TELECOM CUSTOMER SEGMMENTATION

USING DATA MINING TECHNIQUES

A thesis submitted

By

Fikrealem Bayissa Degaga

To

The Faculty of Informatics

Of

St. Mary's University

In Partial Fulfillment of the Requirements

For the Degree Master of Science

In

Computer Science

November, 2019

Acceptance

TELECOM CUSTOMER SEGMMENTATION

USING DATA MINING TECHNIQUES

By

Fikrealem Bayissa Degaga

Accepted by Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the Degree of Master of Science In

Computer Science

Thesis Examination Committee:

Dr. Getahun Semeon

Internal examiner

_____

Signature

Dr. Tibebe Beshah

External examiner

_____

Signature

November, 2019

Declaration

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

Fikrealem Bayissa Degaga

Student

_____

Signature

Addis Abeba

Ethiopia

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

Dr. Million Meshesha

Advisor

_____

Signature

Addis Abeba

Ethiopia

November, 2019

Acknowledgement

First of all, I would like to thank my Almighty GOD for giving me the strength to finishes this research work.

My deepest gratitude goes to my advisor Dr. Million Meshesha for his patience, valuable ideas, supportive advice, and direction, encouragement and kindness to me at every stage of this research work.

I want to thank the Department of Computer Science and all academic staff especially; I would like to thank Dr. Getahun Semeon for their help and assistance during my stay.

Special thanks go to Mr.Tesfaye Ateka (enterprise customer financial institution manager), he consulting as a domain expert, he providing relevant and necessary information to this research work. And also special thanks go to Ms. Meseret Seyoum (IT application support), she is always there for me when I need her help, and especially for her assistance with the data collection and preparation work.

I would like to thank sincerely all my friends those who helped me with their valuable support during the entire process of this thesis, especially, Eskedar Yirga, Sisay Endale ,Solome Samson, and Yared Alibo,

I would like also thanks to my family's dad, moms, brothers, sister who always encourage, love and support me in all my studies, starting from my early school age.

Finally, I would like to thank all of you who support me to complete this research work.

Acronyms

CAAZ=Central Addis Ababa Zone

CRM=Customer Relation Management

CSV=Comma Separated Values

EAAZ=East Addis Ababa Zone

ET=Ethio Telecom

SAAZ=South Addis Ababa Zone

SOHO=Small Office Home Office/ Small and medium enterprise

SSE=Sum of Squared Error

SWAAZ= South West Addis Ababa Zone

WAAZ=West Addis Ababa Zone

Table of Contents

List of Tables

List of Figures

Abstract

The aim of this research is to apply data mining techniques in telecom sectors to build models that can identify the contribution that customer makes to organization profitability based on current relationship with the organization.

The objective of this research is to design enterprise customer segmentation model to Ethio Telecom that is used to identify the high value and behavior of enterprise customer. To meet the objective of the study we use hybrid data mining process model, which consists of six phases to undertake data mining process and to address the business problem. During the understanding of the problem, business practices of ET enterprise section are measured. This is done using interviews with business and technical expert and document analysis.

Data preprocessing is done using different data mining methods. To prepare the data for analysis, we select 162315 records of customer data to conduct this research. After data preprocessing, we get 21126 records with thirteen attributes that are used for data mining task.

This research is conducted using WEKA software version 3.8.2 and three clustering algorithm, namely, k-means, filtered and farthest first are used. Among clustering algorithms, farthest first clustering algorithm has better clustering performance than other cluster algorithms (filtered and k-means). Hence, the model constructed by farthest cluster is used to design a prototype.

The result of this study is interesting and encouraging and confirmed that applying data mining techniques truly support customer segmentation activities at ET. In the future we recommended more segmentation studies by using a possible large amount of customer records and employing other clustering algorithm yield better results.

Keywords: Data mining, Cluster Analysis, Hybrid Process Model, ET (Ethio Telecom)

Chapter One

Introduction

1.1 Background of the study

Telecommunication services play a significant role in the all-round political, economic and social development of a given country. It is also important in a day to day life of society [1].

In Ethiopia, Ethio Telecom (ET) is a government-owned only telecommunications service provider. It provides voice services, internet and, data services to the public throughout the country. Currently, it provides the following major types of service packages such as landline and wireless fixed, mobile, internet and data services for its government, business, and private and other non-government organizations locally and internationally [2].

ET has a huge amount of data generated from telecom services, including call detail, networks and customers [2]. Call detail data describes the call that traverses the telecommunication networks. Networks data, on the other hand, describes the state of hardware and software components in the network. Customer data describes telecom customer's data.

This huge telecom customer data is generated from enterprise and residential ET shops during giving telecom services. Residential customers are customer purchasing services for private use. The residential shop gives different services to customer including, mobile services including GSM and WCDMA, fixed line services and ADSL services, integrated services digital networks, bulk SMS services, internet service and roaming services. All residential customer data are not properly registered by employees of Ethio telecom. Because of this, we can't get full information about the customers' day to day activities and also residential customer database has not well organized customer data and not provides full information about customers. Due to this reason this research work focuses on enterprise telecom customer's data. The enterprise customers are an entity having a business license in Ethiopia or an authorized entity [3]. The enterprise shop gives different services to customers including, fixed line services, ISDN (integrated services digital networks) services, mobile services, bulk SMS services, internet services [4]. Adding to this, the enterprise shop has different activities including, collect money from a customer, selling

mobile devices to customer, selling devices that are used for internet services. Analyzing these enterprise telecom customer data using simple statistics is very difficult and putting similar data in one group and dissimilar data in another group is very difficult. It is also very difficult to make suitable promotion strategies, difficult to analyze customer value and customer loyalty, difficult to allocate resources and to understand the behavior of a customer. Using Data mining techniques has a solution to this problem.

Data mining is a technology that uses various data analysis algorithms to extract hidden business information from a large database, data warehouse, and another data repository to provide customer better services and find more commercial opportunities [5].

Data mining involves the integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis [6].

Data mining can be defined as a new kind of business information processing technology that can extract interesting patterns or knowledge implicated in a large number of incomplete, noisy, and ambiguous data that people do not know in advance but with the potential application [7]. Using data mining techniques segmenting customers provide approaches to better understand their preferences and to more efficiently allocate resources.

Customer Segmentation can be defined as aggregating customers into groups with similar characteristics such as demographic, geographic, psychographic or behavioral traits and marketing to them as a group [8]. It classifies customers into different groups based on their behavior and characteristics. The one group of a customer is different from the other customer groups but within the same groups have great similarity and share the same properties but not totally uniform.

Having, customer segmentation in ET Company is very significant to identify target customers or organization then create a specific marketing plan to communicate effectively with target customers or organization and adding to this market segmentation helpful for ET Company to identify who share similar needs and buying characteristics from enterprise customers.

Segmentation makes management's decisions easier. ET can decide which type of customers they want to target and used to build good customer relationships. It is mainly the selection of the groups of customers that will be recognized, approachable, sufficiently large, stable, perceptive, and will bring profit. Segmentation makes the company to select strong attributes and to make good customer segmentation.

Customer Segmentation requires the collection, organization, and analysis of customer data. With the proper segmentation of a customer's data, it is possible to identify the reliability/loyalty of customers so as to increase the revenue of the organization [9]. According to Tsiptsis [10], Customer segmentation is the process of dividing the customer base into distinct and internally homogeneous groups in order to develop differentiated marketing strategies according to their characteristics. In addition to this, proper segmentation of a customer's data allows to the ET to provide a product within the target audience needs and wants.

Segmentation is essential to ET enterprise division to establish the needs and values of the target customers within each segment, in order for the company to promote into products, brands or services appropriately. By the following ET enterprise, customer segmentation will establish good customer relationship for organization and establishes proper strategies to interact with different customer groups appropriately. In addition to this, the marketer will achieve the necessary knowledge and thereby be able to design a believable and appropriate marketing strategy.

This research work design a prototype that demonstrates how customer segmentation can be applied based on the enterprise customer's data and how to differentiate the high-value customers from low-value customers.

1.2 Statement of the Problem

ET use statistical analysis techniques to identify the contribution that customer makes to organization profitability. The existing customer segmentation way is based on the investment capital of the customer, the number of a permanent employee of the company and the company's branch offices [11]. The expert checks customers database whether the customers are enterprise customer or residential customer and at the same time they cross check the customers database whether the customers key account customers or SOHO (Small Office Home Office) customers

[12]. For classification parameters in this study we use key account and SOHO (small office home office) parameters. In accordance with what they had in their declaration form and also cross check the customers database whether customers are platinum customers, gold customers, silver customers and VIC [13].

The managers using these techniques check the status of the customers, are they inactive state, suspend state, idle state, barring state and deactivated state. These techniques are not effective and efficient and took a considerable amount of time to treat customers according to their behavior. Customer relationship management (CRM) comprises a set of processes and enabling systems supporting a business strategy to build long term, profitable relationships with specific customers [14].

All customers' data stored on well-organized enterprise customer's database but there is no well integrated system or model to solve customer segmentation problem. Adding to this, the ET enterprise database has no well-organized set of rules or procedures that are used for segmenting customers according to their culture, behavior, and characteristics to say whether the customers are potential or low valued customers. As a result, the organization cannot identify high-value customers, what kind of customer has high credit, cannot provide an analytical basis of marketing strategies for developing new business, and cannot categorize the behavior of the customer, difficult to allocate resources and reduce the revenue. Therefore, these techniques are not effective and efficient and took a considerable amount of time to treat customers according to their behaviors.

Data mining technology is appropriate in order to solve the above problem. In the telecommunications industry, there are some attempts by different authors. Subramanian and Prabha [15] conducted a survey on customer relationship management in order to solve customer relationship management problems in enterprise customer data using data mining techniques. Qiuru et al [16] explored Telecom customer segmentation based on cluster analysis using telecom customer data to solve customer segmentation problems using data mining techniques .Customer segmentation so as understand the customers, customer's overall compassion, group characteristics of various valuable customers with different consumption characteristics and credit rating.

In the context of Ethiopia, Henock [17] attempted to apply data mining to support customer relationship management in the case of Ethiopian airlines to explore the behavior of customers and to increase the revenue of the organization using data mining techniques. But, based on the review made there is no work done to support ET customer's segmentation.

This study there for tries to apply data mining techniques for constructing a descriptive model that helps to determine the enterprise customer behavior and also determining the contribution that customer makes to organization profitability based on the current relationship with organization.

Therefore, this research attempted to address customer segmentation problem according to customers' behaviors and customer's value. The study focuses on a clustering customers based on their attributes, which is found in the enterprise customers' database of Ethio Telecom. Finally, the study explores and answers the following major research questions.

- What are the suitable attributes to segment ET enterprise customers?
- Which data mining techniques and algorithms can be used for customer segmentation?
- To what extent the model performs in customer segmentation?

1.3. Objective of the study

1.3.1 General Objective

The general objective of this research is to design a customer segmentation model to identify the behavior of enterprise customers and value using data mining techniques.

1.3.2 Specific objectives

The specific objectives of this research include the following:

- To understand customer segmentation problem based on a review of related works done so far in the area of data mining for the telecommunication industry.
- To prepare data set for model building by selecting, cleaning, constructing and integrating the collected data.
- To select data mining algorithms for descriptive data analysis

- To conduct experimentation for constructing a customer segmentation model.
- To evaluate the performance of the model and select the best model.

1.4 Scope and Limitation of the study

The telecommunications industry was one of the first to adopt data mining technology [18]. This is most likely telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment. In Ethiopia, ET produced a huge amount of high-quality data and produced very large customer data across the country during giving telecom services. Using this huge ET customer's database, the data mining techniques used to develop interesting segments and to extract interesting patterns based on interests, a lifestyle of buyers, priorities and emotions**,** income, religious, education, company size, industry, role, time working for the company. The customer data used for the current research covers four-month data taken from January of 2019 data up to 2019 April. In addition to this, the data sources for gathering customer's data were from the Enterprise database of ET.

In this research, a hybrid process model is adopted to undertake the data mining process and to make large data mining researches. The models of customer segmentation were developed by descriptive models using a clustering algorithm which cluster the subgroup of the dataset to the nearest mean or to the centroid.

During conducting this research there are some limiting factors such as schedule and the financial plan we set to meet the goal of these researches. In addition to this, call detail attributes such as call duration, times of call, the number of different telephone numbers called by the speaker, concentration of call duration and concentration of times of call not involved in the datasets because it is sensitive attributes that involve protection against customer privacy.

Moreover, during conducting these studies we are restricting not to access other data of customers. The data collected for these researches is extracted from an enterprise customer profile database. As a result, it limits the research to integrate data from the different databases of the company branches. Include as limitation, this study is not use residential customer data.

1.5 Significance of the study

The significance of this research is to consider the applicability of data mining techniques in the telecom industry to build models that can groups customers based on their behavior and values. Based on this, the advantage of having customer segmentation in telecom customer data including the following [19]:-

- Improving promotion effect**:** the customer segmentation based on the data mining can be helpful for the enterprises to make suitable promotion strategies, in a suitable time, with suitable products and services, aiming at suitable customers.
- Analyzing customer value and customer loyalty: customer value and customer loyalty are important to the enterpriser's stratagem and management tactics. Enterprises can confirm the rank of the customer according to their expected value and loyalty analyzed by the segmentation model based on data mining.
- Analyzing credit risk**:** risk scoring is an effective way of evaluating certain specific types of customer risk, normally the risk of default.
- Instructing new products: Enterprises can find out the preference of their customers by customer analyzing based on data mining, and make sure that various demand will be realized in the new design.
- Confirming the target market: Customer segmentation based on data mining can make targeted customer group clear and locate the market explicitly.

The main contribution of this research is finding an effective model for customer segmentation. The final output is segmented the customer data into a given number of clusters and categorizes them. Based on this, the following benefit we gain from the finding of this study.

Firstly, the researcher gained skills in conducting research than this study done for academic purpose. Therefore, finding of this study helps another researcher to do further researches in the area.

Secondly, the result of the study helps to ET to be able to manage enterprise customer’s data. As a result, ET is beneficiary by seeing the behaviors of customer groups than the organization can classify new coming customers. In this study, two segmentation group, these groups have their own contribution and WEKA results. The first group of customers’ high-value customers to the

company and their contribution is high then the company creates good relations with this customer in these segments. The company will know what type of product they use, what type of product they interested, we can have invested many financial sources into these segments. The second group of customers is a low-value customer to the company. Under this segmentation type customers are generating low revenue to ET. After seeing this company will build new strategies to have a good relation with them. The Company should build new strategies to increase the revenue of low-value customers. This is done the company should propose some special offer of advantageous price to the customer.

Thirdly, the result of the study can serve as documentation for universities for additional investigation in the area.

1.6. Methodology

The methodology is the procedure that a researcher follows to achieve the above specific and general objectives [20]. Hereunder a description of the step-by-step procedure followed in this study is provided.

1.6.1Research Design

This study follows experimental research to design an optimal customer segmentation model. Experimental research provides a high level of control; it produces results that are specific and relevant to consistency.

For the data mining process, there are different types of standards and methodologies used such as a sample, explorer, modify, model and assess (SEMMA), Knowledge Discovering in Database (KDD), CRoss-Industry Standard Process(CRISP) and hybrid data mining process model. To apply this experimental research, the study follows a hybrid model, developed based on the CRISP-DM model by adopting it to academic research and industries. Hybrid DM process model is selected because of the following reasons [21]:

- It provides a more general, research-oriented description of the steps.
- It introduces data mining steps instead of modeling steps
- It introduces several new feedback mechanism

- It identifies the use of the discovered knowledge for a particular domain in other domains.

The hybrid process model consists of six phases as shown in figure 1.1 below. These phases are understanding of the problem, understanding of the data, preparation of the data, data mining, Evaluation of the discovery knowledge and use of the discovery knowledge [21].

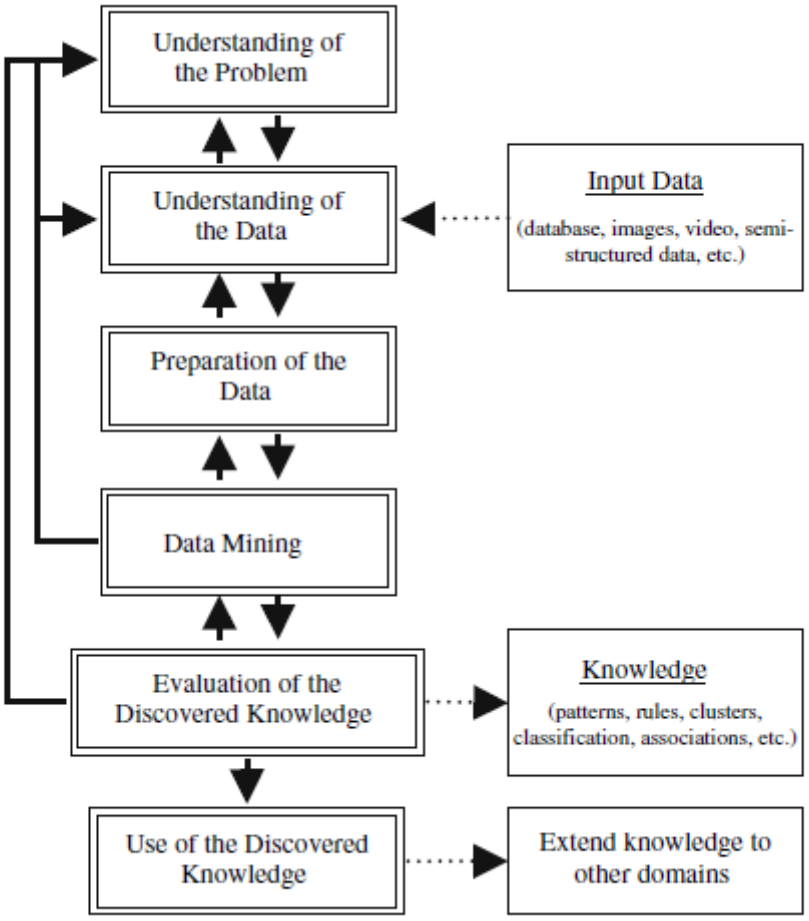


Figure 1.1 Hybrid process models [21].

1.6.2 Understanding of the Problem

This initial step involved working closely with a domain expert to define the problem and determines project goals, identifies key people and learning about the current solution to the problem. It also involves learning domain-specific terminology. Finally, project goals translate

to data mining goals, and the initial selection of data mining tools to be used later in the process is performed.

In this research, we work closely with the domain expert of ET, then we define, identifies, understand and formulate the problem area, then determine to attribute future selection and understand the business process. Based on the knowledge gained from the domain experts about the area of the telecom business, finally, we define the data mining problem. Working together with domain expert we define customer segmentation problem domain in enterprise customers, then we select enterprise customer profile data as the main sources of data collection for this research.

1.6.3 Understanding of the Data

This step includes collecting samples data and decides which data, include format and size, will be needed. Data are checked for completeness, redundancy, missing values and likelihood attribute value. Finally, the steps include verification of the usefulness of the data with respect to data mining goals.

In this research, in order to understand the data, brief discussion on the enterprise customer data are conducted with the domain expert of enterprise section. The discussion includes listing out initial attributes, their respective vale and evaluation of the importance of enterprise customer data.

Oracle database: the enterprise customer data records were very huge data and contain data from different sources and also every enterprise shop activities are stored in this data base and it is saved in damp file. To read this file, Oracle Database is used and retrieves necessary data in collaboration with domain experts. Out of the total 21126 data sets, 58% of them are belongs to the group of Key account customer and 41.43%of them are groups under SOHO. With respect to call detail records where they found up voice usage average holds 29.88%, high holds 33%, and low holds 36.83% respectively, data uploading usage high holds 50%, low holds 49%, respectively, data downloading usage high holds 50%, low holds 49%, respectively and SMS usage high holds 50%, low holds 49%, respectively, zone name holds CAAZ usage high holds 19%, NAAZ holds 10%, respectively, EAAZ usage high holds 12%, SAAZ holds 13%, respectively, Enterprise usage high holds 17%, SWAAZ holds 12%, respectively, WAAZ usage

high holds 13.63%, Customer level name holds gold 50.90% and platinum holds 49.10%, Net business type holds CDMA 12% , fixed line 29%, GSM 20%, LTE 29%, WCDMA 28%, respectively, offer name holds wired line 53%, wireless 47% respectively, segment name holds postpaid 52%, prepaid 47% respectively, investment holds high 60%,40.33% respectively, number of employee holds high 53%, low holds 46% respectively and number of branch holds high 54.11%, 45.89% respectively. To understand the nature of the data, descriptive statics is used.

1.6.4 Preparation of the Data

The main goal of data Preparation is to get appropriate data for the final data set in order to use for the experimentation. These steps concerns decide which data will be used as input for data mining methods in the subsequent input. It involves sampling, running correlation, and significance test, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing value. The cleaned data further processed by feature selection and extraction algorithm (to reducing dimensionality) and by the derivation of new attributes and by summarization of data. The end result is data that meets the specific input requirement for the data mining tool selected in the first step.

In this research, data cleaning done by WEKA attributes selection preprocessing techniques to reduce dimensionality and by derivation of new attributes. The end result of these processes generates datasets for training and testing of the clustering algorithms selected in this study.

1.6.5 Data Mining

In this study, WEKA version, 3.6 DM software is used. It is a knowledge discovery system developed by the University of Waikato in New Zealand that implements data mining algorithms. It implements algorithms for data preprocessing, classification, clustering, and association rules. It also integrates a feature selection and visualization algorithm. For this research work, we choose the K-means algorithm, filtered cluster algorithm, and farthest first algorithm. These algorithms are selected because all they are usually used and also can identify

meaningful natural groups and they can help to group customer into distinct segments that have similar characteristics.

1.6.6 Evaluation of the Discovered Knowledge

Before proceeding to the next steps of the hybrid model, it is important to more thoroughly evaluate the model. The Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impacts of discovered knowledge. The evaluation metrics was used to measure and summarize the quality of the trained classifier when tested with the unseen data. In order to check model performance and effectiveness accuracy, recall, and precision are used. Accuracy or error rate is one of the most common in practices used to evaluate the generalization ability of classifiers. Through accuracy, the trained classifier is measured based on total correctness which refers to the total of instances that are correctly predicted by the trained classifier when tested with the unseen data. General, the accuracy metrics measures the ratio of correct predictions over the total number of instances evaluated. The recall is used to measure the fraction of positive patterns that are correctly classified. Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class [22].

In these researches, to evaluate the results based on the incorrectly clustered instance and ET domain expert judgments, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge was evaluated.

1.6.7 Use of Discovering Knowledge

This is the final step of a hybrid model consisting of planning where and how to use the discovered knowledge. A plan to monitor the implementation of discovered knowledge is created and the entire research documents.

The knowledge gained need to be organized and presented in a way that the customer can use it. In this research, the discovered knowledge is used by integrating the user interface which is designed by Java programming language with a WEKA system in order to show the behaviors and values of ET enterprise customers.

In this research, the Java programming language was used to design a prototype user interface to use the model for customer segmentation. The main reason for selecting java programing language for this research was because of the fact that [23]:-

- Java is easy to learn: - it was designed to be easy to use and is therefore easy to write, compile, debug, and learn than other programming languages.
- Java is object-oriented: - this allows us to create modular programs and reusable code.
- Java is platform-independent: - one of the most significant advantages of Java is its ability to move easily from one computer system to another.
- Freedom of pointers: JAVA is free from pointers hence we can achieve less development time and less execution time.
- Java is Robust: - since Java is both a compiler and interpreter language, it puts a lot of emphasizing on early checking for all possible errors and exceptions.
- Java is Architectural Neutral: - a language or technology is said to be architectural neutral which can run on any available processors in the real world.
- Java is Portable: - A portable language is one that can run on all operating systems and on all processors irrespective their architectures and providers.

The prototype is evaluated following user acceptance testing to see its efficiency, effectiveness, error tolerance, engaging and easy to learn.

1.7 Organization of the thesis

This research is organized into five chapters. The first chapter deals with introducing the data mining, advantage of data mining in customer segmentation and the advantage of having customer segmentation in ET enterprise division. In this chapter objective, Specific objective statement Problem, scope, and limitation of the research and significance of the research discussed in this chapter.

The second chapter is committed to reviewing the literature on data mining concepts and techniques and related research works that intersect data mining and specifically customer segmentation are also reviewed.

The third chapter deals with what kind of techniques followed in order to understand the problem, collect the data, understand the data, analyze the data, construct and evaluate the models. In general this chapter deals with how this project is conducted. The hybrid process

model is used in the methodology part, involves standard research methodology and detailed data mining techniques. The methodology understands the business (domain) and the data as well as to prepare the data for analysis.

The third chapter deals with what kind of techniques followed in order to understand the problem, collect the data, understand the data, analyze the data, construct and evaluate the models. In general, this chapter deals on how this project is conducted. A hybrid process model is used in the methodology part, involves standard research methodology and detailed data mining techniques. The methodology understands the business (domain) and the data as well as to prepare the data for analysis.

Chapter five cluster analyses and experimental results are discussed. In this chapter, we have used the clustering analysis with the k-means algorithm. The distance measure Euclidean Distance is applied to measure the distance from the centroid or mean value. The other clustering analysis conducted in this chapter is an expectation–maximization and farthest first algorithm.

Finally, conclusions and recommendations are given in chapter six.

Chapter Two

Literature Review

2.1 Overview of Data Mining

These days the amount of data stored in the database is growing rapidly and can range in size into terabytes which means the data stored in the database is an enormous amount of data. Using this data the organizations unable to segment customers into different groups according to one or more attributes, unable to understand customers overall composition, unable to understand group characteristics of various valuable customers, unable to understand group characteristics of loss customers, unable to understand consumption characteristics of customers, unable to understand group characteristics of customers with different credit rating and also unable to extract important information for decision-making purpose [19].The knowledge extracted from this database is poor knowledge. Therefore data mining application has a solution for these problems.

2.1.1 What is Data Mining?

Data mining is a technology that uses various data analysis algorithms to extract hidden business information from a large database, data warehouse and another data repository to provide customer better services and find more commercial opportunities [19]. Data mining technique also enables to predict future trends and behaviors and helps organizations to make proactive knowledge-driven decisions and it can solve the problems that traditionally were too much time consuming like preparing databases for finding hidden patterns, finding descriptive information that experts may miss. And, different techniques and algorithms are used to accomplish the tasks of data mining [24]. As a result, data mining analyzes the data from a large database and summarizes that data into useful information. Using these useful information organizations identify valuable customers; predict their future behaviors, which enable them to make positive, knowledgeable decisions.

Data mining is defined as exploration and analysis of large quantities of data by automatic or semi-automatic means to discover meaningful patterns and rules and these patterns allow a company to better understand its customers, and improve its marketing, sales, and customer support operations [25].Data mining study in order to obtain useful information about the characteristics of customers and helps in the decision-making process.

2.1.2 Why data mining?

Computer technology nowadays very rapidly growing and a huge amount of data generated and stored in the database, data warehouse, and another information repository. Using this data extraction of interesting information or pattern from a large database is difficult. The data stored in the information repository is growing rapidly nowadays due to these reasons we can't mine hidden business information from a data repository. It needs technology to perform these tasks. This technology is a new technology and very significant for a human to extract interesting patterns from huge data, convert the data from incomplete, noisy and ambiguous to the meaning full information data, used to extract useful knowledge from a data warehouse and to discovering meaningful new correlations [24]. Data Mining is used by a different organization to provide better services and find more commercial opportunities.

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 [26]. As shown in figure 2.1 the general experimental procedure adapted to data mining problems involves the following five steps [26]: state the problem, collect the data, and preprocess the data, estimate the model (mine the data) and interpret the model and draw a conclusion.

- **State the problem**

Domain - specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement [27]. Understanding the objective of the project is the first step in the data mining project then identifies the data mining problem and develops an initial implementation plan.

- **Collect the data**

This step concerned how the data is collected and generated. The data usually collected from the database, data warehouse and transactions recorded by individuals are the major sources of information. According to Mehmed [26], there are two methods two collect the data these are:

- Designed experiment: the data generation process is under the control of an expert.
- Observational approach (random data generation) the data generation process expert cannot influence.

- **Data pre-processing**

After collecting the data then pre-processing will follow.There are two main tasks in the pre-processing data mining process model. Outlier detection is the first task and unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and sometimes are natural, abnormal values.

Scaling, encoding, and selecting features are the second tasks. Pre-processing steps should not be considered as completely independent from other data - mining phases. Iteration of the data - mining process, all activities, together, could define new and improved data sets for subsequent iterations.

- **Estimate the model**

After define new and improved data set then selecting appropriate data mining techniques is the major task in this phase. From the several techniques choosing appropriate data mining techniques is the best one.

- **Interpret the model and draw conclusions**

Data- mining models should help in decision making and such a model needs to be interpretable in order to be useful. Usually, simple models are more interpretable, but they are also less accurate. Modern data- mining methods are expected to yield highly accurate results using high - dimensional models.

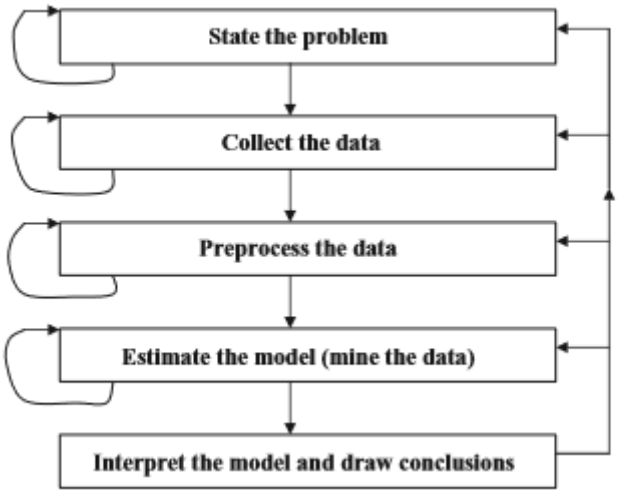


Figure 2.1 data mining process model [26]

Nowadays, there are different data mining process model for the understanding of data mining project and for data mining research.

2.2.1 KDD process model

Data mining is finding a useful pattern in the data and a step in the KDD process consisting of applying data analysis and discovering algorithms that, under acceptable computational efficiency limitation, produce a particular enumeration of pattern over the data [28]. KDD is the process of extraction of previously unknown and potentially useful information from huge storage of data like database, data warehouse. KDD process is the process of using a database along with any required selection, preprocessing, subsampling and transformation of the database [28]. KDD is an iterative process that consists of five data mining stages (see figure 2.2) and they are described below [29].

Data Selection:

In this phase we decide which type of data is going to be used for data mining and also this phase contains creating a target data set or focusing on a subset of variables or data samples, on which discovery is to be performed.

Preprocessing

In this phase, we cleaning target data and to get clean data we apply basic operations such as removal of noise data, collecting the necessary information to model and account noise, decide

strategies for handling missing data, accounting for time sequences information and known changes.

Transformation

In this phase, convert the data from different sources such as database, data warehouse into common new format using data reduction and data categorization methods.

Data mining

In this phase selecting methods to use for searching for patterns of interest in particular representational form and apply classification or clustering techniques to obtain predictive and descriptive models.

Interpretation/Evaluation

In this phase, the mined patterns or knowledge are evaluated based on the given measure and present the result to the user in a meaningful manner using various visualization and GUI strategies.

Combining discovered knowledge

The final step consists of combining the discovering knowledge of the existing system and documenting and reporting to the interest body.

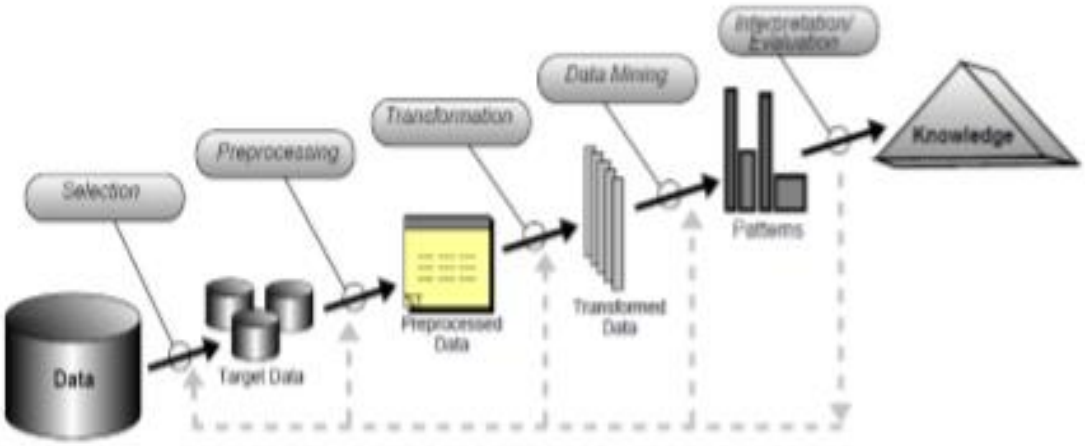


Figure 2.2 KDD process model for data mining [29]

2.2.2 SEMMA DM Process Model

SEMMA stands for (sample, explore, modify, model and assess) is one of the data mining process model and proposed by SAS (Statistical Analysis System) for DM project. It offers and allows understanding, organization, development, and maintenance of data mining projects and it helps in providing the solutions for business problems and goals [30]. A SEMMA life cycle consists of five stages as shown in figure 2.3 [30].

- **Sample:** this first stage optional which focuses on sampling data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.
- **Explore:** the second stage which focuses on the exploration of data. This can helps in gaining understanding and ideas as well as refining the discovery process by searching for trends and anomalies**.**
- **Modify:** this third stage focuses on the modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.
- **Model:** this fourth stage which focuses on the modeling of data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
- **Assess:** this last stage consists of assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

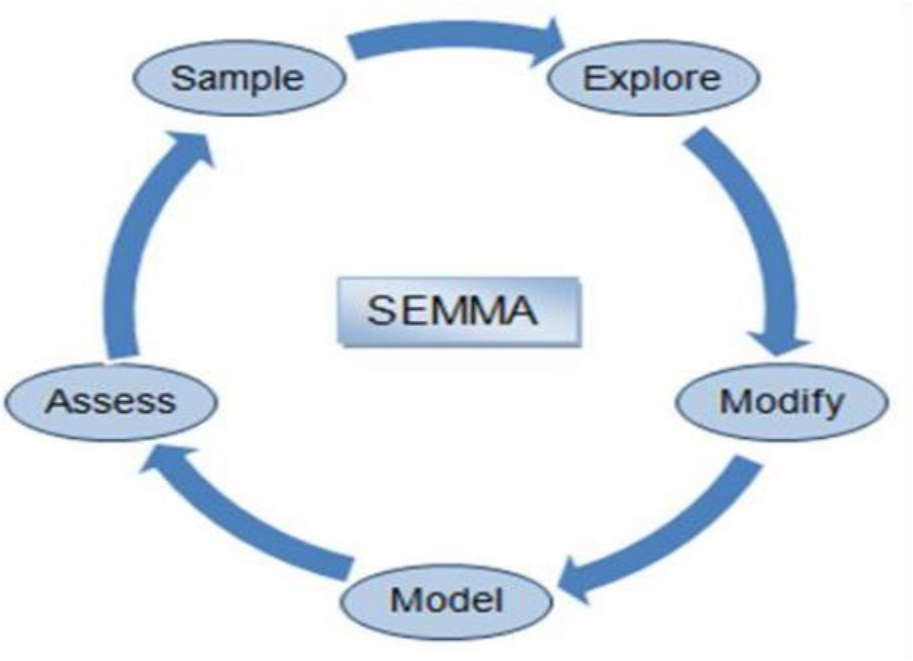


Figure 2.3 SEMMMA data mining process model [30]

2.2.3 CRISP-DM process model

The data mining process model defines the approach for the use of data mining and also define phases, activities, and tasks that have to be performed. CRISP-DM is project proposed a comprehensive process model for carrying out data mining projects. CRISP-DM methodology with its distinction of generic and specialized process models provides both the structure and the flexibility necessary to suit the needs of both groups and useful for planning, communication within and outside the project team, and documentation [31]. CRISP-DM is divided into six phases as shown in Table 2.1 [31].

- **Business understanding**: this phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- **Data understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

- **Data preparation:** The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
- **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase.
- **Evaluation:** What are, from a data analysis perspective, seemingly high-quality models will have been built by this stage of the project? Before proceeding to the final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives.
- **Deployment:** Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

Business Understanding	**Data Understanding**	**Data Preparation**	**Modeling**	**Evaluation**	**Deployment**
Determine business objective	Collect initial plan	Select data	Select modeling techniques	Evaluate Result	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring and maintenance
Determine DM objective	Explore data	Construct data	Building model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Table 2.1 CRISP-DM process model [31]

2.2.4 Hybrid DM Process Model

The hybrid process model has been developed based on the CRISP-DM model by adopting it to academic research [32].

Hybrid Process Model has six phases that are described below [33]:

- Understanding of the problem domain:-the initial phase focus on understanding the project goal, identify the key people, identify the project problem, investigate the current solution to the problem through close discussion with a domain expert, then converting this knowledge into DM problem definition and preliminary plan design to achieve the objective.
- Understanding of the data: - the data understanding phase beginning with data collection then chosen the size of the data and format the data set. In this phase, the data are checked such as checking missing values in the data, redundancy, completeness and checking the usefulness of the data tested with respect to data mining goals.

- Preparation of the data:-In this phase decide which data going to use for DM methods. It consists of activities such as sampling, testing the correlation, significance of the data, cleaning data, checking the completeness of the tuples, handling noisy and missing value. The cleaned data may be further processed by feature selection and extraction algorithms, by derivation of new attributes, and by summarization of data. Finally, the datasets that meet the input requirements of DM tools stated in the first step are selected for modeling purposes.
- Data mining:-In this phase, DM methods and algorithms are selected for predictive and descriptive modeling so as to derive hidden knowledge from preprocessed data.
- Evaluation of the discovered knowledge:-In this phase the result of DM methods evaluate, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. In addition, the approved models are taken and the whole process is revised to pinpoint an alternative solution, in order to improve the results achieved. Finally, the errors arisen in the process are listed and arranged.
- Use of the discovered knowledge:-This is phase is the final stage consists of how the discovery knowledge is used. The knowledge extracted in current knowledge may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.2.5 Comparison of Data mining Process Models

In the data mining process model, there are different standard models such as Cross- Industry Standard Process for Data Mining (CRISP-DM), Knowledge discovery in database (KDD), Sample Explore Modify Model Assess (SEMMA), Hybrid process model. These entire process models have multiple steps executed in a sequence, which often includes loops and iteration. Each subsequent step is initiated upon the successful completion of a previous step and requires a result generated by the previous step as its inputs [34]. All process models have iterative nature of the model and these iterative models consist of many feedback loops and repetitions, which are generated by a revision process. Table 2.2 shows the difference between KDD process models and the SEMMA process model based on the steps they had [35].

Steps	KDD	SEMMA
Areas	Academic	Industrial
1	Data Selection	Data Samples
2	Data preprocessing	Data Explore
3	Data Transformation	Modeling
4	Data Mining	Model
5	Data interpretation/evaluation	Assess

Table 2.2 the difference between KDD and SEMMA process models [33]

Examining it thoroughly, we may affirm that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process since it is directly linked to the SAS Enterprise Miner software. On the second step, the differences between the KDD processes model and CRISP-DM process model based on the steps they had.

The difference between the KDD stages with the CRISP-DM stages is not straight forward as SEMMA situation.

- The Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user.
- The Deployment phase can be identified with the consolidation by incorporating this knowledge into the system.
- The Data Understanding phase can be identified as a combination of Selection and Preprocessing.
- The Data Preparation phase can be identified with Transformation.
- The Modeling phase can be identified with DM.
- The Evaluation phase can be identified with Interpretation/Evaluation.

The CRISP-DM process model is used for the industrial model and this model extremely complete and documented and all stages are duly organized, structured and defined, allowing that a project could be easily understood or revised [34].KDD process model is used for the

academic research model and it more suitable for data mining experts because it is an accurate and complete process model.

On the third step, the differences between the CRISP-DM and Hybrid process model [34].The main difference and extension include:-

- A Hybrid model providing a more general, research-oriented description of the steps
- Hybrid model introducing a data mining step instead of the modeling step,
- Introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
- It subjects the use of the discovered knowledge for a particular domain in other domains.

By considering, the above difference between CRISP-DM and the Hybrid process model, we choose the Hybrid process model for this research because the Hybrid process model developed based on the CRISP-DM by adopting it to academic research and using the discovered knowledge into other domains.

2.2.6 Comparison of process model

Table 2.3 present a comparative analysis of KDD, CRISP-DM, SEMMA, and Hybrid DM process Model.

KDD	SEMMA	CRISP-DM	Hybrid
Pre KDD	----------	Business understanding	Understanding of the problem
Selection	Sample	Data Understanding	Understanding of the data
Pre Preprocessing	Explore		
Transformation	Modify	Data Preparation	Preparation of the data
Data Mining	Model	Modeling	Data Mining
Interpretation/Evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
Post KDD	----------	Deployment	Use of the discovered knowledge

Table 2.3 the difference among DM process model [36]

From the three process model we selected hybrid process model for this research paper. The reason for selecting this model, it provides a more general, research-oriented description of the steps, it introduces data mining steps instead of modeling steps, it introduce several new feedback mechanism and it is identifies the use of the discovered knowledge for a particular domain in other domains.

2.3 Data mining Tasks

Data mining uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. It deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of function involved in data mining. They are descriptive tasks and predictive tasks.

Predictive tasks intended to predict (forecast) a future value or unknown value of a dependent variable (target variable). The model enables to predict the value of one variable from the known values of other variables. Used to predict the future habit or based on the existing data or behaviors. The predictive task includes classification.

Classification is a predictive data mining task aims at predicting (forecasting) a future value or unknown value of a dependent variable and a data type of a dependent variable should be a categorical variable and each value belongs to this variable is called a class label [37] Classification building a model to predict through classifying database records in to a number of predefined classes based on certain criteria.

The model is constructed by analyzing the relationship between the attribute and class of the objects in the training set. Such a classification model can be used to classify further objects and develop a better understanding of the classes of the objects in the database.

Descriptive task concern about finding the correlation between data attributes. Descriptive tasks describe the existing data which means that a model will be created that determine the overall probability distribution of the data and the relationship between the variables. Descriptive tasks include summarization, clustering, and association rules mining. The differences between the two tasks are summarized in table 2.4.

Descriptive Task	Predictive Task
An unsupervised model which discretizes attributes without taking in to account class information.	A supervised model which discretizes attributes while using interdependence between known class labels and attribute values.
The algorithm is provided with a training set of data, which not include the classified values of the target variables. Therefore, a class is blind.	The algorithm is provided with a training set of data, which includes the classified values of the target variables in addition to the predicator's variables. Therefore, a class is aware.
There are no target variables for descriptive, hence DM search for pattern and structure among all the variables.	There are target variables for prediction. No need DM search for a pattern.

Table 2.4 difference between predictive vs. descriptive tasks [6, 49]

This study uses a clustering algorithm to construct a descriptive model for customer segmentation.

2.4 Overview of Clustering

Every DM system developer needs to meet some functionality. This functionality valued to make sure that the systems are performing effectively. Effectiveness means the system fulfills the whole objective of the system. We used to fulfill the whole objective of the system unsupervised clustering techniques for conducting the experiments.

Clustering is a descriptive data mining task that aims at grouping data instances into shared characteristics [38]. It is the process of putting similar data in one group or cluster and dissimilar data in other groups and identifies a set of categories to describe the data. Each category can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or over-lapping categories, Partition, Model-based (a mixture of probabilities) clustering [38]. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity than a cluster of an object is formed. So that object

within a cluster has high similarity, however very dissimilar to object in another cluster. Using this cluster rules are generated.

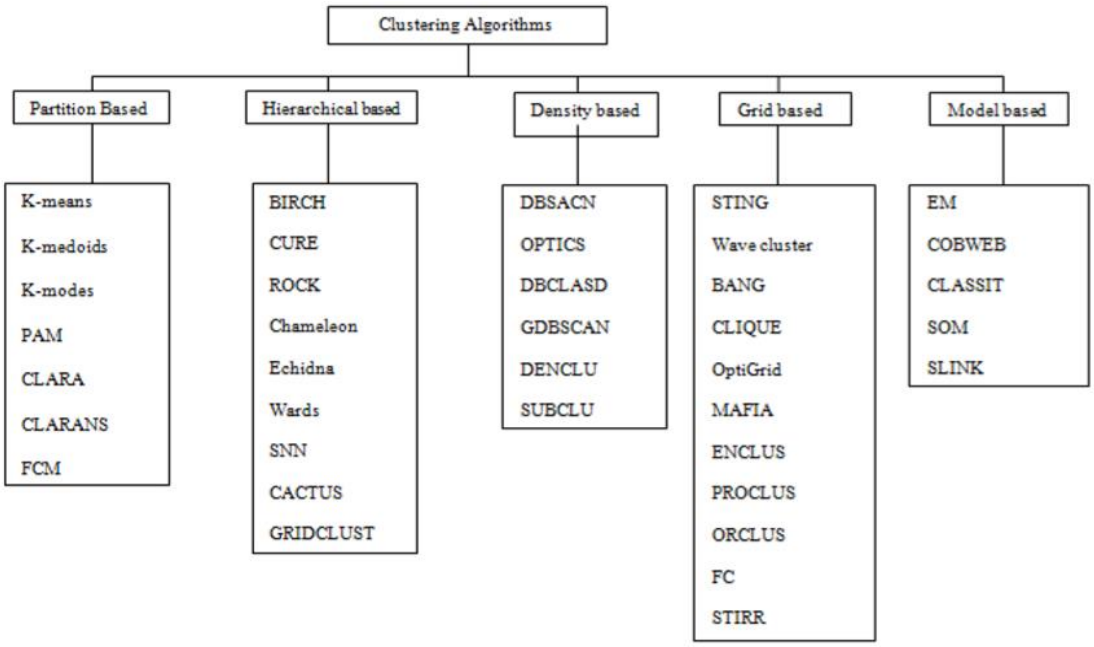


Figure 2.4 Overall clustering algorithms [39]

From various clustering algorithms are partitioning clustering algorithm, hierarchical clustering algorithm, density-based clustering algorithm, and grid-based and model-based clustering algorithms [6]. Clustering algorithm, we use K-means algorithm and filtered clustering algorithm and farthest first clustering algorithm.

2.4.1 Clustering Algorithms

Descriptive modeling finds the relationship between data attributes that describe the relationship between the variables and describe the data. The data within each class form a cluster. The number of clusters is equal to the number of output classes. The clustering technique produces clusters in which the data inside a cluster have high intra class similarity and low interclass similarity. Clustering is mainly classified into hierarchical and partitioning algorithms [38].

The hierarchical algorithms are further subdivided into agglomerative and divisive. Agglomerative clustering treats each data point as a singleton cluster and then successively merges clusters until all points have been merged into a single cluster. Divisive clustering treats

all data points in a single cluster and successively breaks the clusters till one data point remains in each cluster. Partitioning algorithms partition the data set into predefined k number of clusters [6]

2.4.1.1 Partitioning algorithms

It partitions the database D of n objects into a set of k clusters so that it optimizes the chosen partition criterion. Each object is placed in exactly one of the k non-overlapping clusters. K-means algorithm is based on the partition method of clustering. In this, the assignment of objects to k clusters depends on the initial centers of the clusters.

The steps in **K-means algorithm** are as follow [40]:

1) Initialize centers for k clusters randomly
2) Calculate distance between each object to k-cluster centers using the Manhattan distance formula given by below equation 2.4

$$d(x, y) = \sum_i^n |(xi - yi)| \ldots\ldots\ldots\ldots\ldots\ldots (2.4)$$

3) Assign objects to one of the nearest cluster center
4) Calculate the center for each cluster as the mean value of the objects assigned to it
5) Repeat steps 2 to 4 until the objects assigned to the clusters do not change

2.4.1.2Hierarchical algorithms

It creates a hierarchical decomposition of the set of objects either using a top-down approach or bottom-up approach [40]. The agglomerative clustering algorithms use the bottom-up approach and divisive clustering algorithms use a top-down approach. It does not require the number of clusters k as the input but requires a termination condition. Single link and complete link algorithms are examples of the agglomerative hierarchical clustering method.

The step in the single link algorithm as follows [40]:

1) Assign each object to its own cluster(singleton cluster)
2) Calculate the distance from each object to all other objects using Manhattan distance (Eq2.4) and store it in a distance matrix
3) Identify the two clusters with the shortest distance in the matrix and merge them together
4) The distance of an object to the new cluster is the minimum distance of the object to the objects in the new cluster

5) Update the distance of each object to the new cluster in the distance matrix
6) Repeat steps 3 to 5 until the required number of clusters are obtained

It produces non-elliptical shapes but produces long and elongated clusters. The steps in the **complete link algorithm** are as follows [40]:

1) Assign each object to its own cluster (singleton cluster)
2) Calculate the distance from each object to all other objects using Manhattan distance (Eq. and store it in a distance matrix
3) Identify the two clusters with the shortest distance in the matrix and merge them together
4) The distance of an object to the new cluster is the maximum distance of the object to the objects in the new cluster
5) Update the distance of each object to the new cluster in the distance matrix
6) Repeat steps 3 to 5 until the required number of clusters are obtained

The hierarchical algorithm cannot handle more than a few thousand cases effectively. Thus, sampling the cluster population is required. This task is time-consuming and is not an ideal to sample cluster population. Therefore, it is challenging to apply it for business clustering tasks. Yet, another clustering algorithm such as k-means can handle millions of records without sampling. For this research work, we choose the K-means algorithm, filtered cluster algorithm, and farthest first algorithm.

2.4.1.1 K-means Clustering Algorithm

K-means clustering algorithm is one of the most commonly used clustering techniques for customer segmentation research works. K-means is the simplest clustering algorithm because of its simplicity in implementation, fast execution, easily understandable and also the clusters do not have overlapping character. The K-means algorithm has been widely used in customer segmentation, pattern recognition and information retrieval [19].K-means clustering algorithm group's data vectors into a predefined number of clusters, based on Euclidean distance from one another, and are associated with one centroid vector which represents the midpoint of that cluster [41]. This method starts by selecting randomly k points (it can be also examples) to be the seeds for the centroids of k clusters then assign each example to the centroid closest to the example, forming in this way k exclusive clusters of examples. After this calculate new centroids of the clusters. For that purpose average all attribute values of the examples belonging to the same

cluster (centroid) then finally check if the cluster centroids have changed their "coordinates". If yes, start again from step 2. If not, cluster detection is finished and all examples have their cluster memberships defined. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function [42]. The K-means clustering algorithm is shown in figure 2.4.

After loading CSV file (enterprise customers CSV file) we cluster the instance using a K-means algorithm. It selects random k instances and centroids to start the evaluation.

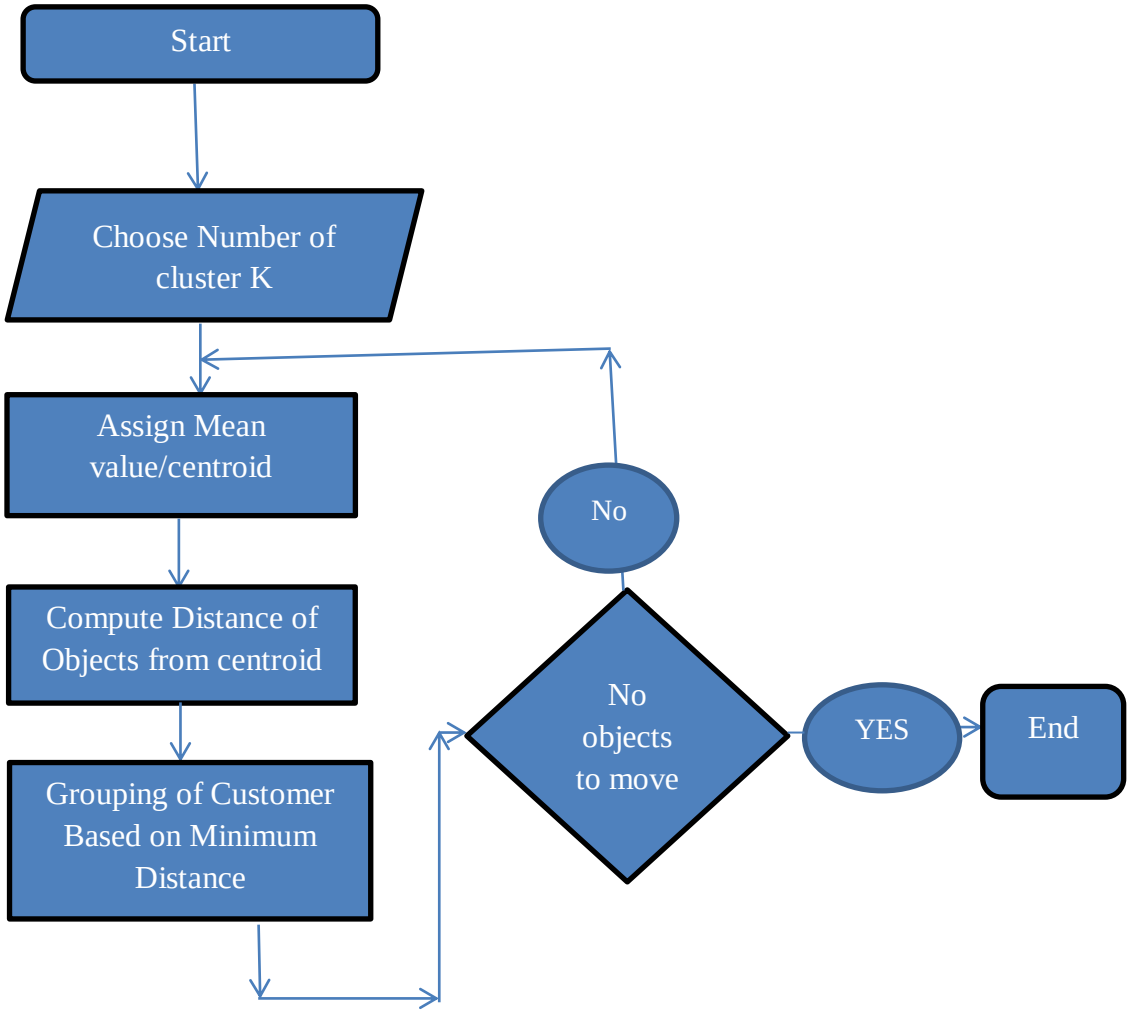


Figure 2.5 flow chart of K-means clustering algorithm [7, 56]

As shown in figure 2.4 k-means clustering algorithm has step by step procedure to cluster a given data set.

Step 1: Choose the value of K

Step 2: Select K objects in a random fashion. Use these as initial sets of K centroids.

Step 3: Assign each object to their closest cluster center according to the Euclidean distance function

Step 4: Calculate the centroid or mean of all objects in each cluster

Step 5: Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-means algorithm describe as follows:

The first step k objects choose the data set for data preprocessing then in the second step K objects are selected randomly. These steps randomly assign k records to be the initial cluster center locations. The third steps assign each object to their closest cluster center according to the Euclidean distance function. The Euclidean distance function is distance connecting two objects determines how the similarity of two elements is calculated and it will influence the shape of the cluster. This distance function [43] is widely used. This distance is given the Pythagoras formula.

By Pythagoras theorem [38] if there are two points' a_1, a_2, a_3 and b_1, b_2, b_3 in three dimensional spaces the correspondence formula is given below in equation 2.5

$$\sqrt{(a1 - b1)^2 + (a2 - b2)^2} \ldots\ldots\ldots\ldots\ldots\ldots 2.5$$

Based on the Pythagoras formula (equation 2.5), the Euclidean distance function [43] is defined in equation 2.6. Using this formula we calculate the distance between each data point and cluster centers the new distance metric as follows:

$$E(x + y) \quad = \sqrt{\sum_{a=1}^{m} (xa - ya)}\, 2 \ldots\ldots\ldots\ldots\ldots\ldots 2.6$$

Where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two input vectors (one typically being from a stored instance, and the other an input vector to be classified) and m is the number of input variables (attributes) in the application. The square root is often not computed in practice, because of the closest instance(s) and will still be the closest, regardless of whether the square root is taken. After computing the distance between the data points and cluster center then the data point is assigned to the cluster center whose distance from the cluster center is a minimum of all the cluster centers.

In the dataset, each data point is allocated to the closest cluster according to the Euclidean distance between every cluster center and every data point. In fourth steps, the data points are assigned recalculate or update the position of Centroids. New cluster center defined by equation 2.7 [44]:

$$vi=(\frac{1}{ci})\sum_{1}^{ci} xi \text{.....2.7}$$

Where, 'ci' denotes the number of data points in i^{th} cluster

The distance between each data point and new obtained cluster centers is recalculated. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

2.4.1.2 Filtered Cluster clustering algorithm

This category of clustering method is for filtering the information or any pattern which is essentially needed. The filtration is carried out based on the keywords that are supplied or some relevant information. The algorithm is based on storing the multidimensional data points in a k-d-tree [45]. A k-d-tree (k-dimensional tree) is a binary tree, which represents a hierarchical subdivision of the point set's bounding box using axis aligned splitting hyper planes [46]. The following algorithm shows how filtered cluster algorithm works by figure 2.8 [46]

```
Filter(kdNode u, CandidateSet Z) {
    C ← u.cell;
    if (u is a leaf) {
        z* ← the closest point in Z to u.point;
        z*.wgtCent ← z*.wgtCent + u.point;
        z*.count ← z*.count + 1;
    }
    else {
        z* ← the closest point in Z to C's midpoint;
        for each (z ∈ Z \ {z*})
            if (z.isFarther(z*, C)) Z ← Z \ {z};
        if (|Z| = 1) {
            z*.wgtCent ← z*.wgtCent + u.wgtCent;
            z*.count ← z*.count + u.count;
        }
        else {
            Filter(u.left, Z);
            Filter(u.right, Z);
        }
    }
}
```

Figure 2.6 Filtered Algorithms [46]

This algorithm is easy to implement, requiring a k-d-tree as the only major data structure. The above algorithm describes in detail [46]:

For each node of the k-d-tree, we maintain a set of candidate centers. This is defined to be a subset of center points that might serve as the nearest neighbor for some point lying within the associated cell. The candidate centers for the root consist of all k centers. We then propagate candidates down the tree as follows: For each node u, let C denote its cell and let Z denote its candidate set.

First, compute the candidate $z^* \in Z$ that is closest to the midpoint of C.

Then, for each of the remaining candidates, $z \in Z \backslash \{z^*\}$, if no part of C is closer to z than it is to z*, we can infer that z is not the nearest center to any data point associated with u and, hence, we can prune, or "filter " z from the list of candidates. If u is associated with a single candidate (which must be z*) then z* is the nearest neighbor of all its data points. We can assign them to z* by adding the associated weighted centroid and counts to z*.

Otherwise, if u, is an internal node, we recourse on its children. If u is a leaf node, we compute the distances from its associated data point to all the candidates in Z and assign the data point to its nearest center.

2.4.1.3 Farthest first Algorithm

The farthest first algorithm is one of clustering algorithm; it is modified of K-means that places the center of each cluster in turn at the point furthermost from the existing cluster center [47]. The first center selects randomly and the second center is greedily select as the points further from the first. The following step used to cluster a given data sets using the farthest first clustering algorithm [48]:

Step 1: Farthest first traversal (D: data set, *k*: integer)

Step 2: Randomly select first center

Step 3: select centers

Step 4: For (i= 2,…,*k*)

Step 5: For (each remaining point) {calculate distance to the current center set;

Step 6: Select the point with maximum distance as new center

Step 7: Apply Euclidean Distance function on each cluster

Step 8: //assign remaining points

Step 9: for (each remaining point)

Step 10: Calculate the distance to each cluster center using Manhattan distance formula

Step 11: put it to the cluster with minimum distance

Step 12: repeat the steps until each cluster remains

The farthest first algorithm has some procedure as a K-means algorithm but, it is a variant of K Means whose objective is to minimize the maximum diameter of any cluster on some set of pints. The first step in the farthest first clustering algorithm is center selected randomly like k means algorithm and the second center is greedily select as the point farthest from the first and each remaining center is determined by greedily selecting the point farthest from the set of the already chosen center.

The algorithm describes below briefly [49]: The algorithm defines initial seeds and then on basis of "K" number of a cluster which we need to know prior and then in farthest first it takes point P_i then chooses next point P_i which is at maximum distance. P_I is centroid and p_1, p_2, ..., p_n are points or objects of dataset belongs to cluster from equation 2.3

$$\min \{\max \text{dist}(P_i,P_1), \max \text{dist}(P_i,P_1),.....\} \quad \text{equation 2.8}$$

The equation, 2.4 calculate the data that is the farthest point from the first point. The above equation 2.3 calculate the Euclidean distance between the two documents di, dj.

$$Euclidean\ Distance\ (di, dj) = \sqrt{\sum_{k=1}^{n}(dik - djk)^2} \quad \text{equation 2.9}$$

The below equation 2.5 calculate the Manhattan distance between the two documents di, dj can be intended as

$$Manhattan\ Distance\ (di, dj = \sum_{k=1}^{n} |dik - djk| \quad \text{equation 2.10}$$

Farthest first actually solves the problem of k-center and it is very efficient for a large set of data. In farthest first algorithm we are not finding mean for calculating centroid, it takes centroid arbitrary and distance of one centroid from the other is maximum figure 2.7 shows cluster assignment using farthest –first [48].

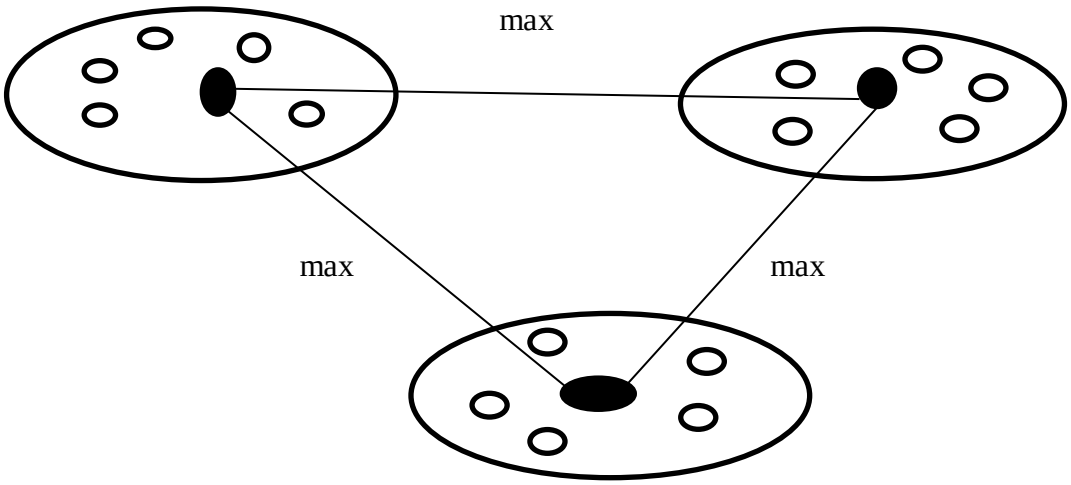


Figure 2.7 Object assignments in cluster [48]

2.4.2 Cluster Interpretation

After the clusters created the result should be interpreted. The three commonly used approaches to understand clusters are [50]:

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

2.4.3 Cluster Result Validity

Cluster validity is a broad and a subject of endless argument since the notion of "good" clustering is strictly related to the domain applications and its specific requirements [51]. Different clusters are obtained in a given dataset and clustering algorithm with different input parameter values. So, there is a need to decide the best clustering that fits the dataset and the business under concern.

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

2.4.4 Cluster Evaluation

The evaluation is important for understanding the quality of the model or technique for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set of models or techniques [52]. The evaluation method used to examine the efficiency and performance of any model. In this study, we perform twelve experiments with k-means, filtered and farthest first algorithms. In order to validate and compare the clustering performance of clustering model done in a way that the attribute value of each cluster in the model is compared to other clustering models with a number of iterations, inter-class similarity error, time to build the model and domain experts judgment

2.5 Application of DM

Data mining has so many applications to help to manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers [53]. It provides a competitive advantage across a broad variety of industries by identifying potential useful information from huge amounts of data collected and stored. An area in which such benefits have to demonstrate includes the bank and financial sector, retail industry, telecommunication industry, biological data analysis, CRM and others. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores (and even how to place them within a store), 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 candidates for development as agents for the treatments of disease.

Application of Data mining in financial data analysis: In the financial sector (bank and financial institution) data mining applications in many different ways. Financial data collected in

the banking and financial industries are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining.

Banks and financial sector can utilize knowledge discovery for various application including [54]

- **Card marketing:** By identifying customer segments, card issuers and acquirers can improve profitability with more effective acquisition and retention programs, targeted product development, and customized pricing.
- **Cardholder pricing and profitability:** Card issuers can take advantage of data mining technology to price their products so as to maximize profit and minimize loss of customers.
- **Fraud detection:** Fraud is enormously costly. By analyzing past transactions that were later determined to be fraudulent, banks can identify patterns.
- **Predictive life-cycle management:** Data mining helps banks predict each customer's lifetime value and to service each segment appropriately (for example, offering special deals and discounts).
- **Detection of money laundering and other financial crimes:** Data mining application helps to detect many laundering and financial crimes such as a large amount of cash flow at certain periods by certain groups of customers.

Application of Data mining in Retail Industry: Through the use of store-branded credit cards and point-of-sale systems, retailers can keep detailed records of every shopping transaction. This enables them to better understand their various customer segments. Some retail applications include [54]:

- **Performing basket analysis:** Also known as affinity analysis, basket analysis reveals which items customers tend to purchase together. This knowledge can improve stocking, store layout strategies, and promotions.
- **Sales forecasting:** Examining time-based patterns helps retailers make stocking decisions. It used to analyze if a customer purchases an item today, when are they most likely to purchase a complementary item.
- **Database marketing:** Retailers can develop profiles of customers with certain behaviors, for example, those who purchase designer labels clothing or those who attend sales. This information can be used to focus on cost-effective promotions.

- **Merchandise planning and allocation:** When retailers add new stores, they can improve merchandise planning and allocation by examining patterns in stores with similar demographic characteristics. Retailers can also use data mining to determine the ideal layout for a specific store.

Application of Data mining in the Telecommunication Industry: The telecommunication industry is rapidly expanding and highly competitive. This creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.

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. Knowledge discovery in telecommunications include the following [54]

- **Call detail record analysis:** Telecommunication companies accumulate detailed call records. By identifying customer segments with similar use patterns, the companies can develop attractive pricing and feature promotions.
- **Customer loyalty:** Some customers repeatedly switch providers, or "churn", to take advantage of attractive incentives by competing companies. The companies can use data mining to identify the characteristics of customers who are likely to remain loyal once they switch, thus enabling the companies to target their spending on customers who will produce the most profit.
- **Fraudulent pattern analysis and the identification of unusual patterns:** Data mining application used in fraud pattern analysis to identify potentially fraudulent users and their atypical usage patterns and to detect attempts to gain fraudulent entry to customer accounts.

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

- **Customer segmentation:** Data mining applicable in customer segmentation used to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis.
- **Manufacturing**: Through choice boards, manufacturers are beginning to customize products for customers; therefore they must be able to predict which features should be bundled to meet customer demand.

- **Warranties:** Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.
- **Frequent flier incentives**: Airlines can identify groups of customers that can be given incentives to fly more.

2.5.1 Application of DM for customer relationship management

Nowadays, new business culture is developing, within it, the economics of customer relationships are changing in fundamental ways, and companies are facing the need to implement new solutions and strategies that address these changes. Data mining tools have a solution for this business issue. Data mining is the process of extracting useful information from huge volumes of data. It finds a useful application in CRM where a large amount of customer data are distributed [55]. Analyzing and understanding customer behaviors and characteristics is the foundation of the development of a competitive CRM strategy, so as to acquire and retain potential customers and maximize customer value. Appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from enormous customer databases, are one of the best supporting tools for making different CRM decisions[54].CRM comprises a set of processes and enabling systems supporting a business strategy to build long term, Profitable relationships with specific customers [56]. It is a very significant technology in every business because all business is customer-centric. Its scope should be the ‘‘personalized’’ handling of customers as distinct entities through the identification and understanding of their differentiated needs, preferences, and behaviors. It consists of identifying, attracting, retaining and developing customers [57]:

Customer identification: This is the beginning of CRM which is also known as customer acquisition. This phase used to identify people who are most likely to become customers or most profitable to the company. Customer identification includes target customer analysis and customer segmentation. Target customer analysis analyzes customer characteristics to seek segments of customers. Customer segmentation is the process of dividing customers into homogeneous groups consisting of customers who are relatively similar within each specific segment then designing and implementing strategies to focus on them.

Customer attraction: This phase follows customer identification. After identifies profitable customer then the organization can direct effort and resources into attracting the target customer segments, which motivates each segment of customers in a different way.

Customer retention: This phase deals with retaining the existing customers and this is the central concern for CRM. Customer satisfaction, which refers to the comparison of customers' expectations with his or her perception of being satisfied, is the essential condition for retaining customers. As such, elements of customer retention include one-to-one marketing, loyalty programs, and complaint management.

One-to-one marketing refers to personalized marketing campaigns that are supported by analyzing, detecting and predicting changes in customer behaviors. Loyalty programs involve supporting activities that aim at maintaining a long term relationship with customers. Specifically churn analysis, credit scoring, service quality or satisfaction forms are part of loyalty programs.

Customer development: This phase deals with maximizing the customer purchase value respectively. This involves consistent expansion of transaction intensity, transaction value and individual customer profitability. Customer development phases help to reveal distinct customer segments, facilitating the development of customized new products and product offerings which better address the specific favorites and priorities of the customers.

2.5.1.1Benefit of CRM

Having customer relationship management has the following benefits [58].

- Share information: With easy and accurate access to a vast array of information, organizations can provide customers with the first-class service, ensuring that the right technician is dispatched with an understanding of the complete history of the customer.
- Increase customer satisfaction: Improve customer service and reduce costs with Web-based tools that enable customers to resolve service issues themselves.
- Make quick, intelligent business decisions: Use standard reports and inquiries to track equipment service details, parts usage, and technician labor. Monitor customer call status,

response times, and technician workload. Analyze customer and equipment call history, service contract profitability, and vital warranty issues.

- Give technicians fast access to maps and directions: Help ensure on-time arrival
- Flexibility: Reverse a contract or credit a customer flat or prorated amount when canceling a contract.
- Provide visibility across the organization to ensure that the right resource is assigned to the right work order.

2.6 Related works

The literature review is done to get more information about the problem and to know which type of works done related to the current study and to know the methods and algorithm other research used.

2.6.1 Foreign Works

Cai Qiuru, et al., [19] explore how to .group telecom customers in the telecom industry to understand consumption characteristics of different customer groups, how to provide an analytical basis of marketing strategies for developing new business, how to stop customer losing and how to achieve the strategic object of profit improvement. To solve this problem the researcher use clustering techniques to segment the customer and to identify the target customer and characterize each group and analyze their properties. Then the researcher uses a K-means clustering algorithm to segment customers in two to groups. The first group is value segmentation and the second group is behavioral segmentation. To meet the objective of the research the researcher applie k-means to group the customers into eight groups based on customer behavior, highly valuable, medium valuable and low value, finally, concluded that customer segmentation increases the effectiveness of resource spending and brings telecom services closer to customers.

The limitation of the study

- This study only considered telecom customer in the telecom industry to understand consumption characteristics of different customer groups. That means, this research did not include the customer geographical information.
- The researcher did not use or clearly stated which data mining process model is used. Such as Crisp-DM, Hybrid, KDD and SEMMA.

According to the investigation made by Rawan and Edward [59], apply data mining techniques to segment mobile network user to target different sectors of customers with efficient marketing strategies and ensure customers retention in light of the intense. In this study, the real data collected 6315 customer mobile telecommunication operation in Jordan. The aim of this study to a case study on customer segmentation in mobile telecommunication markets to meet the objective of the research use self –organizing maps, to detect different usage patterns of mobile users. Then the customers used K-means algorithm to obtain segments with clearly defined borders to help better analyzes common properties underlying these segments. Finally, the researcher concludes that different behavioral segments in this market and highlights the role of data users in modern mobile markets.

The limitation of the study

- The researcher didn't use or clearly stated which data mining process model is used such as CRISP-DM, Hybrid, KDD and SEMMA
- In this study the experiment was conducted with a less rich dataset, the total dataset used in this experiment was only 6315

Aror and Vohra [60] investigate how to improve the revenue of mobile customers and how to categorize the mobile user into the network. To solve this problem the researcher tried customer segmentation using data mining techniques. K-means clustering algorithm used to identify the high consuming customer and low consuming customer and the researcher use call detail records of customer data. By calculating the total score of every call detail records than categorizes customers into premium customer and non-premium customers. After getting the total score then the researcher identifies the revenue maximization cluster. Finally, they conclude that the

clustering algorithm is the best for customer segmentation for the mobile user and also used to identify profitable customers to the telecommunication industry.

The limitation of the study

- The researcher did not use or clearly stated which data mining process model is used. Such as Crisp-DM, Hybrid, KDD and SEMMA.
- The study only considered the customers mobile users. This means the study not include other services user customer in the company.

Begunca, [61] attempted the differences between the behavior of consumers in the market segmentation based on benefits sought/required, to identify behavior features/ characteristics based on benefit sought approach, to identify personal features based on benefit sought and providing a guide for the development of effective segmentation strategies. To do this the researcher used both qualitative and quantitative approaches. Quantitative approach allows statistical data analysis to identify relevant information and k - means cluster analysis is used to identify relatively homogeneous groups of cases based on selected characteristics for this study customer segment into five clusters. Finally, he noted the need for the consumer segment according to the benefit required.

2.6.2 Local Works

Biazen [62] explored how to design a model using data mining techniques for customer segmentation that helps Ethiopian revenue and customs authority organization to treat customers according to their behavior and how to increase the revenue of the organization. The researcher uses quantitative methods to collect and analyze customer's data and qualitative methods to understand the business operation. The researcher used KDD (Knowledge Discovery in Database) process models to produce useful patterns or models. In this study, clustering techniques were used to cluster the subset of the dataset into five groups. Further classification techniques are applied to predict future customer's behaviors. For classification J48 decision tree algorithm and multilayer-perception algorithms have experimented. After conducting various experiments the researcher conclude the 10-fold cross-validation has better classification accuracy than percentage split experimentation. Finally, the researcher concludes that the J48

tree algorithm is the best algorithm to classify the customer value and the clustering algorithm is used for customer segmentation and also used to identify the characteristics of the customers.

Belachew [63] explored the likelihood of applying data mining techniques for predicting profitable customers and segmenting the existing customer to identify best customer segments from bad customer segments. The researcher focused on building a model that helps to classify customers for product and services offering by an organization using a predictive data mining model. To classify customers in the organization has no predefined classes. The researcher uses clustering techniques that result in an appropriate number of clusters. In this study, the data collected from BG MFI (Buusaa Gonofa Microfinance Institution). The study was done on the data set of BG MFI (Buusaa Gonofa Microfinance Institution) customer data. In this work, Weka version 3.7.5 is selected to implement data mining. The methodology of this research consists of four stages such as data collection, data preprocessing, data mining (classification and clustering) and interpretation. In this study, clustering techniques (k-means algorithm) is used to cluster the subset of the data set into five groups. Based on the result obtained through clustering, decision TreeJ48algorithm is employed. Using a J48 decision tree algorithm the researcher conducted a different experiment. Model build by 10-fold cross-validation test mode performs better accuracy. In general, this study helps to identify the potential customer for an institution for better customer relationship management.

Finally, during the review, we noted that most of the international research works focus on customer segmentation, prediction, and telecom fraud detection. However,limited works available in our country to apply data mining for customer segmentation. The output of this research delivers new knowledge, new understanding, and new information to the customer segmentation and related problems. Hence, it helps to improve customer satisfaction, improve company profitability. In addition to this, the output of this research builds a descriptive model to identify the characteristics of enterprise customers or organizations.

2.7 Customer relationship Management

Customer Relationship Management (CRM) has become a leading business strategy in highly competitive business environment. CRM (Customer Relationship Management) is an important business approach that focuses on marketing to each customer individual rather than marketing to a mass of people or firms. It is a customer-centric approach doing business rather than a simple marketing strategy. Its scope should be the ‘‘personalized’’ handling of customers as distinct entities through the identification and understanding of their differentiated needs, preferences, and behaviors. CRM involves all of the corporate functions like marketing, manufacturing, customer services, field sales, and field service who are required to contact customers directly or indirectly.

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

Nowadays, CRM is a business strategy that goes beyond increasing transaction volume. It is primarily a strategic business and process issue rather than a technical issue.

CRM is defined as the business strategy, process, culture and technology that enable organizations to optimize revenue and increase value through a more complete understanding and fulfillment of customer needs.

CRM also defined as “the core business strategy that integrates internal processes and functions, and external networks, to create and deliver value to targeted customers at a profit. It is grounded on high quality customer-related data and enabled by information technology” [64].

On the other hand CRM can be defined as it is a comprehensive approach which provides seamless integration of every area of business that touches the customer namely marketing, sales, customer service and field support through the integration of people, process and technology, taking advantage of the revolutionary impact of the internet.

CRM is also viewed as the strategy of transforming enterprises to become customer centric while still expanding revenue and profit is one of the main strategies in business today. Many business enterprises, therefore, realize the importance of CRM and the potential of these techniques to

achieve and sustain a competitive advantage. To realize CRM success, business and IT executives should implement CRM processes and technologies and promote employee behavior that supports coordinated and more effective customer interactions throughout all customer channels .Therefore, one of the most important processes of CRM is extracting valid, previously unknown, and comprehensible information from a large database and using it for profit.

CRM is illustrated as a combination of people, processes, and technology which provide understanding of customer needs, business strategy support, and build long-term relationships with customers. As a result, companies make every effort to deliver the highest value to customers through better communication, customized promotions, faster delivery, and personalized products and services. To effectively address human behavioral elements appropriate business processes and organizational culture are also required for successful utilization of the integrated technology.

Having CRM has the following benefit [58]:

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

2.7.1 Principles and Tasks of CRM

According to Gray and Byun [65], the overall processes and applications of CRM are based on the following basic principles.

Personalization (treat customer individually):

Personalization deals about treating customers individually so that the content and services to customer should be designed based on customer preferences and behavior.

Acquire and retain customer loyalty through personal relationship:

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

Select customer:

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

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

Discussed about the following principles used for building strong customer relationships:

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

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

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

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

The CRM approach is customer-centric. This approach focuses on the long-term relationship with the customers by providing the customer benefits and values from the customer‟s point of view rather than based on what the company wants to sell.

2.7.1.1 CRM TASKS

According to Gray and Byun [65] the four basic tasks that are required to achieve the basic goals of CRM are:

Customer identification: It refers to the selection and or knowing of customers through marketing channels, transactions, and interactions over time for the purpose of serving or providing value to the customer .

Customer differentiation: This step refers to segmenting customers into different perspective from the company's point of view; because, from the company's point of view each customer has their own lifetime value.

Customer interaction: The main purpose of this step is to analyze the customer's behavior over a long period of time in order to provide the right goods and services at the right time because customer demands are sensitive and it change over time. From a CRM perspective, the customer's long-term profitability and relationship to the company is important. Therefore, the company needs to learn about the customer continually.

Customization / Personalization: "Treat each customer uniquely" is the motto of the entire CRM process. Through the personalization process, the company can increase customer loyalty. The automation of personalization is being made feasible by information technologies.

2.8 customer segmentation

The traditional ways of customer segmentation are mainly categorizing methods based on experiences, statistics or simple partitioning .These segmentation methods segment customer according to simple behavior character or attribute character such as the product category purchased or the region resided in. The new methods of customer segmentation are based on data mining. It is the best solution for extracting meaningful data and information from databases which have a huge amount of data. Segmentation is the process of dividing the customer base into distinct and internally homogeneous groups in order to develop differentiated marketing strategies according to their characteristics.

Customer segmentation is defined as the practice of classifying customer base into distinct groups. In other words, customer segmentation is also described as the process of dividing customers into homogeneous groups on the basis of shared or common attributes [66]. The goal of segmentation is to know the customer better and to apply that knowledge to increase profitability, reduce operational cost, and enhance customer service. Segmentation can provide a multidimensional view of the customer for better treatment strategy.

2.8.1 Applications of Customer Segmentation

According to Balaji and Srivatsa [67], customer segmentation has different applications for a certain organization which is described as follows.

Guiding product development and research:

A comprehensive segmentation solution has been emphasized the fact that individuals have different product needs and usage patterns. Customers in one segment may use a company's full portfolio of products quite frequently, while customers in another segment may only have a need for a single product that is used at irregular intervals (infrequently).Segments that contain less active customers often exposes opportunities to strengthen and broaden these customer relationships by introducing new or re-packaged products that meet a specific customer need. Segmentation provides the means to target research and product development activities with the goal of further stimulating customer demand.

Tailoring marketing programs:

It is true that successful data-driven marketers understand how to communicate with their customers at the right time, right place and with the right message. Distinctive customer preferences and needs represent unique opportunities and challenges that can be pursued by introducing tailored or modified programs for each segment. The make-up of each customer segment and their past and projected behaviors and needs should guide the tailored use of key marketing levers to 44 maximize program effectiveness. Segmentation becomes the focus, supporting program development and ongoing test and learns activities.

Managing customer relationships:

Segmentation also provides an excellent framework to manage the varied needs of customers. Customized customer management and development strategies can be developed for each unique segment. The development plans should include a set of objectives, goals and performance

metrics that are derived from the unique opportunities and challenges present within each customer group. The segment-level plans function as a strategic roadmap, supporting business growth and attempting to maximize the potential of each customer relationship.

Making customer investment decisions:

Segmentation provides a framework to help identify the optimal customer investment strategy for each unique segment. For some segments, the investment may be directed towards further developing customer relationships, while for other segments the investment is made to introduce new products and services that address unsatisfactory customer needs. Ultimately, the key factor driving customer investment decisions has been the expected return on that investment. Segmentation not only helps to determine how much to invest in a customer segment, but how to spend it.

Chapter Three

Problem Understanding and Data Preparation

In this Chapter, we discuss the understanding of the problem, understanding of the data and preparation of the data of the raw data to create suitable data set for data analysis.

3.1Understanding of the problem

This is the step where the researcher attempts to understand the research objectives and requirements from a business perspective and set the data mining goals. Based on the information from ET citizen's charter documents, ET has three missions regarding customers [2]. The first one is to connect Ethiopia through state-of-the-art telecom services by providing high quality, innovative and affordable telecom products and services that enhance the development of our nation and ensure high customer satisfaction. The second one is to build a reputable brand known for its customers' consideration. Finally, the third one is to build its managerial capability and manpower talent that enables ET to operate at an international level. However, to satisfy the customer there is a need to have a clear view of their customers. As a matter of fact, they could not implement proper strategies and actions to gain more advantages in the market. During this research, we study the domain and how domain experts analyze the enterprise customer data.

From a business understanding perspective, currently for analysis of data used simple statistic methods. Using these traditional methods the company waste so much resources with no profit because they work with their entire customers in the same market and also it is difficult to know what is the unique features and need of customers and also difficult to make appropriate predefined strategies for different customer groups. Using traditional methods it is difficult to segment enterprise customers. To solve the existing customer handling problem customer segmentation schemes are the main subject for this company. Four types customer segmentation schema geographic segmentation targets customers based on a predefined geographic border. Differences in interests, values, and preferences vary dramatically throughout cities, states, and countries, so it is important for marketers to recognize these differences and advertise accordingly. Demographic segmentation divides a market through variables such as age, gender,

education level, family size, occupation, income, and more. This form of segmentation is a widely used strategy due to specific products catering to obvious individual needs relating to at least one demographic element. Psychographic segmentation unlike geographic segmentation and demographic segmentation, psychographic segmentation focuses on the intrinsic traits your target customer possesses. Psychographic traits can range from values, personalities, interests, attitudes, conscious and subconscious motivators, lifestyles, and opinions. Behavioral segmentation has similar measurements to psychographic segmentation but focuses on specific reactions and the way customers go through their decision making and buying processes. Attitudes towards your brand, the way they use it, and their knowledge base are all examples of behavioral segmentation. The three segmentation schema consider in this research are customer value segmentation, customer behavior segmentation and geographic segmentation schema. The first segmentation value segmentation focuses on identifying the contribution that a customer makes to overall organizational profitability based on the current relationship with the organization. The second segmentation behavioral segmentation schema is concerning with grouping customers based on their behaviors, such as purchase or services. The third segmentation geographical segmentation provides insight into where a given brand's customers are located and specific location driven behaviors or preferences. There are different types of customers in the ET and it is very interesting to understand the customer characteristics, customer value and segment them in order to make a good relationship with them. ET has criteria's that used to understand customer value and effectiveness to the company. These criteria's are based on the investment capital of customers, the number of permanent employment in the company and the company's branches office. The table 3.1 shows the criteria that are used to segment the enterprise customers.

Customers	Investment capital	Number of permanent Employees	Number of branch office	Value of customer to ET
High value customers	>=5 and 10 million ETB	>= 50 employee	>= two and three branch office with head office and three office	Generates high revenue, high amount of active ET services, Customers who purchase various ET services
Low value customers	<=5 and 10 million ETB	<= 50 employee	<= two and three branch office with head office and three office	Generate low revenue, customers who purchase fewer amount of ET service, low amount of active ET services

Table 3.1 Summary of the existing customer segmentation criteria [68]

The objective of this research is a customer segmentation model to identify the high value of enterprise customers and to identify the behavior of every segmentation group. This is to handle and treat customers accordingly. To accomplish the objective of this research we selected a relevant attribute from the total data collected based on domain expert.

Based on the existing problem domain, the following attributes are considered to be useful for segmenting enterprise customers. zone, customer level name, customer category, net business offer name, segment name, investment capital, number of employees, no of a branch, voice, data _ up, data_ down and SMS (Short Message Services). Among these attributes zone, customer category, net business, offer name, segment name, investment capital, a number of employees and number of branch identifies customer basic customer information, customer behavior and geographical information.

Basic customer information describing customers by their attributes and provides a basis for marketers to communicate with existing customers in order to offer those better services.

Having those attributes in this research concerning with grouping customers based on their behaviors, such as purchase or services, can identifies where is profitable enterprise customers located, identify mostly purchasing products within each segment, and also identify customer internet access within each segment.

Voice, data _ up, data _ down, customer level name and SMS attributes identifies customer value. Customer value identifies the contribution that a customer's makes to overall organization profitability based on current relations with the organization. Those attributes identify how much they use a product.

Zone: Identifies the places where the customers are located, identifies in which zone the profitable customer where live or work, identifies the purchasing behavior of the customer in this zone. They are central Addis Ababa zone (CAAZ), east Addis Ababa zone (EAAZ), north Addis Ababa zone (NAAZ), south Addis Ababa zone (SAAZ), west Addis Aba01ba zone (WAAZ).

Customer level: Identifies the vanity number used by enterprise customers and this number identifies whether customers platinum customers or gold customers. The classification has been done based on how to generate revenue for ET. Currently, in ET there are some customers categorize in platinum and gold category. Under platinum category customer monthly paid for the services are high amount because of this the customers under this group generate high revenue to ET. Although, customer under gold customer level name monthly paid fewer mount than platinum because of this the customer under this group generates low revenue to ET. As a result, these attributes in this research identifies in which group of customers generate high revenue to ET company.

Customer category: Identifies whether the customer's category under Key account customers or SOHO customers. These attribute used in theses research to identify the customer group. If the customer under group key account generate generates high revenue to ET, where as if the customer under SOHO generate less revenue to ET.

Net business: Identifies the services or product type customers are purchasing. Currently, there are different services and product type provided by ET companies to users. These services are CDMA, E1 ISDN, fixed line voice, GSM, LTE, short number, WCDMA. This attribute identifies the customer preferences of customer products. This means which group of customers mostly used products.

Offer Name: Identifies a technology that a customer used for internet access. These attributes describe in detail customer internet access whether it is wired or wireless network. If the customer used a wired network any physical medium consisting of cables and also if the customer used wireless network medium made of electromagnetic waves.

Segment Name: Identifies the payment mechanism for all enterprise customers. Currently, there are two payment mechanisms that are prepaid and postpaid. This attribute identifies the purchasing characteristics of enterprise customers. There are some customers who are prepaid customers and these customers recharge their account before using the services. Postpaid customers are paid after using the services at the end of the month or at the beginning of the month.

Investment Capital: Identifies the customers or organization investment capital. Using these attributes we can identifies what customers made for organization profitability and also company knows what the significance of customer to ET organization.

Numbers of Employees: Identifies how many employees are there in the enterprise customers. Using these attributes we can identify company contribution to the ET Company and also identifies what customer made for organization profitability.

Number of Branch Office: Identifies how many branches the customers has. These attributes identify the importance of customers within origination and also identifies what customer made for organization profitability.

Voice usage: Identifies customer's voice usage in minutes. Voice usage attributes have duration of generated calls (in minutes) during peak hour and off-peak hours. Peak hour and off-peak hours describe a time period with fewer calls that are handled in a busy period. Peak hour refers to the total number of customer calls during the busy hour. This means during peak hour large

volume of subscriber traffic is handled by the platform. Off-peak hours describe a time period with fewer calls than handled in a busy period. Voice usage attributes cover the monthly usage behavior of one customer. This attribute identifies how much voice used by customers both in peak and off-peak hours in each month and also identifies customer behavior in each month.

Data Uploading: Identifies customer's data uploading usage in megabits. Data uploading means data is being sent from customer computers to the internet. These attributes have data usage in megabits during peak hour and off-peak hours. During peak hour large volume of subscriber upload data to the internet. Off-peak hour fewer volume data upload than handled busy period. These attributes identify how much data uploading by customers both in peak hour and off-peak hour each month and identifies customer behavior in each month.

Data Downloading: Identifies customers' data downloading usage in megabits. Data downloading means the process of getting web pages, images, and files from a web server. These attributes also have data usage in megabits during peak hours and off-peak hours. During peak hours large volumes of data customers download data from a web server. During off-peak hour fewer volume data download by the customer. These attributes identify how much data download by customers both in peak hour and off-peak hour each month and identifies customer behavior in each month.

SMS: Identifies the total number of SMS usage sent and receives customers. SMS attributes contain a number of SMS sent and received by a customer during peak hours and off-peak hours. During peak hours a large number of SMS sent and receives by customers. During off-peak hour fewer numbers of messages sent and received by the customer. These attributes identify how many messages sent and received by customers both in peak hour and off-peak hours each month and identifies customer behavior in each month.

3.2 Data Understanding

The data understanding phase starts with an initial data collection, describe data, explore data, verify data with activities in order to get familiar with the data, to identify data quality problem, to discover first insight into the data or to detect interesting subset to form a hypothesis for hidden information [24].

3.2.1 Data Collection

The telecommunication industry generates a huge amount of data during giving telecom services. This data includes customer data, call detail data and network data. The customer data includes customer information and customer location information. Call detail data is generated when every user makes a call and each call have detail information stored in the database and the network data is generated from all network elements about network element status, call setup information. These huge amounts of data handled properly for different purposes like customer segmentation, customer churn prediction, network performance analysis, fraud detection [69].

The major source of the data for this research work is the enterprise customer database of the ET. Through the help of an expert, data regarding enterprise customers is extracted from the enterprise customer database. The enterprise database is huge and contains data from different sources because every enterprise shop activists are stored in this database and also contain customer call detail information. For instance for every customer name, location of customers, type of product purchase, the total amount of customer voice usage, SMS usage, data usage found in this database and also service number for every customer's product type found in this database. The data is collected four-month data taken from January of 2019 data up to 2019 April month of data and the customer’s data around 162315 records. The data extracted from the enterprise customer’s database is huge and using this data to extract knowledge is time-consuming. Therefore, we select for this research work 21126 records out of 162315 and 13 attributes are selected. The attributes ignored from this research work are date, account code, services number, customer name, enterprise attributes, and customer status.

3.2.2 Description of attribute

The description of the data is very significant in order to clearly understand the data. The data sources of this research are the enterprise customer profile database and this database contains information about customers with different attributes. The database handles customer offering date, customer name, customer zone name, customer collection center, customer city name, customer level name, customer type, category name, customer subcategory name, network business type name, offering name, segmentation name, status customer offering product, customer voice usage, customer SMS usage, data usage includes both data uploading and data downloading. The descriptions of this data source with their attribute are described in the tables3.2.

NO	ATTRIBUTE NAME	DESCRIPTION	DATA TYPE	MISSING VALUE
1	ZONE NAME	Area of the customer's where it is found	Numeric	5%
2	CUST LEVEL NAME	Describe the level of customers. They are platinum and gold customer.	Nominal	0%
3	CATEGORY NAME	Whether the customer key account customer or SOHO customers	Nominal	0%
4	NET BUSI TYPE NAME	Describe a service type or product type. The product and services types are CDMA, E1-ISDN, fixed line voice, GSM, LTE, short number and WCDMA	Numeric	10%
5	OFFER NAME	Specifically, describe a technology a customer used for internet access.	Nominal	0%

6	SGMT NAME	Customer segmentation level prepaid and postpaid	Nominal	0%
7	INVESTMENT CAPITAL	Customer or organization asset. Describe customer high, low	Nominal	0%
8	NUMBER OF EMPLOYEES	No of employees in the organization. Describe customer high, low	Nominal	0%
9	NUMBER OF BRANCH	No of the branch the organization has. Describe customer high, low	Nominal	0%
10	VOICE USAGE	Describe voice usage in minutes. The voice usage describe in high, average, low	Nominal	0%
11	DATA UPLOADING	Describe data uploading usage in megabits. The data uploading describe in high, low	Nominal	0%
12	DATA DOWNLOADING	Describe data downloading usage in megabits. The data downloading describe in high, low	Nominal	0%
13	SMS	Describe a number of SMS usages. The number SMS describe in high, low	Nominal	0%

Table 3.2 Describe the selected attribute

3.3 Data preparation

Data quality is one of the biggest challenges for data mining and it mentions the accuracy and completeness of the data. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data [70]. The data preparation phase covers all activities required to construct the final dataset from the initial raw data [71]. Nowadays a large number of data stored in a database and these data are noisy with missing and inconsistent data. In the data preparation phase, it improves data quality. The purpose of data preprocessing is to clean select data for better quality. Data quality is multifaceted issue that represents one of the biggest challenges for data mining. It refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of data being analyzed. The presence of duplicated records, lack of data standards, the timelines of updates human error can significantly impact of the effectiveness of the complex data mining techniques which are sensitive to understand difference that may exist in the data. To improve data quality, it is sometimes necessary to clean the data, which can involves the removal of duplicated records, normalizing the values used to represents information in the database. Select data may be different formats, and then order to use the data needs to convert in to suitable formats.

3.3.1 Data Cleaning

Data cleaning involves different techniques in order to fill in missing values, to handle outliers and to smooth noisy data and to detect inconsistencies and correct the data set [72]. Data cleaning is removing records that had incomplete, missing, duplicated, inconsistent data and irrelevant data under each attribute column. There are different methods used to handle the missing values, such as ignoring the tuples, filling the missing values by using the modal value (for nominal and ordinal variables) and the mean (for continuous variable) [24].In the data extracted from the enterprise customer database used in this research, there are no missing values or noisy data to clean.

3.3.1.1 Handling Missing Values

Missing data is the most common problem that comes up during the data analysis process. Missing values lead to the difficulty of extracting useful information from that data set. Solving

the problem of missing data is of a high Priority in the field of data mining and knowledge discovery. Handling missing values by appropriate methods does not affect the quality of the data. In this thesis the two widely used methods are applied. One is avoiding the missing data and other is data imputation. Avoiding the missing data is not time consuming and same time it is very easy to follow. But there are many drawbacks associated with this method. Deleting records may result in losing some information. If the sample data size is large avoiding some records or attributes may not affect the results, but still we need to keep in mind we are losing something. The absence of information is rarely beneficial. All things being equal, more data is almost always better. Therefore, the researcher considered carefully about how to handle the thorny issue of missing data.

A common method of handling missing values is simply to omit from the analysis the records or fields with missing values. However, this may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting records with missing values would lead to a biased subset of the field value is missing. Replace the missing value with the field mean (for numerical variables) or the mode (for categorical variables) [73]. Therefore, in this research study the investigator tried to handle the missing values by replacing missing value with the field mean, since they are numerical attributes. Table 3.3 summarizes attributes and percentage (%) of missing values associated with each other.

Attribute Name	No. of missing values	% of missing values	Mean value of missing values
Zone name	1000	5%	4
Net business type	2000	10%	6

Table 3.3 Missing values and their percentage

As a result, the missing values of the dataset were handled in accordance with the above Suggestion. The missing value of zone name and net business type attributes were filled by their mean values since they are numeric value type.

3.3.1.2 Data discretization

Discretization transforms numeric (continuous) attributes to nominal (categorical or discrete) attributes. The range of a numeric attribute is divided into intervals and each interval is given a label. Attribute values are replaced by the labels of the intervals into which they fall. Using discretization method can give generalized information which is easier and meaningful to interpret data mining results conducted on different data mining tasks.

Among thirteen attributes voice usage, investment capital, number of employees, number of the branch office, data uploading, data downloading and number of SMS has numeric data types. Therefore, together with the domain expert, we set the threshold value for these attributes. The reason to determine the range of numeric attributes values is the only way to interpret the clustering model output from the business perspective and to understand the behavior and value of customers in each cluster segments. So, the threshold value for the voice usage, investment capital, the number of employees, the number of a branch office, data uploading, data downloading, the number of SMS determine in table 3.4 and numeric values of zone name and net business value determine in table 3.5.

List of Attributes	Threshold Values		
	Low	Average	High
VOICE_USAGE	<15000	15000-25000	>25000
Threshold Values			
	Low		High
INVESTMENT_CAPITAL	90,000 up to 10,000,000		11,000,000 up to 80,000,000
NUMBER_OF_EMPLOYESS	2 up to 50		51 up to 2500
NUMBER_OF_BRANCH_OFFICE	1 up to 4		5 up to 250
DATA_UPLOADING	3066 up to 49997386		49997386 up to 99991707
DATA DOWNLOADING	260385 up to		300084832 up to 599909279
NUMBER_SMS	3000 up to 6000		3000 up to 6000

Table 3.4 List of range of condition by which a cluster result is measured

Attribute value from categorical to numeric value

Attribute of the data set with their numeric description							
Attribute description	Numeric Value						
	1	2	3	4	5	6	7
Zone Name	CAAZ	EAAZ	Enterprise TPO building	NAAZ	SAAZ	SWAAZ	WAAZ
Attribute Description	Numeric Value						
	1	2	3	4	5		
Net Business type	CDMA	Fixed line voice	GSM	LTE	WCDMA		

Table 3.5 Attribute of data set with their numeric description

Attribute of the data set with their description

Attribute of the data set with their description	
Attribute description	Abbreviated attribute
Zone Name	ZN
Customer level name	CLN
Category Name	CN
Net Business Type	NBT
Offer Name	ON
Segment Name	SN
Investment Capital	IC
Number of Employees	NE
Number of Branch	NB
Voice usage	VU
Data uploading usage	DU
Data downloading usage	DW
SMS usage	SU

Table 3.5 Attribute of data set with their description

3.3.2 Data Transformation

Attributes, investment capital, number of employees and number of branches office instance, voice usage, data uploading, data downloading and SMS transformed from nominal to categorical with the help of experts. Zone name and net business type attributes are transformed from categorical value to numeric value.

The pre-transformation the categorical values of attributes zone name CAAZ (central Addis Ababa zone), EAAZ (East Addis Ababa Zone), SAAZ **(**South Addis Ababa Zone), SWAAZ **(**South west Addis Ababa zone)**,** Enterprise (TPO building) WAAZ (West Addis Ababa Zone).

The pre-transformation the categorical values of attributes net business type CDMA (Code division multiple access), fixed line voices, GSM (Global System for Mobile communication), LTE (Long-Term Evolution) and WCDMA (Wideband Code Division Multiple Access).

The pre-transformation the numeric values of the attributes voice usages in minute's instance categories 1 up to 43200 minutes. The new name was given to this instance categorize into three categories, namely as low customers from 1up to 15000 minutes of voice, average customers from 15000 up to 25000 minutes of voice and high customers from 25000 up to 43200 minutes of voice.

The pre-transformation the numeric values of the attributes investment capital instance categorize 90,000 up to 80,000,000 million Ethiopian Birr. The new name given to this instance categorizes in two categories, namely as low-value customers and high-value customers. The low value customers categorize from 90,000 up to 10,000,000 million Ethiopia Birr and high value customers categorizes from 11,000,000 up to 80,000,000 million Ethiopian Birr.

The pre-transformation numeric values of the attributes number of employees instance categorize 2 up to 2500 employees. The new name given to this instance categorizes in two categories, namely as low value, and high value. The low value categorizes from 2 up to 50 numbers of employees within companies and high value categorizes from 51 up to 2500 number of employees within the company.

The pre-transformation numeric values of attributes number of branch office instance categorize 1 up to 205 branch office. The new name given to this instance categorizes in two categories,

namely as low value, and high value. The low value categorizes from 1 up to 4 number of branches the organization has and high value categorizes from 5 up to 250 number of branches the organization has.

The pre-transformation numeric values of attributes data uploading instance categorize 3066 up to 99991707 megabits. The new name given to this instance categorizes in two categories, namely as low value, and high value. The low value categorizes from 3066 up to 49997386 megabits have and high value categorizes from 49997386 up to 99991707 megabits have.

The pre-transformation numeric values of attributes data downloading instance categorize 260385 up to 599909279 megabits. The new name given to this instance categorizes in two categories, namely as low value, and high value. The low value categorizes from 260385 up to 300084832 megabits have and high value categorizes from 300084832 up to 599909279 megabits have.

The pre-transformation numeric values of attributes number of SMS instance categorize 0 up to 6000 SMS. The new name given to this instance categorizes in two categories, namely as low value, and high value. The low value categorizes from 0 up to 3000 number of branches the organization has and high value categorizes from 3000 up to 6000 number of SMS organization have.

3.3.3 Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient to produce the same or almost the same analytical results.

From different data reduction mechanisms we use attribute subset selection to reduce the data set size. Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions). The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has

an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

- ❖ Remove attributes which have no information.
    - Attribute contain date, account code, services number, customer name and enterprise are removed. Those attributes have no connection with customer segmentation and identification of characteristics of customers.
- ❖ Remove attribute which have the same meaning to other attributes.
    - Attribute contain collection center, city name, subcategory name and amount paid are removed. Because those attributes describe the same description with other selected attributes.

Due to theses reason we select thirteen attribute for this research. These attributes are zone, customer level, customer category, net business, offer name, segment name, investment capital, and number of employees, number of branch office, voice usage, data uploading, data downloading and SMS usage attribute.

3.3.4 Data formatting

Data formatting is the activity of changing the data format into a format suitable or understandable by the data mining tool. The datasets should be transformed into a format that is acceptable by WEKA tools. ARFF (Attribute-Relation File Format) is a format used by the WEKA tool. The data set is first saved into a comma-separated format (see figure 3.1) and then saved in ARFF.

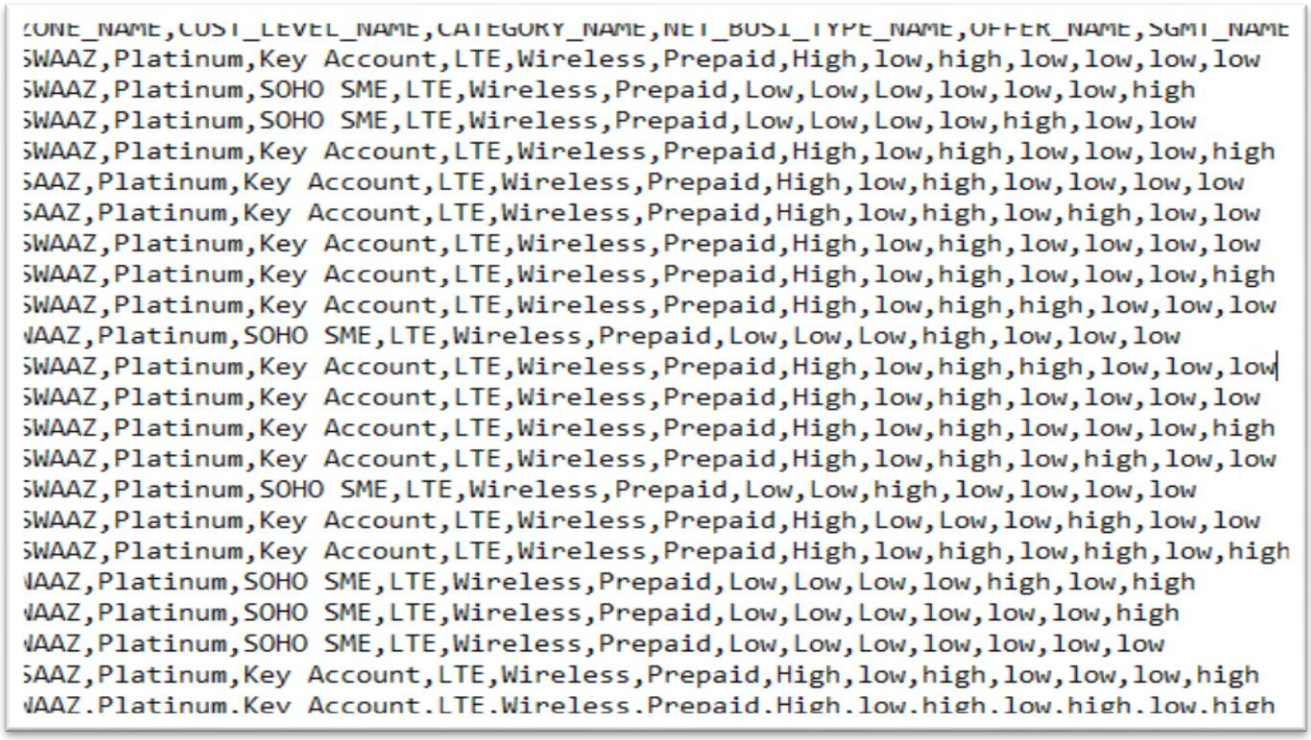


Figure 3.1 Sample data in CSV format

CHAPTER FOUR

Experimental Results and Discussion

This chapter discusses the experiment conducted during the research. As discussed in previous chapters the domain expert in the enterprise section in ET has been doing the analysis using statistical methods on the data which is extracted from the enterprise section database. Using this statistical method the domain expert unable to use the whole data and also unable to reach good analysis results. To solve this problem we use data mining techniques. Using these data mining techniques we discover hidden knowledge from an enterprise customer database which is used to solve customer segmentation problem. For the clustering analysis for this research 13 attributes are selected and the experiments are done using the WEKA tool.

4.1 WEKA Tool

WEKA software is a collection of open sources machine learning algorithm used for preprocessing, classification; clustering and association rule is cover [74]. It is JAVA based tools used in the field of data mining and it uses a text file to describe the data. It can work with a wide variety of data files, the data files must be in "Arff" and "CSV" file format. The figure 4.1 shows the graphical user interface of the users.

Figure 4.1 front views of WEKA tool

Figure 4.2 shows the WEKA explorer window tool. This window shows the name of selected attributes on the explorer window with their instances. The WEKA explorer shows the different tasks performed, such as preprocessing, classification, clustering, association, attribute selection, and visualization.

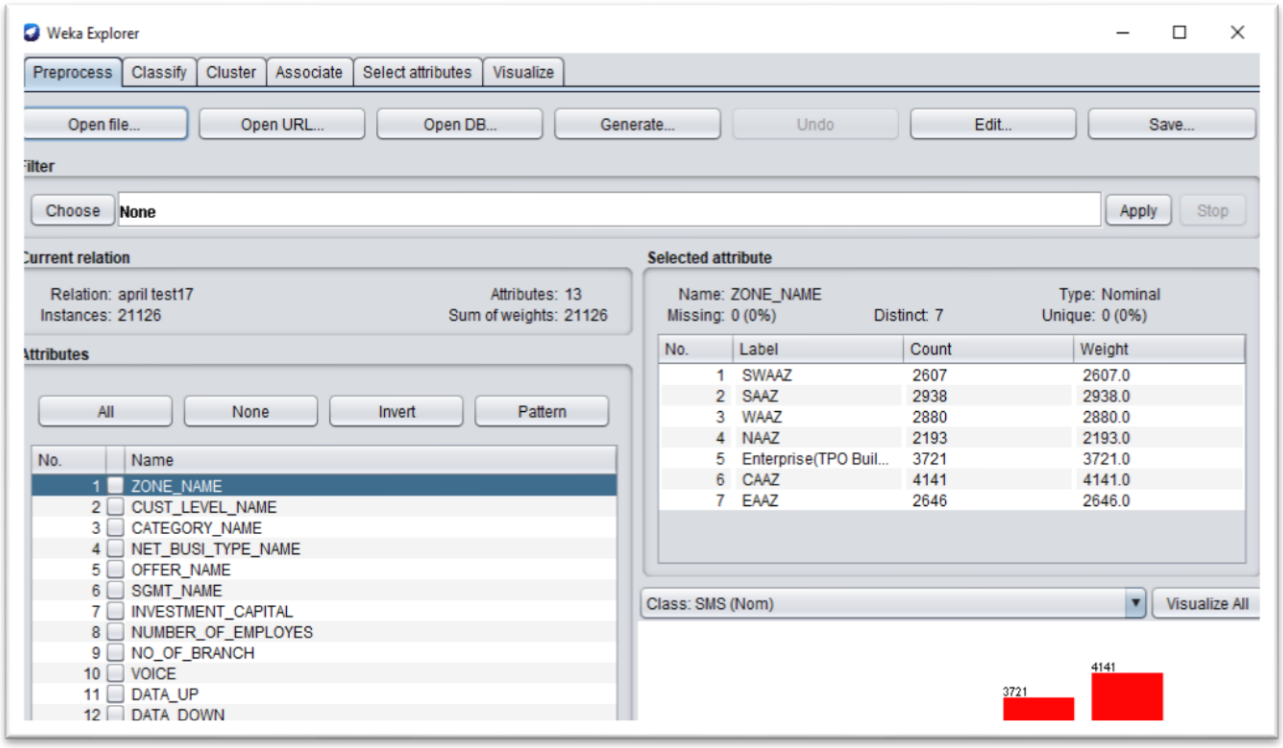


Figure 4.2 Screen Shot of WEKA Explorer which is the initial property of the data

In this study, WEKA is used to conduct different experiments for selecting an optimal clustering model towards designing a customer segmentation prototype.

4.2 Experimental Setup

The most common clustering methods are hierarchical and partitioning methods. During clustering, we use partition methods. Partition clustering methods partition the data objects set into clusters where every pair of objects clusters are either distinct (hard clustering) or have some members in common soft clustering [75]. Partition clustering begins with a starting cluster partition which is iteratively improved until a locally optimal partition is reached. The objectives of cluster analysis are the organization of objects into groups, according to the similarity among them. In other words, clustering is unsupervised data mining techniques, which is applied when the class levels of the training data are unknown. Among the different clustering algorithms in WEKA, K-means algorithm, filtered cluster algorithm and farthest first are selected for experimentation in this study. In this study, we perform twelve experiments with k-means, filtered and farthest first algorithms. In order to validate and compare the clustering performance

of clustering model done in a way that the attribute value of each cluster in the model is compared to other clustering models with a number of iterations, inter-class similarity error, time to build the model and domain experts judgment. Through close discussion with domain expert and advices we set k value 2, 3 for these research works, because to convert this knowledge into data mining problem definition and to achieve the objective of this research and having iteration 10,100 used to see how it changes the variable of each of clusters in these research.

Experiment	Algorithms	Test mode
1	K-means	Where K values 2 and seed size 10
2	K-means	Where K value 2 and seed size 100
3	K-means	Where K value 3 and seed size 10
4	K-means	Where K value 3 and seed size 100
5	Filtered cluster algorithm	Where K value 2 and seed size 10
6	Filtered cluster algorithm	Where K value 2 and seed size 100
7	Filtered cluster algorithm	Where K value 3 and seed size 10
8	Filtered cluster algorithm	Where K value 3 and seed size 100
9	Farthest first cluster algorithm	Where K value 2 and seed size 10
10	Farthest first cluster algorithm	Where K value 2 and seed size 100
11	Farthest first cluster algorithm	Where K value 3 and seed size 10
12	Farthest first cluster algorithm	Where K value 3 and seed size 100

Table 4.1 Experiment with test mode

4.3 Experimental result

Before starting the experimentation it is important to set the threshold value for numeric attributes of the data sets. For this research purpose, we select thirteen attribute based on data-preprocessing method as discusses in chapter three, which are selected attribute to discover hidden knowledge and to segment and to identify the customer behavior.

4.3.1Clustering Result using K-means Algorithm

Using k-means algorithm four experiments conducted by k value to 2, 3 and seed value to 10, 100. The experiment result is shown in table 4.2

Experiments	Algorithm	Test option	Number of iteration	Time taken to build model
1	K-means	2 and seed 10	4	0.38 sec
2	K-means	2 and seed 100	3	0.06sec
3	K-means	3 and seed 10	8	0.16sec
4	K-means	3 and seed 100	5	0.08sec

Table 4.2 Performance result for K-means algorithm

The first experiment is conducted by k value to 2 and default seed value to 10. The experiment has generated a model with a number of iteration 4, time is taken to build the model 0.38 second; within-cluster sum squared error is 74759.The second experiment is conducted by k value to 2 and seed value to 100. The second experiment has generated a model with a number of iteration 3, time taken to build the model 0.06 second; within sum squared error is 74759, a number of iteration 3. The third experiment is conducted by k value to 3 and seed value to 10. The third experiment has generated a model with a number of iteration 8, time taken to build the model 0.16 second; within sum squared error is 67405. The fourth experiment is conducted by k value to 3 and seed value to 100. The fourth experiment has generated a model with a number of iteration 5, time taken to build the model 0.08 second; within sum squared error is 67059.

Generally, from the four experiments conducted before, the model developed with the k vale 2 and seed value 100 test option given better performance of identifying the value and behavior of

the customer. Therefore, among the four experiments of k-means algorithm models built in the forgoing experimentations, k value 2 and seed value 100 is selected

4.3.2 Clustering Result Using Filtered Cluster clustering algorithm

Using k-means algorithm four experiments conducted by k value to 2, 3 and seed value to 10, 100. The Experiment result is shown in table 4.3

Experiments	Algorithm	Test option	Number of iteration	Time taken to build model
1	Filtered	2 and seed 10	4	0.08 sec
2	Filtered	2 and seed 100	3	0.05sec
3	Filtered	3 and seed 10	8	0.13sec
4	Filtered	3and seed 100	5	0.05sec

Table 4.3 Performance result for filtered algorithm

The first experiment is conducted by k value to 2 and default seed value to 10. The experiment has generated a model with a number of iteration 4 , time is taken to build the model 0.08 second; within-cluster sum squared error is 74759.The second experiment is conducted by k value to 2 and seed value to 100. The second experiment has generated a model with a number of iteration 3, time taken to build the model 0.05 second; within sum squared error is 74759. The third experiment is conducted by k value to 3 and seed value to 10. The third experiment has generated a model with a number of iteration 8, time taken to build the model 0.13 second; within sum squared error is67405. The fourth experiment is conducted by k value to 3 and seed value to 100. The fourth experiment has generated a model with a number of iteration 5, time taken to build the model 0.05 second; with in sum squared error is 67405.

Generally, from the four experiments conducted before, the model developed with the k vale 2 and seed value 100 test option given better performance of identifying the value and behavior of the customer. Therefore, among the four experiments of filtered algorithm models built in the forgoing experimentations, k value 2 and seed value 100 is selected.

4.3.3 Clustering Result using farthest first Algorithm

The farthest first clustering algorithm is one of the clustering algorithms we choose for this research work. The farthest first clustering algorithm has some procedures related to the K-means clustering algorithm. In this algorithm, we choose centroids and assign the objects in the clusters but with maximum distance.

Experiments	Algorithm	Test option	Number of iteration	Time taken to build model
1	Farthest First	2 and seed 10	Has no iteration	0.05sec
2	Farthest First	2 and seed 100	Has no iteration	0.03sec
3	Farthest First	3 and seed 10	Has no iteration	0.06sec
4	Farthest First	3and seed 100	Has no iteration	0.02sec

Table 4.4 Performance result for farthest first algorithm

The first experiment is conducted by k value to 2 and seed value to 100. The experiment has generated a model time is taken to build the model 0.05 second. The second experiment is conducted by k value to 2 and seed value to 100. The second experiment has generated a model time is taken to build the model 0.03 second. The third experiment is conducted by k value to 3 and seed value to 10. The third experiment has generated a model time taken to build the model 0.06 second. The fourth experiment is conducted by k value to 3 and seed value to 100. The fourth experiment has generated a model time taken to build the model 0.02 second.

Generally, from the four experiments conducted before, the model developed with the k vale 3 and seed value 100 test option given better performance of identifying the value and behavior of the customer. Therefore, among the four experiments of farthest first algorithm models built in the forgoing experimentations, k value 3 and seed value 100 is selected.

4.4 Comparison of Clustering Algorithm results

Comparative analysis of various clustering algorithm has made. Selecting a better clustering technique for building a model, which performs best in the segment of the customer, is one of the aims of this study. For that reason, three clustering algorithm is selected for the implementation of the clustering model namely; k-means, filtered and farthest first. Then, four experiments conducted for each algorithm, and the obtained results are comparing. For each algorithm, the best performance accuracy is listed in table 4.4 below.

Types of Clustering algorithm	**Time taken (in sec)**	**Number of iteration**	**Sum squared error**
K-mean	0.06sec	3	74759
Filtered	0.05Sec	3	74759
Farthest first	0.02 sec	-	-

Table 4.5 Performance Comparison of the selected models

The farthest first cluster model using has less time to builds the model and has no number of iteration and sum squared error.

This shows that the experiment conducted with the farthest is better for enterprise customer segmentation. The result has validated by using a 21126 data set taken from the enterprise customer database and noticed that datasets are successfully clustered with less time building the model.

The filtered algorithms cluster model has a second better clustering algorithm less time is taken to build the model 0.02 second and has minimum number of iteration.

This shows that the experiment conducted with the filtered is better for enterprise customer segmentation than the k-means clustering algorithm. The result has validated by using a 21126 data set taken from the enterprise customer database and noticed that datasets are successfully clustered with a good time building the model and minimum number of iteration than the k-means clustering algorithm.

The k-means clustering algorithms cluster model a third clustering algorithm and time taken to build the model is 0.06 second and number of iteration 3.

This shows that the experiment conducted with the k-means algorithm less for enterprise customer segmentation than the filtered clustering algorithm. The result has validated by using a 21126 data set taken from the enterprise customer database.

As it is shown in table 4.5, among the three algorithms farthest algorithm using k value 3 and seed 100 has less time to build the model. As a result, according to the result farthest first algorithm chooses as a final model for the study.

4.5 Discussion on finding of the study based on the centroid of the selected algorithm

The selected clustering model experiment is conducted by k value to 3 and seed value to 100. The experiment has generated a model with time taken to build the model 0.02sec. The output of the selected experiment is presented in table 4.5.

Clusters No	Frequency of records (Share %)	ZN	CLN	NBT	ON	SN	IC	NE	NB	VU	DU	DW	SU
1	**9137(43%)**	3	Platinum	3	Wireless	Prepaid	high	high	Low	high	high	high	high
2	**7627(36%)**	5	Gold	2	Wired	postpaid	low	low	high	low	low	low	low
3	**4362(21%)**	7	Platinum	3	Wireless	Prepaid	high	low	high	Average	low	low	low

Table 4.6 Clustering result of the selected experiment

In the selected experiment clusters the customer data into three segments. This result is used to compare the segments which are generated from the clustering algorithm, used to classify the customers according to their cluster groups, identify the behavior of customers and also identify the value of customers. The next steps are to describe each cluster according to tables 4.6.

Cluster 1: The customers who categorize under cluster one are with customers zone name is enterprise (TPO) building. Customers clustered into this category are also with platinum and net business type GSM and wireless product offering. The customer's payment systems in this clustering group are the prepaid payment system. The investment capital is high, a number of employees are also high, and a number of the branch office is low, voice usage high, data uploading usage high, data downloading usage high and SMS usage high. The customers in this

clustering group are high-value customers and generate high revenue for ET. The customer usage in this clustering group is high customer usage.

Cluster 2: The second cluster describes customers with the south Addis Ababa zone and with gold customer's level and fixed-line voices net business type with a product offering of wired. The customer's payment systems in this clustering group are the postpaid payment system. The investment capital is low, a number of employees are low, and a number of the branch office is high, voice usage low, data uploading low, data downloading low and SMS low. The customers in these clustering groups are low-value customers and generated low revenue to ET. The customers in this clustering group are low-value customers and generate low revenue for ET. The customer usage in this clustering group is low customer usage.

Cluster 3: The second cluster describes customers with the west Addis Ababa zone and with platinum customer's level and GSM net business type with a product offering of wireless. The customer's payment systems in this clustering group are the prepaid payment system. The investment capital is high, a number of employees are low, and a number of the branch office is high, voice usage average, data uploading low, data downloading low and SMS low. The customers in these clustering groups are low-value customers and generated low revenue to ET.

4.6 Discussion of Major Findings of experiments

As pointed in problem understanding of the ET enterprise customers high-value customers are those customers who generate high revenue, which uses a high amount of active ET services, those customers who generate under platinum customer level name, those customers who have high investment capital, these means customers have greater than or equal to 11,000,000 million ETB (Ethiopia Birr) investment capital, those customers which has greater than 51 employees, those customers who have greater than or equal 5 branch office and are generate high revenue to ET . Using this attribute we discover discrete segments in there customer base and used to identify people who are most likely to become customer or most profitable customer.

On the other side, low-value customers generate low revenue, which not use actively ET services, those customers who have categorize under gold customer level name, those customers which has low investment capital, these means customers which has less than 11,000,000 million ETB (Ethiopian Birr) investment capital, those customers which has less than 51 employees,

those customers has less than 5 branch office and those customers generate less revenue to ET. The above facts of the business, each cluster has assigned in the following ranking order. Using this attribute we discover discrete segments in there customer base and used to identify people who are most likely to become customer or most profitable customer.

As we know, the farthest algorithm cluster algorithm chooses as a final model for this study.

In the Cluster one, the customers in this cluster are a high-value customer; generate high revenue to ET, because the customer investment capital, number of branch office is high. Customer behavior under this group is voice usage is high, data uploading is high, data downloading is high and SMS usage is high. Based on the above business facts the customer under these groups is a high-value customer. In cluster two, the customers are low-value customers, generates low revenue to ET, because the customer investment capital, number of branch office are low. The customer's behavior under this group is voice usage is low; data uploading low, data downloading low and SMS usage is low. Based on the above business facts the customer under these groups is a low-value customer. In cluster three, the customers are low-value customers, generates low revenue to ET, because the customer number of employees and number of branch office are low. The customer's behavior under this group is voice usage is average; data uploading low, data downloading low and SMS usage is low. Based on the above business facts the customer under these groups is a low-value customer.

According to the centroid values assigned for each cluster, a summary of the finding is given below

- **SOHO:** This is low-value customer groups, this group of customer located at both south addis ababa and west addis ababa zone, the customer level name is platinum, the customer net business type is fixed line and LTE (long term evolution) and the customer under this group used internet access is wireless and customer payment system is prepaid. The customer under this group investment capital and number of employees is low. Number of the branch office is high.The behavior of customers under this group high usage voice in minutes and also low usage data uploading from computer to the internet and the customer under this group low rate data downloading files from a web server and also SMS usage low under this group.

- **KEY ACCOUNT:** the customer group is high-value customer located at enterprise (TPO) building zone, the customer level name is platinum, and the customer net business type is GSM this means the customer under this group mostly used GSM, the customer under this group used internet access is wireless and customer payment system prepaid. The customer under this group investment capital and number of branch high. The Number of employees is low. The behavior of customers under this group is high voice usage, high usage data uploading rates from computer to internet, high data downloading rates from a web server to a computer and also SMS usage high under this group.

4.7 Use of Knowledge

After evaluating the discovered knowledge, the last step is the use of this knowledge to identify the behavior of enterprise customers and also the value of a customer to ET. In this step, the knowledge discovered is integrated into the customer relationship management system for ET customer segmentation.

4.7.1User Interface Design

User interface design is an iterative process involving close relationships between users and designers. Using the centroid generated by the clustering algorithm we design a user interface using Java programming language. The design interface user prototype accepts user queries and groups the enterprise customer according to their query. The development of a graphical user interface in this study was done using Java. This prototype graphical user interface was developed based on the model generated by the farthest algorithm. This prototype clustering model can be used for segmenting enterprise customers on the centroid generated by farthest algorithm .

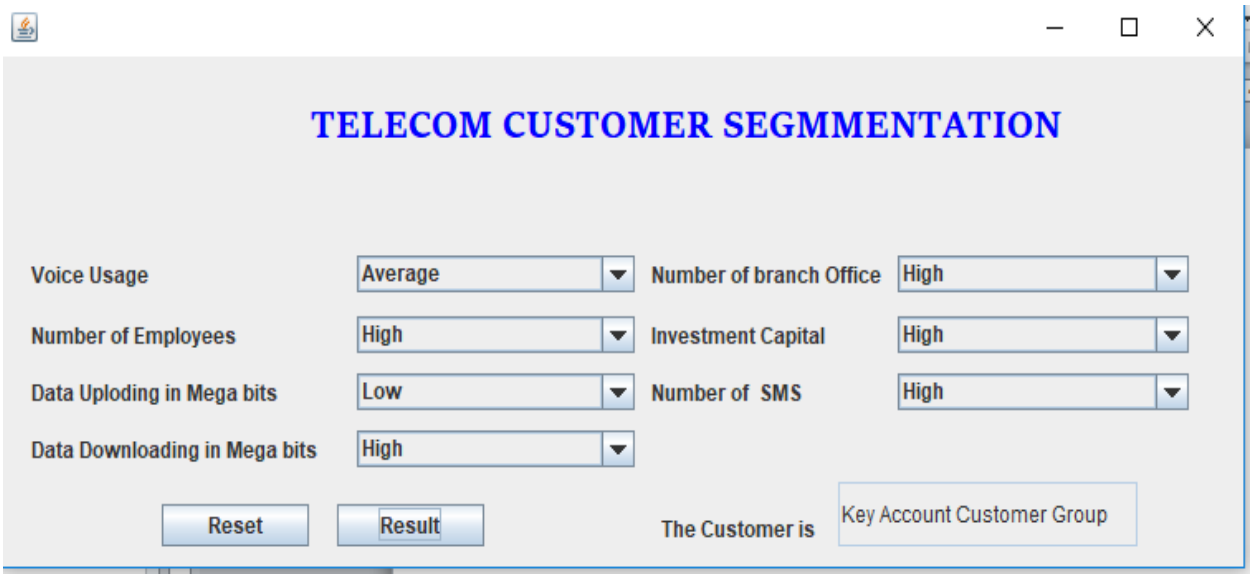


Figure 4.3: Prototype user interface of telecom enterprise customer segmentation

Figure4.3 shows we used thirteen attributes to build a telecom customer segmentation prototype. After receiving this query from the user and it displays the customer segmentation group.

4.7.2 User acceptance testing

User acceptance testing is a testing methodology where the clients or end users involved in testing the product to validate the product against their requirements. The five E's (effective, efficient, engaging, error tolerant, easy to learn) provide the expert with a set of characteristics that can be used to organize and analyze information from users [76]. They offer traceability from initial information-gathering through requirements setting and finally in evaluation.

4.7.3 Effectiveness

Effectiveness is the capability of producing the desired result or the ability to produce the desired output. When something is deemed effective, it means that it has an intended or expected outcome or produces a deep impression.

As discussed in chapter four, we perform an experiment with a k-mean, filtered and farthest first clustering algorithm.

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

4.7.4 Efficiency

Efficiency is the ability to avoid wasting materials, energy, efforts, money, and time in doing something or in producing the desired result. In a more general sense, it is the ability to do things well, successfully, and without waste. 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.

In this research, efficiency is considered as a time taken to cluster enterprise customer records by taking inputs from the user. As discussed in chapter one, currently in ET customer segmentation is done by traditional simple statistical methods which need more time to operate because every operation is done manually.

During our presentation and discussion with domain experts, we conduct sample experiments and compare the efficiency between the current statistical method and our new enterprise customer segmentation method. From these sample experiments, our new enterprise customer segmentation method becomes more efficient and every domain expert agreed up on this.

4.7.5 Engaging

An interface is engaging if it is nice and satisfying to use. The graphic design is the clearest element, even within the same class of interfaces; different users may have widely different needs. What is important is that the design meets the needs of the people who must use the interface. So domain experts satisfy with the interface design and we prepare sample screenshots on the document as shown in figure 4.3.

4.7.6 Error Tolerant

Error Tolerance is concerned about the management of faults originating from defects in design or implementation. In this research, error tolerance is considered as making the error of the experiment free or making its intelligence. During our presentation and discussion with domain experts, we discuss on how to make these experiments intelligence by integrating this experiment with knowledge-based systems and we agreed that some improvements also need.

4.7.7 Easy to learn

Building a product easy to use by the user is one of the non-functional requirements for any product.

4.7.8 Evaluation Result

The researcher developed a user interface prototype to cluster enterprise customer records based on the objective to check the validity of the interface.

Each of the study participants of ET domain experts was asked to give feedback on the acceptability of the segmentation and to rate it on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree). We provide the results in Table 4.6.

Questionnaires	Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency:					
➢ Is the response fast delivered segmentation responses?	90%	10%	-	-	-
➢ Is the customer segmentation saves effort & space?	95%	5%	-	-	-
Effectiveness:					
➢ Is the customer segmentation Reliable?	80%	10%	5%	5%	-
➢ Is the customer segmentation produces a desired result?	75%	15%	5%	5%	-
Error Tolerance:					
➢ Does the customer segmentation tolerate errors?	50%	30%	15%	5%	-
➢ Is the segmentation result satisfactory?	80%	10%	10%	-	-
Engaging					
➢ Is the customer segmentation user interface attractive?	90%	10%	-	-	-
Easy to Learn:					
➢ Is the customer segmentation system	90%	10%	-	-	-
➢ Is the customer segmentation system User friendly?	90%	10%	-	-	-

Table 4.7 Summary of domain Experts response on the telecom enterprise customer segmentation

This study revealed that the majority of domain experts have positive feedback towards the validity of the prototype. This customer segmentation model is much efficient, and it saves their energy and materials while comparing with the way they perform currently which is the simple statistical method. In the case effectiveness, they revealed that this customer segmentation model produces the desired result, but in order to make it perfect, some improvements will be needed. In the case of effectiveness, some domain experts disagree and undecided with the results because they stated that in the case of customer segmentation analysis identifies customer number of employees, number of branch office and investment capital must have to be perfect because it has a significant impact on the organization's revenue.

Most of the domain experts satisfied with the customer segmentation results but some of them strongly make the customer segmentation model more accurate and error tolerable. But we have advice by domain expert integration of the discovered customer behavior of enterprise customers with a knowledge-based system. In some cases, the domain expert disagrees with the customer segmentation model because the model doesn't tolerate errors.

During our discussion with ET domain experts, they understand that our enterprise customer segmentation model is easy to learn. They also agreed that the enterprise customer segmentation model is user-friendly and it is clearer than before.

We had some improvement suggested by the respondent, one respondent from enterprise section employees, explain his thought as follows.

"I found this user interface so interesting in terms of identifying customer value and behavior. Since, we didn't use such data mining techniques before to identify enterprise customer behavior and values. However, I have one suggestion to prototype that is, if other behavior also be included, it will be improved identification of customer behavior and values".

Chapter Five

CONCLUSION AND RECOMMENDATIONS

5.1Conclusion

Nowadays where ET provides telecom services, different data generated as call detail, network and customer. This huge amount of data is stored in the ET database for different purposes. Using these data different report is generated. All these data become more useful when it is analyzed and some dependence and correlation are detected.

ET used traditional simple statistical methods to generate knowledge from this data. These methods are slow and highly subjective.

Using data mining techniques has a solution to this problem by extracting hidden business information from large customer data to provide customer better services and find more commercial opportunities.

The objective of this research is to develop clustering models that identify the behavior of high-value enterprise customer's using data mining techniques. To achieve this research objective, we use enterprise customer profile data including call detail data.

The data mining tasks conducted based on the hybrid data mining process model, which has six major parts, namely understanding of the problem, understanding of the data, preparation of the data, data mining, evaluation of discovered knowledge and use of discovered knowledge. The study conducted using WEKA software version 3.8.2 and three data mining algorithms, such as k-means, filtered and farthest first.

In the data preparation phase, we remove unnecessary and unwanted data. From a total of 162315 records of data, we select for this research work 21126 records. Thirteen attributes are selected for this research. The preprocessed data is saved in the ARFF format which is suitable for DM tasks. Based on the preprocess data set, different clustering experiments have been done with the k-means clustering algorithm by taking k value 2 and 3 and seed size 10 and 100 filtered

clustering the algorithm also is done by taking k value 2 and 3 and seed size 10 and 100 the third clustering algorithm applied for this research is the farthest first clustering algorithm, this algorithm also is done by taking k value 2 and 3 and seed size 10 and 100.

Farthest first algorithm with the value of k value 2 and seed size 100 has a better clustering performance . It enables to cluster corporate customers into a dissimilar cluster of high and low-value customers groups of the corporation and also enables to identify customer behavior in each cluster group. So, this model is chosen for clustering enterprise customers.

In general, the result of this research shows that applying data mining helps ET to know enterprise customer value and behaviors, to understand consumption characteristics of different customer groups, to identify target customer groups, characterizes each group and analyze their properties.

The design interface user prototype can be used for segmenting enterprise customers on the rules generated by the best clustering model. This prototype provides the user to get a better understanding of the system. As a final output of this research, a prototype of an enterprise customer segmentation system is developed. This prototype helps to find the cluster where the enterprise customer belongs and also identifies the contribution that the customer makes to organization profitability.

In undertaking this research, the major challenge we faced is getting completed and relevant data for conducting this study.

5.2Recommendation

Based on the finding of the study, the researcher would like to suggest some important issues which require further study.

- In this research, we used only enterprise customer profile data but further investigation is needed by including other residential customer data. There is also a need to design a well-managed database and data warehouse to simplify data mining.
- In this research, we used only Addis Ababa enterprise customer data. Further study is required to include other regional enterprise customer data for designing country wide ET customer segmentation model.
- This study aimed at building a descriptive model using enterprise customer records. For this research, we used customer voice usage, data uploading usage, data downloading usage, SMS usage attributes. For further research using call duration, times of call, the number of different telephone numbers called by the speaker, concentration of call duration and concentration of times of calls will describe customer behavior.
- The need to integrate the data mining model with a knowledge-based system to design an intelligent system for customer segmentation.
- Customer satisfaction generates higher revenue for Ethio telecom. Data mining techniques for customer segmentation should enable the Authority to identify potential customers. Therefore, to use this capability, awareness on the advantages of CRM among employees at all levels should be created.
- In this research, we use one of the data mining techniques which are clustering techniques. Using these techniques we design a customer segmentation model to identify the high value and behavior of enterprise customers. Further investigation is needed by including classification techniques to classify new ET customers.

Reference

[1] A.ALAMIREW, "repository,smuc.deu.et," 1 november 2013. [Online]. Available: http://www. smuc.et. [Accssed 9 12 2018].

[2] R. Mouly, "An Assessment Of Ethiopian Telecom Customer Satisfaction," Management and Business Research, vol. 10, no. 4, pp. 10-15, 2010.

[3] K.Nibret, "An Assessment of Service Quality Level and Customer atisfaction with broadband internet services," 1 june 2017. [Online]. Available: www.smuc.et. [Accessed 29 august 2018].

[4] Enterprise Product and Service, "Ethiotelecom Enterprise Center," 25 january 2012. [Online]. Available: www.ethiotelecom.et. [Accessed 30 8 2018].

[5] D.G.Efen, "repository,aau.edu.et," 1 june 2015. [Online]. Available: www.aau.edu.et. [Accessed 29 august 2018].

[6] J.Han, data mining techinques and concept, Waltham: Morgan Kaufmann, 2012.

[7] J. Ponce and A. Karahoca, Data Mining and Knowledge Discovery in Real Life Applications, I-Tech Education: Morgan Kaufmann, 2009.

[8] J.Tikmani,S. Tiwari and S.Khedkar, "An Approach to Customer Classification Using K-means," Innovative Research in Computer and Communication Engineering, vol. 3, no. 11, p. 10543, 2015.

[9] D.Merga, "Application of Data Mining for Customer Segmentation in insurance business:The Case of Ethiopia Insurance Corporation", MSC Thesis ,Addis Ababa University,Addis Ababa Ethioopia.

[10] A. C. K. Tsiptsis, Data Mining Techniques in CRM: Inside Customer Segmentation, a thens: A John Wiley and Son, 2009.

[11] Ethio Telecom, "Eth-Em Customer Segmentation," august 2014. [Online]. Available:http://www.ethiotelecom.et. [Accessed 30 8 2018].

[12] Ethio Telecom, "Eth-Em Customer Segmentation," august 2014. [Online]. Available:http://www.ethiotelecom.et. [Accessed 30 8 2018].

[13] Enterprise Product and Service, "Ethiotelecom Enterprise Center," 25 january 2012. [Online]. Available: www.ethiotelecom.et. [Accessed 30 8 2018].

[14] W.Bogale, "a baground paper on telecom and telecom stastic ethiopa," Ethio Telecommuncation Coporation, addis abeba, 2005.

[15] A. Mohod,B.Lilhare,C. Meshram and S. Meshram, "Customer Relationship Management Using Angularjs," advance research and innovative ideas in education, vol. 4, no. 2, pp. 64-67, 2018.

[16] C. Qiuru,L.Ye,X. Haixu,L. Yijun and Z.Guangping, "Telecom Customer Segmentation Based on Cluster Analysis," in Computer Science and Information Processing, Changzhou, 2012.

[17] H.W. Tefera, "Application of Data Mining Techniques To Support Customer Relationship Management At Ethiopian Airlines ," Msc Thesis, Addis Ababa, Ethiopia .

[18] M. Nadaf, "Data Mining in telecommunication," advanced computer theory and engineering, vol. 2, no. 3, pp. 92-96, 2013.

[19] C.Qiuru,L.Ye,X. Haixu,L. Yijun and Z. Guangping, "Telecom Customer Segmentation Based on Cluster Analysis," in Computer Science and Information Processing, Changzhou, 2012.

[20] K. Mehmed, DATA MINING Concepts, Models, Methods, and Algorithm, New Jersey: John Wiley & Sons, Inc, 2011.

[21] K.J. Cios,W. Pedrycz,R. W. Swiniarski and L. A. Kurgan, A Data Mining Knoweledge Discovering Approch, vol. 3, New York: Springer Science and Business Media, 2007, pp. 844-855.

[22] M.Hossin,M.N.Sulaiman, "a rivew on evaluation for data classfication evaluation," international Journal of Data Mining & Knowledge Managment process, vol. 5, no. 2, pp. 1-11, 2015.

[23] Hobart and William Smith Colleges, Introduction to Programming Using Java, David J.Eck, 2016.

[24] D. T. Larose, Discovering Knowledge in Data, New Jersey: John Wiley & Sons, Inc, 2005.

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

[26] M. Kantardzic, DATA MINING Concepts, Models, Methods, and Algorithm, New Jersey: John Wiley & Sons, Inc, 2011.

[27] H. Nasereddin, "New Technique to Deal with Dynamic Data Mining in the," Data Mining Techniques and Applications, vol. 13, no. 3, pp. 806-814, 2012.

[28] U. Fayyad,G. Piatetsky Shapiro and P.Smyth, "knoweledge discovery and data mining :to ward unifiying frame work," in Association for the Advancement of Artificial Intelligence!, redmond,waltham,irrivne, 1996.

[29] H.Abebe, "Application of data mining techinques to customer profile analysis in the Ethiopian Electric Power Corporation," Msc Thesis, Addis Ababa University,Addis Ababa Ethiopia.

[30] R.M. Shah,M. A. Butt and M. Z. Baba, "Predictive Analytic Modeling: A Walkthrough," International Journals of Advanced Research in computer Science and and Software Engineering, vol. 7, no. 6, pp. 424-426, 2017.

[31] P. Chapman,J. Clinton,R.Kerber,T. Khabaza,C. Shearer and R. Wirth, CRISP-DM, tatistical Package for the Social Sciences (SPSS, 2000.)

[32] A. Ahlawat and B. Suri, "Improving Classification in Data mining using Hybrid algorithm," 16 8 2016.

[33] R. W. Swlnlarskl and L.A.kurga, datamining knowelege discovering approch, vol. 40, new york: pringer Science+Business Media, 2007, pp. 5636-5647.

[34] U. Shafique and H. Qaiser, "Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," International Journal of Innovation and Scientific Research, vol. 12, no. 1, pp. 217-222, 2014.

[35] A.Azevedo, "KDD, SEMMA and CRISP-DM: parallel overview," in IADIS European conference on data mining , S.M de Infesta, 2008.

[36] P. Chapman,J. Clinton ,T.Khabaza,T.Reinartz and Rüdiger Wirth , "The CRISP-DM Process Model," 1999.

[37] H. Ziafat and M. Shakeri, "Using Data Mining Techniques in Customer Segmentation," Engineering Research and Applications, vol. 9, no. 4, pp. 70-78, 2014.

[38] M.Bramer, Principles of Data Minin, Portsmouth: Springer-Verlag London Ltd, 2016.

[39] T. Sajana, "A Survey on Clustering Techniques for Big," Indian Journal of Science and Technology, vol. 9, no. 3, pp. 1-12, 2016.

[40] P. Dhandayudam, "an improved clustering algorithm for customer segmentation," Journal of Engineering Science and Technology, vol. 4, no. 2, pp. 695-702, 2012.

[41] T.Sajana,C.M.Sheela and K.V.Narayana, "A Survey on Clustering Techniques for Big," Science and Technology, vol. 9, no. 3, pp. 1-12, 2016.

[42] C. C. Aggarwal, data mining, new york: spring international, 2015.

[43] Manalina and K. M. Prasad , "Effective Clusters Culled out Through Algorithmic," Asian Research Publishing Netwo, vol. 11, no. 9, pp. 5574-5579, 2016.

[44] K. Arzoo, "K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8.2," International Research Journal of Engineering and Technology, vol. 4, no. 4, pp2363-2368, 2017.

[45] J.L.Bentley , "Multidimensional Binary Search Trees Used For Associative Searching," Association for Computing Machinery, vol. 18, no. 9, pp. 509-517, 1999.

[46] N.S. Netanyahu, "An Efficient k-Means Clustering Algorithm:Analysis and Implementation," Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892,2002.

[47] G.Sehgal and K.Garg, "Comparison of Various Clustering Algorithm," international journal of computer science and Iinformation Tecnology, vol. 5, no. 3, pp. 3074-3076, 2014.

[48] S.M.Kumar, "An Optimize Farthest First Clustering Algorithm," in Nirma University International Conference on Engnering , Bhiwani, 2013.

[49] A.Deepshree and H.K.Yogish , "Farthest First Clustering in Links Reorganization," International Journal of Web & Semantic Technology, vol. 5, no. 3, pp. 17-24, 2014.

[50] M.J.A.Berry and G. Linoff, Mastering data mining: The art and science of customer relationship managment, canada : John Wiley and Sons, Inc, 2000.

[51] M. Halkidi and M. Vazirgiannis, "Clustering Validity Assessment:Finding the optimal partitioning of a data set," in *ICDM Conference*, California, 2002.

[52] P. Sharma, "Comparative Analysis of Various Clustering Algorithms Using WEKA," International Research Journal of Engineering and Technology , vol. 02, no. 04, pp. 108-112, 2015.

[53] Two Crows Corporation, introduction to data mining and knowledge discovery, potomac, 2005.

[54] C. Rygielski, J.Wang and D. C. Yen, "Data mining techniques for customerrelationship management," *Technology in Society*, vol. 24, pp. 484-502, 2002.

[55] S.Sowjanya and R.M.Sravan, "Application of Data Mining techniques for Customer Relationship," Engineering Research & Technology, vol. 2, no. 11, pp. 2278-0181, 2013.

[56] Ling and D.C.Yen, "Customer relationship management: An analysis framework and implementation strategies," Journal of computer information system , vol. 4, pp. 82-97, 2001.

[57] M. Shaik,N. Shaik and D. Meghavath, "Data Mining Concepts with Customer Relationship Management," Engineering Research and Applications, vol. 4, no. 7, pp. 2248-9622, 2014.

[58] Microsoft Dynamics GP , "customer relationship managment," 2007. [Online] Available: http://www.microsoft.com/dynamic/gp. [Accessed 5 4 2019].

[59] R. Ghnemat and E. Jaser , "Classification of Mobile Customers Behavior and Usage Patterns using Self-Organizing Neural Networks," Computer Science and Information Technologies,, vol. 5, no. 4, pp. 154-159, 2015.

[60] A.l Arora and Dr. R. Vohra, "Segmentation of Mobile Customers for Improving Profitability Using Data Mining Techniques," Computer Science and Information Technologies, vol. 5, no. 4, pp. 5241-5244, 2014.

[61] A.Begunca, "soft drinks consumer segmentation using benefit sought variables-case study kosovo market," pp. 168-173, 2017.

[62] B. B. Bezabeh, "Knowledge Discovery for Effective Customer Segmentation: The Case of Ethiopian Revenue and Customs Authority," Msc Thesis, Addis Ababa University, Addis Ababa, Ethipia.

[63] B. Reganie, "Application of Data Mining Techinques for Customers Segmentation and Predication:The Case of Buusaa Gonofa Microfinance Institution,"Msc Thesis ,Addis Ababa University, Addis Ababa, Ethiopia.

[64] F.A.Buttle and R.Iriana, "Strategic Operation and Analtical Customer Relationship Management Attribute and Measure ," Computer Science and Information Technologies, vol. 5, no. 4, pp. 24-42, 2006.

[65] P. Gray and J. Byun, "Customer Relationship Management," Msc Thesis , California University, California, U.S.A.

[66] O.Dogan and H.Asan, "Use of Data Mining Techniques In Advance Decision Making Process In A Local Firm ," Computer Science and Information Technologies, vol. 10, no. 2, pp. 24-42, 2015.

[67] S.Balaji and S.K.Srivatsa, "Customer Segmentation for Decision Support Using Clustering and Association Rule based approches ," Computer Science and Information Technologies, vol. 3, no. 11, pp. 2229-3345, 2012.

[68] E.Telecom , "demarcation between key account and SOHO," Ethio Telecom , addis ababa, 2013.

[69] James E.Richard and Peter Thirkell , "Customer relationship management technology impact on business to business customer relationships:Development of a conception model," research get, vol. 25, no. 2, pp. 138-166, 2005.

[70] G.o'scarmarba'n, "A survey of data mining and knowledge discovery process models and methodologies," research get, vol. 25, no. 2, pp. 138-166, 2010.

[71] T.P.Nadeau, Data Mining, new york: Morgan Kaufmann, 2009.

[72] Rana Soudagar, "Customer Segmentation and Strategy defination in segments: in case of an internet service provider in iran," internet provide in iran, 2007. [102] D.T.Larose, Discovering .

[73] R.Houari, A.Bounceur, A.Tari and M.Kechadi, "Handling Missing Data problem with Sampling Methods", research gate, 2014.

[74] A.Dharmarajan and T.Velmurugan, "Lung Cancer Data Analysis by k-means and Farthest first Clustering Algorithm," Indian Journal of science and Technology , vol. 8, no. 15, pp. 2-8, 2015.

[75] J.Ponce and A.Karahoca, Data Mining and Knowledge Discovery in Real, new york : I-Tech Education and Publishing, 2009.

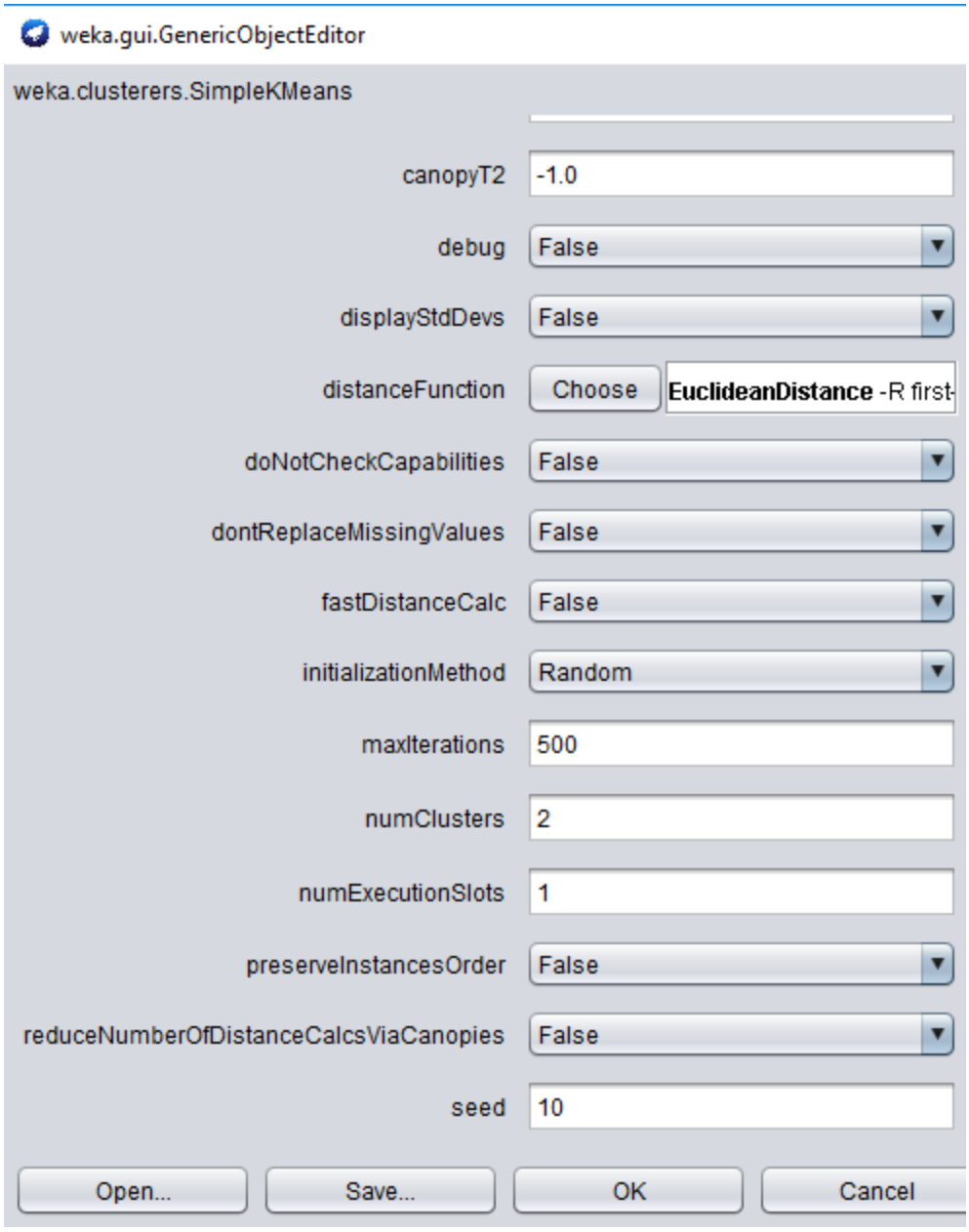
[76] L. Luo, "Software Testing Techniques:Technology Maturation and Research Strategy," Institute for Software Research Internationa:Carnegie Mellon Universityl, Pittsburgh,

ANNEXES

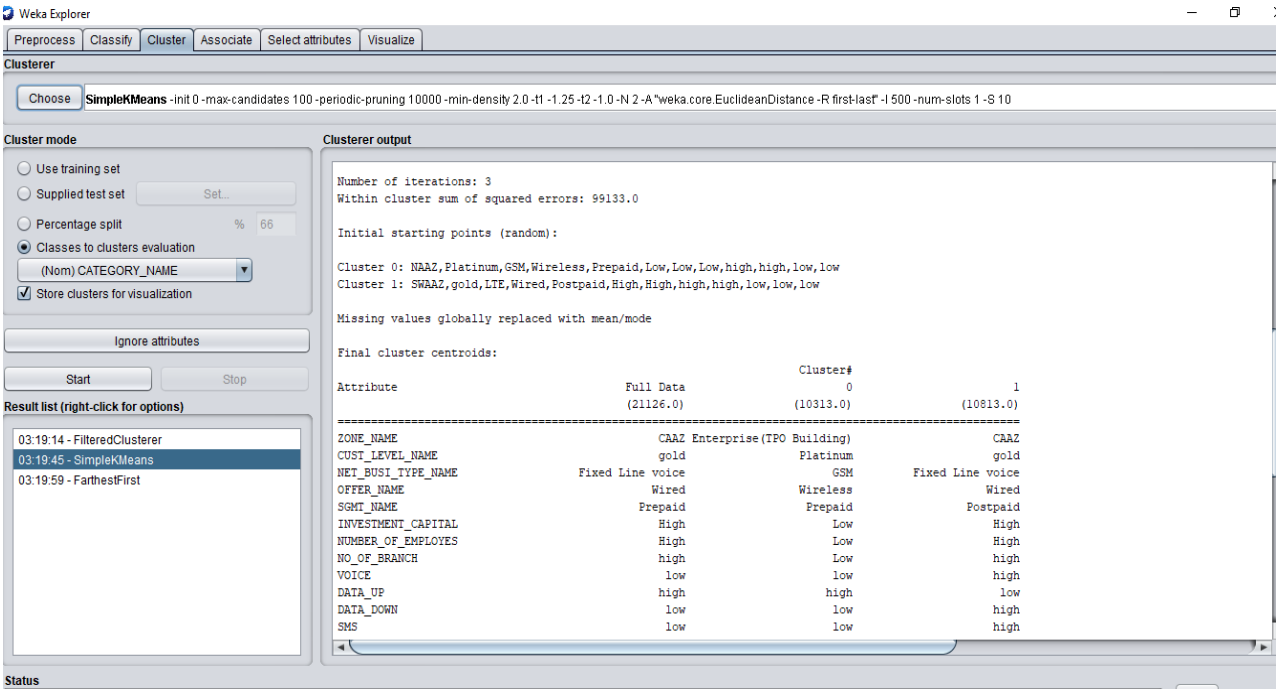
Annexes1:- The original collected sample data

ACTIVATE_DATE	SERV_NO	CUST_NAME	ZONE_NAME	COLLECT_CENTER	CITY_NAME	CUST_LEVEL_N	CUST_TYPE_NAM	CATEGORY_NAM	SUB_CATEGORY_NA	NET_BUSI	OFFER_
2/9/2016	111261212	MIZAN TEPI UNIVERSITY	NAAZ	NAAZ-05-ARADA	Arada sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
21/10/2016	111263221	BANK OF ABYSSINIA legal	CAAZ	CAAZ-001-Leghar	Kirkos sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
6/1/2017	111263398	TRANSSION MANUFACTURI	EAAZ	EAAZ-001-BOLE ZON	Bole Sub City	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
22/08/2016	111263514	ADDIS CAPITAL GOODS FINA	NAAZ	NAAZ-05-ARADA	Arada sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
23/12/2016	111263820	Dashen Bank S.C.	NAAZ	NAAZ-05-ARADA	Arada sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
20/03/2014	111264061	BUNNA INTERNATIONAL BA	Enterprise(TPO BU	Enterprise(TPO Buil	Arada sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
20/10/2016	111265456	G2G IT SOLUTION SHARE CO	NAAZ	NAAZ-05-ARADA	Arada sub city	Platinum	Enterprise	SOHO/SME	Financial Institution	Fixed Line	Fixed L
24/03/2017	111267278	ETHIOPIAN INDUSTRIAL INP	NAAZ	NAAZ-05-ARADA	Arada sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L
27/06/2016	111267720	EOTC-SSD-MAHIBERE KIDUS	NAAZ	NAAZ-12-6KILO	Arada sub city	VIC	Enterprise	Key Account	Retired	Fixed Line	Fixed L

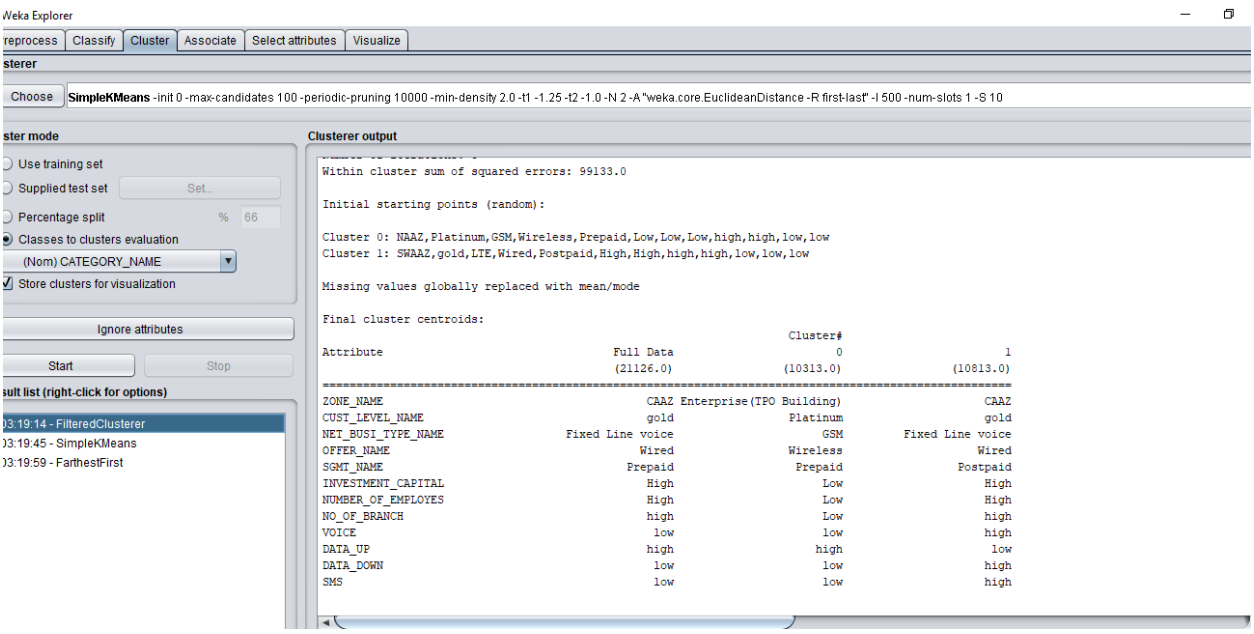
Annexes 2:- Parameter settings of the k-means for conducting the experiments



Annexes 3:- The snapshot running information of k-means algorithm



Annexes 4:- The snapshot running information of filtered algorithm



Annexes 5:- The snapshot running information of farthest first algorithm