DEPARTMENT OF COMPUTER SCIENCE

A DATA MINING APPROACH FOR DETERMINING POWER CONSUMPTION OF ETHIOPIAN ELECTRIC UTILITY CUSTOMERS

MASTER OF SCIENCE DEGREE IN COMPUTER SCIENCE

BY

MOHAMMED ASSEN

JULY, 2019 G.C

ADDIS ABEBA, ETHIOPIA

St. Mary's University
School of Graduate Studies
Faculty of Informatics
Department of Computer Science

A Data Mining Approach for Determining Power Consumption of Ethiopian Electric Utility Customers

A Thesis Submitted to the School of Graduate Studies of St. Mary's University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Science

By
Mohammed Assen

July 2019

St. Mary's University
School of Graduate Studies
Faculty of Informatics
Department of Computer Science

A Data Mining Approach for Determining Power Consumption of Ethiopian Electric Utility Customers

Signed by the examining Committee:

Advisor Million Meshesha (PhD) Signature_____Date_____

Examiner Tibebe Beshah (PhD) Signature_____Date_____

Examiner Getahun W/Mariam (PhD) Signature_____Date_____

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been fully acknowledged.

Mohammed Assen Ali

Student

_____

Signature

Addis Abeba

Ethiopian

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

Dr.Million Meshesha

Advisor

_____

Signature

Addis Abeba

Ethiopian

July 2019

DEDICATION

I would like to dedicate this thesis to all my families especially to my brother Jemal Ibrahim who would like to see the fruits of my effort.

ACKNOWLEDGEMENT

I would like to thank all those who have helped me to accomplish this thesis. First, I gratefully express my deepest thanks to the almighty ALLAH for helping me to accomplish the paper and who added years to my life, Glory to Allah. Next, I would like to extend my deepest thanks to my advisor Dr. Million Meshesha (PhD) for wonderful and unreserved assistance in every step of this study.

My sincere thanks also go to the staffs of electric utility business planning and district manager Engineer kassahun for providing me the business domain knowledge and data resources to conduct this research

I would like to my special thanks also want to put on record my gratitude and indebtedness to my mother Demeku Wolle. She has always encouraged me for higher success. My special thanks go to my beloved family, Zemzem Hassen, Mekin Yemir, Yimer Mekonnen, Erahment Yimer, Erebia Yimer for their unreserved backing and encouragement.

I am also very much indebted to Jemal Ibrahim and Merema Mustefa who stood beside me during this critical time and also my dear wish wife and constant love, prop, understanding and faith in me. Special thanks to my brother and childhood friends Mohammed Mekin, Nuru, Jemal Kebede, Ahmed Sied, Kedir Assefa and Awol Fentaw.

Above all, I thank Almighty Allah who surrounded me with so many wonderful people.

Table content

List of Figures

List of Tables

List of Acronyms and Abbreviation

DM – Data Mining

DMKD – Data Mining Knowledge Discover

EEA - Ethiopian Electricity Agency

EEP – Ethiopian Electric Power

EEPCO - Ethiopia electric power corporation

EEU – Ethiopia Electric Utility

ICS - Interconnected System

KDD – Knowledge Discovery Database

KDP - Knowledge Discovery Process

KVAH – Kilo volt Ampere Hours

KV – Kilo volt

KW – Kilo Watt

KWH – Kilo Watt Hours

SCS - Self-Contained System

SQL – Structure Query Language

WEKA- Waikato Environment for Knowledge Analysis

Abstract

Electric industry is one of most important service provider and back bone of the energy sector in the world. Ethiopia electric utility is the only national organization distributing electric power in our country. Electric power industries are being pushed to and quickly respond to the individual and organization needs and wants of their customers due to the dynamic and highly competitive nature of the industry. According to Energy pedia published in 2016, only 27 % of the population in Ethiopia has access to electricity grid

The aim of this study is designing a predictive model for determining power consumption of Ethiopian electric utility customers using data mining techniques.

This study conducted in Ethiopian electric utility customers to mining big data. The approach followed in this research is hybrid data mining methodology, which being able to be the classification of customer based on power consumption, and to develop a prediction model using classification algorithms. The major steps followed are problem understanding, data understanding, Data Preparation, Modeling, evaluation of knowledge discovering and design user interface to use the discovered knowledge. The data covers from January 2008 to January 2011 E.C for all Ethiopian utility customers data included. The data prepared for mining contain 14 attribute with 85,849 instances.

The study has used four classification algorithms to build predictive model namely: J48, bagging, random tree and PART. The result obtained from the experiments showed that J48 algorithm performed best with accuracy of 96.61% than the other models. In this model the number of correctly classified instances is 82,939 (96.61%) and the number of incorrectly classified instances is 2,910 (3.38%). This study has been classification of prediction power consumption based on new connection of electric utility customers either high and low power consumption.

Hence, based on the findings of this study, the researcher would like to forward recommendations for electric industry to conduct the study further and come up with system that enable to an optimal management of power consumption.

Key word: Data Mining, Ethiopian Electric Utility, Customer Classification, Hybrid Data Mining.

CHAPTER ONE

Introduction

1.1.Background of the study

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 [1]. Data mining is the extraction of useful patterns and relationships from data sources such as databases, texts, number, video, audio, picture and web. It has nothing to do however with SQL, OLAP (online analytical process), data warehousing or any of that kind of thing [1]. Data mining is also 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 [2]. Generally, the goal of data mining is to create models for decision making that predict future to classification of customers based on analysis of past existence data. To effectively exploit the potential of data mining in Ethiopian electric utility customers' database should be first organized into a format that can be used for further data mining process [3].

Data mining is an increasingly popular set of tools for dealing with large amounts of data. These types of huge amount of data are available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision for such type of huge amount of data we need a technique called the data mining which will be transforming in many fields. In Data mining data sets will be explored to yield hidden and unknown predictions which can be used in future for the efficient decision making [4].

The use of data mining technique to apply classification of customers relationship management (CRM) to analysis of those data may lead to a better understanding of the customers and attached contract for organization and employee; identify customers proportion to distribution power need, supporting the offer of new structure or services and identification of risky disbursements. In many cases, several related features need to be simultaneously considered in order to

accurately understand customers [5]. Ethiopian electric utility customers can be classified using the customers' data from data bases to analysis data mining technique which is easy to make a decision to plan the future work.

Data mining came into existence in response to technological advances in many diverse disciplines. There have been tremendous improvements in techniques for collecting, storing, and transferring large volumes of data for such applications as image processing, digital signal processing, text processing and the processing of various forms of heterogeneous data. In other words, all the data in the world are of no value without mechanisms to efficiently and effectively extract information and knowledge from them. Data mining is not just an umbrella term coined for the purpose of making sense of data. The major distinguishing characteristic of data mining is that it is data driven as opposed to other methods that are often model driven [1] . The uses of data mining in Ethiopian electric customers information can be classified into two [6]. These are self-connected and inter connected customers. Self-connected customers are domestic, industrial, commercial, own consumption, retired staff consumption and street light customers. Inter connected customers on the other hand include domestic, industrial, commercial, own consumption, active staff consumption, Retired Staff Consumption, and street light [6].

The tasks of data mining are very diverse and distinct because there are many patterns in large database [7]. In the data mining task for Ethiopian electric utility Customers can be classified by using prediction data mining task because using customers' common attribute to develop a model which can infer predicted common attribute of customers from some combination of other predictor attribute. In the prediction data mining task using classification because the aim of classification in Ethiopian electric utility customers is to classify attribute into several predefined classes. Given a collection of customers common attribute samples, this type of task can be designed to find a model for attributes as a functional value of other attributes.

Data mining task can be classified in two those are verification and discovery of big data set. Verification means goodness of fit, hypothesis of testing and analysis of variance. Discovery data mining task also divided two-part predication and description. In description discovering task is different part like clustering, summarization, linguistic summary and visualization. From prediction task is in data mining better used for classification and regression. Some popular classification methods include Decision tree, J48, Random Tree, PART and Bagging [8].

1.2 Overview of Ethiopian Electric Utility

Electric power industry was introduced to Ethiopian in the late 19th Century, during the regime of Minilik [9]. The first generator was said given to Minilik around the year 1898 to light the palace. Ethiopia electric power corporation (EEPCO) is a national electric power supply and distribution company. It is divided into two-parts those are Ethiopia electric power and Ethiopia electric utility [9, 6]. EEU is a wholly government owned corporation and is responsible for distribution of power for costumer 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 [10]. In addition, globalization and other technological advancement makes difficult to predict about the future. According to this regulation, the EEU 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 [6]. 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, wind power plant and for the SCS mini hydro's and diesel power generators allocated in various locations of the country. EEU is a company responsible for constructed low power transmission, distribution and sales of electricity all over the nation.

EEU has been organized into a corporation in which there are different business units, functional units and departments. The management board at the top is the highest body responsible for the overall functions and general operations of the Ethiopian electric utility. 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 Bill system, ICT, distribution system, marketing and sales, universal electrification, financial and supply chain, human resource and corporate services. Each branch departments have its own functions in the corporation. 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 [11, 6]. 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, platforms and 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 ICT [9].

The vision of EEU is a governmental owned organization striving to realize the vision that reads as energizing Ethiopia's sustainable growth and enabling it to be the power hub of Africa. Likewise, the mission of the organization is to be a world-class utility and contribute towards nation building by ensuring delivery of cost- effective, safe, reliable and high-quality power and to enable interconnections across the African Continent for exporting surplus power. EEU shall strive towards achieving international standards of customer care through sustained capacity building, operational and financial excellence, state-of the-art technologies while ensuring highest standards of corporate governance and Ethics [9].

According to National Bank of Ethiopia 2014/15 annual report, Ethiopian has generated approximately 9.5 billion KWH (a 9.4% increase from 2014/15) of electricity, of which 94.7% came from hydropower resources. Hydro power is exposed to high seasonal variability and climate change related risks. The remaining 5.2% and 0.1% power generated from wind and geothermal sources, respectively. Demand for power is forecasted to grow approximately by 30% per annum. In order to ensure a reliable and climate resilient power supply, in addition to more hydro power projects, the Government of Ethiopia is keen to diversify its energy mix through the exploitation of other renewable energy resources including solar, wind and geothermal [12].

1.3 Statement of the problem

Ethiopian electric utility is the only largest governmental organization in Ethiopian which is responsible to distribute electric power for customers. Ethiopian Electric power is only source of electric power in our country which is generating power from water, wind and geothermal. Ethiopian electric utility is fully monopolized distributed electric consumption by the government with no competition and only governmental energy retail companies to supply electricity for customers [9]. Ethiopian electric utility is distributing of the generated power in interconnect system (ICS) and self- connect system (SCS) distribute to the customers. In 2016 E.C around 27 % of Ethiopian people to distribute electric power for different customers [9]. Ethiopian electric customers data can be classified into two those are self-connected and inter connect customers. Self-connect customers' also different type those are domestic, industrial, commercial, own consumption, retired staff consumption and street light. Inter connect use of customers is lot of customers those are domestic, industrial, commercial, own consumption, active staff consumption, retired staff consumption, and street light. Based on the above customers they have been their own common attributers like district, customer type, tariff and region. Those Ethiopian electric customer databases contain enough data in the organization and it is used to infer missing information using statistical methods which are not supported by modern technology.

The demand of electric power customers increase based on need of good quality service and energy to distribute electric power for every nation and nationality people of Ethiopia. However, Ethiopian electric utility is still unable to succeed this domain because of various reasons including transmission line failures, automatically power reading meter on and off and unbalanced planning of appliance power demand. These drawbacks could consist of problems related to customers' satisfaction, predict future work based on customers demand, difficult to upgrade substation and transmission line, offering special service, change bill and organizational structure in new form and so on.

Besides, there is a case study many local and international Companies involve in providing service like shipping lines, Medical, Electric power, Revenue and customers authority, airline, banks, universities, insurances, telecommunication companies and super markets are potential users of data mining technology for their customer relationship management (CRM). Due to these researches has been and being done in world wide. Customers is the one of area in which

data mining is applied. There are researches that have been done in this area using data mining techniques in Ethiopian organization context by Jembere Abera [13] , Merga Gutema [14] and Biazen Bezabeh [15]. According to a research conducted by Xue Li [16] 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.

The problems of previous research efforts regarding to the customer relationship management and energy industry were not only related to the small proportion of dataset used, but also the data analysis was conducted by using simple statistical techniques (such as logistic regression and verification), applying descriptive data mining techniques, using simple algorithm, and/or lack of standard DM tools. There is high mismatch between data types that models have and currently available on Ethiopian electric utility. Most of the researches were conducted using small dataset and old machine learning algorithms.

Furthermore, the classical data mining techniques such as CART, neural networks and nearest neighbor techniques tend to be more robust to both messier real-world data and also more robust to being used by less expert users [17]. And also because of the use of computer hardware has dramatically increased ability in storing and processing the data makes some of the most powerful data mining techniques feasible today. In Ethiopian electric utility customers information has different attribute from the other organization and historical data of customers attribute. Classification data mining tasks helps the Ethiopian electric utility customers using common attribute easy to make decision and plan future strategic different service for organization, researcher and customers.

This research tries to focus on application of different data mining techniques for constructing a model for determining Ethiopian electric utility customers power consumption. To this end, the study tries to address the following basic research questions: -

- ✓ Which electric utility customers attribute is preferable for determining power consumption?
- ✓ What are the possible techniques to use for preparing customers raw data?
- ✓ Which data mining algorithm can be more suitable for the purpose of classification of Ethiopian electric utility customers?

1.4 Objective of the Study

1.4.1 General Objective

The general objective of this research is to apply data mining tools and techniques in order to design predictive model which can determining power consumption of Ethiopian electric utility customers.

1.4.2 Specific Objectives

In order to achieve the above stated general objective, the following specific objectives are formulated.

- ✓ To review literature on data mining technology and their application in the Electric utility for the purpose of identifying tools, techniques and data mining algorithm that will help in this study.
- ✓ To identify sources of data, collect the required data and prepare data
- ✓ To select the data mining tools and algorithms for prediction and classification of customers in the corporation.
- ✓ To design the prediction model for the classified Ethiopian electric utility customers for better decision making and successful implementation.
- ✓ To test and compare the resulting performances of the prediction and classification models

1.5 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 [18]. For this study the experimental research approach method can be used because of using database data. Experimental research is flexible to account for differences data from database and it is also systematic and appropriate data access to analysis different parameter settings and various experiments through validity testing. WEKA software is used for the study implementation since it is freely available and widely used for research purposes in data mining.

In this section the researcher discussed about the dataset used for this research and applied a hybrid model of six-step methodology.

1.5.1 Research Design

This study follows experimental research designing a predictive model for determining customers power consumption. Experimental research approach is any research conducted with a scientific approach, where a set of variables are kept constant while the other set of variables are being measured as the subject of experiment. Experimental design is plan for assigning experiment unit to treatment levels and statistical analysis classifying with plan [19]. It is important for an experimental research to establish predicted and predictor of a phenomenon, which means, it should be definite that predictor observed from an experiment are due to the predicted. Our research is experimental research design over variables to obtain desired results. Subject or industry is not a criterion for experimental research due to which any industry can implement it for research purposes. Our research is using experimental research design used to hold variable constant, assign experimental unit randomly treatment and data can be using from database.

In this study hybrid data mining model is used to enhance the knowledge discovery process by combining the business and industrial models in data mining projects. The development of hybrid models were adopted from the Cross-Industry Standard Process for Data Mining (CRISP-DM) model as its can be used for business research [2]. These models are research-oriented, which introduces the data mining step instead of the modeling step [20]. The hybrid models more understand domain problem, detailed feedback mechanisms and modification of knowledge discovered for a particular domain which may be applied in last step. There are six steps of hybrid data mining models. These are: problem domain understanding, data understanding, data preparation, data mining, evaluation of the discovered knowledge and knowledge discovery usage.

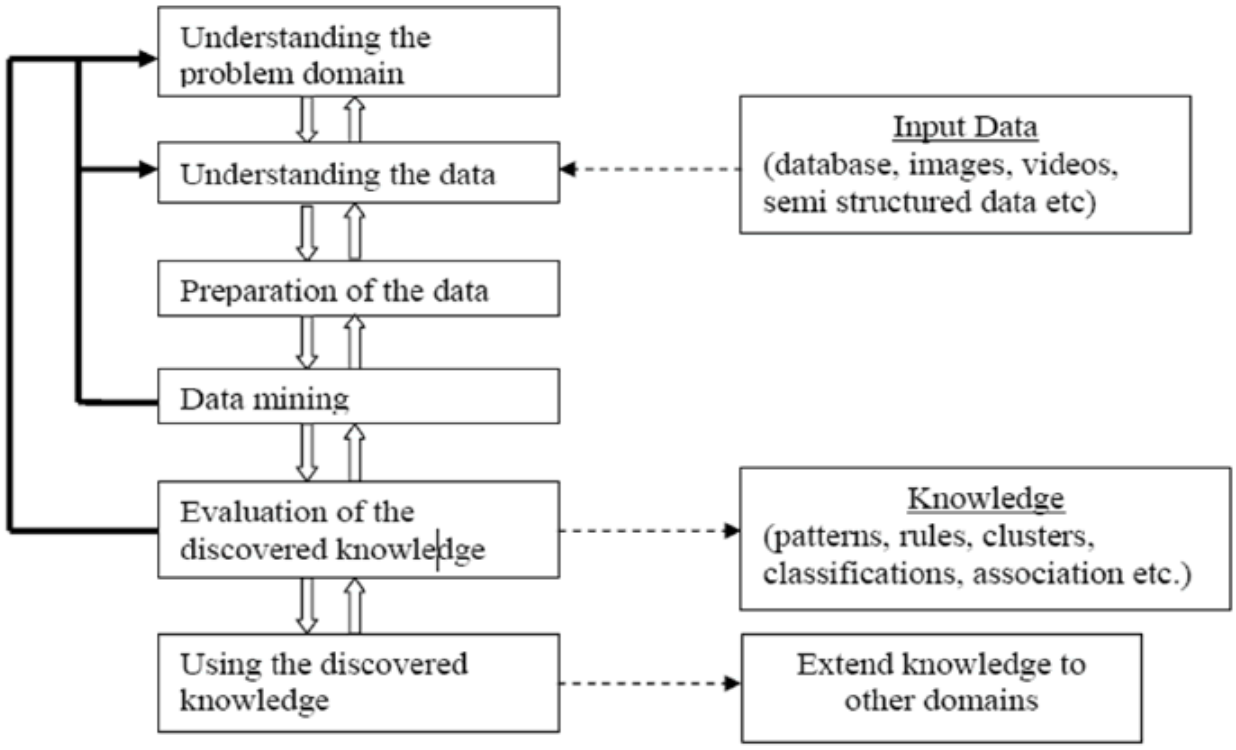

Figure 1.1 Six-step hybrid data mining model [20]

1.5.2 Understanding of the problem domain

The initial step involves task such as the problem definition, research goal determination, identification of key customers and grasping the current solution to the problem through close consultation with the domain experts. The data mining techniques to solve the problem using Ethiopian electric utility current business understand and identify key object, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions is prepared. Finally, Identify Ethiopian electric utility customers' attribute to analysis, initial selection data mining tools and classify the data based on classification of customers to solve the problem in organization, researchers and customers.

1.5.3 Understanding of the data

Data understanding is the second phase of hybrid data mining technique. Data understanding is Seek to better understand data assets and manage from problem understanding attribute. Data

understand the data available and useful for achieving the goal specified, the secondary data collection technique called database analysis has been customers and its power consumption content and structure of the data available was understood. First, describe attribute from problem understanding then select attribute that taken from new connection customer database of Ethiopian electric utility. The initial target dataset of Ethiopian electric utility customers for this research was selected from January 2015 G.C up to January 2019 G.C Ethiopian electric utility customers database. From problem domain to identify basic dependent and independent attribute for classification of customers to construct predict models. This step involves tasks such archive data export from database then collect that data, describe data, explore data, and verify data quality and choosing the size and format of the datasets. To identify the data accurate field and record values for usefulness of the data are verified with respect to the data mining goals. Understand the data based on Ethiopian electric utility database identify the use of data for classification of customers using data mining algorithm. Data can be understood using application software like Microsoft Excel and Microsoft Access easy to analysis [21].

1.5.4 Preparation of the data

After understanding of the data in hybrid data mining, preparation of data is the next step. This is one of the crucial steps to construct dataset used for modeling by Waikato Environment for Knowledge Analysis (hence forth WEKA) software. At this stage, all necessary tasks needed to prepare data mining task is finalized. Data mining techniques, tools and algorithms were decided. The data sets are pre-processed for specific data mining tasks. It consists of tasks such as sampling, testing the correlation and significance of the data, select data, cleaning the data, checking the completeness data, integrate data, handling noisy and missing values based on the needed format. This step also comprises the derivation of new attributes and summarization of the data. The process of data preparation is highly dependent on the specific data mining algorithm and software chosen for the data mining task. This research attempted to prepare the data according to the requirements of the selected data mining software, Weka and selected data mining algorithm. Weka is multi-functional data mining software. The major data mining functions incorporated in the software are data preprocessing, classification, Prediction and analysis input and output [22]. Since the selected data mining software, Weka, does not allow any inconsistency among attribute values and their definition, considerable amount of time was

spent in checking this consistency [21]. Finally, the datasets that meet the input requirements of data mining tools stated in the understanding problem are selected for modeling purpose.

1.5.5 Data mining

After data prepared using data mining methods is applied on the preprocessed data to discover knowledge. This step is application of the selected data mining methods to the prepared electric customers data and testing the generated rules whether they achieve the required classification customers. Beside it is a step of finding hidden, non – trivial and previously unknown information from the data, experimental research approach is applicable on the processed data. WEKA is used for mining the data [23]. This step involves the use of several data mining tools on data prepared. First, the training and testing procedures are designed and the data model is constructed using one of the chosen data mining tools; the generated data model is verified by using testing procedures. The meta-mining generates meta-knowledge from knowledge generated by data mining tools. It is done by dividing data into subsets, generating data models for these subsets, and generation of meta-knowledge from these data models [24].

1.5.6 Evaluation of the discovered knowledge

The results of data mining models are evaluated whether the discovered knowledge is novel and interesting and the results of the models are interpreted with respect to domain expert's knowledge. This step includes interpretation of the customers, cross checking the customers and observing the interestingness and relationship of the discovered knowledge, review the process and another means of discovering knowledge can be assessed [24]. Based on the review another step can be determined. The discovered results are a discussion with domain experts; in this case the domain experts are the data analysts from customers classification-based customers' common attribute from database and derived attribute. The discovered knowledge can be divided in to three parts namely expected and previously known which are rules that confirms user beliefs and can be used to validate the initial approach; the other is unexpected that contradicts customers beliefs and which needs further investigation for its interestingness and the need for taking an action [25]. To evaluate the interestingness of classification rules, the research uses measure based on correlation or lift.

1.5.7 Use of the discovered knowledge

Final step of hybrid model comprises of planning about the usage of discovered knowledge and plan created concerning the implementation of the knowledge discovered and the documentation of the whole research [24]. In this research the discovered knowledge is used by integrating the user interface which is designed by C# programming with a Weka system in order to show the determining power consumption of Ethiopian electric utility customer. Most of users and domain expert is test and agree the interface and rules. Lastly, after evolution of customers data to design the predication of model using hybrid data mining techniques to determining customers power consumption common attribute easy to make decision for organization and customers.

1.6 Scope and limitation of the study

This study mainly focuses on classification of Ethiopian electric utility customers using data mining techniques were examined to develop interesting classified and to extract meaningful patterns based on their customers' common attribute historical data which were gathered from Ethiopian electric utility organization. For this study, around four year (from January, 2015 to January, 2019) historical data were extracted to build classified and predict value-based customer classification model using hybrid data mining techniques. During this study, hybrid data mining was adopted to undertake the data mining process. The data mining task for this research is predication models to classification based on using Ethiopian electric customers' common attribute. Beside this study customer databases contain enough data in the organization and it is used to infer missing information using statistical methods which are not supported by modern technology.

This study is limited to examine the potential of data mining techniques in developing classification of electric customers for prediction model using different machine learning algorithms. As researcher said earlier electric customer is influenced by electric power and its type of customers. In order to assess level of impact of each factor it is better to use different data source which holds attributes in parallel with influencing factors. Here, it is limited to develop the model using only historical data of Ethiopian electric utility customer and doesn't include power data.

1.7 Significance of the study

The aim of this research have contribution for customers, organization and researcher to get different sources of information and easy to make decision. This research is using data mining techniques for classification of Ethiopian electric utility customers to identify basic attribute, to predict customers power consumption based on the value they contribute. Based on these, the subsequent benefits can be for organization, researcher and customers from the finding of this study. There is different literature reviews published in order to describe data mining techniques to classify customer using common attributes. In order to understand Ethiopian electric customers for their organizational business objectives are understanding problem, data collected, cleansed, transformed, integrated and finally prepared for experimenting with the determining prediction and classification algorithms to develop a model.

The output of the research can be used as input for customers attribute data from database. 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 utility customers. As inputs of this research are based on Ethiopian electric utility database of customers attribute. Furthermore, it can be used data mining technique to design prediction model classified customers data based on common attribute. And will help for customers and organization. It can be beneficiaries of the quality service provision.

The finding of this study can be used for organization to make decision easily, success plan for future distribution power for their customers, identifies customers power usage, Electrical plan based on use of Kilo Watt, easy to offer special services increase power and change bill and organization structure in new form of customers. Moreover, the study findings can provide insight for further researches to apply data mining technologies to advance produce and distribute power industry.

Besides, similar to power industry research concept customers data mining approach will be a new insight in electric customers which is a kind of data mining techniques and suitable for power industry, to identify their gap of this study and to show solution for many research problems.

CHAPTER TWO

Literature Review

In this chapter different literature review concerning the concept of data mining methods the applicability of data mining in customers' relation management (CRM) has been discussed. Now a day, large volume of data storage has emerged due to the fact that the increasing use of information technology in Ethiopian electric utility. Concerning the electric customers' data can be store in the database using text and number data type formats. The data stored in text and numbering formats need to be extracted in order to gain information and knowledge so as to help for different decision-making processes. Hybrid data mining technique often known as data mining is a concept which is helpful in extracting the stored data so as to get useful information and knowledge [2].

2.1 Overview of data mining

The historically of data mining is digital data acquisition and the improvement of storage technology has resulted in the growth of huge databases [2]. The need to understand large, complex, information rich data sets is common to virtually all fields of business, science, medical, industry and engineering. In the business world, organization customer data are becoming recognized as a strategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining [4]. This leads to the concept of data mining is often set in the broader context of knowledge discovery in databases using hybrid data mining models [20]. The hybrid data mining model process involves several stages understand Ethiopian electric utility customers problem domain, understanding customers data, preparing customer data, using data mining techniques, evaluation of data discovers knowledge and use to knowledge discovered classification customers [24].

Data mining is one of the most important steps of the knowledge discovery in databases process and is considered as significant subfield in knowledge management. Research in data mining continues growing in business and in learning organization over coming decades. This research

explores the data mining tools which have been developed to support knowledge management process. Data mining has become an essential factor in various fields including business, education, health care, finance, scientific because of the large amount of the data. To analyses this vast amount of data and depict the fruitful conclusions and inferences, it needs specific data mining tools [23]. To bridge the gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as data mining (DM) or knowledge discovery in databases (KDD) has emerged in recent years.

2.1.1 What is data mining?

Different international journal article and reference books provide different conceptual meaning of data mining. According to research conducted by Agrawal and Pratik Agrawal [23] data mining has a long history, with strong roots in statistics, artificial intelligence, machine learning and database research. Data mining is a step in the knowledge discovery from databases (KDD) process that consists of applying data analysis and discovery algorithms to produce a particular enumeration of patterns (or models) across the data. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize and summarize the relationships identified [2].

Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets. These tools can include statistical models, mathematical algorithms and machine learning methods. Consequently, data mining consists of more than collecting and managing data. It also includes analysis and prediction models use of algorithms that improve their performance automatically through experience, such as neural networks or decision trees and induction rules [19].

Data mining is an iterative process within which progress is defined by discovery through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an interesting outcome [26].

Data mining is the search for new valuable and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. The Best results are achieved by balancing

the knowledge of human experts in describing problems and goals with the search capabilities of computers.

According to IEEE Press Editorial Board in 2011 [4] data mining is two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans.

2.2 Data mining process Model

Data mining is a process of discovering various models, summaries and derived values from a given collection of data. A process model is the set of tasks to be performed develop a particular element, as well as the elements that are produced in each task as outputs and the elements that are necessary to do a task as inputs [19].

The data mining process models can be considered as a methodology to support the process which leads to find the information and knowledge. The reason for using the process models is in order to organize the knowledge discovery and data mining research within a common frame work. Besides the process models are helpful to understand the knowledge discovery process and provide a roadmap while planning and carrying out the researches [2].

Data mining is also known as knowledge discovery in database, refers to finding or mining knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. So, many people use the term knowledge discovery in data or KDD for data mining [27]

In Data mining, Knowledge extraction or discovery is done in seven sequential steps as shown in Figure 2.1.

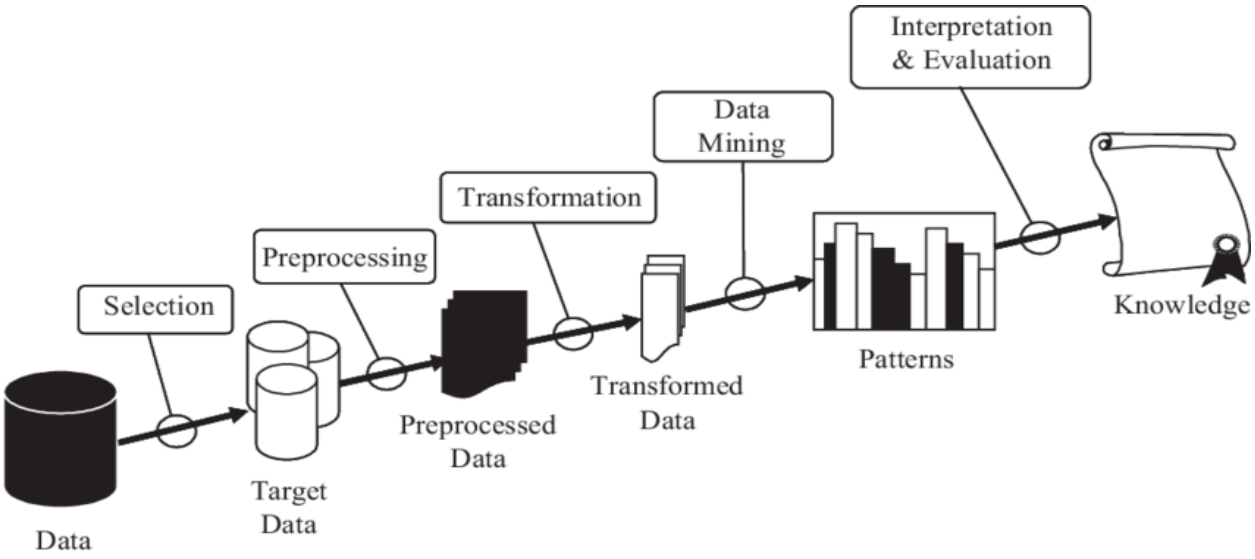

Figure 2.1 Data mining process [27]

- Data cleaning: This is the first step to eliminate noise data and irrelevant data from collected raw data.
- Data integration: At this step, various data sources are combined into meaningful and useful data.
- Data Selection: Here, data relevant to the analysis are retrieved from various resources.
- Data transformation: In this step, data is converted or consolidated into required forms for mining by performing different operations such as smoothing, normalization or aggregation.
- Data Mining: At this step, various clever techniques and tools are applied in order to extract data pattern or rules.
- Pattern evaluation: At this step, Attractive patterns representing knowledge are identified based on given measures.
- Knowledge representation: This is the last stage in which, visualization and knowledge representation techniques are used to help users to understand and interpret the data mining knowledge or result

In addition, the reasons of using the process models which are mentioned in a research study that to ensure the end result, it can be useful for the organization, customers and researchers. Another reason for the need of data mining process models is mentioned by providing support

for managerial processes and to seeking new knowledge [26]. There are three data mining process models those are CRISP, academicals research and hybrid data mining process models.

2.2.1 CRISP-DM (Cross-Industry Standard Process)

CRISP-DM is industry 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 [19]. So, a generic and standardized approach is needed for the use of data mining in order to help organizations, customers and researchers. 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. The CR.ISP-DM (Cross-Industry Standard Process for Data Mining) consists of the following six steps as shows figure 2.2.

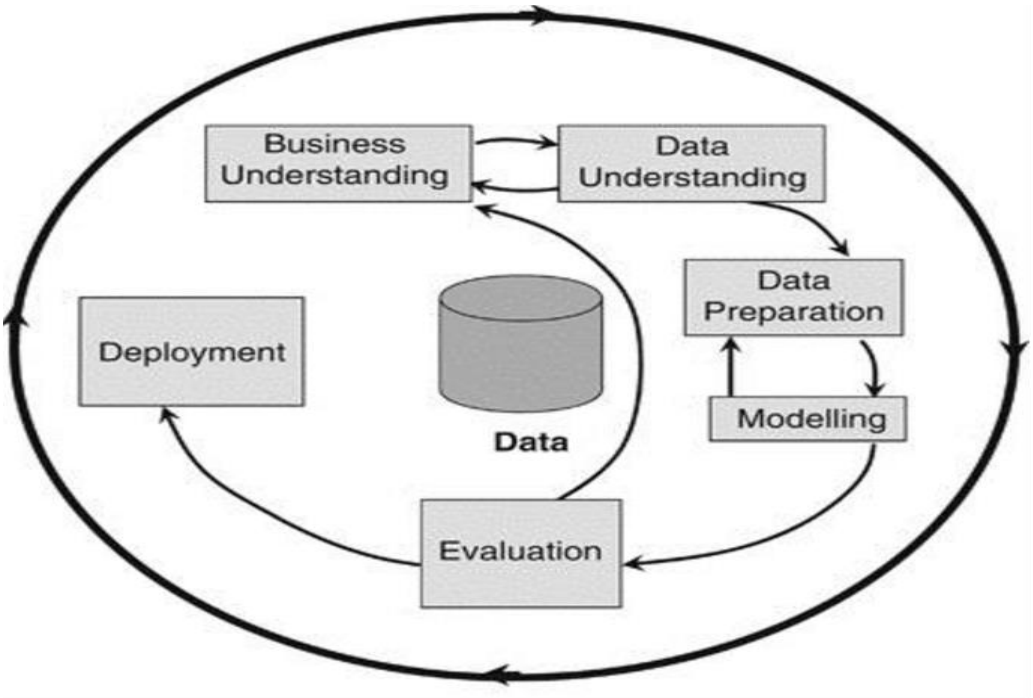

Figure 2.2 Phase of the CRISP-DM models [28]

1. **Business understanding**. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a data mining problem definition, and designs a preliminary research plan to achieve the objectives.
2. **Data understanding**. This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, description, exploration, verification data quality and detection of interesting data subsets.
3. **Data preparation.** This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection, data cleaning, construction of new attributes and transformation data form data set.
4. **Modeling.** At this step, various modeling techniques are selected and applied. Modeling Usually involves the use of several methods for the same data mining problem type and the calibration of their parameters to optimal values. In this step selection of modeling technique, Generation of test design, Creation of models, and Assessment of generated models.
5. **Evaluation.** After one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached.
6. **Deployment.** Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as plan deployment, plan monitoring and generating a report or as complex as implementing a repeatable Knowledge discovers process.

2.2.2 Academic Research Models

Academic Research model the focus point was to provide a sequence of activates that would help to execute a Knowledge Discovers Process in an arbitrary domain. For instance, the well-known model which is developed in 1996 consists of nine steps [29]. This model consists the following nine steps data mining knowledge discovery process [24].

The first step is understanding the application domain, identify the data mining knowledge discovers goals. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge. The second step which is creating a target data set. Here the data miner selects a subset of attributes and data points that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.

Data cleaning and preprocessing is the third step. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes. The fourth step is data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data. In choosing the data mining task step, the data miner matches the goals defined in first step with a particular data mining method, such as classification, regression, clustering, etc.

After data reduction select particular data miner methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate. The other step is exploratory analysis model and hypothesis selection, in this step data mining, interpreting mined patterns and consolidating discovered knowledge.

Finally, in academic research model method using data mining, interpret in the mined pattern data set last to consolidation discovered.

2.2.3 Hybrid Data mining Models

Hybrid data mining model is combination of academic research models and industrial models has led to the development of hybrid models [30]. Hybrid data mining model is to developed six-step knowledge discovering process model [20]. It was developed based on the CRISP-DM model by adopting it to academic research. The main different between the two research process models is providing more general research-oriented description and introducing a data mining step instead of the modeling step. The CRISP-DM model has only three major feedback sources mechanism, while the hybrid model has more detailed feedback mechanisms and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains [24]. The main different and extension include provide more general, research-oriented description of the step and introducing a data mining step instead of modeling step, further description of six steps as shown in figure 2.3.

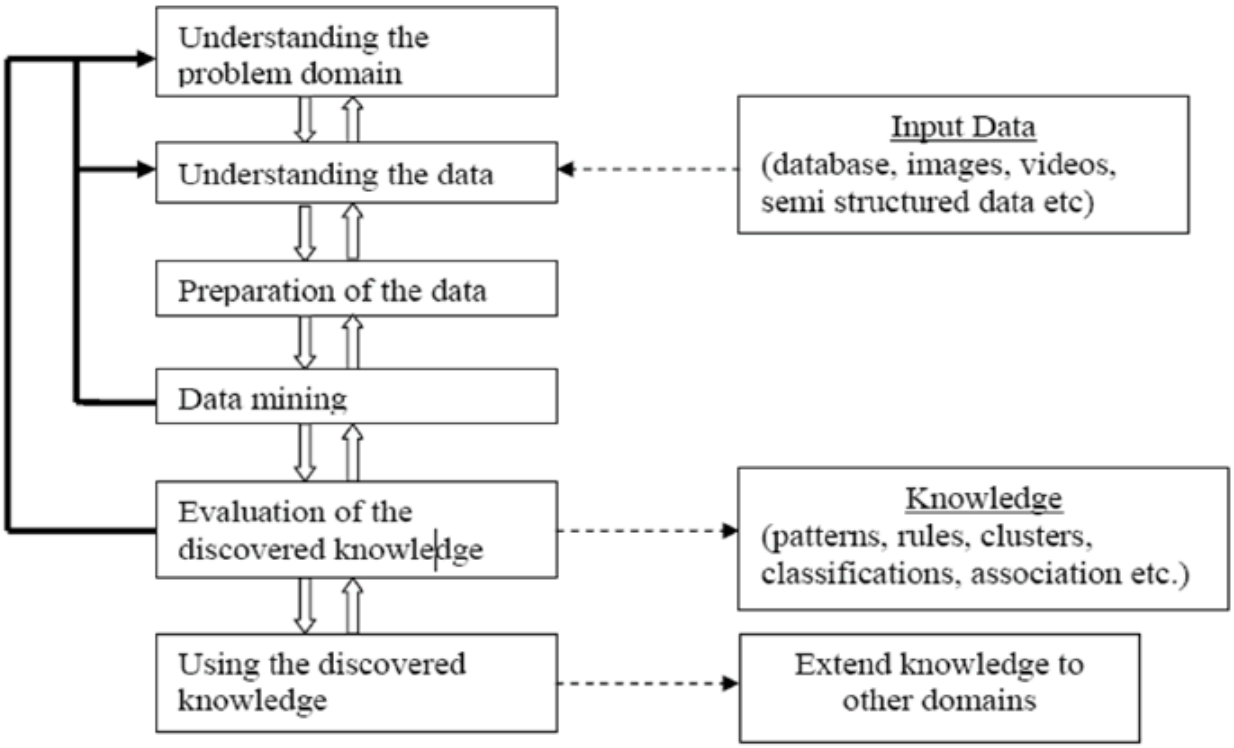

Figure 2.3 Hybrid process data mining models [20]

The following are the description from figure 2.3 of the six steps hybrid data mining models.

Step 1 **Understanding the problem domain.**

In this step one works closely with domain experts to define the problem and determine the research goals, identifies key customers, and learns about current solutions to the problem. It involves learning domain-specific terminology. A description of the problem including its restrictions is done. The research goals then need to be translated into the data mining knowledge discover goals and may include initial selection of potential data mining tools.

Step 2 **Understanding the data.**

This step includes collection of customers common attribute data and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the data mining knowledge discover goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

Step 3 Preparation of the data.

This is the key step upon which the success of the entire knowledge discovery process depends; it usually customers about half of the entire research effort. In this step, we decide which data will be used as input for data mining tools in step 4. It may involve sampling of data, running correlation and significance tests, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms by derivation of new attributes and by summarization of data (data granularization). The result would be new data records, meeting specific input requirements for the planned to be used data mining tools

Step 4 Data Mining

This is another key step in the hybrid knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, such as rough and fuzzy sets, Bayesian methods, evolutionary computing, machine learning, neural networks, clustering, preprocessing techniques, etc. This step involves the use of several data mining tools on data prepared in step 3. First, the customers' data for sampling and testing procedures are designed and the data model is constructed using one of the chosen data mining tools; the generated data model is verified by using testing procedures. The meta-mining generates meta-knowledge from knowledge generated by data mining tools. It is done by dividing data into subsets, generating data models for these subsets, and generation of meta-knowledge from these data models [20].

Step 5 Evaluation of the discovered knowledge.

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

Step 6 Use of the discovered knowledge.

This final step hybrid data mining models it consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other

domains. A plan to monitor the implementation of the discovered knowledge is created and the entire Research documented. Finally, the discovered knowledge is deployed [31].

According to the above description hybrid data mining is better to use in this research because identify electric utility customers problem from organization and users to explain detail feedback mechanism to select, analysis and review data.

Here under in table 2.1 comparison of the three data mining knowledge discovery (DMKD) process models are given bellows.

Six Step CRSIP DMKD	Six step Hybrid DMKD	Nine Step Academic Research Model
1. Business objective determination	1) Understanding the domain problems	1. Understanding application domain, identifying the DMKD goals
2. Data Understanding	2) Understanding the data	2. Creating target data set
3. Preparation of data	3) Preparation of the data	3. Data cleaning and preprocessing
		4. Data reduction and projection
		5. Matching goal to particular data mining methods
		6. Exploratory analysis models and hypothesis selection
4. Modeling	4) Data mining	7. Data mining
5. Evaluation	5) Evaluation of the discovered knowledge	8. Interpreting mined patters
6. Deployment	6) Use of the discovered knowledge	9. Consolidating discovered knowledge

Table 2.1 Summary of DMKD process models [24]

To evaluate the six-step Hybrid process model and compare it with the nine steps Academic research process model and the Six-step CRSIP model we show in table corresponding steps of the three models. The common steps for the three models are domain understanding, data mining, and evaluation of the discovered knowledge.

2.3 Data mining Tasks

Data mining satisfy its main goal by identifying valid, potentially useful, and easily understandable correlations and patterns present in existing data. Data mining functionalizes in two high-level primary goals of data mining in practice tend to be prediction and description [32].

The tasks of data mining are very diverse and distinct because there are many patterns in large database. Different kinds of data mining methods and techniques are needed to find different kinds of knowledge discovered [7]. There are two main types of data mining methods are verification-oriented means system verifies the user's pre-defined hypothesis and discovery oriented in which means the system can finds new rules and patterns autonomously from the data [31].

In other approach data-mining task is divided into two approaches supervised or directed and in unsupervised or undirected the task for direct data mining uses available data. The goal of supervised or directed data mining is to use the available data as like predictive data-mining to build a model that describes one particular variable of interest in terms of the rest of the available. The user selects the target field and directs the computer to determine estimate, classify or predict its value, In case indirect data Ming task goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patters and relationships one they have been found [23].

Data mining tasks are two those are verification and discovery. Verification means goodness of fit, hypothesis of testing and analysis of variance. Discovery data mining task also divided in two-part predication and description. In description discovering data mining task is different part like clustering, summarization, linguistic summary and visualization. From prediction task of data mining is classification, time series analysis and regression. In the above description the data mining task as we can see in figure 2.4 below.

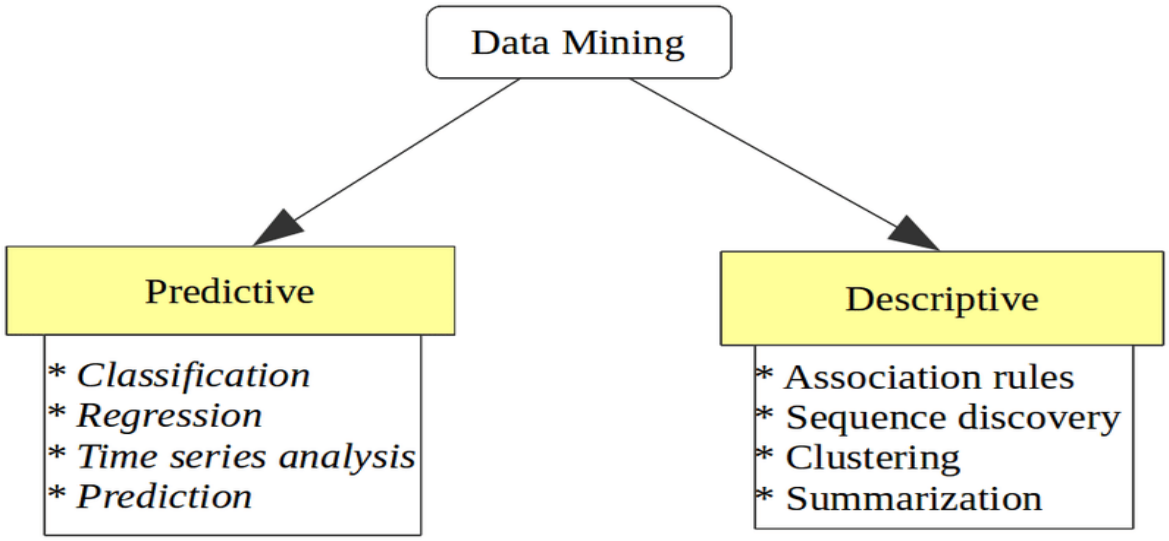

Figure 2.4 Predictive and Descriptive data mining model [33]

2.3.1 Prediction data mining Tasks

In the prediction goal is to develop a model which can infer a single aspect of the data predicted attribute from some combination of other aspect of the data predictor attribute. Prediction requires having labels for the output attribute for a limited data set, where a label represents some trusted "ground truth" information about the output variables value in specific cases [34]. Broadly, there are different types of prediction: classification, time series analysis, regression, and prediction. In classification, the predicted variable is a binary or categorical variable. In Some popular classification methods include decision trees, logistic regression (for binary predictions), and support vector machines. In regression, the predicted variable is a continuous or numerical. Neural networks and support vector machine are regression.

Classification – according to a research which is conducted by [24], the aim of classification is to classify items into several predefined classes. Given a collection of training samples, this type of task can be designed to find a model for class attributes as a function of the values of other attributes. Classification is one of several methods intended to make the analysis of very large data sets effective. To create an effective set of classification rules which answers a query, makes decision based on the query and predicts future plan.

Classification is the most commonly applied data mining technique, which employs a set of pre classified attributes to develop a model that can classify the population of records at large. The data classification process involves learning and classification. In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data sets. The classifier-training algorithm uses these pre-classified attributes to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier [31].

Regression is predictive modeling and analysis is used to make predictions based on existing data by applying formulas. It is a statistical method of data mining. Regression actually used to model between the one or more dependent variables and independent variables. It can be used for building model or classifiers which can analyses the historical data to predict the future trends 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 it is uses existing values to forecast what other values will be [7].
According to a research which is conducted by fayyad, piatetsky & P.Smyth [34]; in prediction the aim is to predict a value of a given continuously valued variable based on the values of other variables, assuming either a linear or nonlinear model of dependency. These tasks are studied in statistics and neural network fields. In this case some well-known regression methods include linear regression, neural network and support vector machine regression.

Prediction is one of prediction data mining task. In prediction the goal is to develop a model which can infer Single aspect of data with combination of some aspect of the data [28] like Similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem type is also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets generally apply [26].

Time-series analysis is a prediction application with one or more time-dependent attributes that usually involves prediction of numeric outcomes such as the future price of individual stock.

This time-series represents a collection of values obtained from sequential measurements. Time series data mining gives natural ability to visualize the shape of data. Time series are very long, considered smooth, as subsequent values are within predictable ranges of one another. Time series are popular in many applications, such as stock market analysis, economic and sales forecasting. It can be useful for observation of natural phenomena like atmosphere, temperature, wind, earthquake, scientific and engineering experiments, and medical treatments [4]

2.3.2 Description data mining tasks

The descriptive model is the data mining model mostly identifies patterns or relationships in data Sets. It serves as easy way to explore the properties of the data examined earlier and not to predict new properties. It focuses on finding human interpretable patterns describing the data. It describes 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 [34]. Description model encompasses task perform as Clustering, Summarization, Association and Sequence.

Clustering is a well-studied and well-known data analysis technique in statistics. Various definitions about clustering method of descriptive modeling exist in literature Clustering is the process of grouping a set of data objects into multiple groups or clusters. So that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures [2].

According to a study which is conducted by [24], the aim of clustering is to identify a set of categories, or clusters, that describe the data. Clustering is particularly useful in cases where the most common categories within the data set are not known in advance; if a set of clusters is optimal, within a category, each data point will in general be more similar to the other data points in that cluster than data points in other clusters.

Summarization is the generalization or abstraction of data. It is technique maps data into subsets with simple descriptions. The summarized date set gives general overview of the data with aggregated information. Summarization can scale up to different levels of abstraction and

can be viewed from different angles. It is a key data-mining concept involving techniques for finding a compact description of dataset. Summarization approaches are basically mean, standard deviation, variance, tabulating, mode, and median. These approaches are often applied for interactive exploratory data analysis, data visualization and automated and automated report generation [33].

The aim of summarization is to find a concise description for a subset of data. Tabulating the mean and standard deviations for all fields is a simple example of summarization. There are more sophisticated techniques for summarization and they are usually applied to facilitate automated report generation and interactive data analysis to find a model that describes significant dependencies between variables [34].

Association is used to discover relationships between variable and object. In these techniques, the presence of one pattern implies the presence of another pattern i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model. For market basket analysis, Commodity management advertising association rules is popular and well researched method for discovering interesting relations between the variable in large data base. It is intended the technique that all possible combinations of potentially interesting product groupings can be explored and to identify strong rule discovered in database. They also play an important role in shopping basket data analysis, product clustering, and catalog design and store layout. This association rules build by programmers and use to build capable of machine learning [33].

According to describe of [7] is discovery of togetherness or connection of object, such kind of togetherness or connection is association rule. An association rule reveals the association relationship among objects, i.e., the appearance of set of objects in a database is strongly related to the appearance of another set of objects. The association rules can be useful for marketing, commodity management, and advertising.

Sequential pattern mining is the machine learning task that addresses the problem of discovering the existing frequent sequence in a given database. This mining task is able to discovering the structured behaviors specifically sequential patterns. This pattern exists when the data which is already to be mined has some sequential nature.

2.4 Classification algorithms

Classification is data mining technique that assigns categories to collection of data in order to aid in more accurate predictions and analysis. Classification is one of several methods intended to make the analysis of very large data sets effective. To create an effective set of classification rules which answers a query, makes decision based on the query and predicts the customers. The main objective of the classification algorithm is to mine, how that set of attributes reaches its conclusion [31]. There are different classification data mining algorithm those are Decision Tree, J48, and Random tree, KNN, CART, Random Forest, Bagging, PART and Naive Bayes.

2.4.1 Decision Tree

Decision Tree are powerful and popular tools for classification and prediction. A decision tree allows the calculation of forward and backward and because of that correct decision will be made automatically. Decision tree is classifier in the form of tree structure which is leaf and decision node. Leaf node is indicating the value of target attribute and decision node also single attribute value with one branch and sub tree for each possible outcome of the test [35]. Decision trees are commonly used decision analysis, which helps to reach a target by identifying a strategy and used for calculating conditional probabilities. Decision tree is easy for the humans to understand and less computation required for classification. They are fast and scalable.

A decision tree is made up of a hierarchical structure of decision nodes which are linked with branches. To determine the root node and after that the next nodes, for each attribute we calculated which one will most exactly classify objects according to the decision variable values. From each decision node, branches going out to each node correspond to the various possible answers to the test. For a continuous variable, the test will be in this form: 'if the value on the variable is higher than a threshold value then takes the first connect, otherwise take the second'. For the categorical variables, one of the outgoing branches of the node is associated with each category of the variable. The final nodes are called leaf nodes and are linked during the training phase to one of the categories of the decision variable. To realize a prediction on a new individual, it is enough to make him traverse the graph to a leaf. The leaf which the new individual reaches will determine the predicted value of the decision variable [36].

A Decision Tree is one of the most popular classification algorithms in current use in data mining and machine learning. Some of decision tree classifiers are J4.8, Random Tree, Random Forest, and others. But here, the focus of the discussion is J48 and random tree as it is one of the algorithms to be used for modeling in this study.

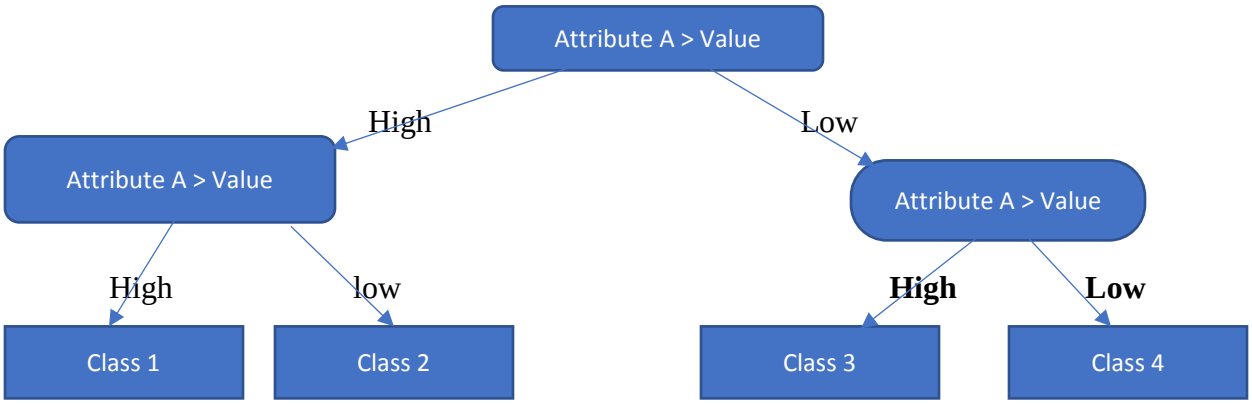

Figure 2.5 An example of decision tree model

2.4.1.1 J48 Decision Tree Algorithm

J48 is actually of prediction machine learning model, which decide the target attribute based on various attribute of available dataset. J48 decision tree classifier uses two phases those are tree construction and tree pruning [31]. Tree constructions are start with the whole data set at the root and check the attribute of data set. Tree pruning is to identify and remove leaf node that reflects noise and outliers to reduce classification error. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation [37].

J48 is an open source Java implementation of the C4.5 decision tree algorithm. A decision tree is actually a predictive machine-learning model, which decides the dependent variable (i.e, Target value) based on various attributes of the available training data set. The internal nodes of a

decision tree denotes varied attributes, the connecting branches of various nodes give us the likely values of the attributes and the terminal node states the classification of the dependent variable. J48 decision tree classifier uses following steps.

A. Starts with the whole data set at the root

B. Check the attribute of the data set and partition them based on the following cases

Case I: - If the attribute value is clear and has a target value, then it terminates the branch and assigns the value as Target value (classification)

Case II: - If the attribute, gives the highest information, then continue till we get a clear decision or run out of attributes.

Case III: - If we run out of attributes or we are presented with ambiguous result, then assign the present branch as target value.

Case IV: - ignore missing values.

2.4.1.2 Random Tree Decision Tree Algorithm

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In standard tree every node is split using the best split among all variables. In a random forest, every node is split using the best among the subset of predicators randomly chosen at that node. The algorithm can deal with both classification and regression problems. Random trees are a group (ensemble) of tree predictors that is called forest. The classification mechanisms as follows: the random trees classifier gets the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of votes. In case of a regression, the classifier reply is the average of the responses over all the trees in the forest. Random trees are essentially the combination of two existing algorithms in machine learning: single model trees are merged with random forest ideas. Model trees are decision trees where every single leaf holds a linear model which is optimized for the local subspace explained by this leaf [38].

2.4.1.3 Random forest

Random forests or random decision forests are an ensemble learning method. It is used mainly used to solve classification, regression problems and also other problems. Random forest is one of the accurate learning algorithms. The basic concept of the algorithm is to build many small decision-trees and then merging them to form a forest. It is computationally easy and cheap process to build many such small and weak decision trees [37]. Its ability to handle thousands of input variables without variable deletion along with quick learning process and its effective method for estimating missing data and maintains accuracy are major sited attributes of this algorithm. The algorithm for random forests uses the common technique of bootstrap bagging.

2.4.2 Naïve Bayes classifier

Neiva Bayes is a supervised learning algorithm which is used for data classification using statistical method. It is a probabilistic classifier that helps to classify the given input over a set of classes using probability distribution [35]. This method goes by the name Naïve because it naively assumes independence of the attributes given the class. Most of classification data mining techniques applying Bayes rule to work out the probability of correct class for given particular attribute. Neiva Bayes is use simple algorithm to implement and its great computational efficiency and classification performance. It gives accurate result for most classification and prediction problems. Additionally, network intrusion detection is support by Neiva Bayes. The limitation of Neiva Bayes is in same case classifier in which the precision of algorithm decreases if the amount of data is less and probability of the class attribute zero the problem will be zero problem. Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes theorem studies comparing classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases [26].

Bayes Basics Theorem

Let X is a data sample (evidence): class label is unknown and let H be a hypothesis that X belongs to class C. Classification is to determine P (H|X), (posteriori probability), the probability

that the hypothesis holds given the observed data sample X. P(H) (prior probability) is the initial probability or the prior probability of each class based on the training tuples and P(X|H) (likelihood) is the probability of observing the sample X, given that the hypothesis holds or conditional probabilities of attributes value for each class, P(X):

$$P(H \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H)P(H)}{P(\mathbf{X})} = P(\mathbf{X} \mid H) \times P(H) / P(\mathbf{X})$$

Probability that sample data is observed and hence from the given training data X, posteriori probability of a hypothesis H, P (H|X), follows the Bayes‟ theorem [26]. Informally, this can be written as posteriori = likelihood x prior/evidence

Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-dimensional attribute vector X = (x1, x2… xn). Suppose there are m classes C1, C2… Cm. Then predict the class label of a tuple using Naïve Bayesian classification and the classification is to derive the maximum posteriori, i.e., the maximal P (Ci|**X**). This can be derived from Bayes' theorem

$$P(C_i \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C_i)P(C_i)}{P(\mathbf{X})}$$

Since P(X) is constant for all classes, only needs to be maximized

$$P(C_i \mid \mathbf{X}) = P(\mathbf{X} \mid C_i)P(C_i)$$

Therefore, one can predicts X belongs to specific class if and only if the probability P (Ci|X) is the highest among all the P (Ck|X) for all the *k* classes. A simplified assumption states that attributes are conditionally independent (i.e., no dependence relation between attributes) and yields:

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

Therefore, due to its lower error rate of predictive accuracy and capacity to provide a standard of optimal decision making, Naïve Bayes was implemented in this research experimentation.

2.4.3 PART Rule Induction

PART is one of rule induction which is the process of extracting useful, if then‟ rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents IF-THEN rules for classification. An IF-THEN rule is an expression of the form Even though the pruned trees are more compact than the originals; they can still be very complex. Hence, generate rules to make a decision tree model more readable, it can be transformed into an IF-THEN decision rule. Decision rules can be generated from a decision tree by traversing any given path from the root node to any leaf. The complete set of decision rules generated by a decision tree is equivalent to the decision tree itself. Rule induction or decision rule classifiers are set of IF-THEN classification. An IF-THEN rule induction is an expression of the form IF condition THEN conclusion. If the condition in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuples [5].

2.4.4 K-Nearest Neighbor algorithm

The k-Nearest Neighbor algorithm uses the concept of learning which is basically a comparison process where a given example to be tested known as test example is compared with the similar training examples in the dataset. Each training example stored in the database consists of n-dimensional pattern space having n attributes and represents a point in the n-dimensional space. For an unknown new test example, KNN searches for the k training examples in the pattern which are closest to that particular unknown example. Then, the searched k-training examples become the k “nearest neighbors” of that particular unknown example. Closeness used here can be defined with the help of a distance metric like Cartesian or Euclidean distance [37]. The algorithm is composed of two basic steps. In the first step k training examples that are closest to the unseen examples are searched and in the second step the most commonly occurring classification for these k examples is chosen.

2.4.5 Bagging Algorithms

Bagging Algorithms can be constructed in such a way that by combining the properties of two or more algorithms. According to research [39] stated that combining the decisions of different models means amalgamating the various outputs into a single prediction. The simplest way to do

this in the case of classification is to take a vote in the case of numeric prediction it is to calculate the average (perhaps a weighted average). Bagging and boosting both adopt this approach, but they derive the individual models in different ways. In bagging the models receive equal weight, whereas in boosting weighting is used to give more influence to the more successful ones just as an executive might place different values on the advice of different experts depending on how successful their predictions were in the past. To introduce bagging, suppose that several training datasets of the same size are chosen at random from the problem domain. Imagine using a particular machine learning technique in order to build a decision tree for each dataset. You might expect these trees to be practically identical and to make the same prediction for each new test instance. But, surprisingly, this assumption is usually quite wrong, particularly if the training datasets are fairly small. This is a rather disturbing fact and seems to cast a shadow over the whole enterprise. Slight changes to the training data may easily result in a different attribute being chosen at a particular node, with significant ramifications for the structure of the sub-tree beneath that node. This automatically implies that there are test instances for which some of the decision trees produce correct predictions and others do not [40].

2.5 Application of data mining techniques

Many foreign and local organizations are using data mining to manage all phases of the customer relationship management such as new customers, increasing power need from existing customers, and retaining good customers. 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, bank, insurance, credit card, education, medical, marketing, transport and military companies are using data mining technique in different purpose. Telecommunication and credit card organization are leaders in applying data mining to detect fraudulent use of their services. Insurance and bank 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. Marketing area to application of data mining technique is 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 electric power sector data mining can be used to classification and prediction of customers based on different attributed.

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. Banking system can be used 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.

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. Telecommunication companies utilize data mining application to improve their marketing efforts, identify fraud, and better manage their telecommunication networks (network fault isolation and prediction).

Educational: Data Mining is a blooming field which provides knowledge from educational Environment data. The goals of EDM are identified as predicting students' learning behavior, emotions and skills. This study improves the educating methods by understanding the ward and to take accurate decisions respectively.

2.6 Related works

Currently, there are some researches that were applied to investigate the application of data mining tools and techniques on customer relationship management and related issues. The following works are reviewed by the researcher to help clarify the significance of the research problem and to show the gap in the current local and international studies. In Ethiopian there are different research using data mining techniques it concerns with customers relation management and in different area. Customers data mining is one of the application data mining. Customers data mining is upcoming field knowledge discovery. Due to widespread growth of Ethiopian electric utility customers classified related use of power consumption, and on their common attribute of the customers can be accurately done though customers data mining.

In the context of other countries, Ziafat H. and Majid Shakeri [18] using data mining techniques in customer segmentation, Konstantinos Tsiptsis [19] apply data mining techniques in customers relationship management inside customers segmentation, Saxena.A.K. [41] Investigated on data mining technology and its application to power system from the power generation plants the electrical energy is transmitted and distributed to end users. Frequent failure of various equipment and the systems has made it impossible to maintain the continuity of supply. The operation and planning of power systems provide a large amount of data and it is difficult to extract the useful information from this large database that is continuously used by operators. Data mining is a process of extracting interesting and previously unknown knowledge from a set of data. The data mining techniques help power systems planner/operator to have smooth system planning, operation and are useful for extracting useful information from the existing data banks. In this research using cluster data mining machine learning technique used in power system is artificial neural networks, but the other data mining techniques are also used in some power system areas with good results using neural network to solving the problem of fault classification in the transmission lines.

A Research which conducted by Mishra and Kumar Routray [42], discussed about popularly adopts the concept customer relationship management (CRM) application and Data mining statistical techniques in large databases in cloud computing environment in order to achieve business analysis. The data mining rules especially association rule mining may be deployed in CRM implementations. The research further describes data mining technologies provided through cloud computing environment is an absolutely necessary characteristic for today's

business to make proactive, knowledge driven decisions for target marketing as it helps them have future trends and customer behaviors predicted under CRM.

According to Pitt, [43] Researcher Electric Load Profiling Using Data Mining is subjected to adaptive clustering algorithms, with the objective of condensing the vast amount of data in the original database into a small number of representatives (end) profiles each pertaining to a particular subset of the domain of the database. The clustering of individual customers' profiles (rather than atoms) is investigated as an extension to clustering of atoms.

A Research which is conducted by Mishra and Kumar [42] discussed about to support and maximizes the organization customers in adaption of customer relationship management and data mining tools in the data intensive cloud computing environment. The main object of this study classifies customers based in their common attributes in customers' relationship management. The classification models were built with J48 decision tree and multilayer perceptron neural net algorithms. From those two classifications model the best classification model was selected by comparing the overall accuracy in classifying high value customers and accuracy in classifying low value customers.

In the Ethiopian context, Jembere and Abera [13], applied data mining techniques to support customers relationship management of Ethiopian Air Lines, Biazen [15], also applied data mining techniques to support customers relationship management of Ethiopian revenue and customers authority. A research conducted by Hailemariam [44], applied data mining techniques to customer profile analysis in the Ethiopian electric power corporation.

According to the study conducted by Gutema [14], the research stated that the customer segmentation models built by using the combination of classification and clustering data mining techniques are necessary for the life Addis district and marketing department of Ethiopia Insurance Company in order to identify the valuable segments of customers and other factors underlying variations of the customers values.

Another research which done by Jembere and Abera [13] focused on using data mining techniques support customers relationship management in the case of Ethiopian airline. The problem of this study there is individual fight activity and demographic information of filling. Based on this deriving new attribute from database then using clustering data mining phase the K-means clustering algorithm was used to segment individual customer record in similar behaviors. The results from this study were encouraging and confirmed the belief that applying

data mining techniques could indeed support CRM activities at Ethiopian Airlines. In the future, more segmentation and classification studies by using a possible large amount of customer records with demographic information and employing other clustering and classification algorithms could yield better results.

CRM Data Mining is concerned with developing new methods to discover knowledge from customers database and can used for decision making. Researcher collected the customers data that have different attribute about their historical and current records data and presents a model based on classification approach to find an enhanced evaluation method for predicting the electric customers power consumption data mining techniques using Decision tree and Naïve Bayes classifier to interpret potential and useful knowledge. The researcher suggests that using different selected classification algorithms based on WEkA.

CRM data mining is improving customers management system one of improvement the activities in the organization. Customer relationship management (CRM) is the overall process of exploiting customer- related data and information, and using it to enhance the revenue flow from an existing customer. As part of implementing CRM, airlines use their frequent flyer databases to get a better understanding of their customer types and behavior. CRM, airline using clustering sub phase the K-means clustering algorithm was used to segment individual customer records into clusters with similar behaviors. In the classification sub-phase, J4.8 and J4.8 PART algorithms were employed. Therefore, as a final output of this research, a prototype of Customer Classification System is developed. The prototype enables to classify a new customer into one of the customer clusters, generate cluster results, search for a customer and find the cluster where the customer belongs, and also provides with the description of each customer clusters [13].

Data mining in Ethiopian revenue and customs authority environment is a powerful tool for organizational business sector. In this study, different characteristics of the Ethiopian revenue and customs authority customers' data were collected from the customs database. Tree Classifiers have main role to developing new methods to discover knowledge from customers database and can used for decision making in CRM system. The tree classifier is J48, decision stump, random forest and random tree algorithm using data mining tools (WEKA) for analysis the Ethiopian revenue and customs authority customers. The researcher deals with a comparative study of various tree classifiers from these he suggests that classification model which was built using J48 decision tree algorithm with default 10-fold cross-validation outperforms 99.95% of

overall accuracy rate. The classification models were built with J48 decision tree and multilayer perceptron neural net algorithms. From these two classifications models the best classification model was selected by comparing the overall accuracy, accuracy in classifying high value customers and accuracy in classifying low value customers [15].

A research conducted by Hailemariam [44], applied data mining techniques to customer profile analysis in the Ethiopian electric power corporation. The research stated that the customer segmentation models built by using clustering data mining techniques are necessary for the Ethiopian electric power corporation analysis customer profile in order to identify the valuable segments of customers and other factors underlying variations of the customers values. According to hailemariam 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 and finally prepared for experimenting with the clustering algorithms to develop a model. Using K-means clustering algorithm was used to segment customer records into clusters with similar behaviors.

Most of the research gaps regarding electric customers classification for prediction is varying in processing method, machine learning algorithm, and datasets to get more accurate result. Different energy industry has stand for different customers distributing of electric power based on interested. And, also energy industry has different environmental structure. Different customers have their own describing attributes. Ethiopian electric utility is inter connect and distribution electric power for different customers type and distributed to different region. Ethiopian electric customers data holding available customers attribute is push in to central system of the organization.

Researches, conducted regarding this area is more of directly pertinent to specific scope. This paper mainly focuses on classification of Ethiopian electric new customers. There is no prior research conducted on classification of predicting electric customer power consumption. During the review of prior research on the related area there is a gap between data types used and currently available data of the Ethiopian Electric utility customers. In addition to this, most of the researches were conducted using either one or two machine learning algorithms. Here, it is proposed to apply several machine learning algorithms to show their effects on classification of customer power consumption prediction, and to add theoretical contribution to the area.

CHAPTER THREE

DATA PREPARTION

3.1 Overview of Ethiopian electric utility

Ethiopian electric utility is fully monopolized distributed electric consumption by the government with no competition and only governmental energy retail companies to supply electricity for customers. The organization had two electric energy supply systems; that are the Interconnected System (ICS) and the Self-Connected System (SCS) [9]. The main energy source of Ethiopian electric utility is ICS for hydro and wind power plant and, the other electric energy supply is SCS, it is mini hydro power plant, geothermal and diesel power generators allocated in various locations of the country. EEU is a company responsible for low power transmission, distribution and sales of electricity all over the nation. According to Energy pedia published in 2016 [9] [45], only 27 % of the population in Ethiopia has access to electricity grid. Ethiopian electric utility infrastructure is increasing due to an extension of the national grid and an increasing number of stand-alone systems and mini-grids. The Ethiopian electric utility (EEU) [6] is one of the public enterprises assigned both to widen and modernize universal access to electricity. As part of its responsibility of expanding the service with suitability and simplicity, it has introduced the prepaid card billing system for electric power consumption.

The purpose of the organization is to engage in the business of distributing and selling electrical energy in accordance with economic and social development policies and priorities of the government and to carry out any other related activities that would enable it achieve its purpose.

The term electric utility as used herein refers to any entity that generates and distributes electrical power to its customers, that purchases power from a power-generating entity and distributes the purchased power to its customers, or that supplies electricity created by alternative energy sources, such as hydro power, diesel, wind power or otherwise, to power generation or distribution entities through the Federal Energy Regulatory Commission (FERC) electrical grid [9]. Electric utility is a company in electric power industry that engages in electricity generation and distribution of electricity for sale generally in a regulated market. Electric utility is Commercial entity that owns and operates equipment and facilities for the transmission, distribution of electric energy which it sells to general public and/or industrial consumers [6].

3.1.1 Electric utility customers and its type

Electric utility customers consist of generating stations that produce electrical power, high voltage transmission lines that carry power long distant from sources to demand centers, and distribution lines that connect individual customers. Customers are usually charged a monthly service fee and additional charges based on the electrical energy in kWh consumed by the domestic or commercial during the month. Commercial and industrial consumers normally have more complex pricing schemes. These require meters that measure the energy usage in time intervals to impose charges based on both the amount of energy consumed and the maximum rate of consumption, i.e. the maximum demand. Frequent reporting also allows the retailer to pass on the spot price to its customers.

Ethiopian electric customers type can be classified into two; those are Self-connect (SCS) and inter connect (ICS) customers. Both customers types are including domestic, industrial, commercial, own consumption, retired staff consumption and street light [6].

- ✓ **Domestic Customers**

Domestic Customers: are customers who consume the power supply for domestic purpose (for house lighting only). It is used to distribute electric current in single-phase transmission line. Domestic electricity consumption is all quantities of electric energy made available for both inter-connect and self-connect system. Most of domestic customers use electric power for lighting and cooking food. Domestic customer is like home, hospital, educational institution, museum and other organization. Ethiopian electric utility domestic customers have unique tariff code. The statues of domestic customer are created service, progress and prepared.

- ✓ **Commercial Customers**

Commercial Customers: are customers who consume power supply for commercial purpose and their power consumption varies according to their need's requirements. Commercial electric customers are those using single and three phase transmission line. Commercial customers are usually service sector businesses, although some manufacturers with low energy demands may also qualify. Commercial customers using small factors like manufacturing, super market, drink, cloth, wood and metal work. The commercial sector is slightly more complicated than domestic because of the unique mix of small business owners with mid to large size businesses. Customers also suffer from the notion that energy conservation equals reduced comfort, a serious obstacle.

Overall, because so few are exposed to the cost of energy in a larger work environment, very few take strides to conserve energy.

✓ Active and retired staff Customers

Active staff customers are user of single-phase transmission line and around 450 kilo watt power consumption. Those customers are employee of Ethiopian electric utility. Active staff customers are using more than bounder power to pay half of other customers. 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

✓ Own Consumption customers

Own consumption customers use both inter connect and self-connect system. Avoid steadily increasing electricity prices and free yourself from dependency on falling feed-in remuneration with your own system. For many years feed-in tariffs were higher than energy supplier power prices. Falling costs thanks to every more efficient system component and increasing power prices have, however, led to what is known as grid parity. Increasing power prices have overtaken decreasing feed-in paid use power, paying for itself thanks to ever more efficient system components and special state subsidies.

✓ Street Light customers

Street light is single and three phase transmission line. The purpose of street lighting is to assist drivers, pedestrians, and cyclists in finding their way in the dark. Many neighborhood groups believe that extra illumination helps prevent crime and business district lighting also may help create a pleasant environment. Access roads are required to allow access along the entire length during construction, operation and maintenance of the power transmission lines. During survey, route line was designed along the existing road in order to reduce the potential impact associated with the construction of new road.

✓ Industrial Customers

Industrial customers are the utility spaces' largest customers, and they use energy in multiple ways. They have process heating where they raise the temperature of components during the manufacturing process such as machine drive. They will heat a boiler to generate steam or hot water use electricity powers their operations. The chemical, petroleum, aluminum, glass, and steel industries are just a few examples of industries which draw a large amount of energy. There are three industrial customers those are used for low voltage, high voltage (15 kilo voltage), and high voltage (132 kilo voltage).

✓ Industrial –High Voltage (132 KV) customers

High voltage customers are using 132 kilo voltage (kv) with three phase transmission line direct access main transmission line only for inter connect system. Those customers use 132,000-volt power and all electric installation cover in that organization. Industrial consumers are quite sensitive both to reliable supply of electricity and to fluctuations in electricity prices. The extension of grid lines can impact both aspects since they can make energy transmission more efficient and reliable. High voltage is used by Largest Companies like cement factory, metal work industry.

✓ Industrial –High Voltage (15 KV) customers

High voltage customers using 15 kilo voltages is three phase transmission line for access inter connect system and self-connect system. High Voltage facilitates communications with major engineering, contractor, and manufacturing companies, and provides a forward analysis on transmission projects by helping to determine critical path or long lead-time material. It also supports the project planning and estimating process by developing a cost/benefit analysis to help customers with material specifications; material costs, and anticipated lead times.

✓ Industrial Low Voltage (LV) customers

Low voltage industrial customers using three phase transmission line the same as with high voltage 15 kV and 132 kilo voltage but low voltage customers use 380 volts. Low voltages customers use power connect from both inter connect and self-connect system. Low voltage usage customers middle factors like manufacturing, mining, agriculture, drink, cloth, wood and metal work. The low voltage electric service is using three phase and 4 electric wire, 50 HZ

frequency and a special low voltage industrial tariff, which helps eligible industrial consumers enjoy lower rates.

3.1.2 Classification of electric utility customers

Ethiopian electric utility customers can be classified in different way based on status and customers type. Those classifications can be used to design predicative model using data mining techniques.

3.1.2.1 Classification by customers status

The status of Ethiopian electric utility customer is three such as create service, connection in progress, prepared and cancelled. Create service for electric customers' status allows customers ready to use electric power and to payment roughly the using amount of electric service in kilo watt for every month, resulting in manageable bills regardless of the season and usage. Electric customers can be use electricity for lighting, heating, cooling, refrigeration, operating appliances, computers, electronics, machinery, and public transportation systems. Connection in progress electric customers' status is either material or power is cannot easy to get for organization and to plans increase distribution Service, necessary electric material like pole, capacitors, transformer and generated power covers with delivering that electricity power.

Cancelled customers status is the disconnection to get electric power service for customers request. A service disconnection typically results from either of the following situations such as a customer very high payment of electric installation, less power, transmission line small capacity to carries voltage.

3.1.2.2 Classified by customers types

The consumption data of customers has the potential to give insights of great importance for utilities and policy makers. Valuable insights can be derived by the knowledge of typical consumption curves of different customer groups and understanding what are the main customers types more use of power consumption. High value organization 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 organization. The less of

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

Ethiopian electric customers data can be classified into two such as self-connect (SCS) and inter connect (ICS) customers. SCS is mini power plant using low power consumption customer like demotic and ICS is also main power plant using high power consumption like industrial customers. Both inter-connect and self-connect customers are also with different types; those are active staff consumption, domestic, industrial, commercial, own consumption, retired staff consumption and street light. Industrial customers are using their transmission line and power consumption is different like High Voltage (132 KV), High Voltage (15KV) and low voltage (LV).

3.1.3 Describing Ethiopian electric utility (EEU) customers

Ethiopian electric utility customers can be describing customers and power. Problem understanding contains the overview of Ethiopian electric utility customers and power. 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.

3.1.3.1 Customers

Customers means any person who is supplied with electricity for his own use by EEU, and includes a person whose supply has been disconnected for the time being; by the Government or by any other person engaged in the business of supplying electricity to the public under the government or any other law for the time being in force and includes any person whose premises are for the time being connected in the purpose of receiving electricity with the works of EEU, as the case may be. Electric utility customers can be use electric power for lighting, heating, cooling, refrigeration, operating appliances, computers, electronics and machinery, someone who has buys electric power. New customers can be describing in different types of attribute in the Ethiopian electric utility. The organization is interested to understand the customer's value and classified in order to make relationship with them. The value or long run profitability of customer to the organization can be evaluated based on variables describing their customers. The customers are increase time to time then organization must be upgrade infrastructure. These

variables are customers' status, customers' types, location and other related variables that can be generated and derive from the previous attribute to see whether these variables may yield better description of customers.

Customers' status: - is to categorization of customers by creating service, progress service, cancelled. Create service customers status allows customers to create electric service. Connection progress electric customers' status is to plans distribution electric service with delivering that electricity power. Cancelled customers status is the disconnection agreement of electric power service to customers. There is no attribute to describe customer status. Customers status is described customers based on either create, progress and canceled electric service.

Customer Type: - is an individual or group of electric customers using same supply phase, voltage, purpose and service. It is useful in providing unique content to a group of customers defined as their type. In Ethiopian electric utility customers is Inter connecting (ICS) and self-connect (ICS) customer type. Inter connect customers is using main energy sources power like hydro and wind power plant system. Self-connect customers also using mini energy source like diesel and mini hydro-power system. Those customers also describe based on using electric power purpose like domestic, commercial, industrial-LV, industrial –HV (15KV), industrial-HV (132KV), Active staff, street light, retired staff and own consumption. Those customers using different electric voltage that can be affect customers, supply phase, power and bill structure.

Location (Regions): - is a geographical place where electric power interconnects with one region in to another region. Electric utilities between regions are many times interconnected for improved economy and reliability. Electrical interconnectors allow for economies of scale, allowing energy to be purchased from large and efficient sources. Utilities can draw power from generator reserves from a different region to ensure continuing, reliable power and diversify their loads. Interconnection also allows regions to have access to cheap bulk energy by receiving power from different sources. For example, one region may be producing cheap hydro-power during high water seasons, but in low water seasons, another area may be producing cheaper power through wind, allowing both regions to access cheaper energy sources from one another during different times. There are 11 regions under Ethiopian electric utility namely Addis Abeba, Gambela, Dire Dawa, Somalia, Amhara, Tigray, Afar, Oromia, Debube Ezeboch, Benishangul and Harer Regions. Those regions have different district, weather condition and office to manage customers.

District: - is an office that manages region to access customers and its power consumption. Districts and region are obliged to unplanned work load, including uncoordinated resources uses in the efforts of solving never ending customers' problem at least on a short-term basis, and ultimately causing an increase in their respective overhead costs. In each region there are different districts to manage transmission line and fix customers problems.

In some location, the season is a key determinant when its customers type use either more or less power consumption. When it's too cold, you turn the heating up, and when it is too warm, you turn the air conditioning on. Which customers can be using more power in cold location and which is not?

Power usage: -is amount of input energy in measure kilo watt required for electric appliance in function. Electricity are usually to distribute electric power for single supply phase for domestic only using 2 wire current carrier and both single and three supply phase commercial customers using for 3 or 4 wire current carrier. Industrial customers use three phase line and can receive it at highest voltages, so supplying electricity to these customers is more efficient and less expensive. High value organization 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 organization. Customers with less value are using low active electric power, high reactive power consumption and generate low revenue. Power usage one of attribute for customer but power usage also varies attributes to describe it. There is difference attribute to describe power usage of customer such as KW metering, supply phase and power factor, installation fee, connect fee.

KW Meter means an equipment used for measuring electrical quantities like energy in KWH, and or KVAH maximum demand in KW and or KVA, reactive energy in KVAR hours etc. including accessories like Current Transformers and Potential Transformers where used in conjunction with such meter and any enclosure used for housing or fixing such meter or its accessories and any devices like switches or fuses used for protection and testing purposes. Meter is interconnected with transmission line and its count amount of electric flow with customer power usage measure in kilowatt. Kilo watt is useful for measuring amounts of electricity used by large appliances for their customer type. Kilo watt-hours are what show up on your electricity bill, describing how much electricity you have used. It will include any seal or

sealing arrangement provided by EEU for avoiding unauthorized use of electricity. This shall also include prepayment meters.

Distribution Line is to carry flow of electric current from transformer in metering. There are two supply phases for Ethiopia electric utility that can be carry electric current from transformer into meter like single and three supply phases. Single phase is using 2 wire and frequency is 50 Hz using demotic and some commercial customers. Three phases also 3 or 4 wire and frequency is 60 Hz to carry large electric voltage using different customers type like industrial, some of commercial, and street light.

Installation Fee is a combination of plan electrical equipment installed from a common electrical supply to fulfill a particular purpose. Connection fee is payment for electric service after electric installation. It is including electric plan for each customer type determine by installation fee.

Customers entity table is to contain customers' description related to Ethiopian electric utility customers. In table 3.1 there are different attribute listed to describe Ethiopian electric utility customers.

No	Attribute	Description
1	Customers status	Describe which customers in monthly use electric power and either service create, paid and cancelled.
2	Customers types	Describe type of customers using their inter-connect and self-connected customers type.
3	Location	Describe location of distribution electric power for customers
4	Power usage	Describe minimum and maximum amount of distribution of electric power for customers.
5	No of customers	Total number of customers in each district for month
6	Connection fee	Frist payment for electric service
7	Supply phase	Describe current carrier of transmission line or wire
8	Metering	To count and manage the amount of electric customer power usage
9	District	Describe to manage customers using particular location.
10	Installation fee	Describe to installation payment of electric power
11	Installation material	Equipment of electric from source to customers destination
12	Date	Number of customer request to get electric service

Table 3.1. List of customer attributes

3.1.3.2 Power

Electric power is the rate, per unit time, at which electrical energy is transferred by an electric circuit. Electric power is the rate of energy consumption in an electrical circuit. Electricity is generated at power plants and moves through a complex system, sometimes called the grid, of electricity substations, transformers, and power lines that connect electricity generation of power and customers. There is different attribute to describe electric power like source (power plant), Transmission line, Load dispatch center (automate) and substation.

Source (Generation): - is production of electric power. Power plant or generating station is an industrial location where electrical power is generated in a large scale. A power plant contains one or more electric generators - machines that convert mechanical energy into electrical. Electric generators used in power plants to produce alternator current electric power. It is an essential device used to supply the electrical power at the time of power outage and offers continuity of daily activities or various business operations. There are various types of sources which are used to generate electric power. Source is one of attribute of power to generate minimum and maximum. In Ethiopian electric energy is generated from the four major types of power plants such as geothermal, diesel, wind and hydro power plant.

Transmission Line means the system consisting of all high pressure cables and overhead lines (not being an essential part of the distribution system of a EEU) transmitting electricity from a generating station to another generating station or a sub-station, together with any step-up and step-down transformers, switch-gear and other works necessary to and used for the control of such cables or overhead lines, and such buildings or part thereof as may be required to accommodate such transformers, switchgear and other works and the operating staff thereof. It shall include any substation and line including 132 KV level and above. Transmission line one attribute of power to describe production power carry from source to destination of customers using various line, frequency, supply phase and voltage level.

Substation: - is an installation at which electricity is received from one or more power stations for conversion from alternator current to direct current, reducing the voltage, or switching before distribution. Electrical substation is the part of a power system in which the voltage is transformed from high to low or low to high for transmission, distribution, and transformation

and switching. There is different type substation such as step up, step down and distribution substation. Step up substations to increase the voltage for transmission and Step-down substations to decrease the voltage and usually serve as a source to a distribution substation. Distribution substations are the ones located closest to customers, making the current safe to be used in homes and businesses. Substation is an attribute of power to describe step up and step-down transmission electric voltage. There is no attribute of substation but it determines by value like increase and decrease voltage based on their customers need.

Load Dispatch Center (LDC): - is the centralized facility for operational coordination and dispatch activities. Load dispatch center is a coordinating agency for regional electricity boards for ensuring a mechanism for safe and secure grid operation. Load dispatch center is an important link between generation and transmission, which coordinates the power requirements of customers. The Load dispatch center is the nerve center for the operation, planning, monitoring and control of the power system. Electricity cannot be stored and has to be produced when it is needed. It is therefore essential that power system is planned and operated optimally & economically. Load dispatch center is one attribute of power to describe optimal power consumption low, high, medium and fair to identify each regional.

Power is one entity of electric power industry to describe about source, infrastructure, substation and its optimization of power. In this table 3.2, there are different attribute to description of power.

No	Attribute	Description
1	Source	Describe amount of generated power from power plant
2	Transmission Line	Describe line to distribution carry electric power from source to LDC and customers
3	Substation	Describe distribution power to increase and decrease voltage
4	Load dispatch center	Describe source power is optimize based on customers location and type.

Table 3.2. List of power attributes

In general, there are two entities to describe Ethiopian electric utility organization; namely, customers and its power. The total attribute of two entities to describe in problem

understanding of Ethiopian electric utility is 16 attribute. These attributes are power and its customers then in this research use more customers related attribute because our object is customer power consumption predication.

3.2. Data Understanding

Data understanding is the second phase of hybrid data mining technique. Data understanding is Seek to better understand data assets and manage from problem understanding attribute. Data understand the data available and useful for achieving the goal specified, the secondary data collection technique called database analysis has been customers and its power consumption content and structure of the data available was understood. First, describe attribute from problem understanding then select attribute that taken from new connection customer database of Ethiopian electric utility. 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.

This study collected data from Ethiopian electric utility organization database. The data contains January 2008 to January 2011 E.C for all Ethiopian utility new customers data included. This step includes initial data collection, description of attribute data and verify data quality.

3.2.1. Initial Data Collected

The data source of this research has been collected from Ethiopian electric utility new connection customer data from data base. The organization customer database contains information about new connection customers information. However, the relevant data to carry out this research have been collected or integrated from of the database. The description of customers' entity for Ethiopian electric utility database and its understanding of problem can be describing with their attribute.

In customers entity there are different attribute to describe Ethiopian electric customers that can be record with in database and also some customers attribute is not record with database. Ethiopian electric utility new connection customers data is to store data only for related to new customers information. In Ethiopian electric utility data base, there are many attributes of new connection customers such as region, district, tariff, work request number, customer name, request date, premise, account number, status, installation fee, connection fee, and date of respond. And other attribute history customers data about amount of power consumption of

customers like active, reactive, revenue and minimum charge but other attribute is same as new connection customer tables. The following table 3.3 is list for Ethiopian electric utility customers database attributes.

No	Attribute	Data type	Description
1	Customers status	String	The customer status data
2	Date report	Date	Generated customer information
3	Region	String	Main location of energy distribution
4	District	String	The particular area electric distribution for customer
5	Tariff code	Number	The unique numbers assign for customers
6	Name of customers	String	The use of customer distribution line
7	Respond year	Date	Customer respond to create electric service
8	Active power	Numeric	The use of active power in KWH
9	Reactive power	Numeric	The use of reactive power in KWH
10	Work Request Year	Date	The customers request service date
11	Connection fee	Numeric	After install of electric power payment
12	Customer name	String	The name of customer
13	Premise	Numeric	The unique id of meter reading
14	Account number	Numeric	The customer's account number
15	System code	Numeric	Identify the customers type code
16	Minimum charge	Numeric	The minimum amount power usage
17	Distribution line	String	Power transmission line of customers

Table 3.3 Customers database attribute

The second entity of this study is power, which is described by source, transmission line, substation and load dispatch center and it come from problem understanding attribute. These attributes are described about flow of power staring from generated to distributed customers. Generated Power is carrying flow of current in different line to optimize the step up and step-down voltage then connects to customers. In this case there is specifically no power data to store in Ethiopian electric utility database like source, transmission line, and substation and load dispatch center, but power attribute substation is same as district and load dispatch center is only

one value it is not importance for this research. Source and transmission line data is not available in data base.

Most of data base attribute is same attribute with in problem understanding of customers like region, district, respond date, status, distribution line, minimum charge, installation fee and connection fee. These attribute directly taken from data base. The other attribute is derived from anther attribute. From new connection customer data like tariff code of customer attribute data base is derive customer type and Supply phase and, from customer name to derive number of customers and also from request date and respond date to derive season. Another attribute derive from history customers data derived power factor charge. Around five attributes derive from new connection and existing customer information either split or convert the given data. The other data base attribute of customer work request date, premises, account number, date of report and system code is used for assign code easy to manage system but not describe their customers. Ethiopian electric utility is only to distributed electric power for customer then this organization can be store data related to customers power usage and new connection information.

According to the above description attribute direct to select and derive attribute from database. Five attributes derive from new connection customers and historical information to description about customers. Next, the attributes within each of the customers and power are listed into a single database. Before doing this, those important attributes from each data base were chosen first. The numbers of attribute collected from problem understanding then selected and derive from data base those are region, district, supply phase, customers type, distribution line, status, number of customers, respond date, season, power factor charge, minimum charge, connection fee and installation fee. The number of records collected from Ethiopian electric utility data base is 300,573.

3.2.2. Data Description

In order to select the best attributes from this initial collected dataset, the researcher evaluates the information content of the attributes using the select attribute technique of WEKA. In data description the selected attributes are describe about data type, amount of data and coding schema. In this research most important attributes that are selected from problem understanding to retrieved from database. These attributes are status, region, district, distribution line, respond date, minimum charge, installation fee and connection fee. The other five attributes are derived

from new connection and historical customers date like customer type, supply phase, number of customers, power factor charge, and season. Other attribute like tariff code, account number, work request number, premise, personal name customer, and system type cannot be used for this research because attribute is not being described customers and power but its assign system. There are 13 attributes selected for this research. Based on the above ideas the study uses for data mining in the following attributes.

- ✓ Region is power distribution areas that have different customers and location. Ethiopian electric utility around 11 regions. This attribute data type is nominal {Addis Abeba, Gambela, Dire Dawa, Somalia, Amhara, Tigray, Afar, Oromia, Debube Ezeboch, Benishangul and Harer}
- ✓ District is distribution and manages electric power from region to customers. Ethiopian electric utility around 427 district office. The attribute type is district is nominal.
- ✓ Customers type is group of customers using same bill structure. Ethiopian electric utility 9 different customers type in Inter connect and self-connect customer. This attribute data type is nominal {active staff consumption, domestic, commercial, own consumption, , retired staff consumption, industrial High Voltage (132 KV), industrial High Voltage (15KV), industrial low voltage (LV), street light }
- ✓ Supply Phase is system of energy interconnected phase. This attribute data type is nominal {ICS and SCS}
- ✓ Status is described about customers' payment activity in monthly. This attribute data type is nominal {create service, progress and cancelled}
- ✓ Numbers of customers are group and individual user to get electric service. The data type customers are numeric only single customers' type.
- ✓ Installation fee is amount payment to installation and electrical plan of electric power for customers. Electrical Plan mean amount of electric power use for electrical equipment of customer it's described about number. This attribute data type is numeric.
- ✓ Distribution line is line that flow electric current carry from transformer into metering. This data type is nominal {Single phase and three phases}.
- ✓ Connection fee is amount of payment in after installation of electric power usage and its first payment. This attribute data type is numeric.

- ✓ Season is customer to use electric power in yearly weather conditions. This attribute data type is nominal {summer, winter, autumn and spring}.
- ✓ Respond year is after request of customer to installation electric power to get electric service or create service date. This attribute data type is date format.
- ✓ Power factor charge is the trigonometric ratio of active and reactive power consumption and difference of power factor below 0.85, and existing power factor charge tariff.

Following table 3.4 provides summery selected attribute that can be get from problem understanding attribute and derived attributes with their description.

No	Attribute	Data type	Description
1	Customers status	String	The customer status data
2	Customers types	String	The customer type data
3	Region	String	Main location of energy distribution
4	District	String	The particular area electric distribution for customer
5	Supply phase	String	Energy source system
6	Distribution line	String	The use of customer distribution line
7	Respond year	Date	Customer respond to create electric service
8	Power factor charge	Numeric	The ratio of active and reactive power consumption
9	Connection fee	Numeric	After install of electric power payment
10	Customer name	String	The name of customer
11	Customer	Numeric	The number of customers
12	Season	String	The weather condition of customer
13	Minimum charge	Numeric	Lowest charge of power consumption

Table 3.4 List of attributes of the data description

3.2.3. Data Quality

Data quality is one of the major concerns in data understanding to qualify attribute values. In data collection to select attribute from data base important for this research, then it's describe the selected attribute, amount of data, and its data type. Next to describe selected attribute then collected data contains missing, incomplete and irrelevant data in data quality. Some important information regarding the customers type, status, region, district, supply phase, customers name, number of customers, request year, respond date, connection fee and paid deposit and some

others are missing. In data quality their different attribute incompleteness, inaccurate, inconsistence and non-standard.

- ✓ Number of customer attribute contains is 193,042 duplicate records that has same type.
- ✓ Status attribute around 4,280 are non-standard, invalidity, inconsistency data quality to remove from the record.
- ✓ Installation fee attribute are 10,520 records is zero, null and negative value and remove from record.
- ✓ Connection fee attribute are 5,238 records is zero, null and negative value and remove from record.
- ✓ Dates of respond around 1,644 records are non-standard, invalidity, inconsistency data quality to remove from the record.
- ✓ Other attribute is fixed to interchangeable and correct missing spelling data.

In general, around 300,573 records are collected from Ethiopian electric utility data base for new connection customer data. From the total record around 21,682 data are removed because of their incompleteness, inconsistency and inaccuracy. Also around 193,042 record removed since there are duplicate instance. The other 85,849 records are complet and used for this research.

3.3. Data Preprocessing

After understanding of the data in hybrid data mining, preparation of data is the next step. This is one of the crucial steps to construct dataset used for modeling by Waikato Environment for Knowledge Analysis (hence forth WEKA) software. At this stage, all necessary tasks needed to prepare data mining task is finalized. Data mining techniques, tools and algorithms were decided. The data sets are pre-processed for specific data mining tasks. Data preparation phase of data mining process includes: data cleaning, data integration and data formatting [26].

3.3.1. Data Cleaning

Data cleaning is one of the activities in data preparation phase and it has to be done before going to derive 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 (for nominal and ordinal variables) and the mean (for

continuous variable). The noise which emerges from incomplete data which is lacking attributes values; this is caused by data generated problems. Similarly, there are attribute values which shows (-) in this case there is no meaning in attribute value simply removed and split for list of record. The other noise is due to the attribute value is list different from in attribute selection table data type list. This is the case removed that record for list of worksheets in MS-Excel. MS-Excel is also used for data preparation. Some attribute is non-standard, incomplete, inaccurate and inconsistent.

Missing value: - In the collected data for this research work there were some missing values like status, respond year, installation fee and connection fee. So these record is removed. The number customer attribute record is duplicate then removed from original data.

The Status, number of customers, installation fee, connect fee, and respond date attributes (see table 3.5) are noise, incompleteness and needs cleaning.

Attribute	Need to be cleaned	Number of records
Status	Status value is empty and zero	4,280
Respond Year	Request year value is year data	1,644
Number of customers	The number of customer value duplicate.	193,042
Installation fee	The value attribute is zero, negative and empty	10,520
Connection fee	The value attribute is zero, negative and empty	5,238

Table 3.5 list of attributes for data cleaning

3.3.2. Derived Attribute

In some of data needs derived to smoothing for mining. Reduce data size by dividing the range of a continuous attribute into intervals against to the standard. Interval labels can be used to replace actual data values.

From new connection customer data tariff code of customer attribute is derived from customer type and Supply phase. Also from customer name to derived number of customers and from respond date to season are derived. Another attribute derived from history of customers is power factor charge. In general around five attributes are derived from new connection and existing customer information as shown in table 3.6.

No	Derived attribute	How the attribute derived
1	Customer type & Supply phase	From new connection customer attribute data in tariff code is split in to two customer type and supply phase. Example ICS-Domestic into customer type domestic and supply phase ICS
2	Number of customers	From new connection customer attribute data in customer name change into number
3	Season	From new connection customer attribute data in Respond date is converted into season Example June, July and august is summer season.
4	Power factor charge	From historical data attribute in active divided by reactive power.

Table 3.6 List derived attribute

3.3.3. Data Discretization

It divides numerical data in to categorical classes that are easier to understand the data. Discretization and concept hierarchy use raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of Abstraction, and are a powerful tool for data mining. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results Concept hierarchies for numerical attributes can be constructed automatically based on data discretization it has some method to use. In this research use bin method, Binning is a top-down splitting technique based on a specified number of bins. Attribute values can be discretized by applying equal-width, or equal-frequency binning, and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by bin medians, respectively [2].

In our case minimum charge attribute have numerical value is converted into categorical class that are easily understand where discretization using the equal-width method. It is method

divides the data into a fixed number of interval of Equal length so all data were categorized in to five interval which labeled (High, Very High, Low, Very Low and Medium). The other is power factor charge attribute have numerical value use to in categorical class that are easily understand where discretization using. These attributes have numerical value use to in categorical two class that are easily understand where discretization is power factor charge greater than 0.85 is high power factor and below 0.85 is low power factor charge.

No	Attribute	Previous value	New value
1	Minimum charge	Numeric	H, VH, M, L, VL
2	Power factor charge	Numeric	High and Low

Table 3.7 List discretization attribute

3.3.4. Dataset Format Conversion

This task will do after data cleaning. After passing the above preprocessing stage the researcher has got 85,849 record data set. The datasets provided to this software were prepared in a format that is acceptable for Weka software. Then the preprocessed dataset in excel is converted to Comma Separated Values (.csv) and Attribute Relation File Format (.raff) to make it compatible with WEKA software.

Finally, for data analysis 14 attributes and 85,849 instances are available. In table 3.8 provide summery of attribute used in this study.

No	Attribute	Data type	Value
1	Customers status	Nominal	List of status of customers
2	Customers types	Nominal	List of customers types
3	Region	Nominal	List of region
4	District	Nominal	List of customer district
5	Supply phase	Nominal	Supply phase (ICS and SCS)
6	Distribution line	Nominal	Single phase and three phase
7	Respond year	Nominal	2015, 2016, 2017,2019
8	Power factor charge	Nominal	High and Low
9	Connection fee	Numeric	Numerical data
10	Installation fee	Numeric	Numerical data
11	Customer	Numeric	Numerical data
12	Season	Nominal	Sumer, winter, spring and autumn
13	Minimum charge	Nominal	VH, H, M, L and VL
14	Power consumption	Nominal	High cons. And low cons.

Table 3.8 Dataset format conversion

```
@relation 'Mohammed _predicted'

@attribute Region {'ADDIS ABABA',OROMIA,TIGRAY,AMHARA,'DEBUBE
EZEBOCH',SOMALI,AFAR,'BENISHANGUL GUMUZ',HARERI,DIREDAWA,GAMBELA}
@attribute District {'CSC_2_YEKA  MICHAEL WUBET BLG','CSC -
KACHISE DISTRICT','CSC - ABIADI DISTRICT','CSC_1_ RAS DESTA
'CSC - SHINISHICHO','CSC - ADWA DISTRICT','CSC - ALAMATA
DISTRICT','CSC - AXUM DISTRICT','CSC - MAICHW
DISTRICT','DISTRICT',CSC_DIKSIS,CSC_MASHA_DISTRICT,CSC_BURJI,CSC_
ROBIT_DISTRICT,CSC_BISTIMA_DISTRICT,CSC_JAMMA_DISTRICT,CSC_MILLE_
CENTER,CSC_ALBKO_DISTRICT,CSC_YEJUBE,'CSC_TEPI
DISTRICT',CSC_COMBOLECHA_SC2,CSC_LUMAME,CSC_GISHEABAY}
@attribute 'Customer Type' {'OWN CONSUMPTION','INDUSTRIAL -
LV',DOMESTIC,'STREET LIGHT','STAFF CONSUMPTION','INDUSTRIAL - HV
(15 kV)','INDUSTRIAL - HV (132 kV)',COMMERCIAL,'RETIRED STAFF
CONSUMPT'}
@attribute 'Supply Phase' {'SCS ','ICS '}
@attribute 'Number of customer' numeric
@attribute 'Distribution Line' {'Three phase','Single phase
','Single phase'}
@attribute 'Respond Year' numeric
@attribute Season {'Autumn ',Winter,Spring,Summer,'Spring '}
@attribute Status {'SERVICE  POINT CREATED',PAID,'CONNECTION IN
PROGRESS','QUOTATION PREPARED'}
@attribute 'Installation fee' numeric
@attribute 'Power factor charge' {'High ',Low}
@attribute 'Minimum Charge' {H,VH,M,VL,L}
@attribute 'Connection Fee' numeric
@attribute 'prediction margin' numeric
@attribute 'predicted Power cons.' {'Low cons','High cons.'}
@attribute 'Power cons.' {'Low cons','High cons.'}

@data
TIGRAY,'CSC - ABIADI DISTRICT',DOMESTIC,'ICS ',2,'Single phase
',2016,'Autumn ','SERVICE  POINT CREATED',0,'High
',M,3324,1,'High cons.','High cons.'
```

Figure 3.1 The ARFF format of the Final Dataset

CHAPTER FOUR

EXPERIMENTATION AND RESULT ANALYSIS

In this chapter, the researcher describes the techniques that have been used in developing a predictive model to classify Ethiopian electric utility customers. In addition to incorporated typical stages that characterize a data mining process. This study has been organized according to hybrid data mining process model, which is described and discussed section methodology in chapter one. To this end, using classification algorithms such as Bagging, J48, Random tree and PART. These algorithms are selected based on classification DM problem; the modeling techniques are selected from the classification modeling techniques and specific requirements of each model have been observed and ascertained that all the selected techniques can adequately support nominal classes and numeric attribute. The classifiers were extract the dataset required for training and testing the models created by the classifiers. For creating predictive model, a total size of 85,849 records was used for training and testing. The validations were done using 10, 15 and 20-fold cross validation, and 66, 72 and 78% split test option.

4.1 Selection of Modeling Technique

Modeling is one of the major tasks which are undertaken the phase of data mining in hybrid data mining methodology. Model selecting one of hybrid data mining technique to select an appropriate model depends on data mining goals. Consequently, to attain the objectives of these research WEKA data mining tool is selected for modeling and classification of Ethiopian electric utility.

For experimentation of this study also selected four classification algorithms can be used for WEKA tool such as Bagging, J48, random tree and PART decision tree. In this study the above algorithms are selected because of easy of understanding, very simple, easy to compatible with weka tool and interpretation of the result of the model. A PART algorithm is common types of rule induction technique which generate a model as a set of rules. In other case the J48 algorithms of decision tree generate a model by constructing decision tree where each internal node is a feature or attribute. Bagging algorithms is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees. Random Tree is a

supervised Classifier; it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In standard tree every node is split using the best split among all variables [46]

4.2 Experimental Design

In this section, it is discussed about how the samples are prepared for modeling, how the predicting accuracy of each model is evaluated and the major tasks to be conducted at each experiment for the selected algorithms. It is necessary to define a procedure to test the model's quality and validity, which means the experimentation plan for training, testing and evaluating is required [47]. Thus, in this experiment 14 attributes were selected during the data preparation phase to train and test the selected classifiers. A total of 85,849 records, extracted from the original dataset. In order to check whether the predicting to classify power consumption the models are affected by the percentages and fold cross validation of partitioning the dataset into training and test sets. Training set is subset to train model and test set is subset to test the training model. The default split training set is 66% then add 6 % for default train set and also default fold cross validation is 10 then to add five with default fold cross validation because of equal value interval for cross validation and split test.

The dataset is systematically partitioned in to three pairs of disjoint sets of training sets and test sets and two-fold cross validation (for validating the reliability of the models):

- 66% Training set and 34% Test set,
- 72% Training set and 28% Test set,
- 78% Training set and 22% test set,
- 10-fold cross validation,
- 15-fold cross validation and
- 20-fold cross validation

Besides, other standard measure including precision, recall, sensitivity and specificity are available. Therefore, the test design specifies that the dataset should be separated into training and test set, and builds the model on the training set and estimate its quality on the separate test set. Process of building predictive models requires a well-defined training and validation protocol in order to insure that most accurate and robust prediction [48]. As above suggestion as in this research use the data set as training and testing. In WEKA Environment has used to set up

and measure the quality, validity and test of the selected model. For purpose of this study K-fold (10, 15 and 20-folds) cross validation percentage than 34-66, 28-72 and 22-78 split test options are used because of its relatively low bias and variations.

4.3 Model Building using WEKA Software

Model building is one of the major tasks that are undertaken the data mining phase in the hybrid methodology of conducting data mining researches. WEKA 3-8-1 supports many types of classification algorithms. Among the classification algorithms that WEKA 3-8-1 supports the J48, PART, random tree and bagging algorithm was used with different input parameters as well as different types of related classifiers. The corresponding algorithm used to extract rules from the decision trees is J48, Bagging, random tree and PART. A J48 decision tree algorithm is the chosen because it is one of the most common decision tree algorithms that are used today to implement classification techniques WEKA. PART rule induction has almost a similar set of parameters with J48 decision tree algorithm that can be adjusted to build better model from datasets.

This research is use WEKA 3-8-1 for conducting a total of 24 experiments, where the first six of the experiments were constructing for J48 algorithm with 10, 15 and 20-fold cross validation and 66, 72 and 78% split test. Secondly, where six of the experiment were constructing for PART algorithm with 10, 15 and 20-fold cross validation, and 66, 72 and 78% split test. Thirdly other six of the experiment were constructing for bagging algorithm with 10, 15 and 20-fold cross validation, and 66, 72 and 78% split test, and the last remaining six of the experiment were constructing for Random tree algorithm with 10, 15 and 20-fold cross validation, and 66, 72 and 78% split test.

There are different experiments carried out by using all the 14 attributes of the records and four data mining algorithms. Analysis of the J48 decision tree predictive model to classify customer power consumption were made in terms of detailed accuracy, precision, recall, F-measure and ROC curve of the classifier based on a confusion matrix of each predictive model resulted of different classes (Low, and High classes in this research).

4.3.1 Modeling procedure

The 24 experiments are conducted for building several models by applying each of the four selected modeling techniques. The rationale for selecting 24 experiments is to address the model building for the default value, for different number of folds of cross validation, and parameter setting procedures.

The experiments are analyzed to compare them in terms of different performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC using four algorithms were used to evaluate and compare the performance of the models.

In classification process outcome is predicted from a given input or future attributes relation to class attributes. For this purpose, the algorithm processes a training set which consists of an attribute, check the algorism feat with given data set and the outcome is called prediction attribute. The input is then analyzed to produce a prediction and how good an algorithm is depends upon the predictions made by the algorithm. Prediction rules are used for knowledge expression in the form of IF-THEN rules, IF part is known as antecedent which consists of a conjunction of conditions and the THEN part is known as consequent which gives the prediction whether satisfies the antecedent or not [31].

4.3.1.1 Model Building using J48 Decision Tree

J48 decision tree is one model building algorithm in data mining technique. The following figure 4.1 show all attribute which are used for the experiment.

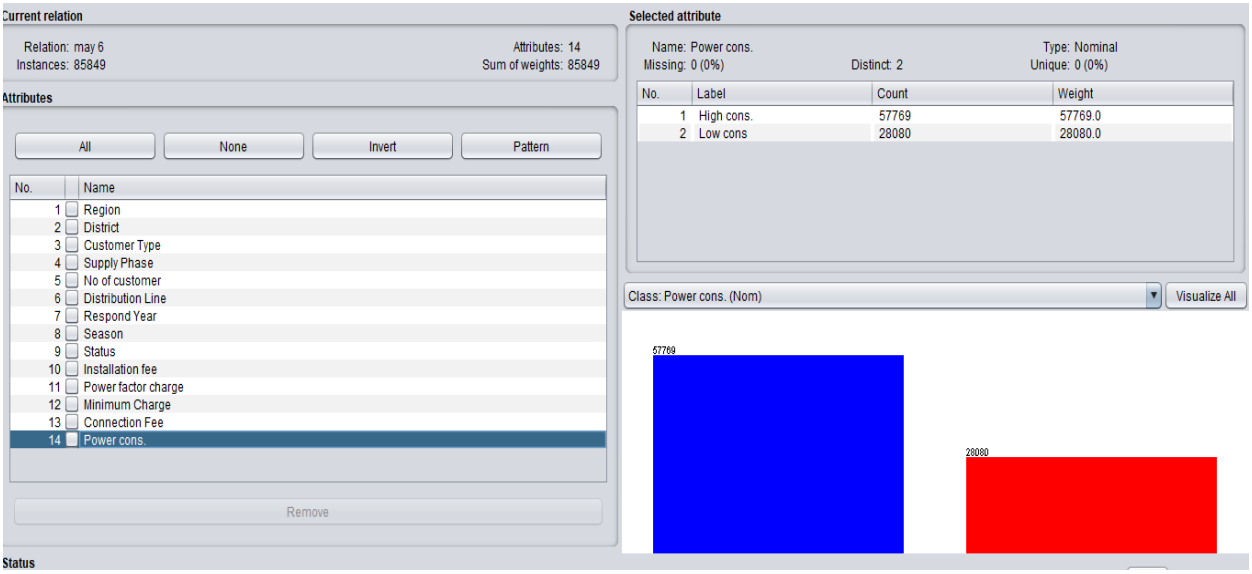

Figure 4.1 Screenshot of all attribute of experiment.

Decision trees are very frequently used for classification, as they are easy to use and understand. The classification accuracy is the proportion of correctly classified instances in a given set of instances. If a decision tree would have been based on all available instances training and testing. Therefore, a decision tree is built on a subset of all available data, called the learning set. Once the modeling techniques, tool and the evaluation criteria were established, then building models with a number of parameters setup which governs the kind of models generated by the process was followed .Since, WEKA Explorer is sensible to its default values, initially, J48 algorithm with its default parameters values has been run on the targeted input dataset and the experiments continued by adjusting those default parameters values in the object editor of WEKA. In this study experiments were conducted using J48 algorithm with K=10, K=15 and K=20-fold cross validation with and 66%, 72% and 78% in split test. The following table 4.1 shows the summary of experimental results of J48 Decision Tree.

Evaluation Criteria	**J48 experimental result**					
	Using 66% test set	Using 72% test set	Using 78 % test set	10-fold cross validation	15-fold cross validation	20-fold cross validation
Number of Leaves	2157	2157	2157	2157	2157	2157
Number of trees	2526	2526	2526	2526	2526	2526
Time taken	2.20	2.47	2.21	2.73	2.61	2.10
Accuracy (%)	96.20	96.24	96.30	96.49	96.58	96.61
Av.TP Rate	0.962	0.962	0.963	0.966	0.966	0.966
Av. FP Rate	0.058	0.057	0.057	0.053	0.052	0.051
Av. Precision	0.962	0.962	0.963	0.966	0.966	0.966
Av. Recall	0.962	0.962	0.963	0.966	0.966	0.966
Av. F-Measure	0.962	0.962	0.963	0.966	0.966	0.996
Av. MCC	0.913	0.914	0.915	0.921	0.922	0.923
Av. ROC Area	0.983	0.984	0.984	0.987	0.987	0.987
Av. PRC Area	0.979	0.980	0.980	0.984	0.984	0.984

Table 4.1 Summary of experimental result of J48 algorithms

KEY: Accuracy: Registered performance of model Average, (TPR) True Positive Rate. (FPR) False Positives Rate, (PR) precision rate, (RR) Recall rate, (MCC) Matthews's correlation coefficient, (ROC) Area: Relative Optical character curve, (PRC) Precision recall.

As it can be observed from the above table 4.1, different experimental results were obtained by applying J48 algorithm with its different test models. Experimental result using 20-fold cross validation is with better accuracy than the other experimental result and its accuracy level was 96.61%.

In this model the number of correctly classified instances is 82,939 (96.61%) and the number of incorrectly classified instances is 2,910 (3.38%). Considering High and Low power consumption class better result, the FT Rate, Time Taken and MCC of this model are 0.051, 2.10, and 0.923 respectively.

4.3.1.2 Model Building using PART Algorithms

This experiment was performed on the PART algorithm it is an alternative representative of decision tree follows the same procedure applied on the previous experiment which are presented above. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. In this experiment were constructing for PART algorithm with K=10, K=15 and K=20-fold cross validation and 66%, 72% and 78% split test. The following table 4.3 present summary of experimental algorithm of PART algorithms.

Evaluation Criteria	**PART experimental result**					
	Using 66% test set	Using 72% test set	Using 78 % test set	10-fold cross validation	15-fold cross validation	20-fold cross validation
Number of Rule	1902	1902	1902	1902	1902	1902
Time taken	0.37	0.31	0.25			
Accuracy (%)	93.6.	94.54	94.16	94.91	94.93	94.94
Av.TP Rate	0.936	0.946	0.942	0.949	0.949	0.949
Av. FP Rate	0.088	0.071	0.076	0.068	0.068	0.068
Av. Precision	0.936	0.946	0.942	0.949	0.949	0.949
Av. Recall	0.936	0.946	0.942	0.949	0.949	0.949

Av. F-Measure	0.936	0.946	0.942	0.949	0.949	0.949
Av. MCC	0.854	0.877	0.867	0.885	0.885	0.885
Av. ROC Area	0.967	0.972	0.969	0.975	0.975	0.976
Av. PRC Area	0.962	0.964	0.961	0.969	0.969	0.970

Table 4.2 Summary of experimental result of PART algorithm

When we compare in table 4.3 summary the performance of the models produced by PART algorithm applying on the targeted input datasets, the generated results obtained from six experimental. PART rule induction model experimental result using 20-cross fold validation is with better accuracy than the other experimental result and its accuracy level was 94.94%.

In this model the number of correctly classified instances are 81,508 (94.94%) and the number of incorrectly classified instances is 4,341 (5, 05%).

4.3.1.3 Model Building using Bagging Algorithms

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging Algorithms can be constructed in such a way that by combining the properties of two or more algorithms combined the decisions of different models means amalgamating the various outputs into a single prediction. The simplest way to do this in the case of classification is to take a vote (perhaps a weighted vote); in the case of numeric prediction it is to calculate the average (perhaps a weighted average). Bagging and boosting both adopt this approach, but they derive the individual models in different ways [58].

In this algorithm with the same procedure, six models are built for different training set samples applying. In this experiment were constructing for Bagging algorithm with K=10, K=15 and K=20-fold cross validation and 66%, 72% and 78% split test. The following table 4.3peresent summary result of bagging algorithm.

Evaluation	**Bagging experimental result**					
	Using 66%	Using 72%	Using 78 %	10-fold cross	15-fold cross	20-fold cross

Criteria	test set	test set	test set	validation	validation	validation
Time taken	0.13	0.08	0.08			
Accuracy (%)	85.72	86.32	87.02	87.75	87.86	87.93
Av.TP Rate	0.857	0.863	0.870	0.879	0.879	0.879
Av. FP Rate	0.203	0.194	0.186	0.170	0.171	0.171
Av. Precision	0.855	0.862	0.869	0.876	0.877	0.878
Av. Recall	0.857	0.863	0.870	0.879	0.879	0.879
Av. F-Measure	0.856	0.862	0.869	0.688	0.778	0.878
Av. MCC	0.670	0.684	0.699	0.721	0.721	0.722
Av. ROC Area	0.917	0.924	0.929	0.936	0.936	0.936
Av. PRC Area	0.917	0.923	0.929	0.934	0.935	0.936

Table 4.3 Summary of experimental result of bagging algorithms

From the above six experiments better result is achieved when 20 fold cross validation with is better accuracy of 87.93 %. In this model the number of correctly classified instances is 75,493(87.93%) and the number of incorrectly classified instances is 10,356 (12.06%). Considering the Average Precision, MCC and PRC area of this model are 0,878, 0.722 and 0.936 respectively.

Among the parameters in Bagging, only the classifier parameter (which assigns an algorithm as a base classifier) results in a better prediction performance and accuracy experiment result. This better result confusion matrix is using Bagging confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records.

4.3.1.4 Model Building using Random Tree

This experiment was performed following the same procedure that are applied for previous experiment. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. In this experiment models are constructed for random tree decision algorithm with K=10, K=15 and K=20-fold cross validation and 66%, 72% and 78% split test. The following table 4.4 is summary of experimented result for random tree algorithm.

	Random Tree experimental result

Evaluation Criteria	Using 66% test set	Using 72% test set	Using 78 % test set	10-fold cross validation	15-fold cross validation	20-fold cross validation
Number of tree Size	133,579	133,579	133,579	133,579	133,579	133,579
Time taken build model	0.33	0.28	0.27	0.40	0.53	0.69
Accuracy (%)	89.69	90.35	91.35	91.70	91.54	91.32
Av.TP Rate	0.897	0.903	0.913	0.917	0.915	0.913
Av. FP Rate	0.132	0.126	0.116	0.110	0.113	0.116
Av. Precision	0.897	0.903	0.913	0.917	0.915	0.913
Av. Recall	0.897	0.903	0.913	0.917	0.915	0.913
Av. F-Measure	0.897	0.903	0.913	0.917	0.915	0.913
Av. MCC	0.765	0.780	0.801	0.811	0.807	0.802
Av. ROC Area	0.893	0.901	0.913	0.916	0.913	0.910
Av. PRC Area	0.870	0.879	0.894	0.896	0.893	0.890

Table 4.4 Summary of experimental result of Random tree algorithms

As it can be observed from table 4.4, there are different experimental results were obtained by applying Random tree algorithms. The summarized results shown above in Table 4.4 illustrate the accuracy level of the generated models in percentage. From the result obtained for the six experiments by applying random decision tree algorithm. Achieved when the training sets is using 10-fold cross validation is with a better accuracy of 91.70 %. In this model the number of correctly classified instances is 78,587(91.70%) and the number of incorrectly classified instances is 7,262 (8.45%). The average TP rate, FP rate, Precision, Recall, F-Measure and MCC of this model are 0,917, 0.110, 0.917, 0.917, 0.917 and 0.811 respectively.

4.4. Evaluation Models

Models evaluation is one of the specific objectives of this study to select the best alternative classification technique for building a model which can be classification of customers based on power consumption to design predictive models. In above discussion, the overall performance of the four selected classification models were compared with each other and evaluated based on their respective results using different evaluation techniques as presented in table 4.5 follows.

Algorithms	Accuracy (%)	TPR	Precision	Recall	F-Measure	MCC	ROC	PRC
J48	96.61	0.966	0.966	0.966	0.966	0.923	0.987	0.984
Bagging	87.93	0.879	0.878	0.879	0.879	0.722	0.936	0.936
PART	94.94	0.949	0.949	0.949	0.949	0.855	0.976	0.970
Random Tree	91.70	0.917	0.917	0.917	0.917	0.811	0.916	0.896

Table 4.5 Comparison of the four algorithms

The test results shown that, decision tree using J48 algorithm performs better than others with 96.61% accuracy.

Before conducting the modeling process, it has been stated in different sections that there are some factors which can affect the performance of the models. These are information about customers, which might serve as useful not include component for classification; the data has some missing values and other quality problems.

From above table 4.5, it can be observed that the weighted average values of TPR, Precision, Recall, F-Measure, ROC and PRC values of the selected models are nearly the same. Based on this narrow different average weight in four algorithms, it is possible to evaluate the performance of the four selected models. Hence, among the four-model generated by using J48 algorithm showed best performance with the other three which is better TPR, Precision, Recall, F- Measure MCC, ROC and PRC.

4.4.1. Confusion Matrix Better Model

Since confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. Here under the confusion matrix of the best selected algorithm is presented.

This algorithm confusion matrix is described in the following way.

```
=== Confusion Matrix ===
   a     b   <-- classified as
56748  1021 |    a = High cons.
 1889 26191 |    b = Low cons
```

The entries in the confusion matrix have the following meaning:

- ✓ 56748 is the number of **correct predictions** of instances to **High** Consumption,

- ✓ 1021 is the number of incorrect predictions of instances to **Low** consumption,
- ✓ 1889 is the number of incorrect predictions of instances to **High** consumption,
- ✓ 26191 is the number of **correct predictions** of instances to **Low** consumption

As shown in the confusion matrix high consumption power data has a better collect prediction the low consumption class.

4.5. Rules generated by J48 decision tree

Based on experiment J48 decision tree with all attribute un-pruned are high accuracy finally generated rule learner with the specified scheme. From these generated rules which are highly predictive are selected 23 rules consider the mandatory attribute by investigator based on the finding and relevant to the domain knowledge or area of specialized people.

The following are some of the rules extracted by J48 decision tree. The listed below some of the rules supposed to be interesting and are selected by domain experts as well as from the literatures. Out of a total rule generated, some rules with greater number of instances classified are selected. The following rule are generated by J48 decision tree algorithm that are concerned with connection fee as best attribute to determine power consumption of customers as high and low.

Rules related to Connection Fee

Rule 1:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Connection Fee <= 3115 AND Connection Fee <= 1566 AND Number of Customer <= 2 AND Connection Fee <= 1300 AND Connection Fee >= 864 AL AND Installation Fee >= -2650 AND Distribution Line = Three phase: High cons. (1412.0/123.0)

Rule 2:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = DOMESTIC: Low cons (2297.0/3.0)

Rule 3:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = INDUSTRIAL - LV: High cons. (2.0)

Rule 4:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = DOMESTIC AND Number of Customer <= 2: High cons. (3714.0/3.0)

Rule 5:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = DOMESTIC: Low cons. (715.0/162.0)

Rule 6:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = DOMESTIC AND Supply Phase = SCS AND Number of Customer <= 5: High cons. (69.0)

Rule 7:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = COMMERICAL: Low cons. (354.0/2.0)

Rule 8:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Respond Year<= 2016: Low cons. (359.0/120.0)

Rule 9:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Respond Year> 2016 AND District = CSC-ABIADI DISTRICT: High cons. (16.0/2.0)

Rule 10:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Respond Year> 2016 AND District = CSC-GUNCHER AND Minimum Charge= M: Low cons. (5.0/1.0)

Rule 11:

IF Connection Fee > 5925 AND Customer Type = INDUSTRIAL -LV: High cons. (3192.0)

Rule 12:

IF Connection Fee > 5925 AND Customer Type = INDUSTRIAL –LV :Low cons. (1460.0/57.0)

Rule 13:

IF Connection Fee > 5925 AND Customer Type = DOMESTIC AND Installation Fee <=4750: High cons. (4454.0/2.0)

Rule 14:

IF Connection Fee > 5925 AND Customer Type = COMMERICAL Number of Customers<=5: High Cons. (410.0)

Rule 15:

IF Connection Fee > 5925 AND Customer Type = COMMERICAL AND Number of Customers>5 AND Installation Fee> 5124: Low Cons. (127.0/5.0)

Rule 16:

IF Connection Fee > 5925 AND Customer Type = STREET LIGHT: High cons. (60.0)

Rule 17:

IF Connection Fee > 5925 AND Customer Type = INDUSTRIAL-HV (15kv): High cons. (35.0)

Rule 18:

IF Connection Fee > 5925 AND Customer Type = INDUSTRIAL-HV (132kv): High cons. (45.0)

Rule 19:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Number of Customer <= 2 AND Distribution Line = Single phase AND Season= Summer: High cons. (5.0/2.0)

Rule 20:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Number of Customer <= 2 AND District = CSC-CHANCHO AND Power factor charge= High AND respond Year> 2015: High cons. (4.0/1.0)

Rule 21:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 Customer Type = DOMESTIC AND Supply Phase=ICS: High cons. (4474.0/18.0)

Rule 22:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = COMMERCIAL: Low cons. (354.0/2.0)

Rule 23:

IF Connection Fee <= 5925 AND Installation Fee <= 1000 AND Customer Type = COMMERCIAL: High cons. (332.0/1.0)

From the above rules that most important variables for building model to classify customer power consumption were Connection fee, installation fee, customer type, district, distribution line, supply phase, season and respond year. Therefore, these attributes play a significant role in classifying records at the higher level of the tree which indicates their statistical significance than other variables occupation. The other attribute is less effect on classification of customer power consumption like number of customers, region, minimum charge, power factor charge and status.

4.6. Use of knowledge and Prototyping

In this research the discovered knowledge is used by integrating the user interface which is designed by C# programming with a Weka system in order to show the determining power consumption of Ethiopian electric utility customer.

After evaluating discover of knowledge, the last step is using this knowledge for the industrial purposes. In this step knowledge discover is incorporated in to performances system and take this action based on the discovered knowledge.

In order to predict of the feature Ethiopian electric utility customer, we analyzed the current Ethiopian electric utility customer to determining the prediction model based on the available data by generate rule selection algorithm. Then we use the generated ruled by the implementing it with a visual basic programming language to predict the feature electric utility customer using data mining approaches.

The prototype is a channel for communication between system and end-user. In order to design the Ethiopian electric utility customer classification for predication of power consumption to be an interactive tool, decision was made by referring the set of rules or command in a simple manner. This prototype is decision support system for organization and customers. In this system that utilize rule-based expert systems, the inference engine must be supplied with the facts and the rules associated with them that are often expressed in sets of "if–then" rules. In this sense, the decision support system requires an extracted knowledge on the part of the decision maker in order to provide the right answers to well-formed questions. On the contrary, the decision support systems customer data mining tools on the part of the decision maker. Instead of the system is designed to find new and unsuspected patterns and relationships in a given set of data. This is assisting electric technical, customers, reduce time of decision making simple and easy to decision and integrated to other service delivery system and the management of electric distribute planning.

The user inserts the new connection customers data screening from system with set of rules, such rules are mandatory attributes and optional. The mandatory attribute is (connection fee, installation fee, customer type, district, respond year) and another attribute category as optional. The user selected and inserts values of attribute from interface screen display all attributes with combo box and text box, then from combo box select value of attribute and from text box insert value of attribute is take integer and three decision option such as prediction, reset and close

button. The following figure 4.2 shows user interface of determining customer power consumption prediction models.

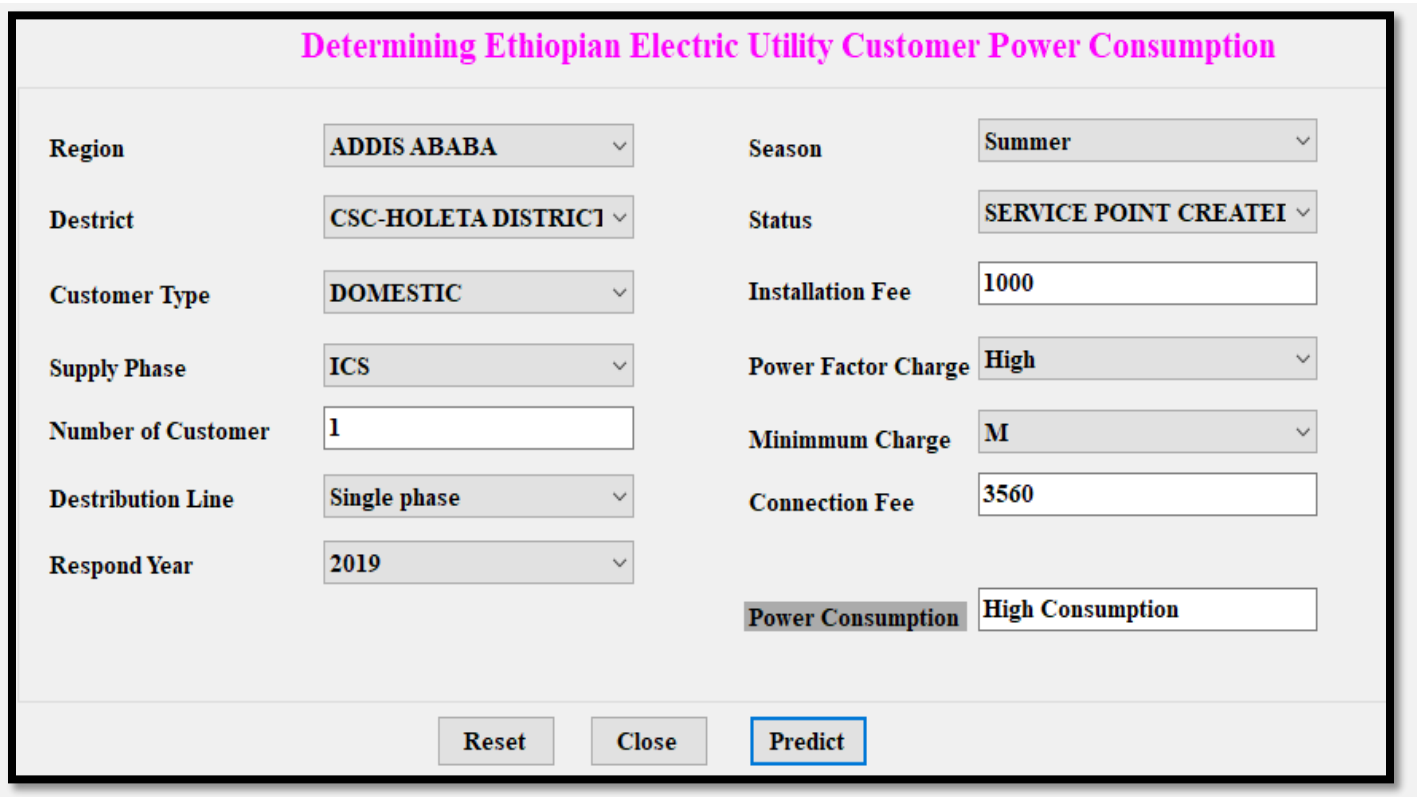

Figure 4.2 prototype of the research

4.6.1. User acceptance testing

After the prototype was developed, the next step was testing and evaluating the system whether the system satisfies the users need and asses‟ performance of the system. The scope of testing and evaluation that to accomplish the significance involved to depend on the complexity, and other core features of the system. As the aim of testing and evaluation of the system is to assure that the system expected what it is required to do.

User Acceptance Testing is used to conduct operational readiness of a product and service. It is a common type of non-functional software testing, used mainly in software development and software maintenance projects. This type of testing focuses on the operational readiness of the system to be supported. It is done when the completed system is handed over from the developers to the customers or users. The purpose of user acceptance testing is rather to give confidence that the system is working than to find errors.

User Acceptance Testing verifies the system's behavior is consistent with the requirements. These tests will reveal defects within the system. The work associated with it begins after requirements are written and continues through the final stage of testing before the user accepts the new system.

This study revealed that from 10 domain experts 9 of them confirm that this determining electric customer classification for prediction model was much efficient, and it saves their energy and materials while comparing with the way they perform currently which is the simple statistical method. But one domain experts do not satisfy with the prediction system especially in error tolerance criteria.

In the case of effectiveness, the domain experts revealed that this prediction model produces a desired result. In order to make the prediction near perfect, there is a need to enhance the performance of the model near 90%. Some (10%) of the domain experts were disagree with the effectiveness result because they stated that in case of determining Ethiopian electric utility customers analysis the reason must have to be perfect because it has a significant impact on the organizations revenue.

Efficiency is the ability to avoid wasting materials, energy, efforts, money, and time in doing something or in producing a desired result. In a more general sense, it is the ability to do things well, successfully, and without waste times. In scientific terms, it is a measure of the extent to which input is well used for an intended task or function (output). It often specifically comprises the capability of a specific application of effort to produce a specific outcome with a minimum amount or quantity of waste, expense, or unnecessary effort.

During our presentation and discussion with domain experts, we conduct sample experiments and compare the efficiency between the current statistical method and our new prediction method.

Effectiveness is the capability of producing a desired result or the ability to produce desired output. When something is deemed effective, it means that it has an intended or expected outcome or produces a deep impression. In this research effectiveness is considered as the accuracy classification of prediction model to electric customer. As discussed in chapter four, we perform experiments with J48 tree algorithm, Random tree, PART from rule induction algorithm and bagging. As a result, J48 algorithm registers better performance of 96.61% accuracy.

During our presentation and discussion with domain experts, our experimental results are discussed, and they give a recommendation to improve this performance.

Error Tolerance is concerned about management of faults originating from defects in design or implementation. In this research error tolerance is considered as making the experiments error free or making it intelligence. As discussed before this experiment registers a better performance of 96.61% accuracy with J48 decision tree algorithm. During our presentation and discussion with domain experts, we discuss on how to make this experiment intelligence by integrating this experiment with knowledge-based systems and we agreed that some improvements also need.

Questionnaires	Strongly disagree	Disagree	Undecided	Agree	Strongly Agree	Average
I think that I would like to use this system frequency			10%	5%	85%	90%
I think the system is easy to use			10%	10%	80%	90%
I think the system is fast to response			5%		95%	95%
I thought that the system is saves energy & materials.					100%	100%
I thought the system does have consistency			30%	10%	60%	70%
Total			11%	5%	84%	89%

Table 4.6 Domain expert respond

Finally, the average usability of the Ethiopian electric utility customer prediction model prototype system according to the evaluation results filled by the participants majority of domain expert 89% agreed that the system prototype has a good and clear informational and functional explanation regarding the objective of research.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

CONCLUSION

The application of data mining technology has increasingly become very popular and proved to be relevant for different market and industries. Particularly in the energy industry, data mining technology has been applied for determining prediction of power consumption. Customer type are main challenge for electric power either high or low power consumption.

In this study, an attempt has been made to determining predict model of classification customers in Ethiopian electric utility, which are having the different customer type using different voltage, supply phase and Line, and those customers type existing data stored in database. The hybrid data mining methodology, which is believed to be the classification of customer type based on power consumption, and to develop prediction model conduct this study. The major steps followed are problem understanding, data understanding, data preparation, modeling, evaluation of knowledge discovering and use knowledge discover.

Before applying the data mining model the data can be collected from Ethiopian electric utility customers data base of data sources and the data required for this study is extracted to different format. A data set with 85,849 total customer data record was used to develop a classification model. Since, this research was intended to fill a gap left by a related research.

As the major goal of the study is prediction models customers power consumption, classification data mining techniques are selected for modeling. The modeling techniques used are: J48, random tree, bagging and PART. The modeling process (experimentation) is conducted in three steps for each modeling technique: first by 10, 15 and 20 fold cross validation and 66%, 72% and 78% data split test mode then around 24 experimental result, the second step is finding better result by changing the number of folds cross validation and split test mode, and third is improving the determining prediction performance of the best models of each algorithm is accuracy. Then from those experimental results, the model which has been built applying J48 algorithm, which shows the highest predicting performance of 96.61%, is selected as the best model satisfying the entire power consumption followed by J48 with other predication modeling techniques respectively.

Generally, the experimentation results obtained from this research work show applicability of data mining technology in developing a predictive model for determining power consumption of Ethiopian electric utility customers.

RECOMMENDATION

This research work has been conducted mainly for academic achievement. However, the researcher strongly believes that the findings of the study can be used by the concerned organizations to further investigate their data quality problems.to choose appropriate data analysis methods, techniques and tools that are currently in use for processing customer power consumption. It is observed in the research work that the J48 modeling technique shows good performance in predicting the power consumption either high and low. The technique can also be applied to give good result in other areas where there is a need to classify a power prediction, Fraud/Non-Fraud, power generation etc.

So, future researches can be conducted in these areas. Hence, based on the findings of this study, the researcher would like to forward the following recommendations:

- ✓ The electric utility needs to have a data warehouse which can accommodate valuable information about their customers so that future data mining researches can easily be conducted to exploit the power data mining technology.
- ✓ Electric utilities are also suffering from electricity theft because a power system can never be 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.
- ✓ 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.
- ✓ In this research we did consider records related to customer information like customer type, location, connection fee and some other related to customer information but not power, transmission line and distribution and also customer profile information. Future research can be undertaken by including these attributes.

Bibliography

[1] A. J. C. et, ""Int. Journal of Engineering Research and Applications"," no. 22, pp. pp.38-41, may 2016.

[2] K. C. &. W. P. &. W. S. a. L. A. Kurgan, Data Mining: A Knowledge, Springer Science Business Media, ed., New York, 2007.

[3] P. Palmerini, ""On Performance of Data Mining: From Algorithms to Management Systems for Data Exploration"," February 2004.

[4] M. K. &. J. w. &. sons, "Data Mining Concepts, Models, Methods, and Algorithms", University of Louisville, 2011.

[5] P.Ozer, "“Data mining Algorithm for classification”," 2008.

[6] E. U. Ethiopia, "Www.EEU.COM," Ethiopia Electric Utility, 2016. [Online].

[7] Y. Fu, "“Data mining Tasks, techniques, and Application, Techniques, and its Applications”," no. 29, 2006.

[8] V. k. &. L. Velide, "“A Data Mining Approach for Prediction and Treatment of diabetes Disease"," International Journal of science Invention, Hyderabad Andhra Pradesh, 2017.

[9] E. P. C. Ethiopia, "WWW.EEPCO.GOV.ET," Ethiopia Electric PowerCorporation, 2013. [Online].

[10] I. Berhanu, "Measuring Customer Satisfaction Of Ethiopian Electric Utility," 2015.

[11] E. E. Utility, "External Relation Management Employee," Addis abeba, 2018.

[12] E. Statistics, "Ethiopian Power sector market," , December 2016 .

[13] D. A. Jembere, "the application of data mining to support customer relationship ethiopian air line," 2003.

[14] M. Gutema, "“Application of Data Mining Techniques for Customer Segmentation in Insurance Business the Case of Ethiopian Insurance Corporation”," 2016.

[15] B. Bezabeh, ". “the application of data mining techniques to support customer relationship management: the case of Ethiopian revenue and customs authority"," 2011.

[16] Z. D. &. X. L. X.lu, "“Electricity market price spike forecast with data mining techniques ”," 2003.

[17] P. T. Lakshay Swani, "Predictive Modelling Analytics through Data Mining," International Research Journal of Engineering and Technology (IRJET), vol. IV, no. 09, 2017.

[18] M. S. Ziafat .H, "“Using Data Mining Techniques in Customer Segmentation”," September 2014.

[19] K. T. A. Chorianopoulos, "Data Mining Techniques in Customer relationship management Inside Customer Segmentation”," 2009.

[20] D. w. &. h. K., "” hybrid data mining technique for knowledge discovery from engineering materials data sets," 2008.

[21] P. S. a. K. Cios, Data Mining Knowledge Discovery Approach, 2007.

[22] S. D. Kulwinder Kaur, "Review of Data Mining with Weka Tool," vol. IV, no. 8, Aguest 2016 .

[23] P. A. H. Agrawal, "“Review on Data mining Tools”," vol. Vol. 1 , no. Issue 2, April 2014.

[24] K. C. &. L. A. kurgan, "” Trends in Data mining and Knowledge discovery”," 2005.

[25] D. P. Vadivu, "“Optimized Feature Extraction and Actionable Knowledge Discovery for Customer Relationship Management”," vol. Int. J. Advanced Networking and Applications Volume, no. 08 Issue: 04 Pages, 2017.

[26] M. &. J. J.Han, "”Data Mining Concept and Techniques Third Edition”," 2012.

[27] D. &. A. Kamath, "Survey on Techniques of Data Mining and its Applications," International Journal of Emerging Research in Management &Technology, vol. 6, no. 2, febuary 2017 .

[28] U. S. a. H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM)," International Journal of Innovation and Scientific Research, vol. 12, no. ISSN 2351-8014, November 2014 .

[29] P. R.S. Pressnab, "“Software Engineering A Practitioners Approach Seven Edition”," 2010.

[30] S. Sharma, "“An integrated knowledge discovery and data mining process Model”," 2008.

[31] R. &. P. S.Ponmani, "”Classification algorithm in data mining Survey”," vol. Vol 6, no. ISSN: 2278-1323, January 2017 .

[32] R. R. Dr. Suresh Jain, "A Survey Paper on Overview of Basic Data Mining Tasks," International Journal of Innovations & Advancement in Computer Science, vol. 6, no. 9, September 2017.

[33] P. Sondwale, "Overview of Predictive and Descriptive Data Mining Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. V, no. 4, April 2015 .

[34] G. &. P. U.fayyad, "”Knowledge Discovery and data mining toward a unifying framework”," 1996.

[35] J. &. L. S.Bavisi, "” Acomparative study of different data mining algorithms”,," International journal of current engineering technology, Vols. Vol.4,No.5, 2014.

[36] S. G. P. a. D. K. I. L. Bhaskar N. Patel, "Efficient Classification of Data Using Decision tree," Bonfring International Journal of Data Mining, , vol. II, no. 1, March 2012.

[37] V. X.Wu, "”Top 10 Algorithm in Data mining”," IEEE International conference on Data mining(ICDM), December 2007.

[38] B. K. R. A. Kumar Mishra, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," International Journal on Advanced Electrical and Computer Engineering (IJAECE), vol. III, no. 4, 2016.

[39] M. S. P. SHRIVASTAVA, "Uses The Bagging Algorithm Of Classification Method With Weka Tool For Prediction Technique," International Journal of Advanced Computational Engineering and Networking, , vol. II, no. 12, December 2014.

[40] M. A. H. Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevie, 2011.

[41] M. &. A. A.K. Saxena, "“data mining Technology And its Application to power systems”," International Journal of Computer Applications (0975 – 8887) , vol. Volume 131 – No.8, December 2015.

[42] S. K. R. a. S. Mishra, "Adaption of Customer Relationship Management and Data mining

tools in the Data intensive Cloud Computing Environment," 2017.

[43] B. D. Pitt, "" Applications of Data Mining Techniques to Electric Load Profiling"," 2000.

[44] Hailemariam, "application of data mining techniques to customer profile analysis in the Ethiopian electric power corporation," 2011.

[45] P. African, "https://www.usaid.gov/powerafrica," Power African, 2019. [Online]. Available: https://www.usaid.gov/powerafrica.

[46] M. S. P. SHRIVASTAVA, "Uses The Bagging Algorithm Of Classification Method With Weka Tool For Prediction Technique," International Journal of Advanced Computational Engineering and Networking, vol. II, no. 12, p. December, 2014.

[47] M. J. Swasti Singhal, "A Study on WEKA Tool for Data Preprocessing Classification and Clustering," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vols. Volume-2, no. Issue-6, p. May, 2013.

[48] A. A. S. Banumathi, "PREDICTIVE ANALYTICS CONCEPTS IN BIG DATA- A SURVEY," International Journal of Advanced Research in Computer Science, vol. 8, pp. September-October, 2017.

[49] T. Adane, ""Mining Insurance Data for Fraud Detection the Case of Africa Insurance Share Company"," 2011.

[50] T. Teklu, "," Identifying "Determinant Factors for Students' Success in Preparatory Schools Using Data Mining Techniques"," June 2017 .

[51] G. G. G. Claudio J. Meneses^, "Categorization and Evaluation ofData Mining," Transactions on Information and Communications Technologies, vol. Volume 19, no. ISSN 1743-3517, 1998.

[52] A.Kumlachew, ""Data Mining Based Hybrid Intelligent System for Medical Application"," 2017.

[53] A. S. S, A. Sultana, G. M. H, M. M. R and G. Ramesh, "Product Demand Forecasting Using Data Mining," International Journal of Engineering Science and Computing, vol. 6, no. 5, 2016.