for Financial Professionals, John Wiley &Sons,Hoboken, NJ.
12. David Hand, HeikkiMannila and Padhraic Smyth (2001). Principles of Data Mining. The MIT Press.
13. David L. Banks and Yasmin H. Said (2006). Data Mining in Electronic Commerce.Journal of Statistical Science, Vol. 21, No. 2.
14. DetlefSchoder and Nils Madeja (2004).Is Customer Relationship Management a Success Factor in Electronic Commerce? Journal of Electronic Commerce Research, Vol. 5, No. 1.
15. Dunham, M. H Stamatis (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.
16. Dunham, M. H Stamatis, Sridhar S (2006). Data Mining: Introductory and Advanced Topics. 1st Edition, Pearson Education, New Delhi.
17. E. Colet (2000). Clustering and Classification: Data Mining Approaches. accessed on October 26, 2011 from http://www.tgc.com/dsstar/00/0704/101861.html
18. Eugene Tuv, Alexander Borisov, George Runger and Kari Torkkola (2009). Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. Journal of Machine Learning Research, Vol. 10.
19. FadzilahSiraj and Mansour Ali Abdoulha (2012). Mining Enrolment Data Using Predictive and Descriptive Approaches.accessed on29/02/2012, fromwww.intechopen.com/download/pdf/13160

20. Fayyad, U. M, Piatesky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery: an overview. Advances in knowledge Discovery and Data Mining, MIT Press, California.

21. Frawley W, Piatesky-Shapiro G and Matheus C. (1991). Knowledge discovery in databases: an overview. Knowledge Discovery in Databases, MIT Press.

22. Gartner Research (2000). Customer Relationship Management (CRM): Perspective. accessed on 20/02/2012 from http://lamarheller.com/technology/crm/whitepapers/crmdataperspective.pdf

23. Halkidi M. and M. Vazirgiannis (2001). Clustering validity assessment: finding optimal portioning of a dataset.In proceeding of ICDM Conference, California, USA.

24. HianChyeKoh and Gerald Tan. (2005). Data Mining Applications in Healthcare. Journal of Healthcare Information Management, Vol. 19, No. 2.

25. Hp-invent, (2012). CRM architecture for enterprise relationship marketing in the new millennium.Technical white paper, accessed on 12/01/2012, from http://h71028.www7.hp.com/enterprise/.../CRMArchitecture_Whitepaper_...

26. In-Tech (2009). Data Mining and Knowledge Discovery in Real Life Applications. Accessed on 23/02/2012, from http://www.intechopen.com/download/pdf/5937.

27. J. Han and M. Kamber (2006).Data Mining Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco.

28. JayanthiRanjan, Ghaziabad, Vishal Bhatnagar (2008). A Review of Data Mining Tools in Customer Relationship Management.Journal of Knowledge Management Practice, Vol. 9, No. 1.

29. Jerome A. Blakeslee (1999). Implementing the Six Sigma Solution. Quality Progress,Vol. 32, No. 7.

30. Jerry W. Thomas, (2007). Market Segmentation. Accessed on 13/10/2012, from http://www.decisionanalyst.com

31. Jonathan Schloo (2012). CRM Planning Guide: Your Roadmap for Success. Accessed on 15/1/2012, from http://www.qiem.com/resources/CRM_Planning_Guide.pdf.

32. K. RavichandraRao (2003). Data Mining and Clustering Techniques. Documentation Research and Training Center, Indian Statistical Institute, Bangalore.

33. K. Sujatha, Dr.N.Pappa and A. Kalaivani (2010). Combustion Quality Estimation in Power Station Boilers using Median Threshold Clustering Algorithms. International Journal of Engineering Science and Technology, Vol. 2, No.7.

34. K. Thangavel, Q. Shen, and A. Pethalakshmi (2006). Application of Clustering for Feature Selection Based on Rough Set Theory Approach.AIML Journal, vol. 6.

35. K.M. Hammouda (2001).Web Mining: Clustering Web Documents A Preliminary Review. Accessed on 25/09/2011, from http://watnow.uwaterloo.ca/pub/hammouda/review-document-clustering.pdf.

36. Klosgen, Wand Zytkow, J (eds) (2002). Handbook of Data Mining and Knowledge Discovery. Oxford University Press.

37. Kurgan, L. A. and Musilek, P. (2006). A Survey of Knowledge Discovery and Data Mining Process Models. The Knowledge Engineering Review, Vol. 21, No.1, Cambridge University Press.

38. L. Chang and X. Bai (2010). Data Mining: A Clustering Application. Accessed on 28/09/2011, from http://www.pacis-net.org/file/2010/P03-10.pdf.

39. Larose, D. T (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc.

40. Letenah Ejigu (2009). Performance Analysis of A Sample Microfinance Institutions of Ethiopia. International NGO Journal Vol. 4, No.5.

41. Lori Bowen Ayre, (2006). Data Mining for Information Professionals. Accessed on 28/09/2011, from http://techessence.info/node/53 .

42. M. Steinbach, V. Kumar, and L. Ertoz (2003). The Challenges of Clustering High Dimensional Data, Springer-Verlag.

43. Mahesh S. (2009). Six Sigma: Concepts, Tools, and Applications. Texas Woman's University press, Denton, Texas, USA.

44. Markus Hegland (2003). Data Mining – Challenges, Models, Methods and Algorithms. Accessed on 17/12/2011, from http://www.datamining.anu.edu.au/publications/2003/dm_script_hegland.pd.

45. Matheus C. J, Chan P. K. and Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. IEEE Transactions on Knowledge and Data Engineering,Vol.5, No.6.

46. Matthew A. Gana (2008). Principles of Marketing, Published by National Open University of Nigeria, Lagos, Accessed on 28/02/2012 fromhttp://www.nou.edu.ng

47. McClusky B. (2000). The Rise, Fall and Revival of Six Sigma Quality: Measuring Business Excellence. The Journal of Business Performance Measurement, Vol.4, No. 2.

48. MeklitMicroFinance Institution, Progynist and AliseiNGO(Organizers) (2004). Trends, Challenges and Other Key Issues in Micro Finance Development in Ethiopia. Workshop Proceedings on December 9 – 10, 2004, Addis Ababa, Ethiopia.

49. Micha'el Addisu (2006).Micro-finance Repayment Problems in the Informal Sector in Addis Ababa, Ethiopian Journal of Business & Development Vol.1 No.2.
50. Microsoft Dynamics GP (2007). Customer Relationship Management. Accessed on 28/02/2012 from http:// www.microsoft.com/dynamics/gp.
51. Mohammad Almotairi (2009). A Framework for Successful CRM Implementation. European and Mediterranean Conference on Information Systems, July 13-14 2009, Izmir.
52. OECD (2009). Competition and Financial Markets. Accessed on 13/11/2011, from www.oecd.org/competition/roundtables.
53. P.Santhi, V.MuraliBhaskaran (2010). Performance of Clustering Algorithms in Healthcare Database. International Journal for Advances in Computer Science, Vol.2, Issue 1.
54. Pavel Berkhin (2012). A Survey of Clustering Data Mining Techniques. Accessed on 25/09/2012, from http://msdn.microsoft.com/en-us/library/ms175595.aspx
55. Paul Gray and JongbokByun (2001). Customer Relationship Management. Accessed on 5/3/2012, from http:// www.crito.uci.edu.
56. Paul Gray and Jongbok Byun (2001). Customer Relationship Management. Accessed on 11/02/2012, from http://www.escholarship.org/uc/item/76n7d23r.pdf.
57. Peter C., Pablo H, Rolf S, Jaap V. and Alessandro Z. (1998). Discovering Data Mining: From Concept to Implementation. Prentice Hall, Upper Saddle River, NJ.
58. Qi Luo (2008). Advanced Knowledge Discovery and Data Mining. IEEE discovery and Data Mining, Workshop on Knowledge Discovery and Data Mining.

59. R. Ali, U. Ghani, and A. Saeed (2000). Data clustering and Its Applications. Accessed on 25/08/2011,fromhttp://www.seminarprojects.com/Thread-data-clustering-and-its-applications-full-report.

60. R. Xu, S. Member, and D.W. Ii (2005). Survey of Clustering Algorithms.*IEEE Transactions on neural networks*, vol.16.

61. RajanishDass, (2012). Data Mining in Banking and Finance: A Note for Bankers.Accessed on 14/01/2012, from http://iimahd.ernet.in/publications/data/Note%20on%20Data%20Mining%20%26%20BI%20in%20Banking%20Sector.pdf.

62. RakeshAgrawal, Heikki Mannila, RamakrishnanSrikant, Hannu Toivonen and A. InkeriVerkamo(1998)12 Fast Discovery of Association Rules.Accessed on 28/02/2012 from http://www.cs.helsinki.fi/hannu.toivonen/pubs/advances.pdf.

63. ReinyIriana and Francis Buttle (2006). Customer Relationship Management (CRM) System Implementations. International Journal of Knowledge, Culture and Change Management, Vol.6, No2.

64. Rene T. Domingo (2010). Applying Data Mining to Banking. Business Management Articles, accessed on 14/02/2012 from http://www.rtdonline.com/BMA/BSM/4.html.

65. S.P. Deshpande and V.M. Thakare (2010). Data Mining System and Applications: A review.International Journal of Distributed and Parallel System(IJDPS) ,Vol. 1, No.1.

66. Shigeki Kozakura et al.(2006). An Interpretation Method for Classification Trees in Bio-data Mining. KES, Part II, accessed on 23/04/2012, from http://www.www.springerlink.com/index/l87072001436w244.pdf.

67. Singapore Institute of Management (SIM) (2002). Data mining and customer relationship marketing in the banking industry. Accessed on 22/04/2012, fromhttp://www.thefreelibrary.com/Data+mining+and+customer+relationship+marketing+in+the+banking...-a087703083 .

68. SPSS (2000). CRISP-DM 1.0: Step-by-Step Data Mining Guide. CRISP-DM consortium. Accessed on 20/05/2012, from http://www.iidia.com.ar/rgm/CD.../TEI-2-CRISP-DM-GdP-material.pdf.

69. SPSS, (2004). Improving Tax Administration with Data Mining. Executive report. Accessed on 20/11/2011, from www.spsslietuva.com/media/collateral/modeling/tax.pdf.

70. SQL Server, (2012).Data Mining Algorithms (Analysis Services - Data Mining).Accessed on 25/09/2012, from http://msdn.microsoft.com/en-us/library/ms175595.aspx

71. Tan Steinbach Kumar (2005).Introduction to Data Mining. Addison Wesley press.

72. Thair, N.(2009). Survey of Classification Techniques in Data Mining. Proceedings of the International MultiConference of Engineers and Computer Scientists Vol. 1

73. Tom Griffin, Tamilla Curtis and Donald Barrere (2000). CRM in Russia and U.S. -- Case Study from American Financial Service Industry. Journal of Technology Research, Vol.1.

74. Two Crows Corporation (1999).Introduction to Data Mining and Knowledge Discovery. Third Edition, U.S.A.

75. Vasile Paul Brefelean (2007). Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. Proceedings of the ITI 29th Int. Conf. on Information Technology Interfaces, Cavtat, Croatia.

76. Wei-Yin, Loh (2011). Classification and Regression Trees. Journal of WIREs Data Mining and Knowledge Discovery, Vol. 1.

77. Witten I. H. and Frank E. (2000). Data mining concepts,.New York, Morgan-Kaufmann.

78. WoldayAmha (2004). The Development Microfinance Industry in Ethiopia: Current Status and the Prospect for Growth, accessed on 07/10/2011, from http://www.www.ipms-ethiopia.org/.../...pdf.

79. X. Cui, T.E. Potok and P. Palathingal (2005). Document Clustering using Particle Swarm Optimization.*IEEE Proceedings*.

80. Zhixian Yi (2008). Knowledge Management for Library Strategic Planning: Perceptions of Applications and Benefits.Journal of Library Management, Vol. 29, No. 3.

APPENDICES

Appendix I: Decision Trees Generated With 10-Fold Cross-Validation Technique

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: cluster_index

Instances: 13057

Attributes: 7

Sex

Age

Maritastatus

Residence

Educationallevel

LoanAmount

Clusters

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Sex = Female

| Age <= 37

| | Residence = Urban

| | | Educationallevel = Junior: very_high_preferred (691.0)

| | | Educationallevel = None

| | | | Age <= 31

| | | | | LoanAmount<= 2250

| | | | | | Age <= 30: high_preferred (428.0)

| | | | | | Age > 30

| | | | | | | LoanAmount<= 1100: high_preferred (11.0)

| | | | | | | LoanAmount> 1100: very_high_preferred (2.0)

| | | | | LoanAmount> 2250

| | | | | | Age <= 29: high_preferred (16.0/1.0)

| | | | | | Age > 29: very_high_preferred (12.0)

| | | | Age > 31: very_high_preferred (192.0)

| | | Educationallevel = Primary: very_high_preferred (355.0)

| | | Educationallevel = Secondary

| | | | Age <= 31: very_less_preferred (401.0)

| | | | Age > 31: very_high_preferred (138.0)

| | | Educationallevel = Higher: very_high_preferred (37.0)

| | Residence = Rural

| | | Age <= 30: high_preferred (3302.0)

| | | Age > 30

| | | | Educationallevel = Junior: high_preferred (131.0)

| | | | Educationallevel = None: high_preferred (678.0)

| | | | Educationallevel = Primary

| | | | | Age <= 31

| | | | | | LoanAmount<= 1100: high_preferred (5.0)

| | | | | | LoanAmount> 1100: very_high_preferred (3.0)

| | | | | Age > 31: very_high_preferred (173.0)

| | | | Educationallevel = Secondary: high_preferred (24.0)

| | | | Educationallevel = Higher: high_preferred (1.0)

| Age > 37

| | Educationallevel = Junior

| | | Residence = Urban: very_high_preferred (237.0)

| | | Residence = Rural: moderately_preferred (92.0)

| | Educationallevel = None

| | | Residence = Urban

| | | | Age <= 41: very_high_preferred (176.0)

| | | | Age > 41: moderately_preferred (576.0/1.0)

| | | Residence = Rural: moderately_preferred (2184.0)

| | Educationallevel = Primary

| | | Residence = Urban: very_high_preferred (283.0)

| | | Residence = Rural

| | | | Age <= 41: very_high_preferred (151.0)

| | | | Age > 41: moderately_preferred (130.0)

| | Educationallevel = Secondary

| | | Residence = Urban: very_high_preferred (105.0)

| | | Residence = Rural: moderately_preferred (15.0)

| | Educationallevel = Higher: very_high_preferred (12.0/1.0)

Sex = Male

| Residence = Urban

| | Educationallevel = Junior: very_less_preferred (218.0)

| | Educationallevel = None

| | | Age <= 32: very_less_preferred (63.0)

| | | Age > 32: less_preferred (62.0)

| | Educationallevel = Primary

| | | Age <= 31: very_less_preferred (75.0)

| | | Age > 31: very_high_preferred (78.0)

| | Educationallevel = Secondary: very_less_preferred (141.0)

| | Educationallevel = Higher: very_less_preferred (22.0)

| Residence = Rural

| | Educationallevel = Junior: less_preferred (279.0)

| | Educationallevel = None: less_preferred (1132.0)

| | Educationallevel = Primary: less_preferred (341.0)

| | Educationallevel = Secondary

| | | Age <= 32: very_less_preferred (48.0)

| | | Age > 32: less_preferred (22.0)

| | Educationallevel = Higher: less_preferred (15.0)

Number of Leaves: 43

Size of the tree: 70

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	13050	99.9464 %
Incorrectly Classified Instances	7	0.0536 %
Kappa statistic	0.9993	
Mean absolute error	0.0003	
Root mean squared error	0.0145	
Relative absolute error	0.1021 %	
Root relative squared error	3.7363 %	
Coverage of cases (0.95 level)	99.9617 %	
Mean rel. region size (0.95 level)	20.0368 %	
Total Number of Instances	13057	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.998	0	0.999	0.998	0.999	1	very_high_preferred
	1	0	0.999	1	0.999	1	high_preferred
	1	0	1	1	1	1	less_preferred
	1	0	1	1	1	1	very_less_preferred
	1	0	1	1	1	1	moderately_preferred
Weighted Avg.	0.999	0	0.999	0.999	0.999	1	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
2642	3	0	0	1	a = very_high_preferred
2	4593	0	0	0	b = high_preferred
0	0	1851	0	0	c = less_preferred
0	0	0	968	0	d = very_less_preferred
1	0	0	0	2996	e = moderately_preferred

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Belachew Reganie

January 2013

The thesis has been submitted for examination with my approval as university Advisor

---

Dr. Gashaw Kebede

January 2013