ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES TO CUSTOMER PROFILE ANALYSIS IN THE ETHIOPIAN ELECTRIC POWER CORPORATION

HAILEMARIAM ABEBE

JUNE, 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES TO CUSTOMER PROFILE ANALYSIS IN THE ETHIOPIAN ELECTRIC POWER CORPORATION

A 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

HAILEMARIAM ABEBE

JUNE, 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES TO CUSTOMER PROFILE ANALYSIS IN THE ETHIOPIAN ELECTRIC POWER CORPORATION

By

HAILEMARIAM ABEBE

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor(s),	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner,	_____	_____

DEDICATION

This research is dedicated to the “Almighty GOD” who is always there for me in all the hard times and challenges of my life.

ACKNOWLEDGMENT

First and foremost I am grateful to the almighty GOD for his kindness and blessings of me with the courage and endurance to successfully complete the research. Next to this, I am highly indebted and grateful to my advisor Ato Getachew Jemaneh for his guidance, suggestions and support throughout the preparation of this thesis. His critical assessment, comments and suggestions have helped me in maintaining the right direction for my study and making it meaningful.

I would also want to put on record my gratitude and indebtedness to my brother Geberma Abebe, who has always encouraged me for higher success, my mother Abeba Adgeh, my family who were with me in this study and Golenta who always asked me about the progress of the paper.

My special thanks also go to staffs of EEPCo, particularly Ato Daniel Tsige, for providing relevant data and other necessary information for this research.

There are many individuals that have really contributed directly or indirectly for the successful accomplishment of this research, and all of them deserve special appreciation and acknowledgement for being with me in all those challenging times of the study.

LIST OF ACRONYMS

ARFF: Attribute-Relation File Format

CRISP: CRoss Industry Standard Process

CRISP-DM: CRoss Industry Standard Process of Data Mining

CRM: Customer Relationship Management

CSV: Comma Separated Values

DBMS: Database Management System

DM: Data Mining

EEA: Ethiopian Electric Agency

EEPCo: Ethiopia Electric Power Corporation

AI: Artificial Intelligent

ICS: Interconnected System

ICT: Information Communication Technology

KDD: Knowledge Discovery in Databases

KVAR: Kilovolt Ampere Reactive

KW: Kilowatt

KWh: Kilo Watt Hour

MW: Megawatt

OLAP: Online Analytical Processing

SCS: The Self Contained System

WEKA: Waikato Environment for Knowledge Analysis

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

ABSTRACT

Data mining is progressively used in information systems as a technology to support decision making activities within business processes. Electric power industries are being pushed to understand and quickly respond to the individual needs and wants of their customers due to the dynamic and highly competitive nature of the industry and customers. Customer Relationship Management (CRM) is the overall process of exploiting customer data and information, and using it to increase the revenue generated from an existing customer and attract new customers by creating good relationship with them accordingly. To implement CRM, electric power industries can use their customer databases to get a better understanding of their customers. And thus, to extract this important customer information from available databases, data mining techniques play a great role.

In this research the applicability of clustering and classification data mining techniques to implement CRM in the Ethiopian Electric Power Corporation (EEPCo) have been explored within the approach of CRISP-DM process model. After understanding business objective of the corporation, customer profiles are collected, cleansed, transformed, integrated and finally prepared for experimenting with the clustering and classification algorithms to develop a model. The final dataset prepared for experimentation consists of 50000 customer records.

The K-means clustering algorithm was used to segment customer records into clusters with similar behaviors. In the classification sub-phase, J48 decision tree and Naive Bayes algorithms were employed. Using the final dataset different clustering models at K values of 4, 5, and 6 with different seed values have been experimented and evaluated against their performances. Consequently, the cluster model at K value of 4 with seed size 1000 has shown a better performance. Finally, its output is used as an input for decision tree and Naive Bayes classification models. First the different classification models with decision tree and Naive Bayes algorithms are experimented with different parameters. Among these, a J48 decision tree model that showed a classification accuracy of 99.894% was selected. The results of this study were encouraging and confirmed the belief that applying data mining techniques could indeed support CRM activities at EEPCo. In the future, more segmentation and classification studies by using a possible large amount of customer records and employing other clustering and classification algorithms could yield better results.

CHAPTER ONE

INTRODUCTION

1.1 Background

Data mining is a new kind of business information processing technology, which can extract interesting patterns or knowledge implicated in a large number of incomplete, noisy, and ambiguous data that people do not know in advance but with potential application (Han and Kamber 2001). It aims to find out 'hidden' correlations among data by extracting, converting, analyzing, and modeling from huge amount of transaction data in business database. Simply it is the process of extracting information in order to discover hidden facts contained in the database using a combination of machine learning, statistical analysis, modeling techniques and database technology in the areas such as decision support, prediction, forecasting and estimating. Generally, the goal of data mining is to create models for decision making that predict future behavior based on analysis of past activity. To effectively exploit the potential of data mining, database should be first organized into a format that used for further the data mining process.

Among a host of recent technology innovations, data mining is making changes to the entire makeup of our skills and comfort zones in information analysis. Not only has introduced an array of new concepts, methods, and phrases, it also departs from the well-established, traditional, hypothesis-based statistical techniques. According to Han and Kamber (2001), data mining is a new type of exploratory and predictive data analysis whose purpose is to describe systematic relations between variables when there are no (or incomplete) a priori expectations as to the nature of those relations.

In recent years, data mining is becoming hot research area and has made some remarkable achievements in business and nowadays, Customer Relationship Management (CRM) is one of the hot issues towards the success of business. Customers are the most

important element for the survival of business organization. As a result, business organizations have found it essential to acquire new customers and retain existing ones by applying data mining techniques and implementing CRM successfully.

CRM is an integration of technologies and business to satisfy the needs of customer during any given interaction with them. It involves acquisition, analysis, and using of knowledge about customer in order to increase business productivity with customer satisfaction and to do it more efficiently. CRM is becoming recognized as business strategy to understand, manage, and maintain customer relationship with the help of ICT. It involves deep analysis of customer behavior using data mining techniques.

CRM comprises a set of processes and enabling systems supporting a business strategy to build long term profitable relationships with specific customers (Ling and Yen 2001). According to Saarenvirta (2002), there are several data mining techniques applied to support CRM. Among those, the most widely used methods are clustering and classification. Classification techniques are used to partition the customer's database into predefined classes, whereas clustering is similar to classification but classes are not predefined.

In industries such as retail trade in which the competition is fierce, it is indispensable to assess preferred customers yielding more profits and form marketing strategies (i.e., business intelligence) to strengthen relationships with these preferred customers by providing unique services. Even in the electric power industry, it is expected to become necessary to form the same marketing strategies as these unrelated industries in consideration of future fierce competition. In order to draw up marketing strategies, it is first indispensable to understand customers by analyzing the customer data using data mining techniques. Since the nature of electric power differs from the products that the retail trade business targets, it is necessary to carry out a type of customer data analysis using data mining techniques to implement CRM.

It is common practice for power utility companies to record customer data, such as administrative facts, contractual data, billing procedures and consumption recordings, in various databases to support their billing activity. Information on customer consumption patterns, as well as their payment transaction patterns, is becoming critical for electric power companies. Buried with this vast amount of data are all sorts of information that could make a significant difference to the ways in which power utilities run their business and interact with their current and prospective customers to gain an edge over their competitors. However, the necessary information that exists within the company databases are too fragmented and complex for a human mind to support efficient conclusions upon. In addition, it is too inaccessible and time consuming to gather, because the information required to make strategic and timely decisions is hidden in complex database systems (Nizar et al. 2006).

In this paper, the researcher explained application of data mining to customer segmentation, as a customer profile analysis for the establishment of business strategies, what is called business intelligence. The potential of data mining is becoming very important and a wide range of companies around the glob has deployed successful applications of data mining (Thearling 2003). However, only few researches have been conducted at the School of Information Science, Addis Ababa University.

The first attempt was made by Gobena in 2000 that was on the application of data mining technology and techniques in the Ethiopian Airlines and this work was extended by Henok in 2002. Askale in 2001 conducted on the application of data mining in the financial industry specifically at the Dashen Bank, Melkamu in 2009 on the applicability of data mining techniques to CRM in case of Ethiopia Telecommunication Corporation Code Division Multiple Access (CDMA) telephone service and Tesfaye in 2002 also conduct on the application of data mining in the Ethiopian Insurance company. Moreover, Shegaw in 2002 has also assessed the potential applicability of data mining technology in the Ethiopian context with particular reference to the health sector. Hence, this research is a continuation of the data mining researches carried out so far, however,

with a different area of application, which is data mining application for CRM in the Ethiopian Electric Power Corporation (EEPCo).

Segmentation of customer helps the corporation (EEPCo) to design and offer different service strategies for different customer segments accordingly. For instance, offering special services to preferred (high value) customers such as discount charge menus and unique services such as face-to-face services by salesperson. For low value customers, on the other hand, attempt to make the business or the service more efficient by giving priority to answering their needs and satisfied them with the services provided.

EEPCo is a national electricity utility established as a public enterprise by regulation No. 18/1997 of the council of ministers. According to this regulation, the EEPCo is mandated to engage in the business of producing, transmitting, distributing and selling electrical energy and to carry out any other activities that would enable it to achieve its stated objectives. EEPCo was named in 1997- after serving in the name of Ethiopian Electric Light and Power Authority which was established in 1956. The corporation had two electric energy supply systems that are the Interconnected System (ICS) and the Self Contained System (SCS). The main energy source of ICS is hydro power plants and for the SCS mini hydro's and diesel power generators allocated in various locations of the country. EEPCo is a company responsible for power generation, transmission, distribution and sales of electricity all over the nation. The corporation manages and operates power-generating facilities, the national transmission and distribution grids, and is also responsible for the supply of electricity to more than 1,473,387 customers (EEPCo official website).

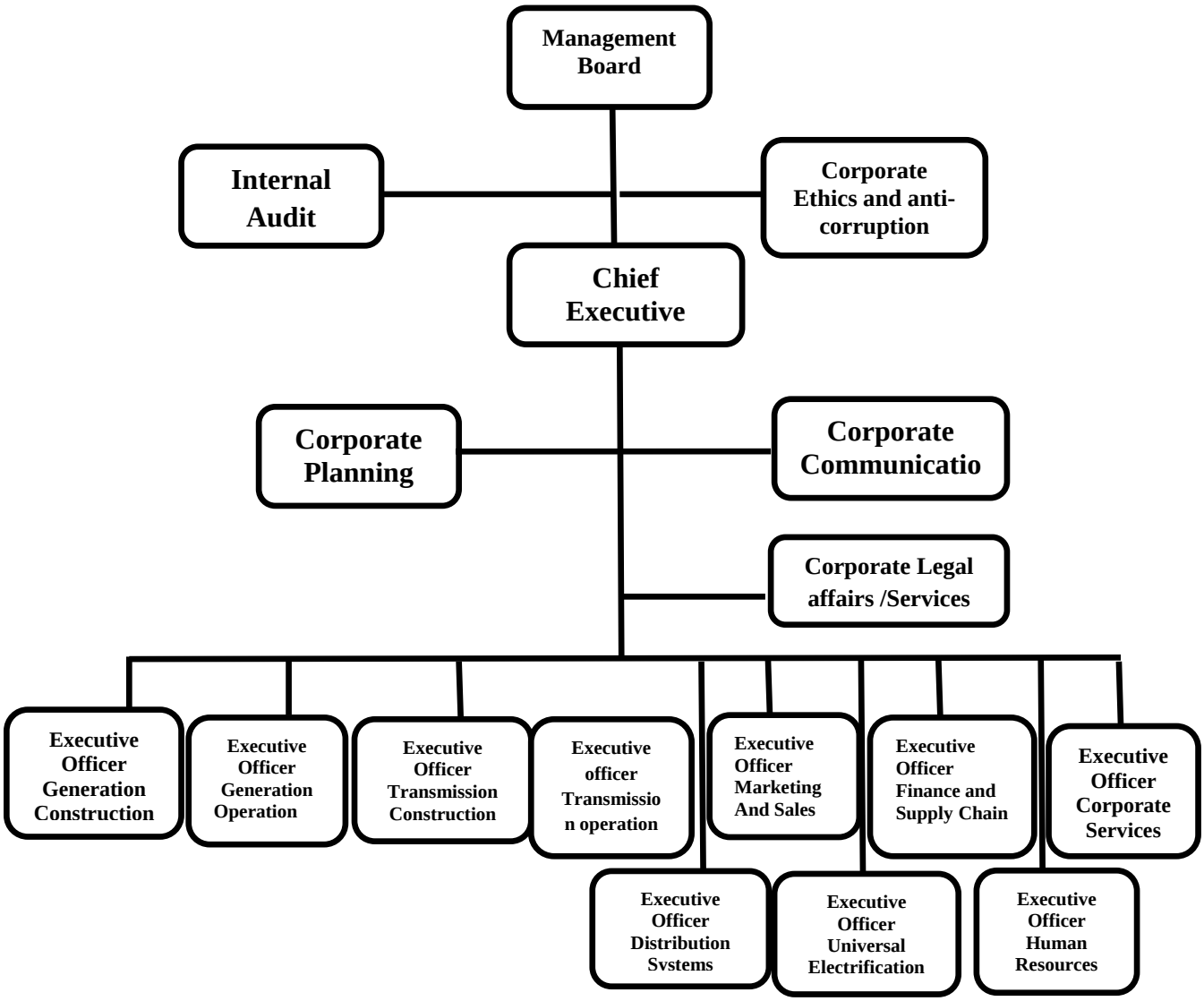


Figure 1.1: Overview of EEPCo Organization Structure (Organizational Chart Document of EEPCo)

Currently, EEPCo has been organized into a corporation in which there are different business units, functional units and departments. The following ongoing discussions describe these different units as depicted in the figure above.

The management board at the top is the highest body responsible for the overall functions and general operations of the corporation. Any business critical decisions are finally approved and decided upon by this group.

Under the management board, there is chief executive officer that is responsible for the overall operations in the corporation next to management board. There are different offices or departments under this executive officer like generation construction, generation operation, transmission construction, transmition operation, distribution system, marketing and sales, universal electrification, financial and supply chain, human resource and corporate services. Each branch departments has its own functions in the corporation.

Generation construction is responsible to the issues related to the construction of electric power from different power sources like water and wind. Generation operation is also an office under the Chief Executive Officer and responsible for managing the operations in the generations of power. Transmition construction is a department that involves in power transmition. Distribution system is a department which is concerned about the distribution of power to customers. Marketing and sales division performs any marketing activities in the corporation. There are different units in the corporate service department. Among them, ICT is the one from which the data about customers is collected. The ICT unit is responsible to make strategic decisions regarding the architecture, services, and platforms, functions of IT. Moreover, it defines the ICT relationship with other business and functional units of the corporation. The system development, network and maintenance, the database development and administration are the tasks of this ICT unit.

There are also other departments in the corporation like universal electrification, financial and supply chain, human resource, corporate ethics and anti-corruption, corporate communication, corporate legal affairs and others.

1.2 Statement of the Problem and Justification of the Study

Organizations that provide good quality service to customers have a competitive advantage that ensure its supremacy over less capable competitors and guarantee its survival. Each organization must strive for quality by taking into consideration its own special features and the demands of its customers.

Many organizations have collected and stored a wealth of data about their current customers, potential customers, suppliers and business partners. However, the inability to discover valuable information hidden in the data prevents the organizations from transforming these data into valuable and useful knowledge (Berson et al. 2000). Data mining techniques could help these organizations to discover the hidden knowledge in the enormous amount of data. Data mining technologies are important to process and extract useful information from customer database for the purpose supporting decision making or implementing CRM.

Companies involve in providing service like electric power corporation, shipping lines, airline, banks, universities, insurances, telecommunication companies and super markets are potential users of data mining technology for their customer relationship management. Due to these researches has been and being done in world wide. CRM is the area in which data mining is applied. There are researches that have been done in this area using data mining techniques in Ethiopian organizations context by Henok (2002) and Kumneger (2006).

EEPCo is a wholly government owned corporation and is responsible for power generation, transmission, distribution and sales of electricity all over the nation. The Electricity Proclamation No. 86/1997 of June 1997 gave way to the establishment of the Ethiopian Electricity Agency (EEA) as an autonomous federal government organization and has become fully operational since the beginning of 2000. The objectives of EEA are ensuring the prevalence of competition and righteousness in the supply and services of

electricity and encourage the promotion of investment in the electricity sector. EEA also regulates operations in the electricity supply sector including licensing and ensuring safety and quality standards (EEPCO official website).

According to Figueiredo et al. (2003), the full liberalization of most of the electricity markets in Europe and around the world creates a new environment were several private retail companies compete for the electricity supply of end users.

From the objective of the government in electric sector and trends of European and other world countries it can conclude that as the demand of electric power consumption increase it is visible that private sectors are expected to establish and expand for distribution of reliable electric power in the future. In addition, globalization and other technological advancement makes difficult to predict about the future. To this end, the company should be ready in advance to have a competitive advantage and exploit its goodwill through planning and implementing of successful CRM data mining application.

The application of data mining technology for EEPCo is important in supporting its effort in customer relationship management. However, currently the corporation is fully monopolized by the government with no competition. As a result, the CRM process of the corporation is not supported by tool and technique that can extract patterns or useful information from customers' database for competitive advantage. The customer databases contain enough data in the corporation and it is used to infer missing information using statistical methods (methods which are not supported by modern technology).

This study has been investigated in the area of business segmentation by considering the future competition in the EEPCo. Which customer is preferable to the corporation, which are not and what measurement should take place in each corporation customer segment.

1.3 Objective of the Study

1.3.1 General Objective

The general objective of this research is to support Customer Relationship Management (CRM) activities at Ethiopian Electric Power Corporation (EEPCo) by employing appropriate data mining techniques on the customers profile database to clusters and classify customers.

1.3.2 Specific Objectives of the Study

- To identify and collect the necessary customer data or profile from the Corporation database.
- To prepare the data for analysis. In other words, extracting the data and transforming it into the format required for the data mining algorithm or the software used to extract or min the pattern.
- To select the data mining tools and algorithms for clustering and classifying customers in the corporation.
- To design and develop the clustering and classification model for the data of EEPCo customers for better decision making and successful CRM implementation.
- To evaluate and deploy the model. And
- Finally, to recommend solutions on the application of data mining techniques in the Corporation customer database.

1.4 Research Methodology

Methodology means the steps or procedures that the researcher follows to achieve the objectives stated. It is a road map that shows the direction how the research is going to be done to reach the end. In this research, the researcher adopted the CRISP-DM process model to achieve the stated objectives. This is because the model has been widely applied for data mining studies. Besides, it is flexible to account for differences i.e. for different business problems and different data. And also it is open-source and industry standard

data mining processes model. Accordingly, this study has followed the following methodologies in order to develop good customer segmentation and classification models for designing and implementing successful CRM.

1.4.1 Review of Related Literature

Relevant literatures on data mining techniques and customer relationship management have been reviewed. Various books, journals, magazines, articles, and papers from the internet pertaining to potential of data mining for CRM in general and in particular successful data mining applications in clustering and classification of Electric Power and other organization customers have been reviewed.

1.4.2 Exploration of the Domain Problem

In order to define and analyze the business problem properly, the primary data was collected by observation and interviewing concerned officers (experts) in the corporation. The offices that the researcher has been conducted the survey are the database administrator and the marketing division of the corporation which is responsible to undertake activities related to marketing. And all the market strategies are designed by this particular office. Then based on the information obtained from these attempts, the overall customer data segmentation or clustering process has been done. The model employed in this research is CRISP-DM process which consists of six phases (Business understanding, Data understanding, Data preparation, Model building, Evaluation, and Deployment of the result).

1.4.3 Identification and Selection of Target Dataset

The data source of this research was the database containing customer records of EEPCo. The database contains information pertaining to customers who use the electric power service provided by the EEPCo. The dataset contains customer contract information and company information (type of industry), amount of electric power energy consumption, load factor, and other customer attributes. Database analysis for the relevance of data to the data mining task based on business objective and organizational needs has been done

with domain expert. The target populations of this research were all electric power customers of EEPCo except own consumption customer, active staff and retired staff. In short, in this phase to understand the customer data secondary data collection technique called database analysis has been employed.

1.4.4 Data Preparation

After the data are collected, orders or tasks such as processing and cleansing has been done in order to make the data more suitable for the particular data mining software, which has been used in the study. This comprises selection of attribute, handling noisy data, handling missing data fields, integrating, transforming and preparing the processed data in a file format acceptable to the tool or software that has been implemented.

1.4.5 Building, Training, Testing and Evaluating the Model

It is necessary to determine what measures to take in response to the type of customers extracted. For this purpose, what kinds of customers belong to each segment should be understood. As a method of extracting this information from each segment, data mining techniques can be applied. In this research, simple k-means is used for cluster and decision tree and Naive Bayes are used for classify customers.

Different experimentations have been done with different parameters. For instance, the clusters at K values of 4, 5, and 6 with seed size 10, 100 and 1000 in each cluster value of K have been carried out. The models of clustering at these different values of K have been evaluated together with the domain experts and finally, the best cluster model, which is at K value of 4 with seed size 1000, was selected and used as an input or cluster index for the J48 decision tree and Naive Bayes classification algorithms. Next, different classification models of the J48 decision tree and Naive Bayes have been experimented through changing some of their parameters. And, after comparing the two algorithms, the overall best model of J48 decision tree algorithm was selected.

After the classification algorithm is trained, it was validated with the 10-fold cross validation learning technique, a technique that uses 90 % of the data set for training and the remaining 10 % for the model for testing. In addition to this, other models have been experimented with 70/30%, 80/20% with different data size. And also full dataset for training and testing. And finally, the process of modeling and results are counterchecked with domain experts in the organizations and the model that showed better performance has been selected. This is because it provides useful information for making optimal customer related decisions.

Generally, feature selection and model building has been done iteratively by modifying the values of the parameters of simple K-means, decision tree and Naive Bayes in order to improve the performance of the model.

One of the critical tasks that have to be performed at this step is also selection of software that supports the data mining techniques to be employed in the study. So, in this research, MS-Excel 2007 is used for the data preparation and preprocessing. For clustering and classification task, WEKA version 3.6.4 has been used.

1.5 Scope of the Study

The scope of this research is restricted to building data mining model for segmenting the customers, interpreting the resulting segment, and generating a classification rules for each segment. The classification model prototype could not be implemented mainly due to time constraints.

1.6 Application of the Study Result

The result of this study can support the routine and strategic decision made by the Ethiopian Electric Power Corporation (EEPCo). By implementing an appropriate customer relationship strategy and by providing an attractive service through the right channel, at the right time and to the right customer, each customer contact in the corporation is more likely to be good or can achieve the goal. This research is considered

to contribute a lot for the corporation in addressing absence of customer relationship management in the future.

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 Ethiopian Electric Power Corporation (EEPCo) in the area of Customer Relationship Management (CRM).

1.7 Thesis Organization

This thesis is consisting of six chapters. The first chapter deals with the general overview of the study including background of the study, statement of the problem, objectives, scope, application and methodology of the research. The second chapter is devoted to literature review of data mining technology. The third chapter discusses about customer relationship management, CRM and data mining and application of CRM in different areas including electric power corporation. The fourth chapter is review of applicable data mining techniques and related research works. The fifth chapter is experimentation. It comprises the CRISP-DM process model steps such as business understanding, data understanding, data preparation, modeling, evaluation and deployment of the result. Results of the experiment are also analyzed and interpreted there. The last chapter which is Chapter Six presents the conclusion that summarize the major points of the research and recommendations forwarded for further research and adjustments in the organization on the ground of the research results.

CHAPTER TWO

LITERATURE REVIEW

DATA MINING CONCEPTS

2.1 Introduction

Huge amount of data is required to generate information. The data can be simple numerical figures and text documents, or more complex information like spatial data, multimedia data, and hypertext documents. To take complete advantage of data stored in files, databases, and other repositories, data retrieval is simply not enough. It requires a technique or powerful tool for analysis and interpretation of such data and that could help in decision-making. These technique or tool is data mining. Data mining is the extraction of hidden predictive and descriptive information from large databases or huge data. It is a powerful technology with great potential to help organizations to focus on the most important information in their data warehouses. Data mining technique also enables to predict future trends and behaviors, and helps organizations to make proactive knowledge-driven decisions and it can solve the problems that traditionally were too much time consuming like preparing databases for finding hidden patterns, finding predictive information that experts may miss. And, different techniques and algorithms are used to accomplish the tasks of data mining.

This chapter discusses the potential of data mining to discover knowledge from huge databases. It also provides a brief historical development of the field. Besides, it presents a review of different data mining methods, common tasks and basic steps for data mining process (CRISP).

2.2 Data Mining

This is an age often referred to as the information age. In this information age, information leads to power and success, and thanks to technologies such as computers, satellites, etc., it is possible to collect tremendous amounts of data/information. Initially, with the advent of computers and mass digital storage, collecting, storing, and counting using computers were started. Unfortunately, these massive collections of data stored on disparate structures very rapidly became increased. This problem has led to the creation of structured databases and Database Management Systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection of data. Today, there is huge amount of information (more information than can handle) from different sources such as from business transactions and scientific data, satellite pictures, text reports and military intelligence and etc. Information retrieval is simply not enough anymore to use this enormous data to extract important information for decision-making purpose. Confronted with huge collections of data, new needs are created now to help us make better managerial choices. These needs are automatic summarization of data, extraction of the “essence” of information stored, and discovery of patterns in raw data. Data mining techniques are the means to do these needs. Technologies such as data warehousing, data mining, and campaign management have greatly assisted companies to gain a competitive advantage. Data mining enables the extraction of hidden predictive and descriptive information from large databases. As a result, business enterprises identify valuable customers, predict future customers’ behaviors, and enable firms to make proactive knowledge driven decisions (Rygielski et al. 2002). Thus, data mining is a means to preserve customers by understanding their needs proactively.

Data mining is defined as exploration and analysis of large quantities of data by automatic or semi-automatic means to discover meaningful patterns and rules and these patterns allow a company to better understand its customers, and improve its marketing, sales, and customer support operations (CRM) (Anand and Kumar 2008). Data mining is

often done to analyze data in order to gain knowledge about the behavior patterns of customers and to identify key relationships that may help in decision making.

2.3 Data Mining and Knowledge Discovery

Data mining is “the principle of sorting through large amounts of data and picking out relevant information” (Anand and Kumar 2008). It is usually used by business intelligence organizations, and financial analysts, but nowadays, it is increasingly used in the science fields to extract information from the enormous data sets generated by modern experimental and observational methods.

Data mining is also known as knowledge discovery. Even though data mining and Knowledge Discovery in Databases (KDD) are frequently treated as synonyms, knowledge discovery is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases, whereas data mining is one part of the knowledge discovery process (Zaiane 1999). The following figure shows data mining is one of the steps in knowledge discovery process.

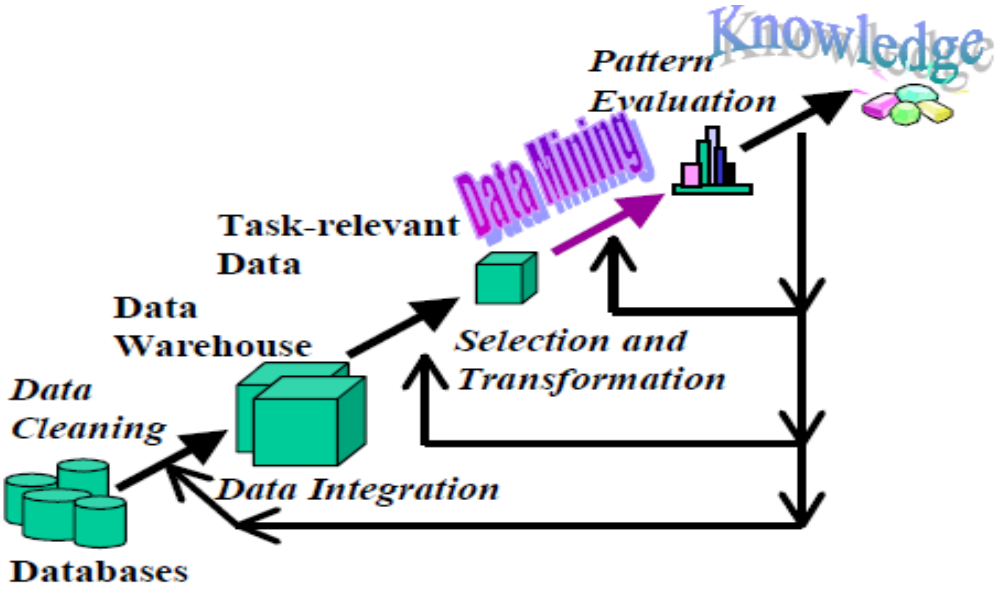


Figure 2.1: Shows data mining as one of the core steps in knowledge discovery process

(Source: Zaiane, O. R. 1999)

KDD refers to the overall process of discovering useful knowledge from stored data, and data mining refers to a particular step in knowledge discovery process. Data mining is a step that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data. Specifically, data mining is the application of specific algorithms for extracting of expected patterns from dataset (Fayyad et al. 1996).

Witten and Frank (2005) also define, data mining step as the process of finding interesting patterns from raw data that are not clearly/explicitly part of the data. These interesting patterns can be used to tell us something new and to make predictions. It also analyzes large observational data sets to find hidden relationships and to summarize the data in the new ways that are both understandable and useful to the user of the data (Larose 2006). Practically, data mining provides tools by which large quantities of data can be automatically analyzed and extracted.

2.4 Historical Development of Data Mining and Knowledge Discovery

The idea of today's data mining techniques started in 1950s when the work of mathematicians, logicians, and computer scientists combined to create artificial intelligence (AI) and machine learning (Buchanan 2006). Its origin lies with the first storage of data on computers continues with improvements in data access, until today technology allows users to navigate through data in real time. Data mining techniques are the result of a long research and product development process. **Table 2.1** below shows the four evolutionary stages of data mining from user's perspective.

The first stage is data collection 1960s. In this stage the data is collected to make simple calculations (summations or averages). Information generated at this stage answered business questions related to figures derived from data collection, such as total revenue or average total revenue over a period of time. Specific application programs were also

created for collecting data and calculations. In addition, according to Dunham (2003), in the 1960s, AI and statistics practitioners were developed new algorithms such as regression analysis, maximum likelihood estimates, neural networks, bias reduction, and linear models of classification. And, the term “data mining” was coined during this decade, but the term was used in a wrong way i.e. for describing the practice of wading through data to finding patterns that had no significance (Fayyad et al. 1996). Also in the 1960s, the field of information retrieval (IR) made its contribution to data mining in the form of clustering techniques and similarity measures. At the time these techniques were applied to text documents only, but they would be utilized later when mining data in databases and other large distributed datasets (Dunham 2003). By the end of the 1960s, information retrieval and database systems were developed in parallel.

In 1971, Gerard Salton published his groundbreaking work on the SMART information retrieval System. This explained a new approach to information retrieval which utilized the algebra-based vector space model (VSM). And, the model would prove to be a key element in the data mining toolkit (Dunham 2003).

The second stage is data access 1980s. In this stage databases are created to store data in a structured format. At this stage, company-wide policies for data collection and reporting of management information were formulated, because every business unit should do the accepted thing to specific requirements or formats. Once individual figures were known, questions that search the performance of aggregated sites could be asked easily.

The third stage is data navigation 1990s. Thanks to multi-dimensional databases, businesses in this stage could obtain either a global view or drill down to a particular site for comparisons with its peers. The introduction of multi-dimensional database enhances business organizations to improve their working strategy.

According to Fayyad et al (1996), in the beginning of the 1990s, the term “Knowledge Discovery” in Databases (KDD) had been coined and the first KDD workshop is held. In

this decade, the availability of huge amount of data created the need to have new techniques for handling these massive quantities of information. Moreover, 1990s saw the development of database warehouses, a term used to describe a large database created from the combination of operational and transactional database data. Along with the development of data warehouses came Online Analytical Processing (OLAP), decision support systems, data scrubbing/staging (transformation), and association rule algorithms were also developed (Dunham 2003; Han & Kamber 2001).

According to Two Crows Corporation (1999), during the 1990s, data mining changed from being an interesting new technology to part of common business practice. The reasons are cost of computer disk storage is decreased, processing power or speed of computers increased, and the benefits of data mining became increasingly important. And, businesses began using of data mining for customer relationship management (for acquiring new customers, increasing revenue from existing customers, and retaining good customers).

Finally, since 2000s, on-line analytic tools provided real-time feedback and information exchange with collaborating business units (Data Mining). This capability is useful when sales representatives or customer service persons need to retrieve customer information on-line and respond to questions on a real-time basis. Information systems can query past data up to and including the current level of business. Often businesses need to make strategic decisions or implement new policies that better satisfy their customers than previous. For example, grocery stores redesign their layout to promote more impulse purchasing. Telephone companies establish new price structures to attract customers into placing more calls. Both tasks require an understanding of past customer consumption behavior data in order to identify patterns for making strategic decision and data mining is particularly suited or appropriate to this purpose. With the application of advanced algorithms, data mining covers knowledge in a vast amount of data and points out possible relationships among the data. It helps businesses address questions such as, “What is likely to happen to Boston unit sales next month, and why?” the stages were

revolutionary because they allowed new business questions to be answered accurately and quickly (Rygielski et al. 2002).

KDD has evolved and continues to evolve, from the intersection of different research fields such as machine learning, pattern recognition, databases, statistics, AI (artificial intelligence), knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in large database. The data mining component of KDD process currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data (Fayyad et al. 1996). Today, these different technologies are mature and coupled with relational database systems (a culture of data integration) and they create a business environment that can take the advantage of knowledge or capitalize on knowledge which is formerly buried within the systems.

Nowadays, data mining is used by a wide variety of industries and sectors including retail, medical, telecommunications, scientific, financial, pharmaceutical, marketing, Internet-based companies, etc. (Fayyad et al. 1996).

Web mining is an area of much research and development activity. There are many factors that drive this activity including online companies wish to learn more about their customers and potential customers, governmental agents tasked with locating terrorists and optimizing services, and the user need for filtered information.

Stages	Business question	Enabling technologies	Product providers	Characteristics
Data collection (1960s)	"What was my average total revenue over the last five years?"	Computer ,tapes, disks	IBM,CDC	Retrospective, static data delivery
Data access (1980s)	"What were unit sales in new Ethiopia last march?"	Relational databases (RDBMS),structured query language (SQL),ODBC	Oracle,Sybase,Informix,IBM,microsoft	Retrospective, dynamic data delivery at record level
Data Navigation (1990s)	"What were unit sales in new Ethiopia last march? Drill down to Boston"	On-line analytic processing (OLAP),multidimensional databases, data warehouses	Pilot,IRI,Arbor,Redbrick,Evolutionary technologies	Retrospective, dynamic data delivery at multiple levels
Data mining (2000s)	"What's likely to happen in Boston unit sales next month (prediction)? Why?"	Advanced algorithms, multiprocessors computers, massive databases	Lockheed,IBM,SGI,numerous startups (nascent industry)	Prospective, proactive information delivery

Table 2.1 Evolutionary stages of data mining technology (Source: Rygielski et al. 2002)

2.5 Tasks of Data Mining

Data mining can be used to accomplish different tasks. But, the task of data mining is depending on the use of the data mining result that means for what purpose the result will be used (Larose 2005). The tasks are classified as follow:

- **Exploratory Data Analysis**: It is simply to exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.
- **Descriptive Modeling**: It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into

groups and models describing the relationships between the variables. In short, to describe the existing data.

- **Predictive Modeling**: To predict the future having or based on the existing data or behavior. The model enables to predict the value of one variable from the known values of other variables.
- **Discovering Patterns and Rules**: It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.
- **Retrieval by Content:** It is finding pattern similar to the pattern of interest in the dataset. This task is most commonly used for text and image datasets.

2.6 Types of Data Mining Systems

Data mining systems can be classified into different categories. The classification is based on: type of data source mind, data model, kind of knowledge discovered, mining techniques used, and also the degree of user interaction involved in the data mining process (Dunham and Sridhar 2006). The classification of data mining systems according to the type of data source mined is based on the type of data handled for mining purpose such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

The classifications of data mining systems based on the data model are relational database, object-oriented database, data warehouse, transactional database, etc.

Classification of data mining systems according to the kind of knowledge discovered; this classification is based on data mining functionalities or methods, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

Classification of data mining systems based on mining techniques used in the data analysis approach are machine learning, neural networks, genetic algorithms, statistics,

visualization, database oriented or data warehouse-oriented, etc. Finally, the data mining system classifications based on the degree of user interaction in the data mining process are query-driven systems, interactive exploratory systems and autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

2.7 Data Mining Methods

Data mining techniques or methods are classifications of data mining systems based on the kind of knowledge discovered. The two main types of data mining methods are verification-oriented (the system verifies the user's pre-defined hypothesis) and discovery-oriented in which the system can finds new rules and patterns autonomously from the data (Fayyad et al. 1996).

Discovery methods are methods that automatically identify or recognize patterns in the dataset. The discovery method consists of prediction methods and description methods. Description-oriented data mining methods are focus on understanding the way the underlying data operates, whereas, a prediction-oriented method aim to build a model that can able to predict newly and unseen data according to the model developed by the sample (training set). However, some prediction-oriented methods can also provide understanding of the data.

Most of the discovery-oriented techniques are based on inductive learning that means the model is constructed explicitly or implicitly by generalizing from a sufficient number of training data. The underlying assumption of the inductive approach is that the trained model is applicable to future new and unseen variable.

The following figure shows the categories (classifications) of data mining methods:

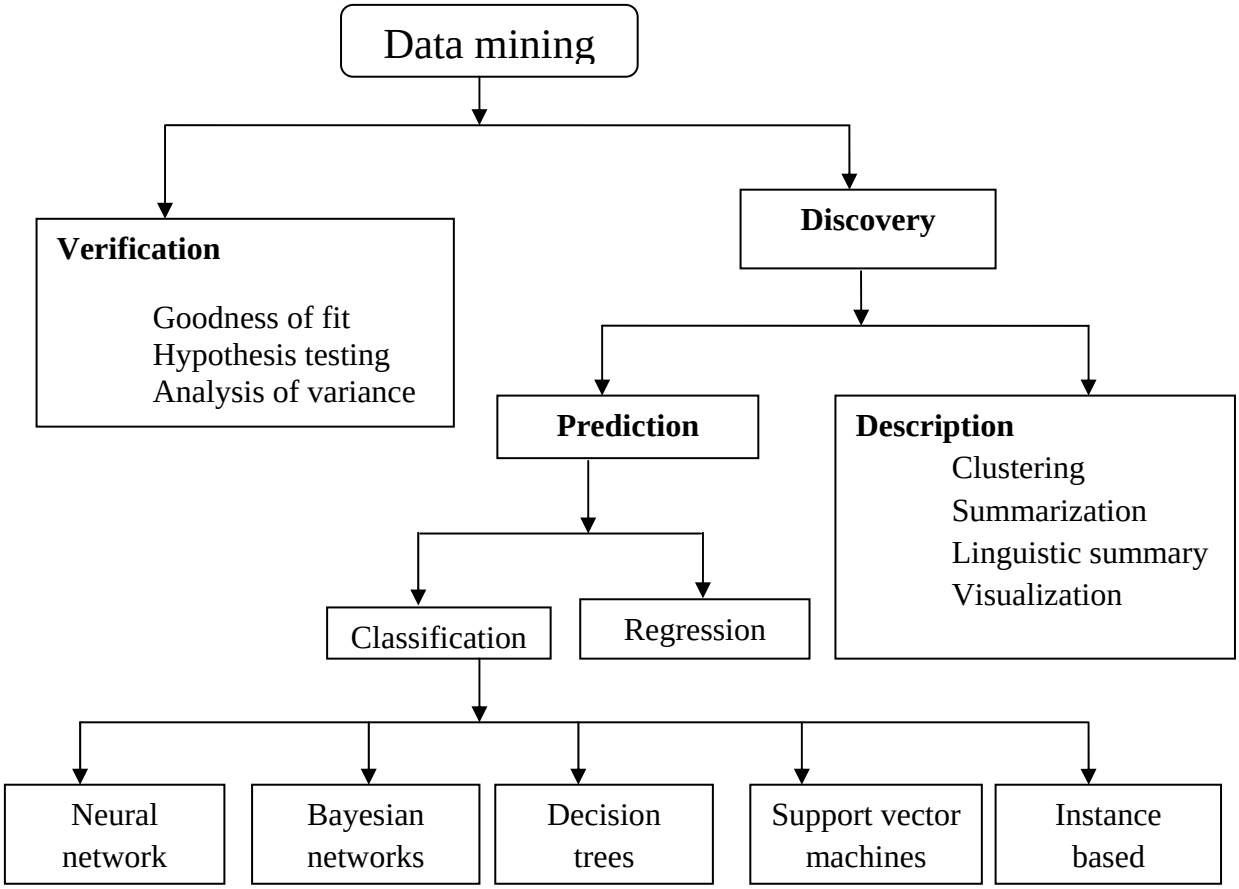


Figure 2.2: Taxonomy of data mining methods (Source: Fayyad et al. 1996)

Verification methods, on the other hand, deal with the evaluation of a hypothesis proposed by an external source (an expert). These methods include the most common methods of traditional statistics like goodness-of-fit test, hypothesis testing, and analysis of variance. These methods have less connection with data mining than discovery-oriented methods because most data mining problems are related to predicting and selecting a hypothesis (out of a set of hypotheses) rather than testing a known one. Moreover, "the focus of traditional statistical methods is usually on model estimation as opposed to one of the main objectives of data mining: model identification".

According to Fayyad et al. (1996), the common data mining methods are classification, regression, clustering, summarization, dependency modeling, and change and deviation detection.

2.7.1 Classification

This method is also known as supervised learning method. The two steps of this method are: supervised learning of a training set of data to create a model, and classifying the new and unseen data according to the model developed by the training data set. Some of the well-known classification algorithms are bayesian classification (based on Bayes Theorem), decision trees, neural networks and backpropagation (based on neural networks), k-nearest neighbor classifiers (based on learning by analogy), and genetic algorithms (Gerald 2002).

According to Gerald (2000), decision trees are the most known top-down approach for classification that divides the data into leaf and node divisions until the entire data set has been analyzed and evaluated. Neural networks are nonlinear predictive tools that learn from a prepared data set and then applied it to new and large sets. Genetic algorithms are like neural networks but incorporate natural selection and mutation. Nearest neighbor utilizes a training set of data to measure the similarity of a group and then use the resultant information to analyze the test data.

2.7.2 Regression

This method is used to make predictions based on existing data by applying formulas. Using linear or logistic regression techniques from statistics, a function is learned from the existing data. The new data is then mapped to the function in order to make predictions (Dunham 2003). According to Witten and Frank (2005), decision trees with averaged values at the leaves are a common regression technique.

2.7.3 Clustering

Clustering involves identifying a finite set of categories (clusters) to describe the data. The clusters can be mutually exclusive, hierarchical or overlapping. (Fayyad et al.1996).

Each member of a cluster should be very similar to the members within its cluster and dissimilar to other clusters members. Techniques used to create clusters on data stored include partitioning method like the k-means algorithm and hierarchical methods which group objects or data into a tree of clusters, such as density-based methods (Han and Kamber 2001).

2.7.4 Summarization

According to Dunham (2003), summarization is also called characterization or generalization. It derives summary data from the stored data or extracts actual portions of the data which briefly characterize (describes) the contents. Summarization maps data into subsets and then applies a compact description for that subset.

The mining methods discussed above form the basis for most data mining activities. Many variations on the basic approaches described above are found in the literature, including algorithms specifically modified to apply to spatial data, temporal data mining, multi-dimensional databases, text databases and the Web (Dunham 2003; Han and Kamber 2001).

2.8 Data Mining Life Cycle (Steps in KDD)

Several researchers, such as Brachmana and Anand (1994), Fayyad et al. **(**1996), Maimon and Last (2000), and Reinartz (2002), and others have proposed different steps or phases of KDD process. In this study, the researcher adopted the well known data mining process model called CRISP-DM.

2.8.1 CRISP-DM Process Model

A data mining process model defines the approach for the use of data mining, i.e. phases, activities and tasks that have to be performed whereas data mining represents a complex and specialized field. So, a generic and standardized approach is needed for the use of data mining in order to help organizations. CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a non-proprietary, documented and freely available data mining

process model created in 1996. It was developed by the industry leaders and the collaboration of experienced data mining users, data mining software tool providers and data mining service providers (Shearer 2000). To develop further and refine this process model and service the data mining community well Special Interest Group (CRISP-DM SIG) was formed. CRISP-DM version 1.0 was presented in 2000 and it is being accepted by business users (Shearer, 2000).

According to Peter et al. (2000), the life cycle (steps) of a data mining (KDD process), consists of six phases. In the transformation of raw data from business transaction and other sources to useful information that could helps to make decisions, each step is built on the previous ones and the sequence of the phases is not rigid, moving back and forth between different phases is always required depends on the outcome of each phase. The main phases are:

- **Business Understanding Phase**: This phase focuses on understanding the data mining (KD process) objectives and requirements from business perspective, then converting this knowledge into a data mining problem definition and this is a preliminary plan designed to achieve the objectives. According to Larose (2006), this phase may also be termed as research understanding phase.

- **Data Understanding Phase**: This phase concerned with data collection and understanding of the data using exploratory data analysis to get familiar with the data, to evaluate the quality of data, and to discover first approaching into the data or to detect interesting subsets or actionable patterns to form hypotheses for hidden information.

- **Data Preparation Phase**: This phase, which is labor intensive, covers all aspects of preparing the final data set, which will be used for subsequent phases, from initial, dirty and raw data. This stage includes operations like dimension reduction (such as feature selection and sampling), data cleansing (such as handling missing values, removal of noise or outliers), data transformation (such as Discretization of numerical attributes and attribute construction) and finally, clean the raw data so that it is ready for the modeling tools.

- **Modeling Phase**: In this phase, various modeling techniques (classification, regression, clustering and summarization) and algorithms are selected and applied or employed accordingly and their parameters are standardized or adjusted to optimal values. Often, several different data mining techniques may be applied for the same data mining problem.

- **Evaluation Phase**: In this stage, the model is methodically evaluated. Review and interpret the mined patterns for quality and effectiveness of the model before deploy it for use in the field. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use or task of the data mining results should be reached or decided.

- **Deployment Phase**: The purpose of the model is to increase knowledge gained from the data, and the knowledge gained need to be organized and presented in a way that the customer can understand and use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

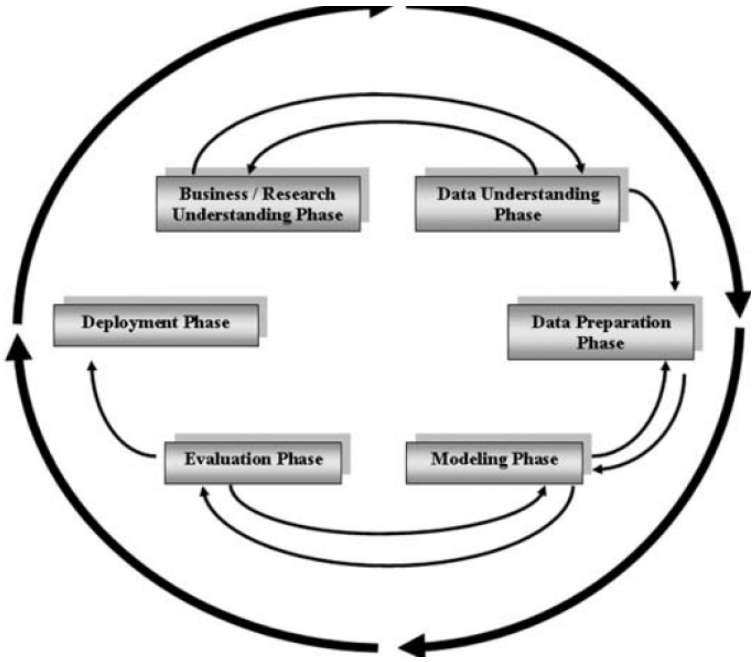


Figure 2.3: CRISP-DM process model (Larose 2006)

2.9 Applications of Data Mining

Today, different organizations are realizing the numerous advantages that come with data mining. And, organizations are using data mining to manage all phases of the customer life cycle (acquiring new customers, increasing revenue from existing customers, and retaining good customers)(Two Crows Corporation 1999). It provides clear and competitive advantage across a broad variety of industries by identifying potentially useful information from the huge amounts of data collected and stored. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies are also interested in applying this technology to reduce fraud. Medical applications are another important area in which data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies involved in financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores as well as to assess the effectiveness of promotions and coupons. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidate for development as agents for the treatments of disease. In power utilities data mining can be used to forecasting power demand of customers.

- **Retail**

The retail industry is realizing that it is possible to gain a competitive advantage by utilizing data mining techniques. Retailers have been collecting enormous amounts of data throughout the years, just like the banking industry, and now have the tool needed to sort through this data and find useful pieces of information. For retailers, data mining can be used to provide information on product sales trends, customer buying habits and preferences, supplier lead times and delivery performance, seasonal variations, customer peak traffic periods, and similar predictive data for making proactive decisions.

Bhasin (2006) says it is universally accepted that many industries (including banking, retail and telecom) are using data mining effectively. Data mining has many uses in industries. Through the use of store-branded credit cards and point-of-sale systems, retailers can keep detailed records of every shopping transaction. This enables them to better understand their various customer segments. Some retail applications include:

- **Performing basket analysis**: Also known as similarity analysis, basket analysis reveals which items customers tend to purchase together. This knowledge can improve stocking, store layout strategies, and promotions.
- **Sales forecasting:** Examining time-based patterns helps retailers make stocking decisions. If a customer purchases an item today, when are they likely to purchase a similar or complementary item?
- **Database marketing:** Retailers can develop profiles of customers with certain behaviors; for example, those who purchase designer labels clothing or those who attend sales. This information can be used to focus cost–effective promotions.
- **Merchandise planning and allocation:** When retailers add new stores, they can improve merchandise planning and allocation by examining patterns in stores with similar demographic characteristics. Retailers can also use data mining to determine the ideal layout for a specific store.

➢ Banking

The banking industry across the world has undergone great changes in the way the business is conducted. In the recent development, the greater acceptance and usage of 'electronic' banking, enables the capturing of transactional data easily and, at the same time, the volume of such data has grown significantly. And, it is beyond human capability to analyze this huge amount of raw data and to effectively transform the data into useful knowledge for the organization. The enormous amount of data that banks have been collected over the years can greatly influence the success of data mining efforts. According to Bhasin (2006), by using data mining techniques to analyze patterns and trends from this huge data, bank executives can predict, with increased accuracy, how customers will react to changes in interest rates, which customers will be likely to accept

new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable by using CRM. Moreover, Bhasin (2006) explains that, there are numerous areas in which data mining can be used in the banking industry, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investments. In addition, banks may use data mining to identify their most profitable credit card customers or high-risk loan applicants. There is, therefore, a need to build an analytical capability to address the above-stated issues and data mining attempts to provide the answer.

The banking industry is widely recognizing the importance of the information it has about its customers. Undoubtedly, it has among the richest and largest pool of customer information, covering customer demographics, transactional data, credit cards usage pattern, and so on. As banking is in the service industry, the task of maintaining a strong and effective CRM is a critical issue. To do this, banks need to invest their resources to better understand their existing and future potential (prospective) customers.

Generally, by using suitable data mining tools, banks can subsequently offer 'tailor-made' products and services to those customers.

- **Telecommunications**

The data mining applications for any industry depend on two factors: the data that are available and the business problems facing the industry.

Telecommunication companies around the world face escalating competition which is forcing them to aggressively market special pricing programs aimed at retaining existing customers and attracting new ones. The telecommunications industry has been one of the early adopters of data mining and has deployed numerous data mining applications. This is most likely because telecommunication companies normally generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a

rapidly changing and highly competitive environment (Weiss 2009). Telecommunication companies utilize data mining application to improve their marketing efforts, identify fraud, and better manage their telecommunication networks (network fault isolation and prediction). However, these companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events such as customer fraud and network failures in real-time. And to respond to these challenges new methods have been developed and existing methods have been enhanced. The competitive and changing nature of the industry, combined with the fact that the industry generates enormous amounts of data, ensures that data mining will play an important role in the future in the telecommunications industry.

According to Rygielski et al. (2002), knowledge discovery or data mining in telecommunications can be applicable for the following purpose:

- **Call detail record analysis**: Telecommunication companies accumulate detailed call records. By identifying customer segments with similar use patterns, the companies can develop attractive pricing and feature promotions.
- **Customer loyalty**: Some customers repeatedly switch providers, or "churn", to take advantage of attractive incentives by competing companies. The companies can use data mining to identify the characteristics of customers who are likely to remain loyal once they switch, thus enabling the companies to target their spending on customers who will produce the most profit.

➢ **Other applications:** Data mining (Knowledge discovery) applications are also emerging in a variety of industries:

- **Customer segmentation**: All industries can take advantage of data mining to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis. One example from industries is electric power industry which is the emphasis of this paper. For example in electric power

industry it can be used for customer segmentation based on power consumption pattern by different techniques.

- **Manufacturing**: Through choice boards, manufacturers are beginning to customize products for customers; therefore they must be able to predict which features should be bundled to meet customer demand.
- **Warranties**: Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.
- **Frequent flier incentives**: Airlines can identify groups of customers that can be given incentives to fly more.

CHAPTER THREE

CUSTOMER RELATIONSHIP MANAGEMENT

3.1 Introduction

The new millennium is in the midst of explosive change witnessing rapidly changing market conditions, volatile equity markets, reconstructed value chains and new global competitors. And, customers themselves are changing – natural customer loyalty is a thing of the past. The concept of Customer Relationship Management (CRM) has taken center stage in the business world for sustainable business advantage. Long-term success requires a great customer relationship management strategy. A technology-enabled CRM strategy to meet customer-focused objectives of the organizations is the vast majority organization's activity. No doubt about that CRM has become a top priority for companies seeking to gain competitive advantage in today's turbulent economy. However, confusion about exactly what CRM is, also confusion in how to best implement it, or even what role it should play in enhancing customer interaction.

The current trend in CRM is to be systematic. Customer transactions are being recorded by more and businesses have a great demand to use this data efficiently for decision making. Data mining (DM), which is the science of finding new, interesting patterns and relationships in huge amount of data, and is a proven solution to analytical CRM (Rinta-Runsala 2006).

3.2 Customer Relationship Management

In the early 20th century, companies concentrated on selling as many products as possible. Suppliers focused on product development, manufacturing capacity, and securing distribution channels, without regard to their consumers need. They did not pay much attention to who bought their products or what their customers needed. They used

classic marketing tactics, i.e., mass marketing – primarily print and broadcast advertising, mass mailings, and billboards (Ueno 2006).

By the middle of the 20th century, however, customers had the power of choice of what they want, because supply had surpassed demand. The period of the passive customer was coming to an end. Companies began to find out who their customers were, what they wanted, and how they could be satisfied. They began to analyze data about their customers and segment them based on their demographics, such as age, gender, and other personal information or attributes. Then they promoted their product or service to a specific subset of customers and prospects. This was called “target marketing”. Each company pays seriously to the “four P’s” (price, promotion, product, and placement), the basic concept of modern marketing, which was first suggested by the expert in the field, Jerome McCarthy, in 1960.

By the middle of the 1980’s, it had become extremely difficult to sell things. Traditional target marketing was not so satisfying under circumstances in which it was so difficult to cultivate new customers that this tactic could not sustain cost efficiency. At this point, the idea of “relationship marketing” gained the confidence of the business sector. This concept was aimed at building long-term relationships with customers and placed a great deal of value on the retention of existing customers rather than the acquisition of new ones. And, the idea of CRM came into existence in the late 1990’s. Although it originated in the United States and to date, has been accepted in a significant number of companies worldwide, there are also some people who have negative opinion to CRM; such views hold that it is difficult to implement successfully and its cost-benefit performance is low, among others. There is no clear definition of CRM; different researchers defined it in different perspectives. For instance, Jeffrey Peel, CEO of Quadriga Consulting, defined it as follows:

“CRM is about understanding the nature of the exchange between customer and supplier and managing it appropriately”. The exchange contains monetary considerations between supplier and customer and also communications. The challenge to all supplier

organizations is to optimize communications between parties to ensure profitable long-term relationships. In short, CRM is a management strategy that unites information technology with marketing. (Ueno 2006)

CRM as a product/market segment was first introduced by Siebel and Oracle companies with many other players, like SAP, PeopleSoft and Microsoft joining later (Srivastava 2006). And, the first set of products mostly support easy management of information for customer facing functions, including contact management, sales force automation, and etc. Applying data mining for better understanding of customers, and its use for relationship management is a recent phenomenon.

CRM has been introduced to understand the role of customers for the strategic position of a company and then to have good relationship with them accordingly. CRM takes a holistic view over customers, it encompasses all measures for understanding the customers and for exploiting this knowledge to design and implement marketing activities, align production and manage the supply-chain. And, also CRM emphasizes on the integration of measures to increase customer lifetime value. It should be noted that CRM is a broadly used term (holistic term), and it covers a wide variety of functions; however, not all of these require data mining.

According to Gray and Jongbok (2001), traditional marketing strategies focused on the four Ps (price, product, promotion, and place) to increase market share (volume of transactions between seller and buyer). However, CRM is a business strategy that goes beyond increasing transaction volume. Its objectives are to increase profitability, revenue, and customer satisfaction. Basically, CRM is a strategic business and process issue rather than a technical issue.

CRM can be defined as the process of predicting customer behavior based on the available data and selecting actions to influence that behavior to benefit the company. In other words, it is “the business strategy that influences an organization’s process, culture and technology to optimize revenue and increase value through understanding and

satisfying the individual customer's need" (TATA Consultancy Service 2009). And, also according to IDC and Gemini (2005), CRM is defined by four elements of a simple framework. These are: Knowing, Target, Sell, and Service. CRM requires the firm to know and understand its markets and customers. This enable to select the most profitable customers and identify those no longer has advantage targeting. CRM also lead to development of the offer i.e. which products to sell to which customers and through which channel. Finally, CRM enables to retain its customers through services such as call centers and help desks.

CRM is essentially a two-stage development concept (Rygielski et al. 2002). The first stage task is to master the basics of building customer focus. This means moving from a product orientation to a customer orientation and defining market strategy from outside-in rather than from inside-out. In other words, the focus should be on customer needs rather than product features. Companies in the second stage are moving beyond the basics; they do not rest on their success but push their development of customer orientation by integrating CRM across the entire customer experience chain, by using technology to achieve real-time customer management, and by constantly increasing their value proposition to customers.

"Modern business encourages a customer orientation approach using CRM as a principle that leads the business organization strategy towards satisfying the customer" (Rygielski et al. 2002). CRM helps business enterprises to understand their customers' value, and to focus on the most profitable customers, and build high-quality relationships with those customers. According to Lee and Park (2005), precise evaluations of customer profitability and targeting or focusing on the most profitable customers are crucial elements for the success of CRM.

Nowadays, companies are changing their business process models and building information technology (IT) solutions that enable to attract new customers, retain existing ones, and maximize the customer's lifelong value (Joe 2000). Furthermore, according to Ruiz et al (2004), companies attempt to deliver the highest value or good service to

customers through better communication, customized promotions, faster delivery, and personalized products and services. However, for successful implementation of CRM it is required to have an integrated and balanced technology, processes, and people (Chen and Popovich 2003).

Generally, CRM is “a business philosophy and set of strategies, programs and systems that focuses on identifying and building loyalty with a retailer’s most valued customers”. Based on the philosophy that retailers can increase their profitability by building relationships with their better customers, the goal of CRM is to develop a base of loyal customers who go to the retailer frequently (Davis 2006).

3.3 Components of Customer Relationship Management

According to Rygielski et al. (2002), customer relationship management is a combination of several components. Before the process can begin, the firm must first possess customer information. Companies can obtain data about their customers from internal customer database or they can purchase data from outside sources.

Another critical component of CRM is an enterprise data warehouse. Most organizations have huge databases that contain marketing, human resource (HR), and financial information. However, the data required for CRM can be limited to a marketing data with limited feeds from other corporate systems. The CRM system analyzes the data using statistical tools, OLAP, and data mining software tools. Firms can use traditional statistical techniques or one of the data mining software tools for CRM process, but marketing professionals need to understand the customer data and their importance to the business. The firm should also employ data mining analysts who will be involved in ensuring that the firm does not lose their original reason for doing data mining. Thus, having the right people who are trained to extract information with these tools is also important. The end result is segmentation of the market, and individual decisions are made regarding which segments are attractive.

The third and last component of a CRM system is campaign execution and tracking. These are the processes and systems that allow the user to develop and deliver targeted messages in a test, learn and interactive environment. Decisions made based on data mining and OLAP are implemented or done through campaign and tracking. Today, there are software programs that help marketing departments handle this complex feedback procedure. Campaign management software manages and monitors customer communications across multiple touch points, such as direct mail, telemarketing, customer service, point-of-sale, e-mail, and the Web (Thearling 1999). Although campaign management software is part of the overall solution, it is primarily the people and processes that contribute to smooth interactions between marketing, information technology, and the sales channels.

On the other hand Gray and Jongbok (2001), proposed the three components of CRM. These are: customer, relationship and management.

CRM is “a process by which a company maximizes customer information in an effort to increase loyalty and retain customers’ business over their lifetimes” (Suresh 2002). According to Suresh (2002), the primary goals of CRM are to:

1. Build long term and profitable relationships with chosen customers by utilizing the data "hidden" in enterprise databases. Examining and analyzing the data can turn raw data into valuable information about customer's needs. By predicting customer needs in advance, businesses can then market the right products to the right segments at the right time through the right delivery channels. Customer satisfaction can also be improved through more effective marketing strategy.
2. Get closer to those customers at every point of contact
3. Maximize your company’s share of the customer’s wallet or file
4. Transform the company into customer-centric organizations with a greater focus on customer profitability as compared to line profitability. The benefit gained from CRM enable companies to calculate or estimate the profitability of individual accounts.

3.4 The CRM Process Cycle

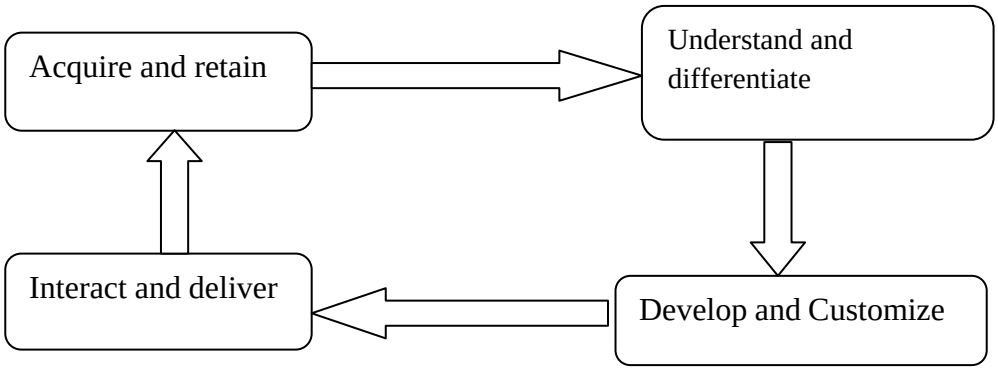


Figure 3.1: The CRM process

As shown in the diagram above, any business organization starts with the acquisition of customers. Acquisition customer data is a vital and the first stage in building customer relationship. For the purpose of customer acquisition an organization is likely to focus its attention on lapsed (failed) customers, former customers, competitor's customers' referrals (recommendations), and on the existing buyers. From these the organizations need to acquire customers and prospective customers and retain valuable customers.

The second stage is understand and differentiating customers. What they value, what types of services are important to them, how and when they like to interact and what they want to buy. True understanding is based on a combination of detailed analysis of customer data and interaction with them. There are several activities that enable to understand customers like profiling to understand demographics, purchase patterns and channel preference, segmentation to identify logical unique groups of customers that tend to look alike and behave in a similar fashion, primary research to capture needs and attitudes, customer assessment to understand profitability, as well as lifetime value or long term potential, etc.

The third stage is developing and customizing, i.e., companies develop products and services based on customers need, because in a customer focused world, product and

channel development has to follow the customers need. Otherwise, it is difficult to retain and attract customers.

The fourth and the last stages are interacting and delivering, this stage is a crucial component of a successful CRM process. Customers interact in many different ways with many different areas of the organization. And, with access to information and appropriate training for customers, organizations will be prepared to progressively increase the value they deliver to customers.

Generally, CRM is an iterative process that turns customer data into customer loyalty through four activities: (1) collecting customer data, (2) analyzing the customer data and identifying target customers, (3) developing CRM programs, and (4) implementing CRM programs. The process begins with the collection and analysis of data about a retailer's customers and the identification of target customers. The analysis translates the customer information into activities that offer value to the targeted customers. Then, these activities are executed through communication programs undertaken by the marketing department and customer service programs implemented by customer contact employees, typically sales associates (Davis 2006).

3.5 Principles and Tasks of CRM

3.5.1 CRM Principles

The overall processes and applications of CRM are based on three basic principles (Gray and Jongbok 2001). These principles are: personalization, loyally and lifetime value. The first one deals with remembering and treating customers individually so that the products and services to customer are designed and provided based on customer preferences and behavior. Acquire and Retain Customer Loyalty through Personal Relationship is the second one and refers to company needs to keep relationships with the customer. This continuous contact with the customer by meeting of customer preferences can create customer loyalty. The last principle is the selection of good customers from 'bad' based

on lifetime value of customers and through analyzing of their respective behaviors. The best customer deserve the most customer care, the worst customer should be dropped.

3.5.2 Key CRM Tasks

What is the benefit of the customer and how can we add the customer's values are the two basic questions that CRM tries to answer (Gray and Jongbok 2001). The basic tasks to address these questions can be categorized into four processes or tasks. These are Customer Identification, Customer Differentiation, Customer Interaction, and Customization / Personalization.

1. **Customer Identification**

The company should know or identify the customer through marketing channels, transactions, and interactions over time to provide value to profitable customers

2. **Customer Differentiation**

Segmenting of customers into different groups because each customer has his/her own lifetime value from the company's point of view and each customer may impose different demands and requirements for the company.

3. **Customer Interaction**

As customer demands change over time and the customer's long-term profitability and relationship to the company is important, the company needs to learn about the customer continually. In short, keeping track of customer behavior and needs is an important task of a CRM program.

4. **Customization / Personalization**

"Treat each customer uniquely" is the slogan of the entire CRM process. Through the personalization process, the company can increase customer loyalty. And, the automation of personalization is being made feasible by information technologies.

3.6 Technologies Used in CRM

The competitive advantage of technology in facilitating tasks in different organizations is unquestionable. And, the application of technology in CRM is the most electrifying, fastest growing and changing the way customers get information about products and services of the organizations and organizations communicate with customers. These technologies include all the equipment, software, and communication links that organizations use to enable or improve their CRM processes. The mostly used technology tools are Sales Force Automation, Call Centres, Data Warehousing, Data Mining and OLAP, Decision Support and Reporting Tools, Electronic Point of Sale (EPOS) (Suresh 2002).

3.7 Data Mining and Customer Relationship Management

Customer relationship management is a broad topic with many layers, one of which is data mining, and which is a method or tool that can help companies in their quest to become more customer-oriented by analyzing their profile (Rygielski et al. 2002).

The customer lifecycle provides a good framework for applying data mining to CRM. The term “customer lifecycle” refers to the stages in the relationship between a customer and a business or the organization. It is important to understand customer lifecycle because it relates directly to customer revenue and customer profitability. Marketers say there are three ways to increase a customer’s value to the organizations: (1) increase their use (or purchases) of products they already have; (2) sell them more or higher-margin products; and (3) keep the customers for a longer period of time. However, the customer relationship is not permanent it changes over time, as the business and the customer learn more about each other. So, customer lifecycle is a framework or outline for understanding customer behavior. In general, according to Rygielski et al (2002), there are four key stages in the customer lifecycle.

1. **Prospects**—People who are not yet customers but are in the target market in the future

2. **Responders**—Prospects who are interested in a product or service provided by the organization.
3. **Active Customers**—People who are currently using the product or service of the organization.
4. **Former Customers**—May be "bad" customers who did not pay their bills and who are not appropriate customers because they are no longer part of the target market, or those who may have shifted their purchases to competing organization products.

Data mining can predict the profitability of prospect customers as they become active customers, how long they will be active customers, and how likely they are to leave. In addition, data mining can be used over a period of time to predict changes in details. It will not be an accurate predictor of when most lifecycle events occur. Rather, it will help the organization identify patterns in their customer data that are predictive. Data mining plays a critical role in the overall CRM process, which includes interaction with the data warehouse in one direction, and interaction with campaign management software in the other direction. In the past, the link between data mining software and campaign management software was mostly manual. It required that physical copies of the scoring from data models be created and transferred to the database. This separation of data mining and campaign management software introduced considerable inefficiency and was prone to human error. Today, the trend is to integrate the two components in order to gain a competitive advantage (Thearling 1998).

3.8 Applications of CRM in the Electric Power Industry

CRM plays important role in different companies like telecommunication, bank, airline, retail and etc. It plays a great role in industries having huge amount of data. CRM in electric power industry has many applications. According to Kitayama et al. (2002), some of the applications are the following:

- **Customer Segmentation**: when an electric power company analyzes their customer profile, they can understand the needs of customers and concentrating

on preferred customers. And segmentation will be performed into a better way in response to the profits and needs for the electric power companies.

- **Service Menu Planning**: CRM provides measures to deliver good service menus for preferred customers. In strictly speaking, offering discount charge and added-value services to what the company considers preferred customers (those who generate high revenue by consume high active power and low reactive power).

- **Income Analysis**: CRM also enables to analyze how both electric power income and related income change, and investigate the contents of the services.

- **Service Response Analysis**: CRM measures whether customers are satisfied with the service provided or not and reflects the results in the service menu planning.

3.9 Customer Relationship Management in the Ethiopian Electric Power Corporation

In a computation environment, Customer Relationship Management (CRM) is very important. CRM involves assessing customers yielding profits and constructing and adjust the relationships with these customers through the implementation of ideal measures directed at the customers. In addition to market-driven approaches looking at what market segments the traditional marketing serves, CRM is also a customer-based approach that continuously provides products aiming at improving customer satisfaction (Kitayama et al. 2002).

The Ethiopian Electric Power Corporation (EEPCo) is a corporation that allowed national monopolies in the background of scale merit. It was unnecessary for electric power utilities to deeply understand customers, since the Electric Utility Industry Law imposed on these utilities the legal obligation to supply electric power to customers. However, in order to effectively oppose new entrants in liberalized sector in the future, it is necessary

for this government monopolize utilities to recognize or identify preferred customers based on their profiles (e.g., amount of electric power energy consumption, load factor and etc) and not merely view all customers uniformly.

As it is evident in its vision, mission and objective the corporation has a strong commitment for realizing customer satisfaction. However, in reality strong criticisms are repeatedly forwarded from many customers and the private presses that the electric power corporation services provided now are not to be sufficient for customers as the current dynamic environment requires. It is not difficult to see that the reason for such strong criticism is the fact that the strategic polices are not implemented well through integration of people, process, technology, and other resources. Tacking this fact into account seriously, EEPCo has been designing and implementing various programs that enhance its operational efficiency and effectiveness. One of the core projects in the way of implementation in this regard is the business process re-engineering (BPR).

CHAPTER FOUR

DATA MINING METHODS FOR CUSTOMER CLUSTERING AND CLASSIFICATION

4.1 Introduction

Data mining techniques are most useful in information retrieval. Some of these techniques are classification, association rules, clustering, regression and etc. The task of customer segmentation is done mostly using clustering and classification techniques in data mining.

Classification is a very important and frequently used technique in data mining. It is a process of finding a set of models that express and distinguish data classes or concepts into pre-defined class. While clustering is a technique in which data is clustered into groups that are somehow similar in characteristics. Clustering is often confused with classification, but there is a difference between the two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are formed by the algorithm autonomously (classes are formed without pre-defined classes) (Rao 2003).

4.2 Clustering Techniques and Algorithm

Clustering in data mining is a discovery process that groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized (Fekadu 2004). It is to find groups that are very different from each other, and whose members are very similar to each other.

Clustering technique is also known as unsupervised learning. In unsupervised learning, the problem is to group 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 cluster the dataset 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.

➢ Basic Clustering Step

✓ Preprocessing and Feature Selection

Most clustering models assume that n-dimensional feature vectors represent all data items. This step therefore involves choosing an appropriate feature, and doing appropriate preprocessing and feature extraction on data items to measure the values of the chosen feature set. It will often be desirable to choose a subset of all the features available, to reduce the dimensionality of the problem space. This step often requires a good deal of domain knowledge and data analysis.

✓ Clustering Algorithm Selection

Clustering algorithms are general schemes, which use particular similarity measures as subroutines. The choice of clustering algorithms depends on the desired properties of the final clustering output and time and space complexity. A clustering algorithm attempts to find natural groups of data based on some similarity. The clustering algorithm also finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids.

✓ Result Validation

This step is justification of the algorithm result by the specified validation criteria. And, if not, iterate back to some previous stage and train and test the algorithm repeatedly.

✓ Result Interpretation and Application

The last step in clustering technique is interpretation of the result and applying of it for desired purpose. Some of the typical applications of clustering include data compression

(representing data samples by their cluster representative), hypothesis generation (looking for patterns in the clustering of data), hypothesis testing (e.g. verifying feature correlation or other data properties through a high degree of cluster formation), and prediction (once clusters have been formed from data and characterized, new data items can be classified by the characteristics of the cluster to which they would belong).

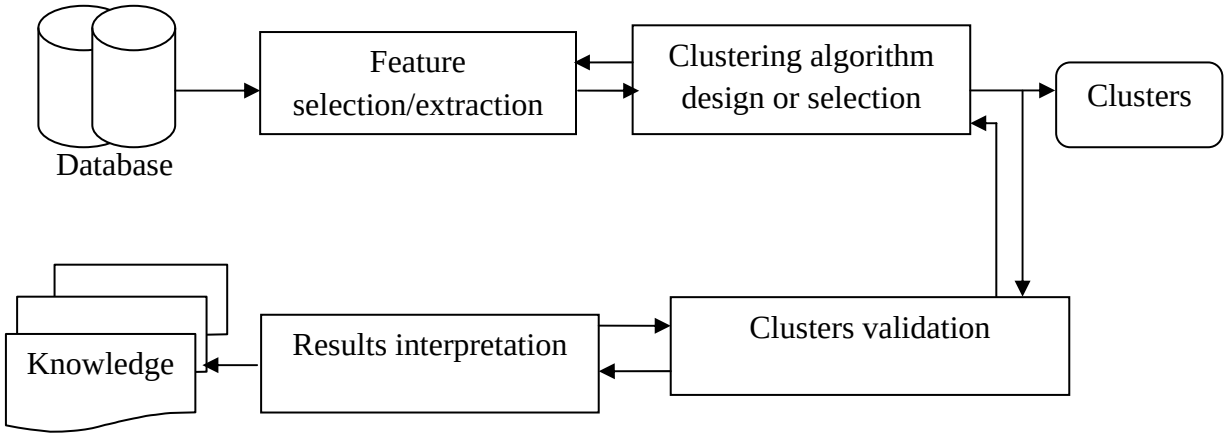


Figure 4.1: Clustering procedure. (Source: Xu and wunsch (2005) survey of clustering algorithms)

Clustering techniques are broadly divided into hierarchical and partitioning (Rao 2003). Hierarchical clustering is further subdivided into agglomerative and divisive. Agglomerative, starts with the points as individual clusters and, at each step, merges the most similar or closest pair of clusters. Divisive starts with one, all-inclusive cluster and, at each step, splits a cluster until only singleton clusters of individual points remain. In other words, it begins with all patterns in a single cluster and performs splitting until a stopping criterion is met. In this case, there is a need to decide, at each step, which clusters to split and how to perform the split. Hierarchical algorithms form a tree like structure either in a bottom up (agglomerative approach) or top down (divisive approach).

In contrast to hierarchical techniques, partitioned clustering techniques do not construct a tree like structure, rather create a one level partitioning of the data points. In other words, a partition clustering algorithm obtains a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique. A

problem accompanying the use of a partition algorithm is the choice of the number of desired output clusters (Jain et al. 1999). If *K* is the desired number of clusters, then partitioned approaches typically find all *K* clusters at once. There are a number of partitioned clustering techniques, K-means algorithm is one and very popular in data mining. It is based on the idea that a center point can represent a cluster. In particular, for K-means we use the notion of a centroid, which is the mean or median point of a group of points. It should be noted that a centroid almost never corresponds to an actual data point.

The K-means clustering algorithm which is used in this research is used for customer segmentation by various researches even in power industry. Other partition (non-hierarchical) algorithms are Squared Error, Graph-Theoretic, Mixture-Resolving and Mode-Seeking. Partition methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally too expensive.

According to Jain et al (1999), there are different approaches to clustering data, and they described the taxonomy of clustering as shown below.

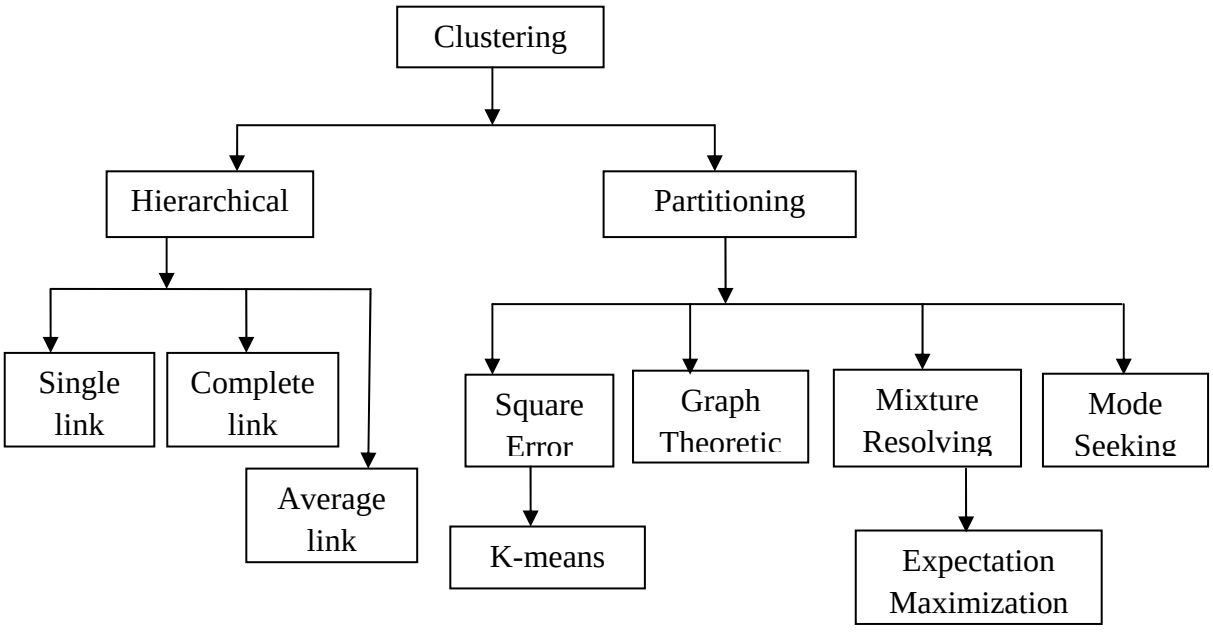


Figure 4.2: Taxonomy of clustering approaches. (Source: Jain et al. 1999)

Hierarchical algorithms are more flexible than partition algorithms. For example, the single-link clustering algorithm works well on data sets containing non-isotropic clusters including well-separated, chain-like, and concentric clusters, whereas a typical partition algorithm such as the *k*-means algorithm works well only on data sets having isotropic clusters (Jain et al.1999). However, the time and space complexities of partition algorithm are typically lower than that of hierarchical algorithm.

According to Jain et al. (1999), the availability of vast clustering algorithms in the literature confused users attempt to select an algorithm suitable for the problem at hand. Some of the criteria that used to compare clustering algorithms are based on: (1) the manner in which clusters are formed, (2) the structure of the data, and (3) sensitivity of the clustering technique to changes that do not affect the structure of the data. It should be noted that other researchers also defined additional clustering algorithm selection criteria.

4.2.1 The K-Means Method

The K-means clustering algorithm is a non-hierarchical one and it was developed by Macqueen in 1967. The k-means algorithm is the simplest and the most commonly used algorithm (Macqueen 1967). It is popular and the most applied in practice because it is easy to understand, implement, and its time complexity is $O(n)$, where n is the number of patterns. This algorithm works based on the concept of distance and partitions the data set into predefined number of clusters (K). It initially assigns points randomly for each cluster and calculates cluster centroid for each cluster. At this point, the cluster centroids are the same as the value of randomly selected/formed vector for each cluster. Then, a point (object, case, record) in the data set is taken and assigned to a cluster having the closest centroid to the object and cluster centroids are updated for the change (iteration) based on the distance of the point from cluster centroids. This process continues until all data points are assigned to given cluster. At the segmentation, formation (partition of the data set into groups) is completed and, the analysis and interpretation of each cluster can be done with domain expert. K-means algorithm is primarily suitable to clustering a data

set containing variables with numeric values. It is also possible to adjust the algorithm to fit to a data set with categorical attributes (Angoss 2003).

As also described by Mingoti and Lima (2006), the clustering algorithm starts with *k* initial points of clustering one for each cluster, then all the N (total) objects are compared with each point by means of the Euclidean distance and assigned to the closest cluster point. The procedure is then repeated over and over again. In each stage the point of each cluster is recalculated by using the average vector of the objects assigned to the cluster. The algorithm stops when the changes in the cluster points from one stage to the next are close to zero or smaller than a pre-specified value. Every object is assigned to only one cluster. However, a major problem with this clustering algorithm is that it is sensitive to the selection of the initial partition (Jain et al. 1999) and the accuracy of the K-means algorithm or procedure is very dependent upon the choice of the initial points (Mingoti and Lima 2006). To obtain better performance the initial points should be very different among themselves. But K-means clustering algorithm generates initial cluster randomly. If random initial starting points close to the final solution, K-means has high possibility to find out the cluster center. Otherwise, it will lead to incorrect clustering results (Cheung 2003).

And, also According to Saarenvirta (1998), another challenging thing in this clustering process is determining of the ‘optimal’ value of clusters (K). This value may also depend on the organizations capacity to manage how money customer groups at a time. Thus, the optimal value of K can then be obtained by segmenting the dataset repeatedly into acceptable values and comparing the results together with the business analyst or domain expert in the organization.

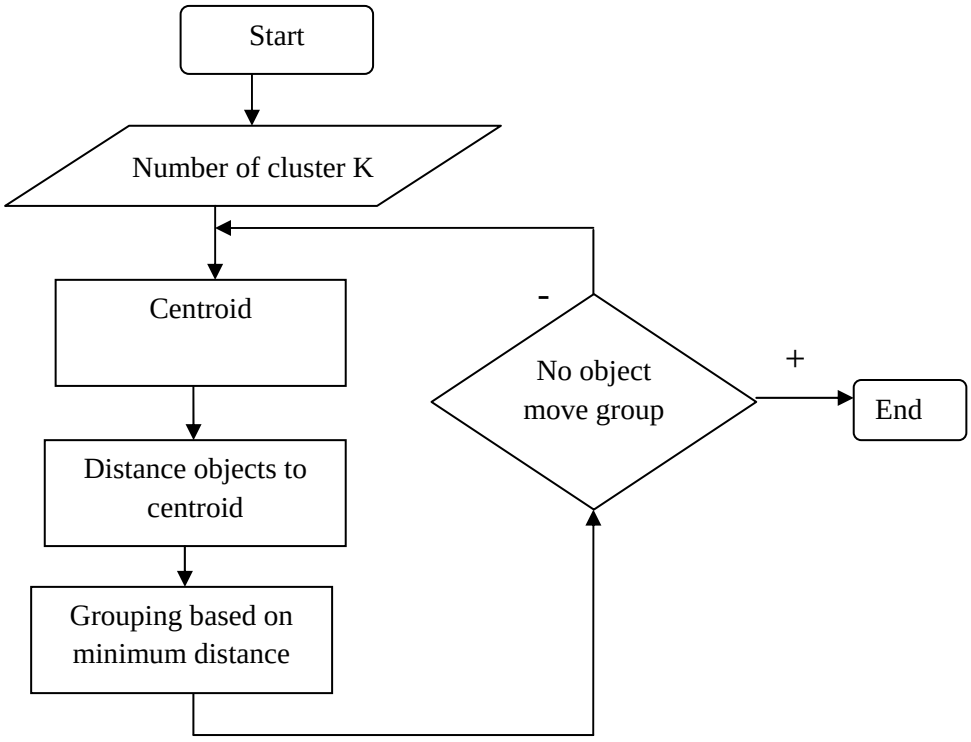


Figure 4.3: The K-means clustering process

Although the usability of K-means clustering algorithm is limited by its weakness (the clustering result is heavily dependent on the selection of the initial centroids points and the number of clusters (k)), it is famous and used in various fields due to its simplicity, high speed, and efficiency in clustering large data sets. It is a partition clustering method that separates data into *k* mutually extreme groups. As discussed above, by iterative partitioning, K-means minimizes the sum of distance from each data to its clusters. It is also a basic framework for developing numerical or conceptual clustering systems because various possibilities of distance and prototype choice.

4.2.2 Cluster 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 cluster label as the target variable and using it to generate rules explaining how to assign new records to the correct cluster.

4.2.3 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 in a given dataset and clustering algorithm with different input parameter values. So, there is a need to decide the best clustering that fits the dataset and the business under concern.

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

4.3 Classification Techniques and Algorithm

According to Rao (2003), classification is the most important and popularly used technique in data mining. It is a process of finding a set of models or pre-defined conditions that describe and distinguish data classes or concepts.

Supervised learning method is alternative term to express the classification technique. In supervised learning (classification), we are provided with a collection of labeled (pre-classified) patterns and the problem is to label a newly encountered, yet unlabeled pattern. The given labeled (training) 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) rules, decision tree, neural networking, bayesian networks etc.

4.3.1 Decision Tree

"A decision tree is predictive modeling technique used in classification, clustering, and prediction tasks. It uses a divide and conquers technique to split the problem search space into subsets" (Dunham 2000). A decision tree is a classifier expressed as a recursive partition of the instance space. Decision tree is used in data mining to classify objects into values of the dependent variable based on the values of independent variables. According to Fekadu (2004), there are two main types of decision trees. Theses 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 tree used to predict continues variables (variable 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.

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. A typical decision tree is shown in the figure below

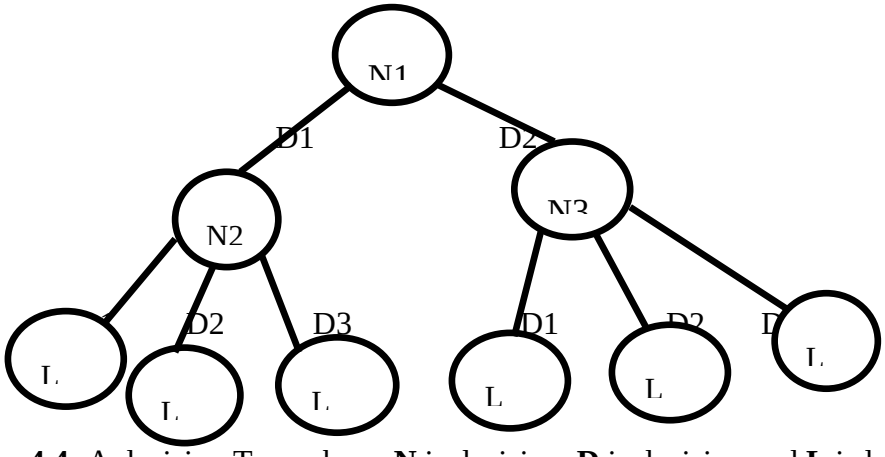


Figure 4.4: A decision Tree where, **N** is decision, **D** is decision and **L** is leaf

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.

4.3.2 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 (Han and Kamber 2001). 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. But, this is not always true.

There are different basic methods of attribute subset selection which includes the following techniques, where the stopping criteria for those different techniques may vary (Denekew 2003).

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.

The “best” (and “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.

4.3.3 How to Prune Decision Tree?

Constructing trees without limiting the depth of the tree can take longer and became meaningless. The problem can be controlled by rules that can stop or limit trees growth. One of the common stopping rules is simply limit the maximum depth up to which a tree may grow. Another stopping rule is to establish a lower limit on the number of records in a node and by saying “don’t splits” below this limit. According to Han and Kamber (2001), an alternative from using stopping rules or criteria is to prune the tree. In prune method, the tree is allowed to grow to its full size and then, using built-in heuristics or user intervention, the tree is pruned back to the smallest size but that does not compromise accuracy.

There are different decision tree algorithms like Id3, J48graft, AD tree, C4.5, J48 etc. J48 algorithm is WEKA’s improved implementation of C4.5 algorithm.

The process of the J48 algorithm to build a decision tree is as follows:

1. Choose an attribute that best differentiates the output attribute values.
2. Create a separate tree branch for each value of the chosen attribute.
3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
4. For each subgroup, terminate the attribute selection process if:
    a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.
    b. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.
5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

4.3.4 Bayesian Network Classifiers

Bayesian networks are graphical models which are very useful for representing variables (as nodes of the graph) and the probabilistic relationships between them (as connections, or edges of the graph). By knowing the value at one of the nodes in a Bayesian network, one can infer the value of other nodes in the network. Bayesian network classifiers are used in many fields and one common class of classifiers are Naive Bayes classifiers. The induction of classifiers from data sets of pre-classified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. One of the most effective Bayesian network classifiers, in the sense that its predictive performance is competitive classifiers, is the Naive Bayesian classifier (Friedman et al. 1997). This classifier learns from training data the conditional probability of each attribute *Ai* given the class label *C*. Classification is then done by applying Bayes rule to compute the probability of *C* given the particular instance of *A*1,…., *An*, and then predicting the class with the highest posterior probability.

According to Elkan (1997), on many real-world datasets naive Bayesian learning gives better test set accuracy than any other known method, including backpropagation and C4.5 decision trees. Also, these classifiers can be learned very efficiently.

Bayesian networks can have different advantages. Among those, some of them are provide probabilistic output, can operate with limited sensor data availability, more flexible relative to engineering development then traditional expert systems, used for both data qualification (state recognition) and anomaly reasoning, can operate in a central or distributed run-time environment either shore-side or ship-board.

The reason why use bayesian networks is Bayesian inference methods have proven to be valuable for knowledge-based data mining applications, and are based on a causal (explanation based) modeling framework. Because relationships between variables in a Bayesian network are defined probabilistically, trends can be detected and analyzed over a continuous scale, rather than in a Boolean fashion.

4.4 Criteria for Evaluating and Selecting of Data Mining Software

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

- **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 the software**.**

- **Functionality*:*** The inclusion of a variety of capabilities, techniques, and methodologies for data mining. This 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**,** data type flexibility**,** algorithm modifiability**,** data sampling**,** reporting**,** and model exporting**.**

- **Usability*:*** 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 includes 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 include 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.

4.5 Related Research Works

There are different researches that have been done on the applications of data mining in electric power corporation in different country context. And, this section reviews the techniques used in those studies in determining the customer classes in an electricity market.

In the Ethiopian context, Gobena in 1999 applied data mining technology and techniques in the Ethiopian Airlines, Henok extended Gobena's work in 2002. Askale in 2000 applied data mining in the financial industry, specifically at the Dashen Bank; Melaku Girma in 2009 used data mining techniques to support CRM at Ethiopian Telecommunication Corporation Code Division Multiple Access (CDMA) telephone service, and Tesfaye in 2002 also undertook research applying data mining in Ethiopian Insurance Company. Moreover, Shegaw in 2002 has assessed the potential applicability of data mining technology in the Ethiopian context with particular reference to the health sector. In the context of other countries, Nizar et al. (2006) investigated on determining the best load profiling methods from data mining techniques to classify, detect and predict non-technical losses (faulty metering and billing errors) in the electric power sectors, as well as to gather knowledge on customer behavior and preferences so as to

gain a competitive advantage in the market. They also make a comparative analysis on three clustering methods (K-Means, Cobweb and EM) for the purpose of clustering/classifying the electricity customers into several groups of clusters. From their results, it is shown that the three clustering techniques gave different results in terms of number of clusters. However, the performance on each clustering technique also gave a significant impact in terms of cost. And, in the testing of performance, simple K-means clustering technique performs better than others in the datasets.

Another research that has been done in the electric power industry is on the title "Clustering Algorithms and Self Organizing Maps to Classify Electricity Customers" by Chicco et al. 2003. According to them, in competitive electricity markets, the electric power providers have been given new degrees of freedom in formulating dedicated tariff to be applied to properly defined customer classes. For this reason, they may take advantages from identifying the power consumption patterns of their customers and grouping together customers exhibiting similar load diagrams using data mining techniques. Having this idea, Chicco et al. (2003), investigated their paper on the effectiveness of various unsupervised clustering algorithms such as modified follow-the-leader, k-means, fuzzy k-means and two types of hierarchical clustering and the Self Organizing Map (SOM) to cluster or group together customers having a similar electrical power consumption behavior. And, the results of the research on capability (performance) assessment show that two algorithms – the modified follow-the-leader and the hierarchical clustering are the most capable and perform better than the other algorithms in terms of adequacy (performance). In other words, both algorithms are able to provide a highly detailed separation of the clusters, isolating customers with uncommon behavior. Finally, an overall evaluation of the clustering algorithms leads to consider the modified follow-the-leader as the most efficient one, on the basis of both clustering adequacy and computational speed.

There is also a paper done by Valero et al. (2004) on the capacity of Self-Organizing Map (SOM) algorithm as a methodology to cluster the electric power costumers in Spain. This algorithm enables to extract the pattern of customer behavior from historic load demand

series. Even though there are many ways of data analysis from load demand curve to get different input data to feed the algorithm, the researchers in this paper proposed two methods to improve customer clustering such as the use of frequency-based indices and the use of the hourly load curve. After conducting the experiment, they conclude that the hourly load profile training provides a clear map with different clusters and places similar customers in the same group. The index success ratio shows the identification success capacity. They finally, conclude that the identification of new customer is of a high quality in both alternatives: hourly load profile and frequency-based indices. However, the other treatments are also very good and they obtain a quite high rate of success. The Self-Organizing Maps (SOM) appears as an interesting clustering algorithm for electrical customer's segmentation, with a broad application field. And they recommend that the electrical market segmentation could be studied using the Self-Organizing maps, training the network with a great quantity of daily load data.

There are different researches conducted in determining the customer classes in an electricity market particularly in Taiwan, Portugal, and UK with different technique and most of the researchers proposed K-means clustering technique because of the performance of K-means is better than other clustering techniques (low time and space complexity).

However, nothing was done in our country to cluster and classify the electric power customers by power consumption pattern. And, in this study the researcher investigated customer segmentation and classification in the Ethiopian Electric Power Corporation (EEPCo) for successful CRM implementation. Because the power consumption pattern and tariff structure of different countries may not be similar. So, there is a need to conduct clustering/classification data mining techniques in the Ethiopian context. Thus, in this research the researcher clusters electric power customers and to use the clustered output for classification purpose. The technique that was employed is chosen based on scalability to larger dataset, time and space complexity and result presentation for interpretation. In addition, recommendation of different researchers has been considered.

Simple K-means algorithm has been used for clustering because this algorithm is proposed by many researchers for electric power customer profile analysis. In addition, Simple K-means can show the good performance, both in precision and speed. For example, according to Nizar et al. (2006), K-means technique performs better than Cobweb and EM. According to McQueen (1967), the *k*-means is the simplest and most popular and commonly used algorithm, because it is easy to implement, its time complexity is *O (n)*, where *n* is the number of patterns and it is order-independent.

And, for classification purpose the researcher employed decision tree classification techniques. The reason to use decision tree is that it is more comprehensible or understandable than other classifiers for decision makers. Moreover, decision tree is most often used method in classification and estimation of customers. To compare the decision tree classification algorithm performance, the researcher also employed Naive Bayes classification algorithm.

4.6 Customer Segmentation in Electric Power Industry

Like other industries the idea of customer segmentation in the electric power industry is applicable. However, in electric industry, first, it is important to judge the value of customers based on the amount of electric power use and other important attribute values. Different industries may use different method of measurement for their customer values. For example, in the retail business a measurement method of the value of customers is the RFM analysis approach, which carries out analysis focusing on frequency, and monetary value from customer purchasing data. The electricity power industry products differ from retail industry products and the response to customers in relationship is also different. Because of this, the RFM analysis approach from the retail industry cannot be applied as is, as a method of analyzing customers profile in electric power industry (Kitayama et al. 2002). Here, in this study the researcher used yearly electric power consumption pattern to cluster and classify customers in the corporation and to use it for successful CRM marketing strategies. The customers of EEPCo are categorized as domestic, commercial, street light, industrial, active staff, retired staff and own consumption customers.

CHAPTER FIVE

EXPERIMENTATION

5.1 Introduction

In this chapter the researcher describes the data mining goals, sources of data and the techniques that have been used in preprocessing and model building phases. It 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.

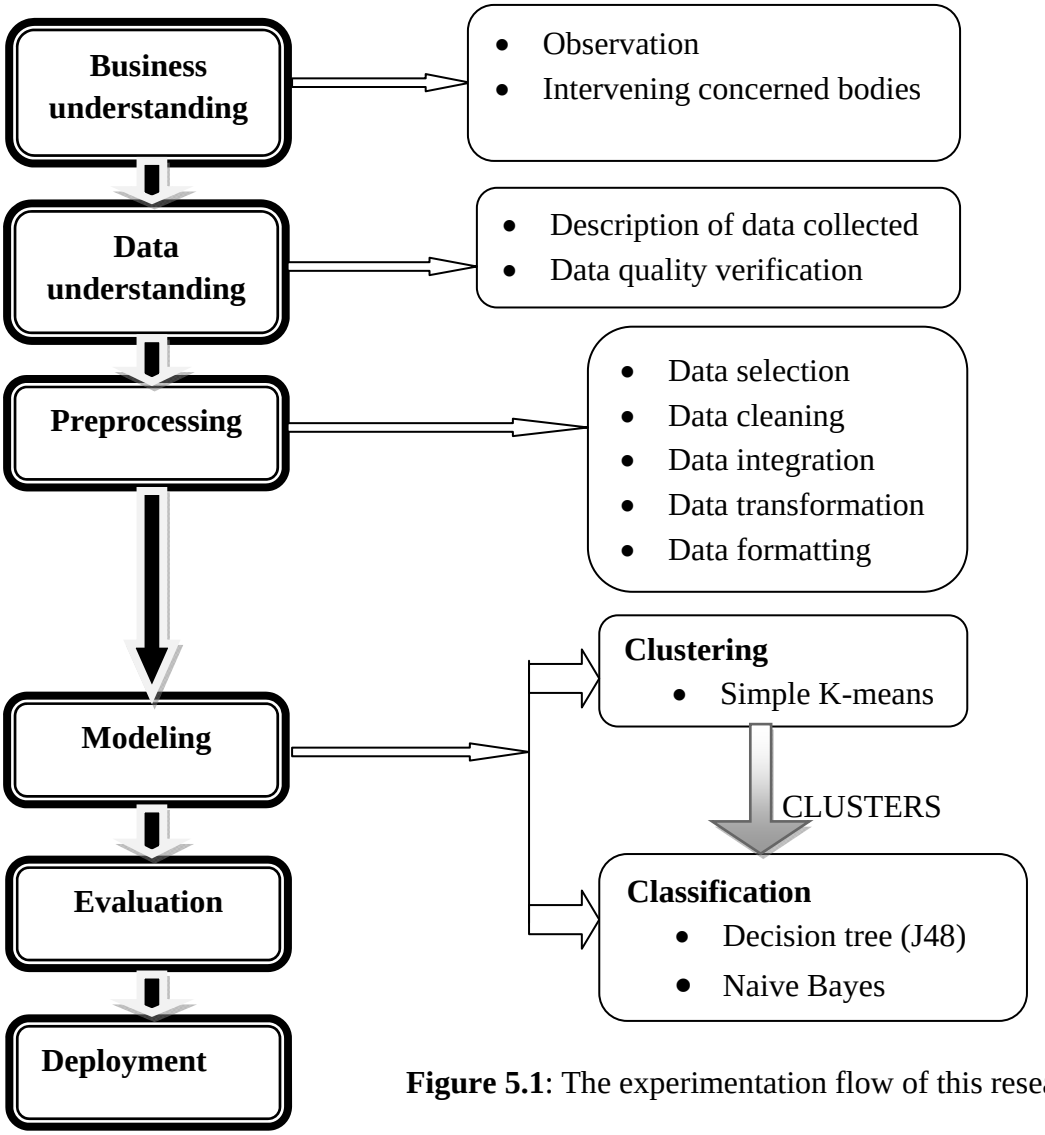


Figure 5.1: The experimentation flow of this research

According to Thearling (2000), creating and identifying market segments or customer groups based on their value to the organization is the first task in order to implement successful CRM. Since no research has been carried out before, the data from the corporation database has been used for the two phases of this research, customer clustering and classification. Accordingly, the main objective of this research was to provide a model that clusters customers and then classifies new customers with respect to the cluster index (important dimensions of customer's behavior). In other words, to segment customers based on their behavior and develop a model that assigns new customers into appropriate cluster. And, this enables the organization to identify the most profitable customers and to treat them according to their needs (CRM).

5.2 Understanding of the Business

Based on observation and interviewing data collection techniques, the researcher understands the business or domain of this research. This phase focused on understanding the research objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

There are different types of customers in the Ethiopian Electric Power Corporation (EEPCo). And, the corporation is interested to understand the customer's value and segment them in order to make relationship with them accordingly. The value or long run profitability of customer to the corporation can be evaluated based on some variables. These variables are power consumption (active and reactive), revenue generated, power demand, power factor and other related variables that can be generated to see weather these variables may yield better data mining results.

Hence, high value corporate customers are customers that are using a high amount of active electric power, low reactive power consumption and high revenue generated. Customers characterize medium active consumption, medium reactive consumption and generates medium revenue are medium value customers to the corporation. Finally, low

value customers are customers that are using low active electric power, high reactive power consumption and generate low revenue to the corporation.

The variables that determine customer value in the corporation have been used to create the customer segments, and then classification rules. This research enables the corporation to design and offer different service strategies for different customer segments accordingly. The most appropriate data mining techniques which are clustering (or segmentation) and classification have been used for this purpose.

5.2.1 Selection of Data Mining Tool

Different researchers develop a framework to evaluate and select an appropriate data mining tools. However, the evaluation and selection of an appropriate data mining tool for this research was done based on certain criteria. The researcher had to first set criteria for tool selection and the criteria used in this research to select one tool from the other were the following:

- Platform Variety (Does the software run on a wide-variety of computer platforms?)
- Performance of the tool in terms of speed and quality
- Algorithmic Variety (inclusion of various clustering and classification algorithms)
- The data mining tasks that the tool is intended for
- The compatibility of the tool to the operating system at hand (MS window)
- The possible formats for the data that is to be analyzed

Although there are different software that fulfill the criteria stated above, in this research the tool or software used is WEKA version 3.6.4 by which many data mining researches and projects have been done and implemented.

WEKA tool contains several data mining techniques including classifiers, clustering algorithms, association rules and functions. Decision trees, production rules, NN, the

apriori algorithm, bayes classifier, linear regression, logistic regression and the K-Means algorithm are some of the more popular techniques included in WEKA tool. WEKA also has nice graphical features and many preprocessing capabilities. Moreover, Waikato Environment for Knowledge Analysis (WEKA) is freely downloadable open source software.

5.3 Data Understanding

In this phase, to understand the data available and useful for achieving the goal specified, the secondary data collection technique called database analysis has been employed and the content and structure of the data available was understood. First, raw data was initially collected or taken from customer database of EEPCo. And, a careful analysis of the data relevancy and its structure was done together with domain experts by evaluating the relationship of the data with the problem at hand in the organization and the particular data mining task to be done. The following section describes the initially collected nature of the data and its structure.

5.3.1 Description of the Initial Data Collected

As indicated above, the major data source of this research was the customer database of the EEPCo. The corporation customer database contains information about customers in different tables. However, the relevant data to carry out this research have been collected or integrated from five tables of the database. The description of this data source tables with their attribute and data type is depicted in **Table 5.1, Table 5.2, Table 5.3, Table 5.4** and **Table 5.5,** respectively.

Attributes	Data type	Description
NAME	String	Name of the customer
NIS-RAD	Number	Service number
ACC-NO	Number	Old account number

Table 5.1 sumcon table

Sumcon table is a table that contains customer information related to contract information. In this table, there are different variables with their values, and from those variables the initially selected for the data mining task of this research were: name of customer or organization, service number and account number. These attributes were selected from this table based on domain expert's advice and values for this research objective.

Attributes	Data type	Description
CTE-CSMO	Number	Consumption (meter) constant
YEAR	Number	Year of power consumed
DIF-LECT	Number	The difference between actual and previous consumption
CSMO	Number	The difference between actual and previous multiplied by meter constant (active consumption)
CYCLE	Number	The month in which the meter reading is read
RC-CO	Number	The difference between total power supplied and active consumption multiplied by meter constant (Reactive consumption)

MAX-DEM	Number	Maximum power demand supplied by the corporation in the month (Maximum demand)
PFC	Number	The highest load supplied by corporation to the customer in the month.(Power factor)

Table 5.2 apmedadida.co table

The monthly customer's power consumption data is stored in the corporation database. And, thus the apmedadida.co table contains this information. Although there are many variable in this table, the initially selected variables for the data mining task of this research were: year, cycle or month, maximum demand, active consumption, reactive consumption, power factor and consumption (meter) constant. More of the attributes important for this data mining research objective has been taken from this table, because this research objective is focusing on customers power consumption patterns and relationship with the service provider (EEPCo in this case).

Attributes	**Data type**	**Description**
COD-TAR	Number	Code tariff
DESK-TAR	String	Description of the tariff
TIP-SER	String	Service type
CODE-TAR-OFI	Number	Official code for the tariff

Table 5.3 mtarifas table

The mtarifas table in the corporation customer database contains information related to tariff system and code of tariff with their description. The selected variable from this table for this research purpose was official code for tariff. This official code for tariff was used to identify the customer type in the corporation. In other words, there are different types of customer in the corporation, like industrial customer, commercial customer,

domestic customer, street light customer, active staff customer and retired customer. So, the researcher was used this official tariff code to identify the customer type.

Attributes	Data type	Description
NUM-REC	Number	Bill number
NIS-RAD	Number	Service number
F-FACT	number	Billing date
IMP-TOT-REC	Number	Total amount of the bill
PFC-CH	Number	Power factor charge

Table 5.4 recibos table

At the end of each month customers are expected to pay their monthly power consumption charge. And, this data is recorded in recibos table, which contain customer's monthly billing information in the corporation customers' database. This table contains variables, such as billing number, billing date, billing month, billing year, total billing amount, etc. However, the initially selected attributes for data mining task in this research are total amount of bill and power factor charges.

Attribute	Data type	Description
R-REC	String	Region
S-SER	String	Service center

Table 5.5 Unicon Table

The unicon table in the customer database contains information related to areas and locations in which the service is provided to the customer. Although there are many

variables in this table the initially selected variables for this data mining research task were: region and service center.

The names of attributes in the tables were abbreviation, and to make it clear in the experimentation the researcher used the description of attributes as the name of attribute.

The data collected for analysis in this data mining research was customer's power consumption pattern of twelve month (one year), from March 2010 to January 2011.

5.3.2 Data Quality Verification

As the data is produced electronically the reliability and completeness of records are relatively good. However, the data collected may contain missing, incomplete and irrelevant data. Hence, in many of these cases, the data which is missed and irrelevant for data mining task to be carried out should be adjusted at preprocessing step and the irrelevant data should be removed. WEKA version 3.6.4 has the capability to preprocess such operations.

5.4 Data Preparation

In the observational setting, data are usually collected from the existing databases, data warehouses, and other data sources. However, preprocessing is a challenging and time taking task, especially in large databases. It is not only time consuming to specify a preprocessing operations but also to apply it on a large databases. The data in the real world is dirty they may contain incomplete, noisy (outlier), inconsistence and irrelevant data. "No quality data, no quality mining results". So, this stage involves a serious of steps to provide the final dataset from raw data for modeling purpose. According to Han and Kamber (2001), data preprocessing phase of data mining process includes: data cleaning, data integration, transformation, data selection and data formatting.

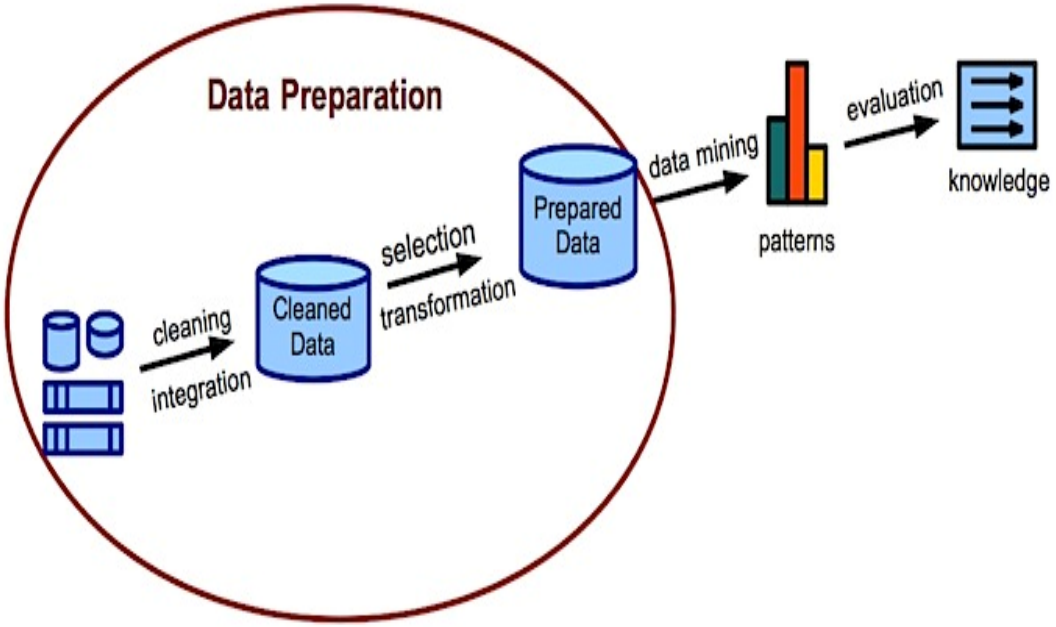


Figure 5.2: Data preparation steps

5.4.1 Data Cleaning

Data cleaning is one of the activities in data preparation phase and it has to be done before going to drive new attributes from the basic ones. Data cleaning is removing of records that had incomplete, missing, duplicated, inconsistent data and irrelevant data under each attribute column. There are different methods used to handle the missing values, such as ignoring the tuples, filling the missing values by using the modal value (for nominal and ordinal variables) and the mean (for continuous variable). And, in this research the missing data was filled by using WEKA preprocessing facility “replace missing values with modes and means from the training data” and removed manually, duplicated attribute values that do not vary at all or that vary too much had been also removed. There are two strategies for dealing with outliers: detect and eventually remove outliers as a part of the preprocessing phase, or develop robust modeling methods that are insensitive to outliers. In this research, removing of incomplete and outlier data had been done manually as part of the preprocessing phase. Active staff, retired staff and own consumption customer groups were deleted from the dataset because the domain experts advise that this consumptions are used in offices and by the staffs of the corporation. So,

they are not important for the data mining objectives of this research, rather they are outliers. The expert also advises domestic customers are active power customers.

5.4.2 Data Integration

The data contained in different tables of customers profile database are raw data, which is not in the form of appropriate for the business goal to be addressed and the corresponding data mining task to be done in this research. So, this data integration step is important. Data integration is the activity of merging data from multiple sources into coherent data store. In other words, integration of multiple databases data cubes, or files together. The integration of data in this research was made by using the service number attribute, which represents each customer as a key attribute. The data sources tables for data integration were: Sumcon table, which contains customers contract information, apmedida.co table contains customers monthly power consumption, recibos table containing billing information that customers generated in each month with other information, mtarifas table containing tariff code with their descriptions, and finally unicom table that contains data related to regions and locations in which the services is provided to the customers.

The customer data collected and integrated from different tables of the database and ready for data mining techniques to be undertaken in this research looks like the following.

Field name	***Data type***	***Description***
Region	String	A region in which the service is provided to the customer.
Service center	String	Service centers to the corporation are centers from which customers of electric power corporation can access the corporation service.
Tariff code	Number	Official code for the tariff. This code was used to identify customer type (com, industry, street etc).

Service No	Number	The number to identify the customer uniquely.
Account No	Number	The customers account number.
Name	String	The name of the customer or organization.
Year	Number	The year that the customer is provided the service.
Cycle	Number	The month in which the meter consumption is read by the suppliers (EEPCo in this case).
Power factor	Number	The trigonometric ratio of active and reactive power consumption.
Maximum Demand	Number	The highest load supplied by corporation to the customer in the month.
Active consumption	Number	Active consumption is the final electricity consumption of the customer recorded in the premises (measured in kilowatt hour (kWh) meter) of the customer applied directly for production/services purpose but excluding the losses indirectly incurred to the supplier (EEPCo in this case).
Reactive Consumption	Number	Reactive consumption is electricity consumption of the customer recorded in the premises (measured in kVAR meter) but it is uneconomical use of the scarce resources on the customer side. It is the difference between the electricity supplied and the electricity converted into useful power.
Active constant	Number	The actual corresponding values of customers active electricity consumptions.
Reactive constant	Number	The actual corresponding values of customers reactive electricity consumptions.
Power factor charge	Number	The multiplication of MD, difference of power factor below 0.85, and existing power factor charge tariff.
Active reactive	Number	The division of active power consumption by reactive power consumption.

Total revenue	Number	The total amount of revenue that customers generated to the corporation.
Powed	Number	The multiplication of power factor by power demand.

Table 5.6: List of all attributes with their corresponding type and description

After combining all relevant tables in the corporation customers' database for this research purpose, the next task to be done was driving relevant attributes from the basic once.

5.4.3 Data Transformation

Data transformation is transforming the data from low level or primary data into higher level concepts. In this activity normalization techniques can be applied. Since normalization may improve the accuracy and efficiency of data mining algorithms, it is better to apply this activity in the preprocessing phase. There are also other preprocessing activities like data reduction, data discretization and data construction (drive attributes from basic attributes). All Bayes network algorithms implemented in Weka assume the following for the dataset: all variables are discrete finite variables and no instances have missing values. And, since the dataset of this research was continuous variables the researcher Discretize them to make appropriate for the algorithm. The missing values were filled by "replace missing values" filter facility in data cleaning activity of data preparation. Data creation involves the creation of new variables by combining existing variables to form ratios, difference and etc. (Saarenvirta 1998), and this can improve the data mining algorithms. In this research, the active reactive and powed attribute were derived from existing basic attributes. The active reactive attribute was obtained by dividing active power consumption with the reactive power consumption and powed attribute was obtained by multiplying power factor with monthly power demand.

5.4.4 Data Selection

From 85,709 electric power customer records taken from the corporation database, a total of 50000 records were used for this research. After removing outlier and irrelevant data

from the originally collected data, there were only 50000 records. So, the researcher used this total number of records for data mining task in this research. The clustering task would be better if it is done by total records in the corporation customer database but due to outliers (irrelevant data) in the data and processing time, the experimentation of the research was conducted by 50000 preprocessed customer records only. However, since the size of the sample is relatively high it explores and generalizes about the data it needs relatively high number of records.

5.4.5 Data Formatting

Data formatting is the activity of changing the data format into a format suitable or understandable by data mining tool (algorithm).

For WEKA to analyse datasets, it needs a format that users can input so that it can understand the structure of the data. The ARFF format is what WEKA uses and is very simple. First, the data should be prepared in MS-excel and then saved with the format Coma Separated Value (CSV*)* and the file is opened with MS-word then the header should be labeled. The header defines the name of the dataset along with the set of attributes and their associated types. Finally, the data is saved into ARFF format which is suitable for the data mining tool or algorithm.

5.5 Modeling

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

5.5.1 Selection of Modeling Techniques

A good segmentation model can divide customers into homogeneous group on the basis of shared common attribute. Clustering technique is a direct data mining technique

(where there are no predefined classes to be predicted) and the instances are to be divided into natural groups. After the clusters are identified, new customers should be classified to one of these cluster indexes. In this research, automatic cluster detection (simple K-means clustering algorithm), decision trees and Naive Bayes are the selected data mining techniques for customer segmentation and classification, respectively. The selection of the techniques was made because of widely applicability for segmentation/classification problem and well implemented in the selected data mining tool. For clustering purpose, simple k-means algorithm is selected as it is a very good general purpose clustering algorithm and is recommended for most situations. In addition, simple K-means is good in handling discrete and numeric attributes.

Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "income"). WEKA version 3.6.4 software provides filters facility to accomplish all of these preprocessing tasks and this is the reason that Simple K-Means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The WEKA Simple K-Means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

For classification purpose, the decision tree algorithm was selected as it is powerful, flexible, easy to understand, powerful visualization feature and well implemented in the selected data mining tool. Moreover, it can handle large amount of variables with either continuous or discrete dependent variable. There are a number of classification algorithms available in WEKA version 3.6.4 like KD tree, BF tree, J48, J48graft and others. However, the algorithm used in this research was J48. The reason to choose J48 algorithm are, it is most widely used algorithm, it has a visual description of the output, and it has the benefits of C4.5 algorithm. Naive Bayes classifier was also selected to compare the classification accuracy with the decision tree classifier.

5.5.2 Test 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 a 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 of the sample size (100%) or available datasets has been used to train and develop better clustering model. However, in case of decision tree and Naive Bayes 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 good accuracy performance from the two algorithms 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 expert in the corporation involved in the process of segmenting and interpreting the segment results. Therefore, the final segmentation provides good knowledge for designing and implementing appropriate CRM strategies in the corporation.

5.5.3 Model Building

The model building process of this research also consists of three activities namely: attribute selection, clustering, select the better clustering model, and finally building different decision tree and Naive Bayes classifier models and select the one that shows good accuracy.

The attribute selection activity involves selection and identification of best variables or attributes for segmenting and identifying customers from business perspective in which the organization involves and developing clustering/classification model in the Ethiopian Electric Power Corporation (EEPCo). Clustering is formation of clusters based on the selected attributes. Moreover, interpretation and analysis of cluster results were also made in this section (clustering). Next to this, there is selection of the clustering model

that shows better clustering performance based on the evaluation criteria. The final section is building and choosing the best classifier model using decision tree and Naive Bayes algorithms. This last section involves developing of rules by taking the final selected clusters as dependent variables.

5.5.3.1 Attribute Selection

Most machine learning algorithms are designed to learn the most appropriate attributes to use for making their decisions (Witten and Frank 2005). However, algorithms have their own limitations, such as considering insignificance attribute as crucial one and ignoring the most important attribute as irrelevant. Due to this limitation, the attributes for customer segmentation should be selected after they have been tested for their goodness for clustering customers in the organization from business perspective. So, different models were built having different values of 'K', different attributes number and different iteration to identify attributes which have high information content to cluster customers and that would enable to develop a better clustering model according to organizations business objective. Moreover, the domain expert in the corporation plays a great role in the selection of best attributes for developing clustering and classification model because, the expert is familiar with the importance of the attribute to identify the customers in the corporation from business perspective. So, there were variables to be discarded. However, it does not mean that this attributes have no importance rather this variables were believed to provide very little useful information for the problem at hand and since clustering is unsupervised learning they may reduce the accuracy of the clustering algorithm.

All attributes listed in the dataset were used for attribute selection experiments. The numbers of attributes in different experiments were different. And, the other parameters, the number of iteration (*i*) refers to the maximum number of times the algorithm reads the dataset to create clusters was also different. The range of iteration in the experiment runs from 10-10,000. One thing that should not be forgot is that as the number of iteration increases, the longer the algorithm run and the result would be more accurate. However, there is no change at all in WEKA software.

After comparing the different models built for the different values of k, having all of the attributes in the dataset with different iteration, the relevant attributes for the problem or objective of this study were selected. The selection of attributes had also considered other things like domain expert's opinion (belief), information gain attribute evaluation result and others.

The final attributes selected for the next step which is clustering customers were: power factor, power demand, active power consumption, reactive power consumption, active reactive, customer type (tariff code), service center and total revenue that customers generated to the corporation. As explained above, the reason to exclude other initially selected attributes like region, name of customer, account number, year, cycle and others they believed to have less significance for this research objective at hand, rather they reduce the clustering algorithm performance.

5.5.3.2 Clustering of Customers

This step involves clustering the dataset based on selected attributes and this clustering model should enable to develop the decision tree and Naive Bayes classifier models.
According to Qiu et al. (2004), clustering does not require pre-defined classes rather the records are grouped based on self-similarity. In other words, clustering is undirected knowledge discovery- no target variables (or dependent variables) are defined. And, the clustering result interpretation is up to the user, researcher and domain expert.

The challenging task in the K-means clustering algorithm is 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 segments (*K*) can range from 4 to 10 and also depend on the organizations capacity to mange various clusters properly. However, the optimal number of cluster size (the value of k) is obtained through trial of different clustering experiments by adjusting default parameters.

In this research, the researcher after discussed with the domain expert in the EEPCo, the value of *K* was set to be four to six. Hence, different clustering models were built at three different values of *k* (four, five and six) as well as different seed size using WEKA version 3.6.4. And, finally, the best clustering model based on specified criteria was selected.

5.5.3.3 Assessment of Clustering Customers Models

Since there is no actual quantitative definition of a good clustering output, assessing the clusters based on certain crucial attribute is reasonable. So, it was important to discus with experts at the EEPCo. The discussion was focused on assessing the most influential attributes from selected ones for clustering customers in the EEPCo. And, thus the expert discussed and stated the most important variables that are used to select or identify preferable customers in the corporation.

According to the domain expert, active power consumption, reactive power consumption, and revenue generated were given a high weight from selected attributes. As result, the analysis and interpretation of each and every cluster in the experimentation was highly depending on these attributes. However, it does not mean that other attributes have no importance, rather to express the weight given to this attributes by the domain expert in the corporation to cluster customers.

As explained above, as the number of iteration increase, the accuracy of the clustering algorithm is also increased. However, in WEKA version 3.6.4 there is no change at all i.e. different values of i (number of iteration) did not bring any change in the distribution of the segments. Due to this, the researcher compares the nine (where K=4, 5, and 6 with randomly selected seed size 10, 100 and 1000) clustering models without considering the number of maximum iteration as it couldn't affect the clustering result.

The comparison of the clustering model was done in a way that the average attributes values of each cluster in a model are compared to other cluster models, the number of iteration that the algorithm reads the dataset given to cluster, inter-class similarity error

(similarity error exists within a cluster) and the domain expert's judgment. There is no clearly stated boundary or description that demarks power customers according to their different attribute values in the corporation. So, to map the particular attribute average value into five discrete values the setting of range in each attribute was made together with domain expert's who is familiar with the values of each attribute in the corporation. WEKA's preprocessing window helps to display min/max value for each attribute. The five discrete values are very high, high, medium, low and very low. The reason to determine the range of attribute values is that it is the only means to interpret the clustering model output from business perspective and to understand the types of customers clustered in each cluster/segment.

The following abbreviations of variables or attributes are used for experimentation analysis

APF- Average power factor

ARC - Average reactive power consumption

AD - Average demand

AAR - Average active reactive

AAC - Average active power consumption

AR - Average revenue

The following abbreviations are used for analysis of the five discrete ranges.

VH – very high

L - low

H – High

VL - very low

M – Medium

Attributes	Very High	High	Medium	Low	Very Low
Power Factor	>=200	199>=1	0.99 > 0.05	0.04 >=0.005	<0.005
Demand	>=500	499>=100	99 >= 50	49>=1	<1
Active Consumption	>=1834367	1834366>=10000	9999>=3000	2999>=1000	<1000
Reactive Consumption	>=22569477	22569476>=589247	589246>=100000	99999>=20000	<20000
Active Reactive	>=58374	58373>=100	99>=50	49>=10	<10
Revenue	>=10000	9999>=7000	6999>=2000	1999>=100	<100

Table 5.7: List of range of conditions by which a cluster result was assessed

5.5.3.3.1 Experiment 1

segmentation (K=4 and seed size 100)							
Cluster #	Freq. of records (share in %)	APF	AD	AAC	ARC	AAR	AR
1	9291 (19%)	1.3853	96.1	71052.102	63909.80	106.37	495109744
2	5406 (11%)	0.7351	101	3671.5956	3427.072	1.4224	2420.0807
3	3137 (6%)	0.7607	17.9	16031.631	11269.98	2.1013	2714.7422
4	32166 (64%)	0.7177	13.8	2403.641	2514.719	1.4234	2535.1167
Total	50000(100%)						
Cluster #	Freq. of records (share in %)						
1	9291 (19%)	H	M	H	L	H	VH
2	5406 (11%)	M	H	M	VL	VL	M
3	3137 (6%)	M	L	H	VL	VL	M
4	32166 (64%)	M	L	L	VL	VL	M
Total	50000(100%)						

Table 5.8: Cluster description based on average values of attributes for K=4 and seed size 100.

The upper part of **Table 5.8** shows the attributes' average values in each segment (1-4) and bottom part of the table shows the corresponding mapping of these values into five discrete values or ranges.

After the average value of each attribute in each cluster has been replaced with the corresponding discrete values, a description for each segment of cluster has been done as summarized in **Table 5.9**. And, the ranking of cluster is determined based on the basic attributes and the profitability of custumers to the corporation.

Cluster	*Description*	*Possible Rank*
1	High power factor, medium demand, high active consumption, low reactive consumption, high active reactive and very high revenue generated.	1st
2	Medium power factor, high demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	2nd
3	Medium power factor, low demand, high active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	3rd
4	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	4th

Table 5.9: Cluster summary and corresponding ranks based on basic attributes for K=4 and seed size 100

As indicated in column three of the above table, it is possible to give ranks to each cluster based on the average values of attributes to provide information in each cluster, and by considering the values of customers to the business. As it can be clearly seen in **Table 5.9**, the first Cluster is ranked first. This is due to the fact that customers in this cluster are generated very high revenue by using high active consumption and low reactive consumption.

The second Cluster is ranked second as customers in this segment are medium active consumption, very low reactive consumption and medium revenue generated customers. The third cluster is Cluster 3 it contains customer that are medium power factor, low demand, high active consumption, very low reactive consumption, very low active reactive and medium revenue generated customers. The last cluster is Cluster 4 which

contains medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and medium revenue generated customers.

This experimentation with K=4 and seed=100 seem to have created dissimilar clusters or segments of customers. However, it couldn't differentiate sufficiently among the high, medium, and low value customers as stated in the business understanding phase of this research. As shown in the above table, Cluster 1 contains high value customers with the rank of first. Cluster 2, 3 and 4 contains medium value customers but no Cluster which contains low value corporate customers.

5.5.3.3.2 Experiment 2

segmentation (K=4 and seed size 1000)							
Cluster #	Freq. of records (share in %)	APF	AD	AAC	ARC	AAR	AR
1	7472 (15%)	1.5095	78.606	66554.50	46295.27	107.85	615638421
2	4977 (10%)	0.7072	49.012	8422.483	7301.268	2.484	6461.0749
3	35732 (71%)	0.7256	22.589	2953.562	2754.694	1.3351	1986.6478
4	1819 (4%)	0.875	168.34	89527.10	136265.8	100.30	37830.544
Total	50000 (100%)						
Cluster #	Freq. of records (share in %)						
1	7472 (15%)	H	M	H	L	H	VH
2	4977 (10%)	M	L	M	VL	VL	M
3	35732 (71%)	M	L	L	VL	VL	L
4	1819 (4%)	M	H	H	M	H	VH
Total	50000 (100%)						

Table 5.10: Cluster description based on average values of attributes for K=4 and seed size 1000.

Cluster	*Description*	*Possible Rank*
1	High power factor, medium power factor, high active consumption, low reactive consumption, high active reactive and very high revenue generated.	1st
2	Medium power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	3rd
3	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and low revenue generated.	4th
4	Medium power factor, high demand, high active consumption, medium reactive consumption, high active reactive and very high revenue generated.	2nd

Table 5.11: Cluster summary and corresponding ranks based on basic attributes for K=4 and seed size 1000

As shown in the experiment summary **Table 5.11** above**,** Cluster 1 is ranked first because customers in this cluster are generated very high revenue by consuming high active power and low reactive power. Cluster 4 is ranked second and it contains customers generated very high revenue, consume or use high active power, and medium reactive power. Cluster 2 is the third and it contains customers' generated medium revenue, medium active consumption, and very low reactive consumption. Finally, Cluster 3 is ranked last as the customers in this cluster generated low revenue by consuming low active power, and very low reactive power.

Although this experimentation result shows dissimilar clusters according to the corporation business objective (clustering as high, medium and low value customers) stated in the business understanding phase of CRISP-DM process model in this research, another clustering model experiment have been done with (K=5 and 6 with seed=10, 100

and 1000) in the quest for a better segmentation model with minimum algorithm iteration and small similarity error between instances in a cluster.

5.5.3.3.3 Experiment 3

segmentation (K=5 and seed size 100)							
Cluster #	Freq. of records (share in %)	APF	AD	AAC	ARC	AAR	AR
1	7908 (16%)	1.5042	105.1	79510.4	70296.05	122.86	58169679
2	3168 (6%)	0.7360	91.63	4673.62	3297.417	1.3301	1947.1317
3	6360 (13%)	0.7068	48.10	11524.3	11670.30	4.5687	15094.211
4	10853 (22%)	0.7235	39.43	3245.64	3058.951	1.3173	2010.4951
5	21711 (43%)	0.725	4.09	2702.48	2523.409	1.3446	1980.4929
Total	50000 (100%)						

Cluster #	Freq. of records (share in %)						
1	7908 (16%)	H	H	H	L	H	VH
2	3168 (6%)	M	M	M	VL	VL	L
3	6360 (13%)	M	L	H	VL	VL	VH
4	10853 (22%)	M	L	M	VL	VL	M
5	21711 (43%)	M	L	L	VL	VL	L
Total	50000 (100%)						

Table 5.12: Cluster description based on average values of attributes for K=5 and seed size 100.

The table above shows the experimentation result where the number of cluster (K) is five and seed size 100. And, the summary of clusters is the same like that of the previous one. The explanation of each cluster is summarized in the table below as follows.

Cluster	*Description*	*Possible Rank*
1	High power factor, high demand, high active consumption, low reactive consumption, high active reactive and very high revenue generated.	2nd
2	Medium power factor, medium demand, medium active consumption, very low reactive consumption, very low active reactive and low revenue generated.	5th
3	Medium power factor, low demand, high active consumption, very low reactive consumption, very low active reactive and very high revenue generated.	1st
4	Medium power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	3rd
5	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and low revenue generated.	4th

Table 5.13: Cluster summary and corresponding ranks based on basic attributes for K=5 and seed size 100

In this cluster experimentation, where the value of K is five and seed size 100, Cluster 3 and 1 contain high value customers to the corporation. Cluster 3 contains customers' generated very high revenue, high active consumption, very low reactive power consumption, and thus ranked first. Cluster 1 is the second since it contain customers generated very high revenue, high active consumption, and low reactive consumption. Cluster 4 is the third and it contain customers use medium active power, very low reactive consumption, and medium revenue generated customers. And, Cluster 5 is the fourth and it contain customers' generated low revenue, by consuming low active power and very low reactive power. The last cluster is Cluster 2 it contain customers' generated

low revenue by consuming medium active power and very low reactive power. Finally, Cluster 3 and 1 contain high value corporate customers, Cluster 4 contains medium level customers and Cluster 2 and 5 contain low value customers. In different literatures it is found that to make fair the distribution of instances or records in all clusters or segments trying it with different seed size.

5.5.3.3.4 Experiment 4

segmentation (K=5 and seed size 1000)							
Cluster #	Freq. of records (share in %)	APF	AD	AAC	ARC	AAR	AR
1	7472 (15%)	1.509	78.60	66554.508	46295.276	107.85	61563942
2	4977 (10%)	0.707	49.01	8422.483	7301.2685	2.484	6461.0749
3	26623 (53%)	0.724	8.614	2753.7228	2582.5337	1.3395	1977.9703
4	1819 (4%)	0.875	168.3	89527.100	136265.88	100.30	37830.544
5	9109 (18%)	0.729	63.43	3537.6363	3257.8727	1.322	2012.0094
Total	50000(100%)						
Cluster #	Freq. of records (share in %)						
1	7472 (15%)	H	M	H	L	H	VH
2	4977 (10%)	M	L	M	VL	VL	M
3	26623 (53%)	M	L	L	VL	VL	L
4	1819 (4%)	M	H	H	M	H	VH
5	9109 (18%)	M	M	M	VL	VL	M
Total	50000(100%)						

Table 5.14: Cluster description based on average values of attributes for K=5 and seed size 1000

This table is similar to that of previous except the seed size is 1000. The explanation of the table is described below in the summary table.

Cluster	*Description*	*Possible Rank*
1	High power factor, medium demand, high active consumption, low reactive consumption, high active reactive and very high revenue generated.	1st
2	Medium power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	4th
3	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and low revenue generated.	5th
4	Medium power factor, high demand, high active consumption, medium reactive consumption, high active reactive and very high revenue generated.	2nd
5	Medium power factor, medium demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	3rd

Table 5.15: Cluster summary and corresponding ranks based on basic attributes for K=5 and seed size 1000.

In the above cluster summary, Cluster 1 is better than others since it contain customers' generated very high revenue by consuming high active power from the corporation and low reactive power to the corporation. The next (the second cluster) is Cluster 4 because it contain customer generated very high revenue, used high active power and medium reactive power to the corporation. The third one is Cluster 5 as it contains customers use medium active consumption, very low reactive consumption and generated medium revenue. The fourth cluster is Cluster 2 it contain medium active consumption, very low reactive consumption, and medium revenue generated customers. Cluster 3 is the fifth and it contains customers' generated low revenue, low active consumption, and very low reactive consumption. Cluster 1 and 4 contains high value customers to the corporation.

Cluster 2 and 5 contain medium value customers and Cluster 3 contain low value customers of the corporation.

This clustering run also created dissimilar clusters according to business objectives of the corporation. It identifies high, medium, low value customers. However, as compared to the previous clustering model with value of K=4 and seed size 1000 in another comparison ways of clustering models (i.e. algorithm iteration and similarity-error within a cluster), this cluster run does not segment customers as the former does (with minimum algorithm iteration and small similarity-error within a cluster).

5.5.3.3.5 Experiment 5

Segmentation(K=6 and seed size 100)							
Cluster #	Freq. of records (share in %)	APF	AD	AAC	ARC	AAR	AR
1	7908 (16%)	1.5042	105.15	79510.4	70296.0	122.86	5816967946
2	3177 (6%)	0.7371	91.555	3641.40	3283.70	1.3306	1957.4125
3	6360 (13%)	0.7068	48.108	11524.3	11670.3	4.5687	15094.2113
4	10009 (20%)	0.7212	39.838	2621.56	2691.42	1.2741	1985.629
5	19262 (38%)	0.7201	4.3085	1841.30	1995.13	1.2793	1962.5025
6	3284 (7%)	0.7602	10.525	9823.86	6890.90	1.8525	2159.657
Total	50000 (100%)						
Cluster #	Freq. of records (share in %)						
1	7908 (16%)	H	H	H	L	H	VH
2	3177 (6%)	M	M	M	VL	VL	L
3	6360 (13%)	M	L	M	VL	VL	VH
4	10009 (20%)	M	L	L	VL	VL	L
5	19262 (38%)	M	L	L	VL	VL	L
6	3284 (7%)	M	L	M	VL	VL	M
Total	50000 (100%)						

Table 5.16: Cluster description based on average values of attributes for K=6 and seed size 100

The summary of the above experimentation table is as follows.

Cluster	*Description*	*Possible Rank*
1	High power factor, high demand, high active consumption, low reactive consumption, high active reactive and very high revenue generated.	1st
2	Medium power factor, medium demand, medium high active consumption, very low reactive consumption, very low active reactive and low revenue generated.	4th
3	High power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and vey high revenue generated.	2nd
4	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and low revenue generated.	5th
5	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and low revenue generated.	5th
6	Medium power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	3rd

Table 5.17: Cluster summary and corresponding ranks based on basic attributes for K=6 and seed size 100

In this clustering experiment, Cluster 1 is ranked first because it contains very high revenue generated customers with high active consumption and low reactive power consumption. Cluster 3 is the second as it contain customers generated very high revenue, medium active consumption, and very low reactive consumption. Cluster 6 is ranked the third and it contains medium revenue generating customers, medium active consumption,

and very low reactive power consumption customers. Cluster 2 is ranked the fourth as it contains customers generating low revenue, medium active consumption, and very low reactive consumption. Cluster 4 and 5 are the fifth and they contain similar customers groups that are low active power consumption, very low reactive consumption and generated low revenue for the corporation.

Although there are two clusters containing similar customer groups, this clustering experiment shows better segmentation of customer. It clusters customers according to business objectives of the corporation. However, the number of iteration that the algorithm runs and similarity-error within a cluster in this clustering (where K=6 and seed size 100) is also greater than the previous one (where value of K=4 and seed size 1000).

5.5.3.3.6 Experiment 6

Segmentation (K=6 and seed size 1000)							
Cluster #	Freq. of records (share in %)	APF	AD	AAC	ARC	AAR	AR
1	7472 (15%)	1.5095	78.606	66554.5	46295.27	107.85	615639421
2	4977 (10%)	0.7072	49.012	8422.48	7301.268	2.484	6461.0749
3	3873 (7%)	0.756	12.226	9250.82	6611.013	1.8261	2163.2434
4	1819 (4%)	0.875	168.34	89527.1	136265.8	100.30	37830.544
5	8545 (17%)	0.7286	64.845	3100.71	2997.418	1.294	2005.9554
6	23314 (47%)	0.7195	8.8235	1853.50	2025.107	1.2686	1950.2345
Total	50000(100%)						

Cluster #	Freq. of records (share in %)						
1	7472 (15%)	H	M	H	L	H	VH
2	4977 (10%)	M	L	M	VL	VL	M
3	3873 (7%)	M	L	M	VL	VL	M
4	1819 (4%)	M	H	H	M	H	VH
5	8545 (17%)	M	M	M	VL	VL	M
6	23314 (47%)	M	L	L	VL	VL	L
Total	50000(100%)						

Table 5.18: Cluster description based on basic attributes for K=6 and seed size 1000

Cluster	*Description*	*Possible Rank*
1	High power factor, medium demand, high active consumption, low reactive consumption, high active reactive and very high revenue generated.	1st
2	Medium power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	4th
3	Medium power factor, low demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	4th
4	Medium power factor, high demand, high active consumption, medium reactive consumption, high active reactive and very high revenue generated.	2nd
5	Medium power factor, medium demand, medium active consumption, very low reactive consumption, very low active reactive and medium revenue generated.	3rd
6	Medium power factor, low demand, low active consumption, very low reactive consumption, very low active reactive and low revenue generated.	5th

Table 5.19: Cluster summary and corresponding ranks based on basic attributes for K=6 and seed size 1000.

In this experiment, Cluster 1 is ranked first since it contain customers generate very high revenue, high active consumption and low reactive consumption. Cluster 4 is the second it contain customers that are a high active consumption, medium reactive consumption and very high revenue generated customers. Cluster 5 is the third and customers in this cluster are medium active consumption, very low reactive consumption and generated medium revenue. Cluster 2 and 3 are the fourth as they contain similar customer groups

that are medium active consumption, very low reactive consumption, and generated medium revenue. Cluster 6 is the fifth which contain customers consuming low active electric power, very low reactive consumption and low revenue generated customers. Generally, Cluster 1 and 4 contain high value customers. Cluster 2, 3 and 5 contain medium value customers, and Cluster 6 contains low value customers to the corporation. However, in this experiment there is also a problem that is creating of two clusters for similar customer groups.

Almost all clustering experiments conducted are good in segmenting customers according to business objectives of the corporation (identification of customers as high, medium, and low values to the corporation). And, thus to select the best clustering model from those the researcher used other cluster model comparison methods. This are domain experts decision about number of clusters, number iteration that the algorithms run to read the dataset and square-error between instances within a cluster.

5.5.3.4 Choosing the Best Clustering Model

In finding the best clustering experiment, the domain expert also plays a great role, because the expert can understand each segmentation needs and business know-how in the corporation.

As explained above, nine experiments were done to come up with the appropriate clustering model and six of them are presented and discussed. And finally, the clustering model that satisfied the good clustering model criteria more than any of other eight was selected. The best set of clusters may be simply defined as the one that shows some expected pattern in the data (Berry and Linoff 2000).

Although separation among the clusters and cohesion within clusters are generally accepted measures of cluster result validity, it is a broad issue and subject of endless arguments because the idea of “good” clustering is strictly related to the application domain (data mining objective) and its specific requirements. Thus, the comparison of clusters result validity in this research has been done in relation to average attributes

values in each cluster, algorithms iteration, similarity error within a cluster and the domain expert's judgment.

According to the explanation of expert's in the EEPCo, if a customer has the following characteristics he/she is considered as having a higher probability to be the preferred customer of the corporation: high active power consumption, low reactive power consumption and generates high revenue.

On the other hand, customers are considered as having a low probability to be a preferred customer of the corporation if the condition looks like the following: low active power consumption, high reactive power consumption and low revenue generated.

As observed above, in an attempt to improve the distribution of instances in different segments, different seed values (10, 100 and 1000) with different values of K (4, 5 and 6) have been tried and after a number of experiments, better cluster model has been obtained. The seed value at 1000 and value of K=4 gives better distribution of instances in the segments. Moreover, based on the additional advices from the domain expert about the number of clusters the segmentation where, the value of K=4 and seed size 1000, is good. In this experiment, dissimilar clusters were formed and the number of iterations that the algorithm runs to read the dataset and the error between inter-class in this model is minimum than in other clustering experiments.

Number of K	Seed size	Number of iteration	Within cluster sum of squared errors
4	10	8	4989.5122900585175
4	100	16	6810.134002643614
4	**1000**	**2**	**1413.13066102217504**
5	10	17	4990.157044583125
5	100	24	1833.1237031769874
5	1000	14	1441.12852929592384
6	10	11	2211.481287668616458
6	100	21	1833.1232805205277
6	1000	20	1444.128073685160492

Table 5.20: Comparison of clustering model experiments

5.5.3.5 Classification Model Building

According to Qiu et al. (2004), clustering serves as a starting point for supervised data mining techniques or modeling called classification. The task that was done and reported in the clustering sub-phase allowed is to get the class labels (cluster indexes) and then the classification process to be carried out.

In the classification, the output of clustering model is used as an input for the purpose of classification. The classification model should allow the attribution of a new consumer to a certain cluster, based on the rules generated by the classification model. Therefore, the rules must be intelligible and, for that to occur, normalized shape indicators were used as attributes in the classification model. The algorithms selected for classification purpose were J48 decision tree and Naive Bayes those can classify an instance to already identified cluster indexes. The researcher tested the algorithms with different parameters and record numbers to improve the classification accuracy. Finally, compared and select

the best classification model from the two algorithms. The classification model selected finally generates rules that enable to place or appoint a new record to one of the clusters.

5.5.3.5.1 Decision Trees Model Building

Experiment	Number of records	minNumObj	Number of attributes		Number of leaves	Size of tree	Test mode	Time taken	Accuracy
			Inputted	Used					
1	50000	2	6	6	57	113	All for training	65.61	99.984%
2	50000	10	6	6	46	91	All for training	69.47	99.918%
3	12500	2	6	6	54	107	All for training	8.11	99.944%
4	12500	10	6	6	32	63	All for training	12.02	99.568%
5	50000	2	6	6	57	113	70/30%	62.5	99.873%
6	50000	10	6	6	46	91	70/30%	61.42	99.74%
7	12500	2	6	6	54	107	70/30%	12.33	99.84%
8	12500	10	6	6	32	63	70/30%	14.64	99.52%
9	**50000**	**2**	**6**	**6**	**57**	**113**	**10 fold cross-validation**	**67.61**	**99.894%**
10	50000	10	6	6	46	91	10 fold cross-validation	64.2	99.834%
11	12500	2	6	6	54	107	10 fold cross-validation	16.39	99.608%
12	12500	10	6	6	32	63	10 fold cross-validation	15.81	99.304%

Table 5.21: Input parameters and the resulting decision trees output parameters

As can be observed in the table above (**Table 5.21**), the researcher conducted different classification experiments with different parameters of J48 decision tree algorithm.

From the experiments it can be concluded that although it significantly pruned the size of the tree, as the minNumObj parameter value increases, the accuracy of the classification algorithm was decreased. This is true in all test modes or options of classification. The accuracy of classification algorithm in large dataset is also better than in a small dataset.

From all test modes or options, using all dataset for training resulted better classification accuracy i.e. 99.984%. However, using all the dataset for training has its own limitations, that means there may be bias of classification. So, to validate the model there should be dataset that used to test the model developed. And, from the all experiments carried out, the researcher selected the algorithm with 10 fold cross-validation test option with the default parameter values i.e. the 9th experiment and the confusion metrics is as follows.

Actual	**Predicted**				**Total**	**Score (accuracy rate)**
	Cluster 1	Cluster 2	Cluster 3	Cluster 4		
Cluster 1	7461	0	0	11	7472	99.85%
Cluster 2	4	4972	1	0	4977	99.89%
Cluster 3	0	3	35729	0	35732	99.99%
Cluster 4	20	14	0	1785	1819	98.13%
Total	7485	4989	35730	1796	50000	**99.894%**

Table 5.22: Summary of the confusion matrix with default parameter (10 fold cross-validation)

Using the default values of the parameters the experimentation has been conducted and resulted in a decision tree containing 113 nodes (size of tree) and 57 leaves.

As shown in the confusion matrix above, the accuracy of this learning scheme is 99.894 percent, which indicates that, out of the total number of records supplied, 49947 (99.894%) records are classified correctly, while the remaining 53 (0.106%) records are classified incorrectly. Besides, 7461 (99.85%), 4972 (99.89%), 35729 (99.99%), and 1718 (98.13%) records are classified correctly in each cluster i.e. Cluster 1, Cluster 2, Cluster 3, and Cluster 4, respectively. The decision tree generated from this model is attached in **Appendix 2.**

Although this decision tree model has shown a good accuracy, it is somewhat resulted in too lengthy tree to drive all relevant rules. So, there is a need to adjust some of the J48 default parameters values to minimize the tree size and number of leaves. And, as observed in **Table 5.21**, the researcher tested different J48 decision tree experiments with different parameter values of minNumObj (the minimum number of instances per leaf). However, as the value of minNumObj parameter increased, the accuracy of the algorithm was decreased. The reason for this is that, records in a given leaf could be in different class and there could be attributes that could further split the records in the same node into disjoint classes.

Classifiers should be trained before they are reliably used on new data. Of course, it stands to reason that the more instances the classifier is exposed to during the training phase, the more reliable it will be as it has more experience. However, once trained, one would like to test the classifier too, so that he/she is confident by the model that it works successfully. For this, yet more unseen instances are required.

A problem which often occurs is the lack of readily available training/test data. These instances must be pre-classified which is normally time-consuming. A nice method to avoid this issue is cross-validation. It works as follows:

1. Separate data into fixed number of partitions (or folds)
2. Select the first fold for testing, whilst the remaining folds are used for training.
3. Perform classification and obtain performance metrics.
4. Select the next partition as testing and use the rest as training data.
5. Repeat classification until each partition has been used as the test set. And finally
6. Calculate an average performance from the individual experiments.

The experience of many machine learning experiments suggest that using 10 partitions (ten fold cross-validations), and in this research yields almost the same error rate as if the entire data set had been used for training the model.

Thus, the J48 decision tree model built using 10-fold cross-validation with default parameters values has shown better classification accuracy than splitting of the dataset for training and testing of the model and any other experiment by adjusting minNumObj parameter values. Hence, 10 fold cross validation with default parameter has been chosen due to its better overall classification accuracy than the different decision tree models built in the previous experimentations. This classification model also tested with new randomly taken 100 datasets and its performance was 96.75%. A tree generated from this model is depicted in **Appendix 3.**

5.5.3.5.2 Bayesian Network Classification Model

After experimenting with J48 decision tree algorithm with many parameter values, the best model that has shown better overall classification accuracy has been chosen. And, to compare the result of the J48 decision tree classification model, different Naive Bayes experiments have been carried out.

The same attributes that are used to build the decision tree models, are also used in this Naive Bayes modeling experiments. With all preprocessing takes in place, the experimentation proceeded with the different Naive Bayes models by changing the default parameter values as shown in the table below.

The test split options are randomly taken with ten ranges which are 70%, and 80%. The percent refers to the portion of the data that is allocated for training and the rest is for testing.

Size of the dataset	Test options	Accuracy
50000	10 fold cross-validation	99.48%
25000	10 fold cross-validation	99.488%
12500	10 fold cross-validation	99.136%
50000	70/30%	99.5%

25000	70/30%	99.5467%
12500	70/30%	98.72%
50000	Full training and testing	99.48%
25000	Full training and testing	99.532%
12500	Full training and testing	99.16%
50000	80/20%	99.45%
25000	80/20%	99.52%
12500	80/30%	98.92%

Table 5.23 The Naive Bayes classifier parameters with their values and performances.

As shown in the above experiment table, in 50000 and 25000 data size 70/30% test split option performs better the accuracy was 99.5% and 99.5467%, respectively. In 12500 data size, taking all dataset for training test split option perform better (99.16%).

Despite different Naive Bayes models was built by changing parameters' values, the over all classification performance of Naive Bayes classifier is less than that of J48 decision tree classification algorithm.

5.5.3.5.3 Comparison of Decision Tree and Bayesian Network Models

The results of the two algorithms are compared each other by their overall classification accuracy (performance). And, as can be clearly shown in the table above (**Table 5.21**), the overall performance of the decision tree model was 99.894% with 50000 datasets and 10 fold cross-validation. However, the classification accuracy of the Naive Bayes model with this data size and parameter was 99.48%. In Naive Bayes classifier the highest classification accuracy was achieved in 25000 datasets and 70/30 % test split option.

The J48 decision tree has shown better classification performance with total of dataset and 10 fold cross-validations. Hence, it is really reasonable to conclude that the J48 decision tree model is the best classifier model for implementing of CRM applications in the organization. There are different rules applicable or generated for the selected decision tree. This sample rules are annexed as **Appendix 4.**

5.6 Evaluation

After an optimal model is built and selected, critical assessment or evaluation against business goal to be achieved in the research is very decisive.

Appropriate CRM strategies and programs can be designed and implemented if customers are segmented into meaningful groups according to the organizations need, and this is the business goal to be achieved. The clustering process was focused on customer's value to the business. This customer value was defined by the three selected basic attributes: active power consumption, reactive power consumption, and revenue that customers generated to the corporation. The underline idea of this CRM research is grouping customers having similar power consumption pattern in the same group and the groups formed are different from each other. And, the result of data mining process was to facilitate and provide a way of reaching data mining solutions for good market segmentation and successful CRM implementation in the organization. The clustering result selected (where k=4 and seed size 1000) was good since it clusters customers having similar power consumption patterns in the same group and the groups formed were different from each other. It clusters customers as high, medium and low value customer according to the organization need. Besides, the selected clustering model segmented customers within minimum algorithm iteration and the squared error within a cluster is also small. So, this selected cluster index was used as a class to classify new customers into their appropriate cluster. In classification sub-phase of this research different experiments has been carried out using J48 decision tree and Naive Bayes algorithms with different input parameter values. These decision tree and Naive Bayes models are compared based on their overall classification accuracy. And, finally from all

classification experiments 10 fold cross-validation with default parameter of J48 decision tree performs better than the rest of others by decision tree and Naive Bayes. Its overall classification accuracy with 50000 dataset was 99.894%. And also its performance with new dataset was 96.75%. Hence, it is selected as the best classification model for this research.

5.7 Deployment of the Result

The purpose of the data mining process is to increase the knowledge gained from the data stored. And, deployment is the last step of data mining process, which means using the data mining result of clustering and classification. The knowledge gained from data need to be organized and presented in a way that the organization can understand and use it for successful CRM. To make this result applicable, integration of resources like people, business processes, and technology, are required. Moreover, the integration of resource is based on the information or result obtained from the segmentation model. And therefore, after integrating the necessary adjustments by group of domain experts the result of this study can be deployed for market decision making and successful CRM purpose in the corporation and the corporation will be beneficiary from the research result.

Basically customers are identified as good corporate customers and provided special service if they are high active power consumers, low reactive power customers and generated high revenue. On the other hand customers are identified as low value customers if they are low active power consumer, high reactive power consumer and generated low revenue. The following are sample rules generated based on corporation important attributes

If revenue is greater than 22800, active consumption is less than or equal to 10950, power factor is greater than 0.1589, reactive consumption is less than or equal to 15720 and active reactive is less than or equal to 5.7 then the customer is classified in **cluster 1.** If active consumption is less than or equal to 18560, revenu is less than or equal to 13240, active reactive is less than 5.684211, reactive consumption is less than or equal to 15680, demand is less than or equal to 157 and power factor is less than 0.1571 then the

customer is classified in **cluster 3.** If active consumption is greater than 31640, reactive consumption is greater than 26746, active reactive is less than or equal to 5.671429, revenue is less than or equal to 13240 and power factor is greater than 0.3204 then the customer is classified in **cluster 1. (See Appendix 4)**

The result of this research is best to identify and create a long lasting relationship with customers based on important attributes. Active consumption, reactive consumption, revenue customers generated, power factor and maximum demand are important attributes to identify customers and create and design marketing strategy. Active consumption is the consumption that customer applied directly for production/services purpose. As this consumption increase customers are liable to pay high revenue to the corporation. Reactive consumption is the uneconomical use of supplied power on the customer side. This consumption creates traffic overload on power supply system and enforce the corporation to generate (produce) additional electric power. Power factor is the trigonometric ratio of active and reactive power consumption. If customers' active consumption is less than reactive consumption, the power factor value becomes small. The bench mark for power factor in the corporation is 0.85. So, if customer power factor is less than the bench mark they are liable to pay power factor charge because this customer reacts high power to the corporation. Maximum demand is the highest load supplied by corporation to the customer in the month. This consumption can be above or below the demand in the contract agreement. When customers are used below minimum demand agreement, they are liable to pay minimum demand charge. On the opposite when customers are used above maximum demand in the agreement, the corporation encouraged them because they are high revenue generating customers. Revenue is the active power consumption charge that customers pay per month excluding service and penalty charges. The importance of attributes for customer market segmentation in the corporation for the future may change according to the policies of the government in the electric power sector. Hence, this model can be deployed with further modification and evaluation according to policies in the corporation to identify and provide special services to corporate customers. In addition to these, the model can have a paramount help in the marketing strategy formulation.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

With recent advances in computer technology large amounts of data could be collected and stored in different organizations. But the question is what should be done with this collected and stored organization data, because all this data could become more useful when it is analyzed and some dependencies and correlations are detected. This can be accomplished with machine learning algorithms. As a means of analyzing, detecting relations and extracting useful information and knowledge from a huge amount of data, the new generation of computerized methods known as Data Mining (DM) or Knowledge Discovery in Databases (KDD) has emerged in recent years.

The application of data mining techniques has increasingly become very important for different sectors that have a huge data, such as retail trade, airline industry, banking, telecommunication, electric power industry and healthcare sectors.

This study focused on the application of data mining techniques in the area of CRM for customer segmentation at the Ethiopian Electric Power Corporation (EEPCo). Different literatures have been reviewed concerning data mining techniques and CRM. The data mining task was conducted based on the CRISP-DM process model (approach), which was carried-out in six major parts, namely: business understanding, data understanding, data preparation, model building, evaluation and result deployment.

The business goal to be achieved in this research was to group electric power customers into similar groups based on their electric power consumption pattern. The criterion to segment customers was based on customers' value to the business, which is measured with three basic attributes. This basic attributes were active consumption, reactive

consumption, and total revenue generated. With this three basic and other attributes the two data mining tasks (clustering and classification) were employed. Based on the preprocessed dataset, different clustering experiments have been done with the K-means clustering algorithm (4, 5 and 6 are values of k and other different input parameters). The model where K=4 and seed size 1000 had shown better clustering performance. It enables to cluster corporate customers into dissimilar clusters of high, medium and low value customer groups of the corporation with minimum iteration of algorithm and minimum square error between records in a cluster, 2 and 1413, respectively. So, this experiment or model has been chosen for cluster index.

The classification experiment was done with decision tree (J48) and Naive Bayes algorithms. Different experiments were also done in classification with different parameters of the two algorithms. And finally, after compared the two algorithms, the classification experiment that showed better accuracy was selected. The selected decision tree parameter was 10 fold cross-validations. With this parameter the classification accuracy of the decision tree algorithm was 99.894%. The accuracy of the algorithm was better than other data mining research results done before in Ethiopian context in other organizations, and the reason may be the size of dataset used in this research was large, and this was the recommendation of different researchers to train and test the algorithm with large dataset to improve the accuracy of the algorithm.

In undertaking of this research, considerable time of the research has been spent on the data preparation phase, and consulting the domain expert's in the organization on the interpretation and selection of the appropriate model developed, which were built for clustering and classification tasks of data mining.

6.2 Recommendations

It is the researcher's belief that the contribution of this research work could be a good experience for a competitive study in customer-oriented business organizations in the future. And also it is the researcher's belief that the findings of the research would encourage business-oriented organizations to work on the application of data mining

techniques for successful CRM implementation, and as a result gain a competitive advantage in the global market.

Finally, the researcher makes the following recommendations based on findings of the study.

- The Ethiopian Electric Power Corporation (EEPCo) should encourage data mining research on its customer profile databases. This could be achieved by a data mining team formed from the business experts of the corporation, the information technology professionals in the organization and external data mining professionals.

- Undertake further data mining research by using other data mining techniques, because business problems often become diverse and complex, and data mining results could be improved through the use of other clustering and prediction techniques, such as SOM, EM, NN etc. and making comparison between the accuracy of the models.

- The inclusion of many customer attributes, as much as possible, should be given high attention, and more comprehensive models should be built by using large training and testing datasets taken from the relevant customer databases of more than one year customer profile in the organization.

- Staffs of the organization should be trained to improve the service provided to the customer with the help of communication technologies and to implement the proposed CRM strategy successfully.

- It would be very useful to have a professional with expertise in data mining and some knowledge about the business in the organization to implement successfully the result of research projects in data mining.

- Since the nature or behavior of the customer is dynamic, it would be important to conduct this kind of research continuously and the result output would be used to design, implement and improve CRM strategies and programs at all levels of the organization continuously.

- The data preprocessing/preparation phase usually take much of the time for a data mining research. This time can be reduced if the data warehouse is developed in the corporation. EEPCo does not have a data warehouse. Due to this reason, much of the research time has been spent on the data preprocessing/preparation. So, the researcher strongly recommends EEPCo to develop customers' data warehouse, which contain detailed customers' information about power consumption, contract information and other relevant data clearly without missing and outlier data. Developing data warehouse is also useful for statistical analysis in addition to data mining purpose.

- The result of the clustering model was evaluated by the domain expert in the corporation. So, the probability of evaluating the clustering model correctly depends on the level of domain expert's business understanding and experience in the organization. In addition, extensive time of the research has been spent in interacting with and consulting the domain experts to create clear understanding between the researcher and domain expert and to evaluate attributes and models developed from business perspective. This consulting time creates task overload on domain expert. So, the researcher recommends that it would be easy and less time consuming task with high probability of evaluating the model correctly, if there is a knowledge base system that automatically evaluates the correctness of the models generated. Hence, there is a need to integrate the knowledge base system with the data mining tools that evaluates data mining clustering models consistently.

- Power utilities are also suffering from electricity theft because a power system can never be 100% secure from it. Electricity theft can be in the form of fraud (meter tampering), stealing electricity (illegal connections), billing irregularities and unpaid bills. The financial losses resulting from this electricity theft are critical to electric power organizations. So, other researchers can investigate on this area of detecting non technical losses due to faulty metering and billing errors.

REFERENCES

Anand, V. P. and P. Kumar. 2008. *Data mining as a tool for building and managing better customer relationships*. IES Management College and Research Centre.

Angoss software corporation knowledge studio user manual: Knowledge studio user manual. Viewed 20 February 2011, home page: http://www.angoss.com.

Askale, M. 2001. *The application of data mining techniques in supporting loan disbursement activities at Dashen Bank S.C.* Addis Ababa University, Unpublished Master's Thesis.

Berry, M.J.A. and G. Linoff .2000. *Mastering data mining: The art and science of customer relationship management*. Canada: John Wiley and Sons, Inc.

Berson, A., S. Smith, and K. Thearling. 2000. *Building data mining applications for CRM*. USA: McGraw-Hill.

Bhasin, M. L. 2006. *Data mining: A competitive tool in the banking and retail industries.* Mazoon College, Muscat, and Sultanate of Oman, Pp. 588-594.

Buchanan, B.G. 2006. *Brief history of artificial intelligence: history of artificial intelligence.* Viewed 19 February 2011, Homepage: http://www.aaai.org/AI Topics/bbhist.html.

Chen, I. J. and K. Popovich. 2003. *Understanding Customer Relationship Management (CRM) people, process and technology*. Business Process Management Journal, 9, Pp. 672-688.

Cheung, Y. 2003. *K*-Means: A new generalized K-Means clustering algorithm*. Pattern recognition Lett, 24, Pp. 2883-2893.

Chicco, G., R. Napoli, and F. Piglione. 2003. *Application of clustering algorithms and Self Organizing Maps to classify electricity customers.* Bologna power tech conference, June 23th -26th, bologna, Italy, Pp. 1-7.

Collier, K., B. Carey, D. Sautter, C. Marjaniemiand. 1999. *A Methodology for evaluating and selecting data mining software*. Proceedings of the 2nd Hawaii International Conference on System Sciences, Pp. 1-4.

Davis, J. 2006. *Retailers use technology to thwart would-Be thieves*. San Diego Tribune: viewed 5 March 2011, homepage: http://www.fashion era.com/ Trends_2006/2006_spring_fashion_trends_returns_consumer_fraud.htm.

Decision support solutions: Compaq. Object relational data mining technology for a competitive advantage. Viewed 26 February 2011, homepage: http://www.tandem.com/brfs.

Denekew, A. 2004. *Application of data mining to support Customer Relationship Management (CRM) at Ethiopian Airline*. Addis Ababa University, Unpublished master's thesis.

Dunham, M. H. 2000. *Data mining techniques and algorithms*. South Methodist University: prentice hall.

Dunham, M. H., S. Sridhar. 2006. *Data mining: Introductory and advanced topics*. New Delhi: Pearson Education, Inc.

Dunham, M.H. 2003. *Data mining introductory and advanced topics*: Upper Saddle River, NJ: Pearson Education, Inc.

Elkan, C. 1997. *Naive bayesian learning.* Harvard University. Viewed 27 February 2011, homepage: http://www.nbl.com.

Ethiopian Electric Power Corporation. 2002. *Annual report*: Addis Ababa, EEPCo. Viewed 12 February 2011, homepage: http://www.eepco.com.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. *From data mining to knowledge discovery in database.* AI magazine, American Association for Artificial Intelligence, Pp. 37-44.

Fekadu, M. 2004. *The application of data mining to support customer relationship management at Ethiopian Telecommunication Corporation*. Addis Ababa University, Unpublished Master's Thesis.

Figueiredo, V., F. Rodrigues, R. Pinto, and Z. Vale. 2003. *Data mining decision support tool to study electricity retail contracts*. Portugal: Knowledge Engineering and Decision Support Group, ISEP – Polytechnic Institute of Porto, Pp. 1-2.

Friedman, N., D. Geiger, M. Goldszmidt.1997. *Bayesian network classifiers.* Netherlands: Kluwer Academic Publishers, 29, pp.131–163.

Gerald, B. 2002. *Data mining: Annual review of information science and technology.* American Society for Information Science and Technology, 36.

Gobena, M. 2000. *Flight revenue information support system for Ethiopian Airlines.* Addis Ababa University, Unpublished Master's Thesis.

Gray, P. and B. Jongbok. 2001. *Customer relationship management*. 10, Pp. 245-332.

Guan, Y., A. Ghorbani, and N. Belacel. 2005. *K-means+: An autonomous clustering algorithm.* Canada*:* New Brunswick University, Pp. 1-3.

Halkidi, M., and M. Vazirgiannis. 2001. *Clustering validity assessment: Finding the optimal partitioning of a dataset*. California, USA: In Proceedings of ICDM Conference.

Han, J., and M. Kamber. 2001. *Data mining: concepts and techniques.* San Francisco: Morgan Kaufmann.

Henock, W. 2002. *The application of data mining to support customer relationship Management at Ethiopian Airlines*. Addis Ababa University, Unpublished Master's Thesis.

IDC and C. Gemini. 2005. *Four elements of customer relationship management*. Cap Gemini White Paper.

Jain, A.K, M.N. Murty, and P.J Flynn. 1999. *Data clustering a review*. ACM Computing Surveys, 3, Pp. 265-316.

Joe, P. 2000. *Customer Relationship Management (CRM) in financial services.* Journal of European Management, 18, Pp. 312-327.

Kitayama, M., R. Matsubara, and Y. Izui. 2002. *Application of data mining to customer profile analysis in the Electric Power Industry*. New York, USA: Presented at Proceedings of Winter Meeting of the Power Engineering Society, Pp. 632-634.

Kumneger, F. 2006. *Application of data mining techniques to support CRM in Ethiopian Shipping Lines*: Addis Ababa University, Unpublished Masters Thesis.

Larose, D. T. 2005. *Discovering knowledge in data: An introduction to data mining*: New Jersey: John Wiley & Sons, Inc.

Larose, D. T. 2006. *Data mining methods and models*. New Jersey: .John Wiley & Sons, Inc.

Lee, J. H., and S. C. Park. 2005. *Intelligent profitable customers' segmentation system based on business intelligence tools*. Expert Systems with Applications, 29, Pp. 145-152.

Ling, R., and, D. C. Yen. 2001. *Customer relationship management: An analysis framework and implementation strategies.* Journal of Computer Information Systems, 41, Pp. 82–97.

McQueen, J. 1967. *Some methods for classification and analysis of multivariate observations*. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Melaku, G. 2009. *The applicability of data mining techniques to customer relationship management the case of Ethiopian telecommunication corporation Code Division Multiple Access (CDMA) Telephone Service.* Addis Ababa University, Unpublished Master's Thesis.

Mingoti, S. A., and J. O. Lima. 2006. *Comparing SOM neural network with Fuzzy C-Means, K-Means and traditional hierarchical clustering algorithms*. European Journal of Operational Research, 174, Pp. 1742–1759.

Nizar, A. H., Z. Y. Dong and J. H. Zhao. 2006. *Load profiling and data mining techniques in electricity deregulated market.* Published at IEEE.

Peter, C., C. Julian, K. Randy, T. Khabaza, T. Reinartz, C. Shearer, and W. Rudiger. 2000. *CRISP-DM step-by-step data mining guide.* U.S.A: SPSS Inc, CRISPWP-0800, viewed 25 February 2011, homepage: http:// www.crisp-dm.org/.

Qiu, M., S. Davis, and F. Ikem. 2004. *Evaluation of clustering techniques in data mining tools*. Issues in information system, 1.

Rao, R. 2003. *Data mining and clustering techniques*. India: Bangalore, Pp. 2-11.

Rinta-Runsala, E. 2006. *Bringing data mining to customer relationship management of every company*. Finland: VTT technical research center, Pp. 1.

Ruiz, J. P., J. C. Chebat and P. Hansen. 2004. *Another trip to the mall: A segmentation study of customers based on their activities*. Journal of Retailing and Consumer Services, 11, Pp. 333-350.

Rygielski, C., W. Jyun-Cheng, D. C. Yen. 2002. *Data mining techniques for customer relationship management*. Technology in Society, 24, Pp. 484-493.

Saarenvirta, G .1998. *Mining customer data, a step by step look at a powerful clustering and segmentation methodology*. IBM database magazine.

Shearer, C. 2000. *The CRISP-DM model: The new blueprint for data mining*. Journal for data warehousing, 4, Pp. 13-22.

Shegaw, A. 2002. *Application of data mining technology to predict child mortality pattern (Case of Butajira Rural Health Project (BRHP)*. Addis Ababa University, unpublished Master's Thesis.

Srivastava, J. 2006. *Data mining for customer relationship management.* Viewed 22 February 2011, home page: http://www.dcrmconf.

Suresh, H. 2002. *Customer relationship management an opportunity for competitive advantage.* India: PSG Institute of Management, Pp. 3-5.

TATA consultancy service. 2009. *Innovation in customer relationship management*. India: Pp .3. Viewed 22 February 2011, homepage: http://www.tcs.com

Tesfaye, H. 2002. *Predictive modeling using data mining techniques in support of insurance risk assessment.* Addis Ababa University, Unpublished Master's Thesis.

Thearling, K. 1998. *Increasing customer value by integrating data mining and campaign management software*. Exchange Applications White Paper, viewed 28 February 2011, homepage: http://www.crmforum.

Thearling, K. 1999. *Data mining and CRM*. *Zeroing in on your best customers*. DM Direct, viewed 2 March 2011 http://www.dmreview.com/editorial/dmreview/printaction.cfm?

Thearling, K. 2003. *An introduction to data mining.* Viewed 7 March 2011, http://www3.shore.net/~kht/text/dmwhite.htm.

Two Crows Corporation. 1999. *Introduction to data mining and knowledge discovery*. Viewed 15 February 2011. http://www.twocrows.com.

Ueno, S. 2006. *The impact of customer relationship management*. USJP Occasional Paper. Harvard University, Pp. 2-3.

Valero, S., M. Ortiz, J. Francisco, G. Franco, and A. Gabaldon. 2004. *Characterization and identification of electrical customers through the use of Self-Organizing Maps and daily load parameters.* Spain: Published at IEEE, Pp. 1-7.

Weiss, M. 2009. *Data mining in the telecommunications industry*. USA: Fordham University. Viewed 23, February 2011 http://storm.cic.fordham.edu/~gwess/papers/kluwer/04-telecom.pdf.

Witten, I. H. and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*. San Francisco: Elsevier.

Xu R. and D. Wunsch II. 2005. *Survey of clustering algorithms.* University of Missouri-Rolla, Rolla, IEEE transactions on neural networks, 3, Pp. 646.

Zaiane, O. R. 1999. *Principles of knowledge discovery in databases:* University of Alberta. Pp. 1&3.

APPENDICES

Appendix 1: The Original Collected and Integrated Sample Data

Region	Service Center	Tariff	Service No	Old Account No	Name	Year	Cycle	Power Factor	Maximum Demand	Active Consumption	Reactive Consumption	Active Constant
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	3	0.9165	2.039	2200	960	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	4	0	2.039	0	0	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	5	0.8634	2.103	3080	1800	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	6	0.8805	2.266	1560	840	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	7	0.871	2.352	1560	880	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	8	0.8448	2.405	1200	760	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	9	0	2.485	0	1040	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	10	0.8093	2.56	2480	1800	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	11	0.8849	0.023	760	400	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2010	12	0.8898	0.028	1560	800	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2481077	889704800400	TEEHENCIAL SCHOOL ALEM KETE	2011	1	0.8455	0.052	1520	960	40
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	3	0.8288	33.83	800	540	0.36
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	4	0.7818	34.01	1003	800	0.36
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	5	0.8509	36.03	972	600	0.36
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	6	0.7268	44.52	2751	2600	0.36
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	7	0.5178	41.61	3391	5602	0.36
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	8	0.9117	42.18	3017	1360	0.36
ADDIS ABABA EA	CSC - ALEM KETEMA DI	46	2911404	889724000300	ATO BIZUAYEHU DESALGN	2010	9	0.8621	33.97	1584	931	0.36

Appendix 2: Sample of the Decision Tree Generated With 10-Fold Cross-Validation Technique

```
=== Run information ===

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

Relation:     cluster index

Instances:    50000

Attributes:   7

              Power factor

              Demand

              Active consumption

              Reactive consumption

              Active reactive

              revenue

              Cluster

Test mode:    10-fold cross-validation


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


J48 pruned tree

------------------


 Active consumption <= 18560

|   revenue <= 13240

|   |   Active reactive <= 5.684211

|   |   |   Reactive consumption <= 15680

|   |   |   |   Demand <= 157

|   |   |   |   |   Power factor <= 0.1571: cluster2 (309.0)
```

| | | | | Power factor > 0.1571: cluster3 (35732.0)

| | | | Demand > 157

| | | | | Demand <= 271: cluster2 (685.0)

| | | | | Demand > 271: cluster1 (84.0/1.0)

| | | Reactive consumption > 15680

| | | | Reactive consumption <= 26720

| | | | | Demand <= 183

| | | | | | Power factor <= 0.8914: cluster2 (678.0)

| | | | | | Power factor > 0.8914

| | | | | | | Power factor <= 22: cluster2 (16.0)

| | | | | | | Power factor > 22: cluster4 (2.0)

| | | | | Demand > 183

| | | | | | Demand <= 270: cluster2 (14.0)

| | | | | | Demand > 270: cluster4 (4.0)

| | | | Reactive consumption > 26720

| | | | | Power factor <= 0.1574: cluster4 (84.0)

| | | | | Power factor > 0.1574

| | | | | | Demand <= 152: cluster1 (451.0)

| | | | | | Demand > 152: cluster4 (12.0/1.0)

| | Active reactive > 5.684211

| | | Active reactive <= 9.428571: cluster2 (908.0/2.0)

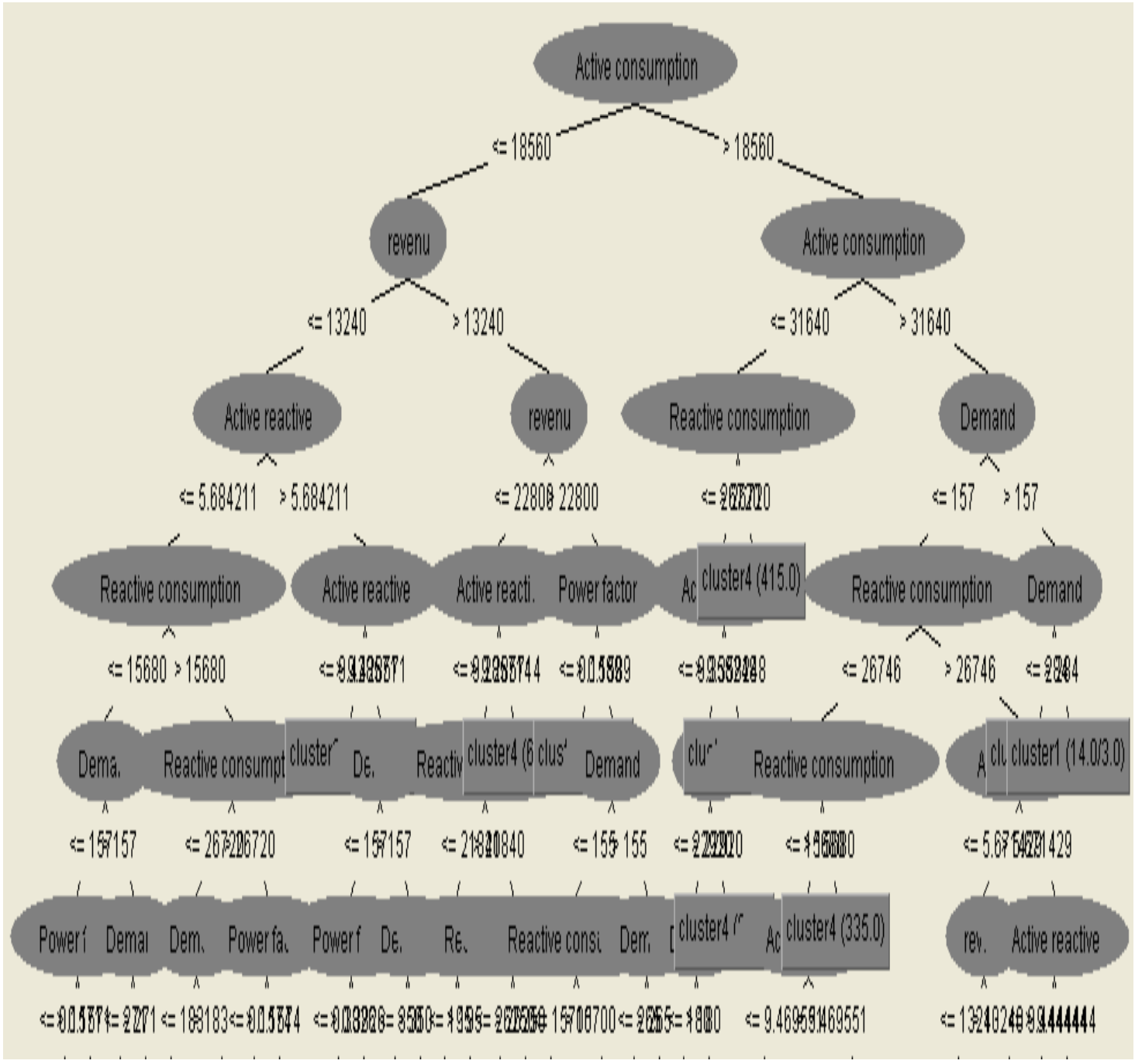
| | | Active reactive > 9.428571

| | | | Demand <= 157

| | | | | Power factor <= 0.3826

| | | | | | Power factor <= 0.1574: cluster4 (8.0)

Appendix 3: The partial overview of decision tree



Appendix 4: Sample Rules to Predict New Instances into Their Corresponding Cluster

Rule #1: if active consumption is less than or equal to 18560, revenu is less than or equal to 13240, active reactive is less than 5.684211, reactive consumption is less than or equal to 15680, demand is less than or equal to 157 and power factor is less than 0.1571 then the customer is classified in **cluster 3.**

Rule #2: if revenue is greater than 22800, demand is greater than 152 and less than or equal to 265 then the customer is classified in **cluster 4**

Rule #3: if revenue is greater than 22800, active consumption is less than or equal to 10950, power factor is greater than 0.1589, reactive consumption is less than or equal to 15720 and active reactive is less than or equal to 5.7 then the customer is classified in **cluster 1**

Rule #4: if active consumption is greater than 31640, demand is greater than 157 and less than or equal to 284 then the customer is classified in **cluster 4**

Rule #5: if active consumption is greater than 31640, reactive consumption is greater than 26746, active reactive is less than or equal to 5.671429, revenue is less than or equal to 13240 and power factor is greater than 0.3204 then the customer is classified in **cluster 1**

Rule #6: if active reactive is greater than 9.428571, demand is less than or equal to 157, revenue is less than or equal to 13200, active consumption is less than or equal to 18574 and power factor is greater than 0.3826 then the customer is classified in **cluster 1**

Rule #7: if reactive consumption is less than or equal to 26720, active consumption is less than or equal to 31640, revenue is less than or equal to 22800, active reactive is less than or equal to 9.439807, demand is less than or equal to 271 and power factor is less than or equal to 0.9987 then the customer is classified in **cluster 2**

Rule #8: if active consumption is greater than 18600 and less than or equal to 31640 then the customer is classified in **cluster 4**

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

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

_____

Advisor