ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL & COMPUTATIONAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

MINING CUSTOMS DATA FOR CUSTOMER SEGMENTATION: THE CASE OF ETHIOPIAN REVENUES AND CUSTOMS AUTHORITY (ERCA)

A THESIS SUBMITTED TO THE SCHOOL OF INFORMATION SCIENCE ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

SEWAGEGN TENAW

ADVISOR

TEMTIM ASSEFA (Ph.D)

JUNE, 2018

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL & COMPUTATIONAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

MINING CUSTOMS DATA FOR CUSTOMER SEGMENTATION: THE CASE OF ETHIOPIAN REVENUES AND CUSTOMS AUTHORITY (ERCA)

BY

SEWAGEGN TENAW

Name and signature of members of examining Board

Chairperson:	_____	_____	_____
	Name	Signature	Date
Advisor:	_____	_____	_____
	Name	Signature	Date
Examiner:	_____	_____	_____
	Name	Signature	Date
Examiner:	_____	_____	_____
	Name	Signature	Date

DECLARATION

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

_____

Sewagegn Tenaw

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

_____

Temtim Assefa (Ph.D)

ACKNOWLEDGEMENT

First of all, I praise the almighty GOD and his mother ST. MARY for enabling me to complete the masters program at AAU. Thank you Holy Mother of GOD! Thanks again and again for giving me the strength, courage and the opportunity to do this.

I would also like to express my deepest gratitude to my advisor Dr. Temtim Assefa for his guidance, suggestions and support throughout the preparation of this thesis. His critical assessment, comments and suggestions have helped me in maintaining the right direction for my study and making it meaningful. Thank You So Much!

I would also like to express my deepest gratitude and indebtedness to my brothers Ato Goshu Tenaw, Ato Wubetie Tenaw and Ato Tariku Goshu and my mother w/ro Yeshalem Mengistie who have always encouraged me for higher success.

Finally, I would like to forward my special thanks to the Ethiopian Revenue and Customs Authority employees. Many thanks go particularly to the eCMS project manager , Ato Ababu Emiru for his kind collaboration, and my deepest thanks goes to Aytenew Sewnet, Tesfaye Fenta, Tekalign Tamru and Kedir Yasin for their supportive suggestions on ERCA's business issues.

LIST OF ACRONYMS

CRM	=	Customer Relationship Management
CS	=	Customer Segmentation
ERCA	=	Ethiopian Revenues and Customs Authority
ASYCUDA	=	Automated System for Customs Data
SIGTAS	=	Standard Integrated Government Tax Administration System
TIN	=	Tax Identification Number
KDD	=	Knowledge Discovery for Database
SQL	=	Structured Query Language
IR	=	Information Retrieval
OLAP	=	Online Analytical Processing
MIS	=	Management Information System
AI	=	Artificial Intelligence
SAD	=	Single Administrative Document
CHAID	=	Chi-squared Automatic Interaction Detection
CART	=	Classification And Regression Trees
SVM	=	Support Vector Machine
SOM	=	Self-Organized Maps
VDM	=	Visual Data Mining
ANN	=	Artificial Neural Network
NN	=	Neural Network
ICT	=	Information Communication Technology
WEKA	=	Waikato Environment for Knowledge Analysis
CSV	=	Comma Separator Value
ARFF	=	Attribute Relational File Format for data processing
CRISP-DM	=	CRoss-Industry Standard Process for Data Mining
DM	=	Data Mining
ITMD	=	Information Technology Management Directorate
SSE	=	Sum of Squared Error
DBMS	=	Database Management Systems
IT	=	Information Technology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Customer relationship management (CRM) is the overall process of exploiting customer data and using it to increase the revenue generated from an existing customer and attract new customers by creating good relationship with them accordingly. CRM links business processes across the supply chain from back office functions through all touch points, enabling continuity and consistency across customer relationships. On the other hand, customer segmentation is the process of dividing customers into homogeneous groups, where customers within each group are similar to each other and share common attributes. In order to analyze CRM sophisticated data, one needs to explore the data by using different aspects. Data mining is one the newly emerged technology in this endeavor. Data mining finds and extracts hidden knowledge in corporate data warehouses.

In this study the applicability of clustering data mining technique to support CRM activities for ERCA has been explored using the CRISP-DM process model approach. After understanding business objective of the authority, different characteristics of the ERCA customers' data were collected from the ERCA's ASYCUDA++ database. Once the customers' data were collected, the necessary data preparation steps were conducted and finally the dataset consisting of 65535 records/instances were employed to develop a clustering model.

To segment customers, the k-means clustering algorithm was employed. During the cluster modeling, different experiments were conducted using different cluster numbers (k=4, 5, 6) with different seed values (10, 100, 1000). Then, the one which performed the best was selected. Thus, the model where k=4 and seed size 100 had shown better clustering performance. It enables to cluster the authority's customers into dissimilar clusters of high, medium and low value customer groups with minimum iteration value of 6 and minimum sum of square error (SSE) within clusters value of 2142.82.

The results of this study have shown that the data mining techniques are valuable for customer segmentation. Hence future research directions are pointed out to come up with an applicable other data mining technique

CHAPTER ONE

INTRODUCTION

1.1 Background

Now a days, most of the organizations have been changing their approach from product centric to customer centric. In customer-centric approach, an organization considers its customers as an asset and focuses on acquiring, attracting and retaining many customers [56]. The attention of many service-providing industries has been targeting on customer orientation and awareness creation. While competition is increasing, there is a need to have a better understanding on customers' behavior and to quickly respond to their individual needs and preferences [44]. Thus, the objective of customer oriented industries have come up with increasing the share of wallet for each individual customer and save costs by focusing on more targeted promotions and awards to increase the profitability of the company and loyalty of the customers [11].

Customer Segmentation is the process of dividing customers into homogeneous groups, where customers within each group are similar to each other and share common attributes [20]. The segmentation of customers into unique groups is often based on multi-dimensional customer information such as observed customer purchase potential, product usage behavior, demographic characteristics, product/ service preferences, attitudes and needs [35]. Customers are the most important property of an organization: customer is considered as a king. Organizations can't survive without satisfying and retaining their customers in any business prospects. As far as customer is concerned, the main goal or objective of every organization is to understand each customer's: behavior, preferences, needs and it is easier to make business with them [21][55].

According to Almotairi [2], Customer Relationship Management (CRM) is a newly emerged concept in the fields of Information Technology (IT) and today's competitive business that aims to strengthen the relationships between an organization and its customers. CRM could be stated as, managing all interactions of customers information and organizations to all stages of organizations relationships with its customers [21]. In addition, CRM is a process by which an organization maximizes its customers satisfaction in an effort to increase loyalty and retain customers over their lifetimes [6][28].

Customer Relationship Management (CRM) organizes resources around customer relationships rather than product groups and promotes activities that maximizes the value of lifetime relationships [41]. From an operational standpoint of view, CRM links business processes across the supply chain from back office functions through all touch points, enabling continuity and consistency across customer relationships [41]. As far as Customer Relationship is concerned for an organization, the main theme of Customer Relationship Management (CRM) is knowing well about its customers will enable to serve them better and keep them loyal forever [32].

According to Ngai and Chau [39], the application of data mining tools in CRM is an emerging trend in today's global economy. Understanding and analyzing customers' behaviors and characteristics is the foundation of the development of a competitive CRM strategy, so as to attract and retain potential customers and maximize customer value. Applying appropriate data mining tools, which are good for extracting and identifying useful information from huge customer databases are the best supporting tools for making different CRM decisions [39]. Many organizations are using data mining that helps to manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers [6][49][56].

Currently, the technology of databases have been becoming more and more advanced where huge amount of data is required to be stored in the databases in turn the accumulation of big data and information overload [45]. So, this wealth of information hidden in those datasets has to be realized by business people as a useful tool for making business strategic decisions. To extract the targeted or required useful information from these sophisticated datasets, data mining has attracted more attention as it promises to extract valuable information or knowledge from the raw data that businesses can use to increase their profitability through a profitable decision-making process [26].

Data mining is the process of extracting interesting patterns from large amounts of data and also about solving problems by analyzing data already present in datasets [25]. It provides tools for automated learning from historical data and developing models to predict future trends and behaviors. The essence of data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. In general, data mining tasks can be classified into two categories: descriptive and predictive [45]. Descriptive data mining

models extract useful patterns that can describe the mined or extracted data and explore its properties and characterize the general properties of the data in the database. Predictive data mining tasks perform inference on the current data in order to make predictions, in other words, predict unknown values or future trends or behaviors depending on other variables or historical values presented in the mined or extracted database [49].

Classification is a data mining technique which maps the target data into the predefined groups or classes. Classification is known as a supervised learning technique because the class label information is predefined before the examination of the target data have been employed [30]. Classification is mainly used for predictive model which makes prediction about unknown data values by using the known values. Even though, classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training patterns, that the classifier uses to model each group [25].

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster [26]. Clustering is known as unsupervised learning because the class label information is not known in advance. For this reason, clustering is considered as a form of learning by observation, rather than learning by examples. Unlike classification, clustering do not rely on predefined classes and class-labeled training examples rather applied on the whole dataset [4]. The results of cluster analysis may contribute to the definition of a formal classification scheme [25].

1.3 Statement of the problem

In today's competitive world, organizations that provide good quality service to customers have to be competitive to ensure its supremacy over less capable competitors and guarantee its survival. Each organization has to strive for quality of service by taking into consideration its own special features, the demands of its customers and the emerging dynamic technological influences.

The Ethiopian government is striving to cover its annual expenditure from local financial resources and ERCA has been given the major responsibility in this endeavor. ERCA is playing a significant role in supporting the growth and development plan of the country. In practice, Ethiopian Revenues and Customs Authority has been facing many challenges. One of the main challenges of the authority is customer dissatisfaction [12]. Poor customer handling and service delivery tends to customer dissatisfaction and also this leads to abuse of rules and regulations of the organization and fosters engagement on the act of smuggling and fraud. ERCA is the victims of such types of practices. To tackle this problem, the authority has been using different mechanisms to satisfy its customers but because of lack of properly established model to segment customers, the problem still exists.

Currently, ERCA has been using statistical analysis and assessment techniques to identify high and low valued customers in terms of revenue generation and transaction volume. The Authority checks and controls whether the customers discharge their responsibility in accordance with the rules and regulations or not, that has been done based on the data captured in the two huge databases: domestic tax and customs databases. However, the existing techniques are not effective and efficient to treat customers according to their characteristics. The importers and exporters have been encountering different problems while processing import and export activities so as to deliver goods to domestic and international market on time. Due to this reason, the customers viewed ERCA's procedures as an obstacle that is blocker for the movement of international trade facilitation [6][47].

The Ethiopian Revenues and Customs Authority (ERCA) collects duties and taxes yearly in billions of birr from its customers. The authority has planned to register or incorporate all business enterprises and individuals those are engaged in the business sectors or activities in the tax network. Even though, ERCA has been using a well organized customers databases, there is no such an integrated model being applied to segment customers so as to maximize access to service delivery and revenue generation. In addition, the Authority doesn't have well-organized set of rules or procedures that are used for segmenting customers who are engaged in the import and export activities based on their revenue generation, transaction volume and their characteristics to segment customers as high or low valued customers. As a result, the authority

is unable to collect the expected revenue and also couldn't control the act of smuggling and fraud [6][47].

There are studies conducted in the area of application of data mining techniques on customer segmentation and customer relationship management: Henok [28] and Denekew [13] studied on application of data mining to support CRM at Ethiopian Airlines. Belachew [5] conducted a study on application of data mining for customer segmentation and prediction at BG MFI. Tenkir [47] studied on customs management system on ERCA to assess the customs valuation system currently applied in Ethiopia and to identify the problems in the practical process which is more of focused on the functional process. Dandi [11] conducted a study on application of data mining for customer segmentation in insurance business: the case of Ethiopian Insurance Corporation. Hailemariam [23] worked on application of data mining techniques to customer profile analysis in the Ethiopian Electric Power Corporation. Belete [6] studied on knowledge discovery for effective Customer Segmentation: the case of Ethiopian Revenues and customs Authority.

All the studies mentioned above are focused mainly on classification data mining techniques to classify customers. Belete [6] tried to reveal customer classification in his research work at ERCA and he identified the gaps; data integration, attribute selection limitation, algorithm selection limitation particularly clustering algorithms, less data complexity and sum of squared-error.

Thus, this study focused on the application of data mining techniques on grouping of customers based on the revenue generation and transaction volume by using the available datasets which is found in the customers' database of ERCA and to generate a model that identifies potential and low valued customers. In the data mining process, the research mainly focused on clustering algorithms which were not fully employed by the previous researches. The study also attempted to address the gaps identified and recommendations pointed out by Belete [6].

This study attempts to answer the following research questions:

- *What are the most significant attributes that help to segment ERCA's customers?*

- *Which data mining technique is more appropriate to identify determining factors for Customer Segmentation?*
- *Which algorithm best suits for the selected attributes to create a clustering model?*
- *What are the criteria used or employed to evaluate the model?*

1.4 Objective of the study

1.4.1 General Objective

The general objective of this study is to mine customs data, segment potential customers of Ethiopian revenues and Customs Authority and to develop a clustering model for the better Customer Relationship Management.

1.4.2 Specific Objective

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

- To review related literatures and previous works in the area to have better understanding about the work and to know others contributions in the area.
- To identify sources of ERCA's customer data that are required for the customer segmentation
- To construct ERCA target dataset used for data mining tools following major data preparation and pre-processing steps.
- To select appropriate data mining tool and algorithms to be used for developing the customer clustering model.
- To conduct experimentation to evaluate the accuracy of the model.
- To draw recommendation based on the findings of the study.

1.5 Scope of the study

The scope of this research is restricted only on customers' data who in engaged import/export activities and tax payers transactions via Ethiopian Revenues and Customs Authority customers database. The research is also restricted on building a data mining model for clustering or segmenting the customers, interpreting the resulting clusters or segments and developing a clustering model.

1.6 Significance of the study

The study could be helpful to support the routine and strategic decision making tasks of the authority and it could open an opportunity to implement an appropriate customer relationship management strategy and to provide attractive customer service through the right channel, at the right time and to the right customer with each customer contact. In addition, the study will be vital for the organization to properly cluster or segment the customers and treat them accordingly and to have improved service delivery. If the organization implements proper Customer Relationship Management with the help of the newly emerging technologies, the satisfaction of the customers will be increased and the authority's revenue collection volume will be maximized. As a result, the challenges those are currently facing for the authority like smuggling and fraud will be minimized.

The contribution of the result this study will open an opportunity for the authority to revise its existing customer related procedures and to integrate information technology, customer relationship and management.

1.7 Organization of the Thesis

This research paper is organized into five chapters. The first chapter deals with the general overview of the study including background of the study, background of Ethiopian Revenues and Customs Authority (ERCA), statement the problem, the general and specific objectives of the study, the scope of the study and significance of the study.

Chapter two, discusses literature review mainly on customer segmentation, customer relationship management and knowledge discover in database (KDD) or data mining technology. In this chapter, different data mining techniques, such as clustering and classification with their respective algorithms are reviewed.

Chapter Three, discusses methodology of the study; Introduction, research design, CRISP-DM process model. In this chapter, the six CRISP-DM process models; business understanding, data understanding, data preparation, modeling, evaluation and deployment have been presented.

Chapter Four, emphasizes on the experimentation phase of the study, which mainly discusses different stages of clustering technique experimentation processes to interprets the results of the clustering experiments and build a clustering model.

Finally, Chapter Five dictates conclusion and recommendations; this chapter presents the conclusion of the result of the study and provides recommendation forwarded for further research and adjustments in the organization on the ground of the research results.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Currently, a new business culture has been developing significantly. 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 this change [44]. The concepts of mass production and mass marketing, first created during the Industrial Revolution era, are being replaced by new ideas in which customer relationships are the central business issues. Now a days, Organizations are concerned with increasing customer value through analysis of the customer lifecycle. The tools and technologies of data warehousing, data mining, and other customer relationship management (CRM) techniques brought new opportunities for businesses to work on the concepts of relationship marketing. So, the emerging of data mining technology has created an opportunity for the implementation of customer relationship management [12][20].

In recent years, the rapid development in digital data acquisition and storage technology has resulted in the growth of huge databases and the amount of data kept in databases is growing in alarming rate. At the same time, the users of these data are requiring to extract or mine targeted and useful information from these huge databases and storage. In doing so, simple Structured Query Language (SQL) is not adequate to support these increased demands. Information retrieval (IR) is also not adequate anymore for decision-making. With the huge amounts of data collections, there is a need for new technologies to extract hidden knowledge from huge datasets. These needs are automatic summarization of data, extraction of the core useful information and the discovery of patterns in raw data [25]. In this regards, one of the newly emerging technologies is data mining which is also known as knowledge discovery in database (KDD) [6][20][26].

In the era of knowledge society, information overload and big data are becoming a challenging phenomena. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, novel, valid patterns and relationships in large datasets. These tools can include statistical models, mathematical algorithms and machine learning methods. Moreover,

data mining also functions analysis and prediction based on algorithms such as k-means clustering, expected maximization, apriori, neural networks and decision trees [5][25][49].

2.2 Customer Relation Management

Overview

According to Gupta and Aggarwal [21], Customer Relationship Management (CRM) refers to the methodologies and tools to manage customer relationships in an organized approach. Information communication technology (ICT) in particular, the World Wide Web, is creating opportunity to build better relationships with customers than has been previously possible in the offline world.

According to Microsoft, Customer Relationship Management (CRM) is "a customer-focused business strategy designed to optimize revenue, profitability, and customer loyalty. By implementing a CRM strategy, an organization can improve the business processes and technology solutions around selling, marketing and servicing functions across all customer touch-points (for example: Web, e-mail, phone, fax, in-person)". The overall objective of CRM applications is to attract, retain and manage a firm's profitable ("right") customers.

Customer Relationship Management (CRM) is an important business approach that focuses on marketing to each customer individually rather than marketing to a mass of people or firms. It is a customer centric doing business rather than a simple marketing strategy [44]. CRM involves all of the corporate functions like marketing, manufacturing, customer services, field sales and field service who are required to contact customers directly or indirectly. Its objective is to return to the world of personal marketing to increase profitability, revenue and customer satisfaction [20].

In traditional marketing strategies, the main concern was to increase the volume of transactions between seller and buyer, based on the four Ps (price, product, promotion, and place) to increase market share. Performance of marketing strategies was measured through volume of transactions. Now a days, CRM is a business strategy that goes beyond increasing transaction volume. It is primarily a strategic business process issue rather than a technical issue [20]. It is an era of company loyalty to the customer in order to obtain customer loyalty to the company. The reason is that consumers are more knowledgeable than ever before and because the customer

is more knowledgeable, companies must be faster, more agile, and more creative than a few years ago [5][20].

Communicating, valuing changes in organization culture through rewards and ensuring that all employees understand the importance of adopting customer-centric behaviors as they strive to develop stronger customer relationships is important for CRM [5].

According to Rodpysh, Aghai and Majdi [43], four lifecycles of Customer Relationship Management are mentioned below:

Customer Identification

Customer relationship management (CRM) begins with the identification of customers. Customer identification refers to the collection of information about customers through marketing channels, transactions, and interactions over time for the purpose of serving or providing value to the customer [6][39][43].

Customer Attraction

The second phase is discussion with customer relationship management. Customers' satisfaction for the expectations and perception is essential to protect customers. And the organization preserves elements that includes marketing person to person pointing to the marketing person who analyze with support. Loyalty programs include support activities with the long-term relationships with customers especially analyze customers shunned, ranking credit, the quality of services or satisfaction from the loyalty programs [6][39][43].

Customer Retention

This third phase is the main issue of Customer Relationship Management discussion. Satisfaction of customer expectations with customer perceptions of satisfaction is the basic condition for maintaining customer retention and marketing person to person marketing analysis is supported. Detect and predict changes in customer loyalty programs; including activities aimed at supporting long term relationships with customers turn away; especially analysis, credit rating, service quality or satisfaction from the loyalty program [6][39][43]

Customer Development

This phase includes a plurality of transactions, transaction value and customer profitability. Customer development involves the consistent expansion of transaction intensity, transaction value and individual customer profitability. The elements of customer development includes analysis customer lifetime value, cross selling, up selling. Customer lifetime value analysis is defined as the prediction of the total net income a company can expect from a customer . Up selling and cross selling refers to promoting activities which aim at augmenting the number of associated or closely related services that a customer uses within a firm. [6][39][43].

2.2.1 Principles and tasks of CRM

According to Gray and Byun [20], the overall processes and applications of CRM are based on the following basic principles.

Treat customer individually: Remember customers and treat them individually. CRM is based on philosophy of personalization. Personalization means the "content and services to customer should be designed based on customer preferences and behavior" According to Hagen cited in Gray and Byun [20], personalization creates convenience to the customer and increases the cost of changing vendors prosperous.

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

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

According to Peppers, et al. (2000) cited in (Gray and Byun [20], the four basic tasks that are required to achieve the basic goals of CRM are:

Customer Identification: To serve or provide value to the customer, the company must know or identify the customer through marketing channels, transactions and interactions over time.

Customer Differentiation: Each customer has its own lifetime value from the company's point of view and each customer imposes unique demands and requirements for the company.

Customer Interaction: Customer demands change over time. From a CRM perspective, the customer's long-term profitability and relationship to the company is important. Therefore, the company needs to learn about the customer continually. Keeping track of customer behavior and needs is an important task of a CRM program.

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.

2.3 Customer Segmentation

Overview

According to Jansen [30], Customer segmentation is a preparation step for classifying each customer according to the customer groups that have been defined and it is essential to cope with today's dynamically fragmenting consumer marketplace. Segmenting means putting the population in to segments or groups according to their affinity or similar characteristics.

According to Upadhyay, Vidhani and Dadhich [51], segmentation is the process of making smart decisions about how various different attributes which are a part of the customer profile relate to give important patterns and groups or segments among the customers.

Customer segmentation is the practice of classifying organization's customer base into distinct groups and this separation of customers into unique groups is often based on multidimensional customer information such as; observed customer purchase and product usage behavior, demographic and life stage characteristics, or even self-reported product / service preferences, needs and attitudes [36].

As Ziafat and Shakeri [55] stated, according to the segmentation criteria used, there are various types of segmentation methods. Particularly, customers can be segmented according to their value, socio-demographic and life-stage information, their behavioral need/attitudinal and loyalty characteristics. And the type of segmentation used depends on the specific business objective and

the organization's target. Different criteria and segmentation methods are appropriate for different situations and business objectives.

The goal of segmentation is to know the customer better and to apply that knowledge to increase profitability, reduce operational cost, enhance customer service and customer satisfaction. Segmentation can provide a multidimensional view of the customer for better treatment strategy [6].

2.3.1 Application of Customer Segmentation

According to Mike [36], different applications of customer segmentation are described as follows:

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 of driving customer investment decisions has been the expected return on that investment. Segmentation is not only helps to determine how much to invest in a customer segment, but also how to spend it [6][36].

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 [6][36]

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 maximize program effectiveness. Segmentation becomes the focus, supporting program development and ongoing test and learns activities [6][36].

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). According to Mike [36], 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 [6].

2.4 Overview of Data Mining Technology

Data mining is the analysis of large observational data sets to find unsuspected or unknown relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. In addition, data mining is the technique or mechanism of finding useful information that enables us to find the relationship among the huge data collection which will be very helpful for making prediction about the future by using different algorithms. The relationships and summaries derived through a data mining process are often referred to as models or patterns [5][30].

According to Han and Kamber [25], data mining refers to extracting or "mining" knowledge from large amounts of data sets. The term is actually a misnomer. Notice that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term may not reflect the emphasis on mining from large amounts of data. Thus, such a misnomer that carries both "data" and "mining" became a popular choice and have been attracting the attentions' of the scholars and organizations.

Data mining is the process of discovering insightful, interesting and novel patterns or hidden knowledge, as well as descriptive, understandable and predictive models from large-scale datasets [56]. It is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions and descriptions [49].

Data mining refers to extracting or mining knowledge from large amounts of datasets and involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing and spatial data analysis. Data mining applications can use a variety of parameters to examine the data. Those include association, sequence or path analysis, classification, clustering and forecasting [1].

According to Han and Kamber [25], the major reason that data mining has attracted more attention of the organizations in the information industry in recent years is due to the availability of huge amounts of data and the progressive need for changing such data into useful information and knowledge. The authors also mentioned that data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of functionalities such as data collection, database creation and data management (including data storage, retrieval and database transaction processing).

As a discipline of business intelligence, data mining got its role with the development of computing technologies in 1990's, not only in the storage segment, but also in the developing modern systems for relational database system [11]. Now a days, the number of organizations relied their operations on processing of huge amounts of data in real time is increasing. Such data are a large information potential that needs to be liberated one way or another . One of the latest huge data extracting technologies is data mining. [8]

2.5 Historical Development of Data Mining and Knowledge Discovery

The idea of today's data mining techniques has been started since 1950s when the work of mathematicians, logicians and computer scientists combined to create artificial intelligence (AI) and machine learning [13]. Its origin lied on the first storage of data on computers continues with improvements in data access, until today technology allows users to navigate through data in real time. Data mining techniques are the result of a long research and product development process.

According to Rygielski and Yen [20], the four evolutionary stages of data mining from user's perspective are mentioned bellow:

1960s the data collection stage: In this first stage, individual sites had collected data used to make simple calculations such as summations or averages. Information generated at this step answered business questions related to figures derived from data collection sites, such as total revenue or average total revenue over a period of time. Specific application programs were created for collecting data and calculations [20][24].

1980s the data access stage: In this second step, databases are created to store data in a structured format. At this stage, company-wide policies for data collection and reporting of management information were formulated. Because every business unit should do the accepted thing to specific requirements or formats. Once individual figures were known, questions that search the performance of aggregated sites could be asked easily [20][24].

1990s the data navigation stage: In this third stage, thanks to multi-dimensional databases, businesses in this stage could obtain either a global view or drill down to a particular site for comparisons with its peers. The introduction of OLAP, warehouses and multi-dimensional database enhances business organizations to improve their working strategy [20][24].

2000s the data mining stage: At this stage, since, on-line analytic tools provided real-time feedback and information exchange with collaborating business units (Data Mining), this capability is useful when sales representatives or customer service persons need to retrieve customer information on-line and respond to questions on a real-time basis. Information systems can query past data up to and including the current level of business. Often businesses need to make strategic decisions or implement new policies that better satisfy their customers than previous. For example, grocery stores redesign their layout to promote more impulse purchasing. Telephone companies establish new price structures to attract customers into placing more calls. Both tasks require understanding of past customer consumption behavior data in order to identify patterns for making strategic decision and data mining is particularly suited or appropriate to this purpose [20][24].

Nowadays, data mining is being used by a wide variety of industries and sectors including retail, medical, telecommunications, scientific, financial, pharmaceutical, marketing, Internet-based companies, service giving organizations etc [23][18].

2.6 Data Mining Tasks

The tasks of data mining can be modeled as either Predictive or Descriptive in nature [15]. A Predictive model makes a prediction about values of data using known results found from different data while the Descriptive model identifies patterns or relationships in the data and explores the properties of the data examined. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties.[25] Descriptive data mining focuses on finding patterns describing the data that can be interpreted by humans and produces new non trivial useful information based on the available datasets. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. Predictive model data mining tasks include classification, prediction, regression and time series analysis. The Descriptive model task encompasses methods such as clustering, summarizations, association rules, and sequence analysis [45].

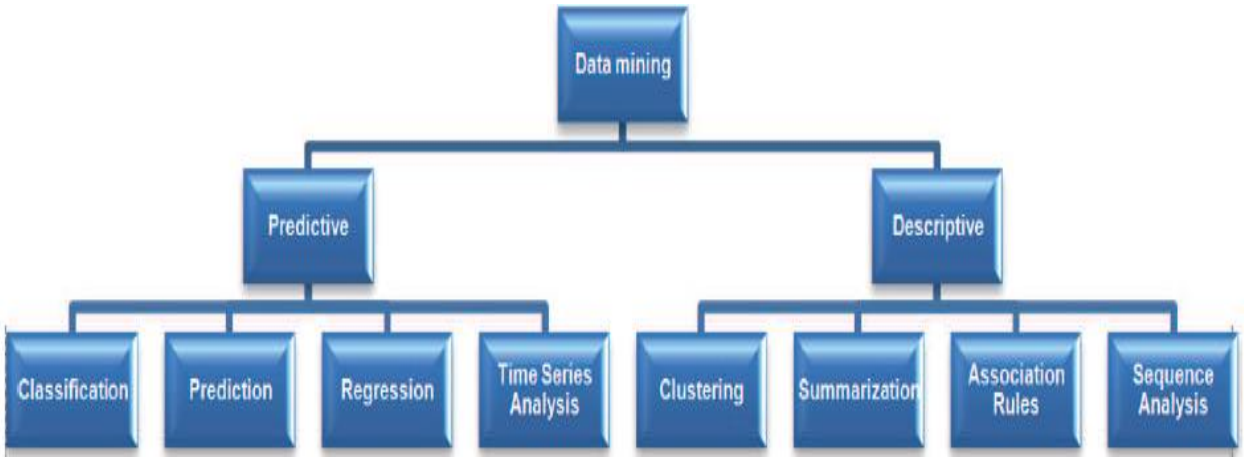

Fig. 2.1 Data mining tasks and models [45].

2.6.1 Predictive Modeling

This model permits the value of one variable to be predicted from the known values of other variables. Predictive modeling is a supervised learning functions and the event in which a model is made or chosen to predict an outcome accurately. Businesses have been grown and science is a

lot more complex, therefore data mining requires more computerized method of doing this operation. Software and computer applications have made data mining seem an effortless job, because it does not require a direct hands-on approach [25]. A predictive model makes a prediction about values of data using known results found from different datasets [5].

In the predictive modeling, classification and regression represent the largest part of problems to which data mining is applied today and creating model to predict class membership (classification) or value (regression). Classification is used to predict to what group a case belongs to the class variables and regression is used to predict a value of a given continuous valued variables based on the values of other variables by assuming a linear or non-linear model of dependency [26]. Logistic regression is used for predicting a binary variable. It is generation of linear regression; the binary dependent variable cannot be modeled directly by linear regression. Logistic regression is a classification tool while used to predict categorical variables such as whether an individual is likely to purchase or not, and a regression tool while used to predict continuous variables such as the probability that an individual will make a purchase [25]

2.6.1.1 Classification

Classification is the process of finding a model, which describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown [25]. The derived model is based on the analysis of a set of training data (i.e. data objects whose class label is known). Classification problems aim to identify the characteristics that indicate the group to which each case belongs to each other [49]. This pattern can be used both to understand the existing data and to predict how new instances will behave. There are different algorithms that are used for classification purpose such as decision tree, neural network, naïve bayes, etc [46].

Decision Tree

According to Hajizadeh, Ardakani and Shahrabi [22], decision trees are powerful and popular tools for classification and prediction. In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. Decision trees are also useful for exploring data to gaining insight into the relationships of a large number of candidate input variable to the target variable. Decision trees are a powerful first step in modeling process even when building the final model using some other techniques.

Hajizadeh, Ardakani and Shahrabi [22] also described decision tree as, it is a model which consists of a set of rules for dividing a large heterogeneous population into smaller more homogeneous groups with respect to a particular target variable. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories or to classify the record by assigning it to the most likely class.

According to Han and Kamber [25], decision tree induction is the learning of decision trees from class-labeled training instances. A decision tree is a flowchart like tree structure, where each internal node (non leaf node) denotes a test on an attribute. And each branch represents an outcome of the test and each leaf node (terminal node) holds a class label. The topmost node in a tree is the root node.

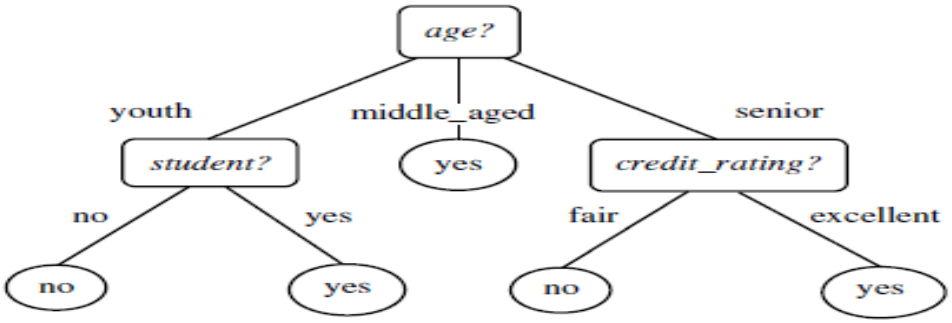

Fig. 2.2 Decision Tree flowchart [25].

According to Two Crows Corporation [49], decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), J48, JRip, Quest, and C5.0.

Decision Tree Pruning method

According to Han and Kamber [25], decision tree models pruning method is a means of solution to overcome the over fittings like outliers and noises in the data set. Pruning can increase the degree of relevance of the model and its applicability. It uses statistical techniques to remove irrelevant branches of the tree model. One can choose pruning trees before (pre-pruning) or after

the model is built. The former works by altering the parameter of the run information to stop splitting the training subset at a given node while the later (post-pruning) eliminate the branches of sub trees from full grown tree. Hence, the branches are replaced and labeled with leaf of most frequent class sub-tree being replaced. Post pruning is mostly used method than the pre pruning to build decision tree model. The pruned tree of Decision tree model produces more coherently to classify the test data [26].

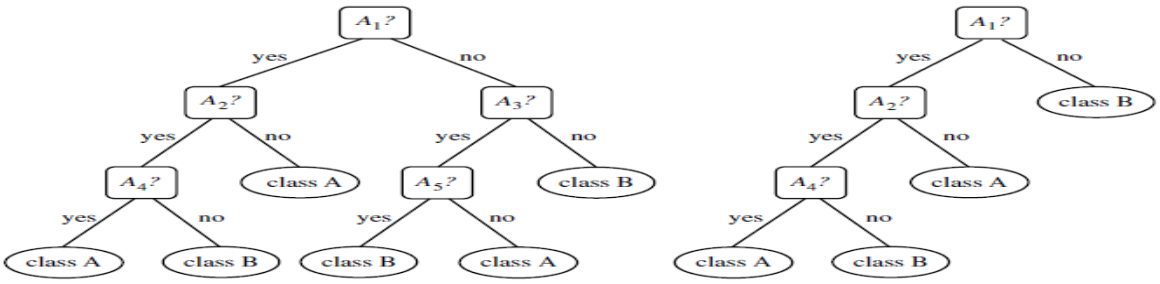

Fig. 2.3 An un pruned and a pruned decision tree [25].

Naive Bayes Algorithm

Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [40]. Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based, is rarely true in real world applications [27].

According to Han and Kamber [25], the Naïve Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. Each tuple is represented

   by an n-dimensional attribute vector, X = (x1 , x2,….xn), depicting **n** measurements made on the tuple from **n** attributes, respectively, A1,A2,….. An.

2. Suppose that there are **m** classes, C1, C2,……Cm. Given a tuple, **X**, the classifier will predict that **X** belongs to the class having the highest posterior probability, conditioned on **X**. That is, the Naïve Bayesian classifier predicts that tuple **X** belongs to the class **Ci** if and only if

   P(Ci |X) > P(Cj |X) for $1 \leq j \leq m, j \neq i$.

   Thus we maximize P(Ci |X). The class Ci for which P(Ci |X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem $P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$

3. As P(X) is constant for all classes, only P(X| Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = …. = P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = |Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in D.

4. Given datasets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the Naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that are no dependence relationships among the attributes). Thus,

$$P(\boldsymbol{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

We can easily estimate the probabilities P(x1 | Ci), P(x2 | Ci), ……., P(xn | Ci) from the training tuples. Recall that here xk refers to the value of attributes Ak for tuple X. For each attribute, we look at whether the attribute is categorical or continuous.

5. In order to predicate the class label of X, P(X| Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is class Ci if and only if

$$P(X|Ci)P(Ci) > P(X|Cj)P(Cj) \text{ for } 1 \leq j \leq m, j \neq i.$$

In other words, the predicted class label is the class Ci for which P(X | Ci)P(Ci) is maximum.

Neural Network Algorithm

As Han and Kamber [25] stated, the neural networks field was originally stimulated by psychologists and neurobiologists who sought to develop and test computational analogs of neurons. It is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input instances. Neural networks involve long training times to train the input instances [26]. They require a number of parameters that are typically best determined empirically such as the network topology or "structure."

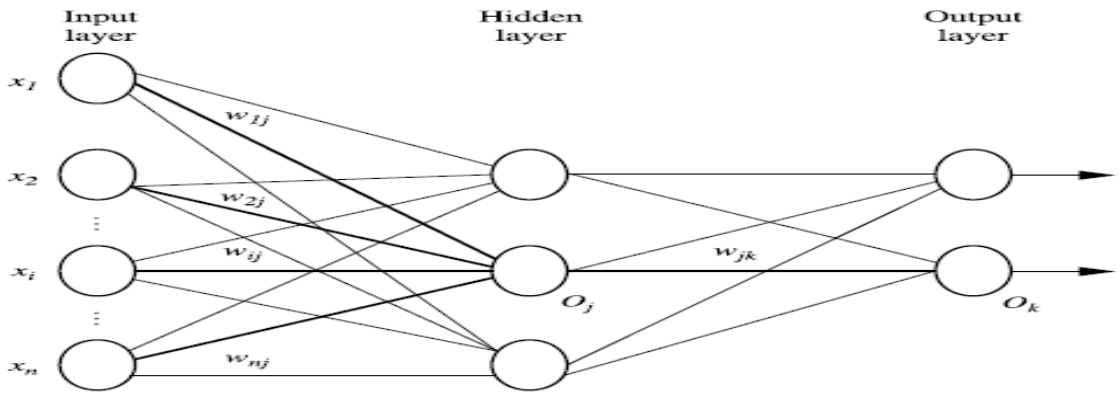

Fig 2.4 Multilayer feed-forward neural network [25].

According to Hajizadeh, Ardakani and Shahrabi [22], neural network is a computational technique that benefits from techniques similar to ones employed in the human brain. It is designed to mimic or imitate the ability of the human brain to process data and information and comprehend patterns. It imitates the structure and operations of the three dimensional lattice of network among brain cells. Technology is inspired by the architecture of the human brain, which uses many simple processing elements operating in parallel to obtain high computation rates. Similarly, the neural network is composed of many simple processing elements or neurons operating in parallel whose functions are determined by network structure, connection strengths and the processing performed at computing elements or nodes.

According to Han and Kamber [25], neural network has the following advantages:

- Neural Networks are well suited for continuous-valued inputs and outputs, unlike most decision tree algorithms.
- Neural Networks have been successful on a wide array of real-world data, including hand written character recognition, pathology and laboratory medicine, and training a computer to pronounce English text.
- Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process. In addition, several techniques have been recently developed for rule extraction from trained neural networks.

Even if neural networks have a number of advantages, they have been criticized for their poor interpretability. For example, it is difficult for humans to interpret the symbolic meaning behind the learned weights and of "hidden units" in the network. These features initially made neural networks less desirable for data mining [26].

2.8.1.2 Prediction

Prediction forecasts the missing or unavailable values rather than class label, numeric prediction and continuous value functions. It is one of the examples of predictive modeling and it can be viewed as the building and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of an attribute [25]. According to Siraj and Abdoulha [45], the purpose of Prediction model is to determine the future outcome rather than current behavior.

In predictive models, the values or classes that have been predicted are called the dependent or target variables and the values used to make the prediction are called the predictor or independent variables. Prediction has attracted considerable attentions by giving the potential implications of successful forecasting in a business context. Once a classification model is built based on a training set, the class label of an object can be forecasted based on the attribute values of the object and the attribute values of the classes. Prediction is more often referred to the forecast of missing numerical values, or increase/decrease trends in time related data. The major idea in prediction is to use a large number of past values to consider probable future values [26].

2.6.2 Descriptive Modeling

According to Siraj and Abdoulha [45], descriptive data mining modeling is unsupervised learning functions. These functions do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness etc of the data. It presents the main features of the data or a summary of the data. Data randomly generated from a "good" descriptive model will have the same characteristics as the real data [5].

Descriptive modeling techniques; such as summarization, association rule, sequence analysis and clustering which produce classes (or categories) are not known in advance. Summarization maps data into subsets with associated simple descriptions. Basic statistics such as mean, standard deviation, variance, mode and median can be used as summarization approach [15]. Association Rule is a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. The investigation of relationships between items over a period of time is also often referred to as sequence analysis. Sequence Analysis is used to determine sequential patterns in data. The patterns in the data sets are based on time sequence of actions, and they are similar to association data, however the relationship is based on time. In market basket analysis, the items are to be purchased at the same time. On the other hand, for sequence analysis the items are purchased over time in some order [15].

2.6.2.1 Clustering

According to Hajizadeh, Ardakani and Shahrabi [22], clustering is a tool for data analysis, which solves classification problems. Its objective is to distribute cases (people, objects, events etc.) into groups, so that the degree of association can be strong between members of the same cluster and weak between members of different clusters. In clustering, there is no pre classified data and no distinction between independent and dependent variables. Instead, clustering algorithms search for groups of records (the clusters composed of records similar to each other). The algorithms discover these similarities.

Clustering divides a database into different groups and the goal of clustering is to find groups that are very different from each other and whose members are very similar to each other [49]. Techniques for creating clusters include partitioning (often using the k-means algorithm) and hierarchical methods (which group objects into a tree of clusters). In clustering, there are no

predefined classes. The records are grouped together on the basis of self-similarity. It is up to the user to determine what meaning, if any, to attach to the resulting clusters [22][25].

K-means Algorithm

K-means clustering algorithm is unsupervised learning functions which is one of the most widely used clustering algorithms in various disciplines, particularly for large datasets. K-means starts with an initial cluster solution which is updated and adjusted until no further refinement is possible. Each iteration refines the solution by reducing the within-cluster variation. In the K-means algorithms, the "k" represents the number of *k* clusters to be formed which comes from the fact that users should specify in advance. The "means" part of the name refers to the fact that each cluster is represented by the mean of its records on the clustering fields, a point is referred to as the cluster central point or centroid or cluster center [34].

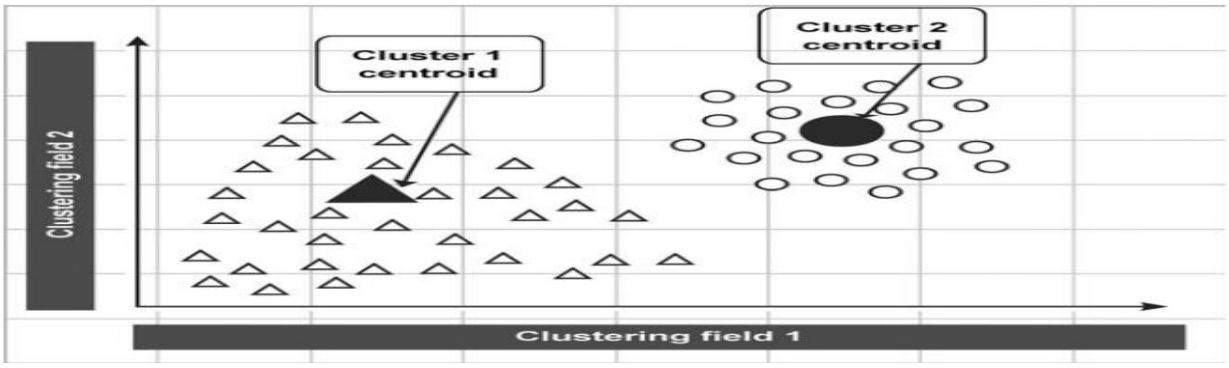

Fig. 2.5 Graphical representation of the cluster centers or centroids [34].

K-means algorithm uses the Euclidean distance measure and iteratively assigns each record in the derived clusters. It is a very quick algorithm since it does not need to calculate the distances between all pairs of records. Clusters are refined through an iterative procedure during which records move between the clusters until the procedure becomes stable [34].

Hierarchical Clustering Methods

A hierarchical clustering technique is unsupervised functions that works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be classified as either *agglomerative* or *divisive*, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down approach. According to Han & Kamber [25], the quality of a pure hierarchical clustering method suffers from its incapability to perform adjustment once a merge or split

decision has been executed. That means, if once a particular merge or split is done, the method cannot backtrack and correct it [26].

Agglomerative hierarchical clustering: This clustering method is a bottom-up (merging) approach that starts by placing each object in its own cluster and then merges these atomic clusters into a bigger cluster. The process continues until all of the objects are arrived at a single cluster or until certain termination conditions are satisfied [25].

Divisive hierarchical clustering: This clustering method is top-down (splitting) approach that performs the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster and subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own. The process continues until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold [25].

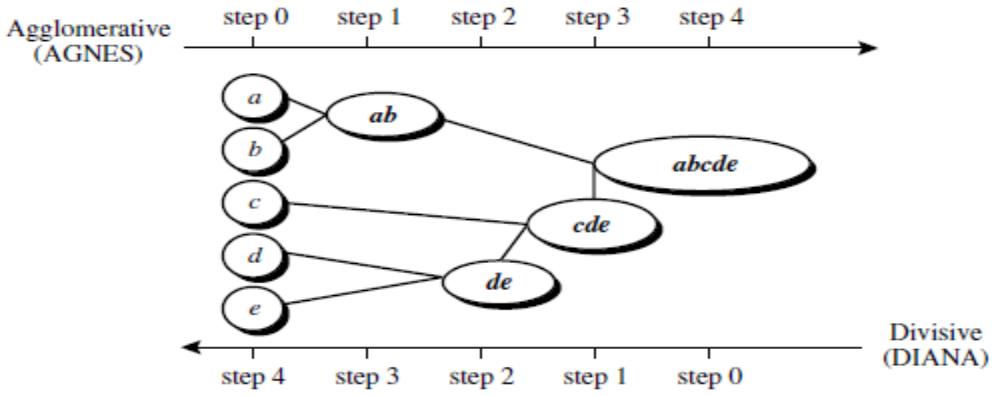

Fig. 2.6 Agglomerative and divisive hierarchical clustering [26].

Expected Maximization

EM (Expectation-Maximization) clustering algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. The algorithm can be viewed as an extension of the *k*-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. In other words, there are no strict boundaries between clusters. Therefore, new means are computed based on weighted measures [25].

According to Han and Kamber [26], EM starts with an initial estimate or "guess" of the parameters of the mixture model which is collectively referred to as the *parameter vector*. It rescores the objects iteratively against the mixture density produced by the parameter vector. Then the rescored objects are used to update the parameter estimates. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster. The algorithm is described as follows:

Make an initial guess of the parameter vector: This involves randomly selecting *k* objects to represent the cluster means or centers (as in *k*-means partitioning), as well as making guesses for the additional parameters [26].

Iteratively refine the parameters (or clusters) based on the following two steps:

a) **Expectation Step:** Assign each object ***xi*** to cluster *Ck* with the probability

$$P(\boldsymbol{x_i} \in C_k) = p(C_k|\boldsymbol{x_i}) = \frac{p(C_k)p(\boldsymbol{x_i}|C_k)}{p(\boldsymbol{x_i})},$$

where $p(\boldsymbol{xi}|Ck) = N(mk, Ek(\boldsymbol{xi}))$ follows the normal (i.e., Gaussian) distribution around mean, *mk*, with expectation, *Ek*. In other words, this step calculates the probability of cluster membership of object ***xi***, for each of the clusters. These probabilities are the "expected" cluster memberships for object ***xi***.

b) **Maximization Step**: Use the probability estimates from above to re-estimate (or refine) the model parameters. For example,

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x_i}P(\boldsymbol{x_i} \in C_k)}{\sum_j P(\boldsymbol{x_i} \in C_j)}.$$

This step is the "maximization" of the likelihood of the distributions given the data.

DBSCAN Clustering Methods

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of *density-connected* points. Density-based clustering methods have been developed to discover clusters with arbitrary shape. DBSCAN grows clusters according to a density-based connectivity analysis [25][26].

DBSCAN searches for clusters by checking the e-neighborhood of each point in the database. If the e-neighborhood of a point ***p*** contains more than *MinPts*, a new cluster with ***p*** as a core object is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters. The process terminates when no new point can be added to any cluster [25][26].

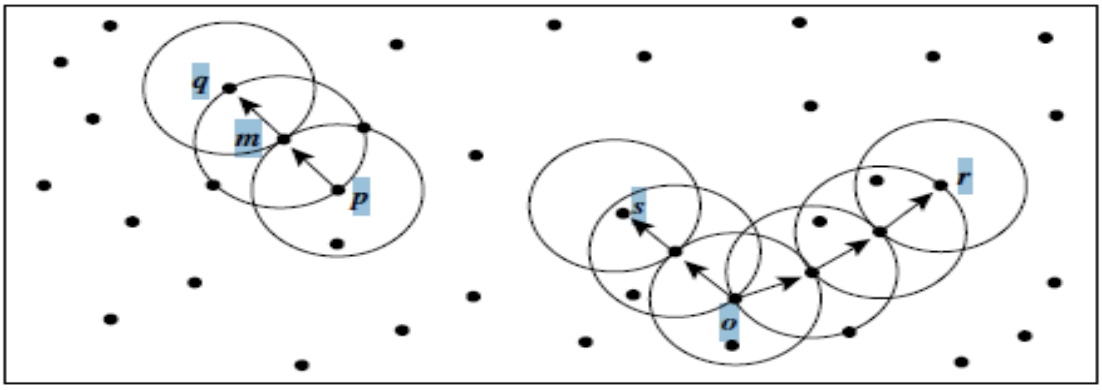

Fig. 2.7 Density connectivity in density-based clustering [25].

Kohonen Network/Self-Organized Maps (SOM)

Kohonen networks are special types of neural networks used for clustering technique. The network typically consists of two layers. The input layer includes all the clustering fields, called input neurons or units. The output layer is a two-dimensional grid map, consisting of the output neurons which will form the derived clusters. Each input neuron is connected to each of the output neurons with "strengths" or weights. These weights (which are analogous to the cluster centers referred in the K-means procedure) are initially set at random and are refined as the model is trained [34].

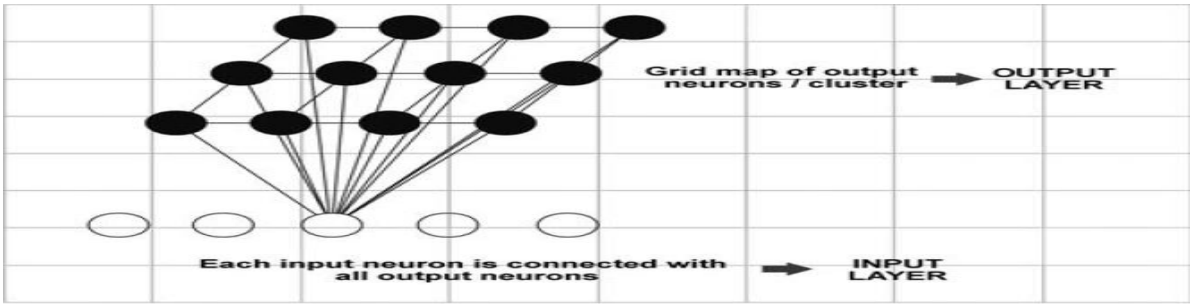

Fig. 2.8 Self-organizing map graphical representation [34].

2.6.2.2 Association Rules

In data mining, association rule is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules mining is another key unsupervised data mining method, after clustering, that finds interesting associations like relationships and dependencies in large sets of data items. The items are stored in the form of transactions that can be generated by an external process, or extracted from relational databases or data warehouses. Due to good scalability characteristics of the association rules algorithms and the ever-growing size of the accumulated data, association rules are an essential data mining tool for extracting knowledge from data. The discovery of interesting associations provides a source of information often used by businesses for decision making [25].

According to Agrawal [1], association and correlation are usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one.

2.7 Data Mining and Knowledge Discovery Process Model

The idea of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. So the term data mining has mostly been used by statisticians, data analysts and the management information systems (MIS) communities. It has also gained popularity in the database management systems field. The phrase knowledge discovery in databases was incubated at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the artificial intelligence (AI) and machine-learning fields [18].

As Fayyad [18] explained, knowledge discovery in databases (KDD) refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from the data sets. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, data transformation, incorporation of appropriate prior knowledge and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data.

The KDD process is interactive and highly iterative, user involved, multistep process; comprising a number of phases requiring the user to make several decisions. KDD employs methods from various fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems and data visualization [18]. It is said to employ a broader model view than statistics and strives to automate the process of data analysis, including the art of hypothesis generation [5].

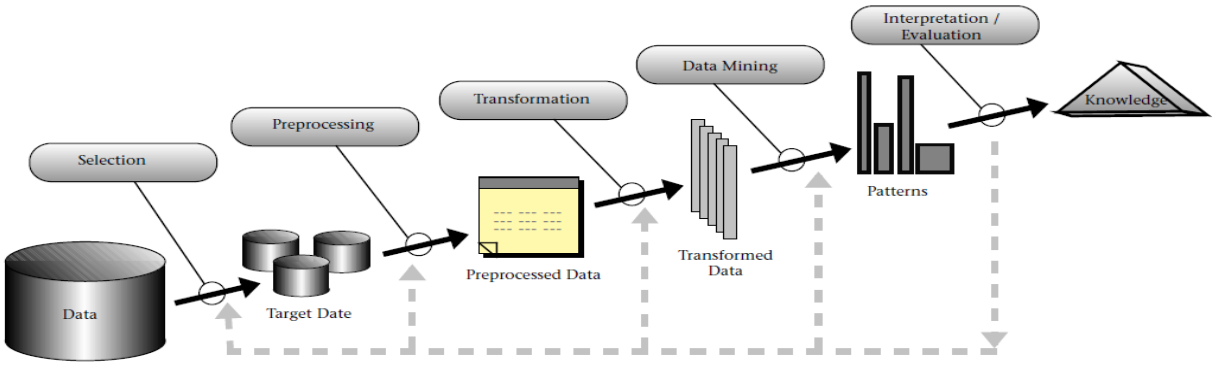

Fig 2.9 An Overview of the Steps that compose the KDD Process [18].

As described by Brachman and Anand (1996) cited in [18], the KDD process is an interactive process which consists of the following steps:

Developing and Understanding the application domain: This step includes learning the relevant prior knowledge and identifying the goal of the KDD process from the customer's point of view.

Creating target data set: Selecting a data set or focusing on the subset of variables or data samples, on which the discovery is to be performed.

Data cleaning and Preprocessing: basic operations such as the removal of noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes are to be performed.

Data reduction and projection: finding useful features to represent the data depending on the goal of the task.

Choosing the data mining task: here the data miner matches the goals defined in step 1 with a particular data mining method, such as classification, clustering and regression, etc.

Choosing the data mining Algorithm(s): Selecting method(s) to be used for searching patterns in the data. This includes deciding which models and parameters may be appropriate and matching a particular data mining method with overall criteria of the KDD process.

Data mining: searching for patterns of interest in a particular representational form or a set of such representations: classification rules or trees, regression, clustering and so forth.

Interpreting Mined patterns: here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.

Consolidating discovered knowledge: incorporating this knowledge into another system for further action or simply documenting it and reporting the interested parties. This also includes checking and resolving potential conflicts with previously extracted knowledge.

2.8 Review of related literature

Emilia [17] studied on Behavioral Segmentation of Telecommunication Customers. The objective of this thesis was first of all to perform a comparative study of different segmentation methods. comparison was done on two unsupervised and one supervised learning algorithm. K-means, self organizing map (SOM) and Support Vector Clustering were identified as the most commonly used methods for customer segmentation. The first part of the thesis was for critical analysis and the second part was an empirical study, which was performed to verify one of the clustering methods by conducting behavioral customer segmentation on a telecommunication network transaction data.

Zhiyuan [56] studied Visual Customer Segmentation and Behavior Analysis. The objective of the study was explore three contributions:

The primary contribution of this thesis was to propose a toolset based upon Visual Data Mining (VDM) techniques to conduct customer segmentation. VDM techniques are capable of providing

a better solution through embedding users in data exploration and the model building process. But the researcher mentioned that the application of visualization techniques is regarded as a practical problem still open to research.

The second main contribution of this thesis was to explore both the static structure and the temporal dynamics of the customer base visually. Customers' behavior and demographic characteristics are often not static but change constantly. Visualizing dynamics in data are also identified as one of the most challenging problems.

The third contribution lies in demonstrating the usefulness of the models in a real-world setting. The data used in this thesis, including customer demographic, lifestyle information, and their purchase history, are from a retailer's customer data warehouse including more than one million customers. Industry data like this are difficult to access and utilize for research purposes.

Belachew [5] conducted a research on Application of Data Mining Techniques for Customer Segmentation and Prediction: the case of Buusaa Gononfa Microfinance Institution. The researcher mentioned that due to lack of appropriate tool to segment the customers for Buusaa Gononfa Microfinance Institution, the institution couldn't identify the potential customers and measure their profits.

The objective of the study was to build a model that can help to classify the customers for the institution. The methodology selected by the researcher for this study was CRISP-DM. The researcher used clustering techniques to segment customers in to appropriate number of clusters. K-means clustering algorithm was used to clusters which exhibit similar characteristics. Then after, the researcher used J48 algorithm and built a predictive model to predict potential customers.

Finally, the predictive model had 99.5% accuracy. The researcher concluded that the result gained from the predictive modeling is encouraging. He also recommended the institution should to develop integrated warehouse to apply data mining techniques.

Belete [6] conducted a research entitled as "Knowledge Discovery for effective customer segmentation: the case of Ethiopian Revenues and Customs Authority". The objective of this study was to design a model using data mining techniques for customer segmentation that helps the organization to maintain the potential CRM.

The methodology used by the researcher was Cios et al. (2000) KDD process model. To segment customers, the researcher used the K-means clustering algorithm. The classification modeling was built by using J48 decision tree and multilayer perceptron ANN algorithms with 10-fold cross-validation and splitting (70% training and 30% testing) techniques.

At the end, among the derived models, a model which was built using J48 decision tree algorithm with default 10-fold cross-validation shows better performance which is 99.95% of overall accuracy rate. The researcher recommended that inclusion of additional customer attributes, integration of customs and revenue database, comparison of the k-means clustering algorithm with other clustering algorithms, increasing the number of records, reducing within cluster sum of squared error could generate a better result.

Dandi [11], the attempt was done on Application of Data Mining Techniques for Customers Segmentation in Insurance Business: The Case of Ethiopian Insurance Corporation. The objective of the study was to apply data mining techniques for customer segmentation based on customers' value in Ethiopian Insurance Corporation.

To meet the objective of the study, the CRISP-DM methodology, which involves six steps was adopted to undertake the data mining process. In clustering phase, to build the customer segmentation models, k-means clustering algorithm was employed. Then in classification phase, J48 decision tree algorithm. The result of the research pointed out that the customer segmentation models built by using the combination of classification and clustering data mining techniques are necessary for the Life Addis District and Marketing department of EIC in order to identify the valuable segments of customers and other factors underlying variations of the customer's value.

Another research entitled as "Application of data mining technology fraud protection on domestic tax collection: the case Ethiopian Revenues and customs Authority" was conducted by Daniel [12]. The objective of the study was to explore the potential applicability of the data mining technology in developing models that can detect and predict fraud suspicious in tax claims with a particular emphasis to Ethiopian Revenue and Custom Authority.

The researcher used the six-step Cios et al. (2000) KDD process model for this study. K-Means clustering algorithm was employed to find the natural grouping of the different tax claims as

fraud and non-fraud. The resulting cluster was then used as an input for developing the classification model. The classification task of this study was carried out using the J48 decision tree and Naïve Bayes algorithms in order to create model that best predict fraud suspicious tax claims.

The final result of this research indicated that the model developed using the J48 decision tree algorithm has showed highest classification accuracy of 99.98%. This model was then tested with the 2200 testing dataset and scored a prediction accuracy of 97.19%. The results of this study have showed that the data mining techniques are valuable for tax fraud detection. The researcher recommended that by applying different classification techniques, incorporate additional related attributes and consider other sections of the authority needs to be studied.

Denekew [13] conducted a study entitled as "The application of Data Mining to support customer relationship management at Ethiopian Airlines". The objective of this research was to assess the application of data mining techniques to support CRM activities at Ethiopian Airlines based on the Ethiopian Airlines' frequent flyer database. The researcher conducted the study based on the gaps identified by the previous researcher Henok [28].

In the data mining process, the researcher used two major phases: first phase, data preparation and formatting. Second phase is model building. In the clustering sub phase the K-means clustering algorithm was used to segment individual customer records into clusters with similar behaviors. In the classification sub-phase, J4.8 and J4.8 PART algorithms were employed to generate rules that were used to develop the predestined model that assigns new customer records into the corresponding segments.

At the end, the researcher developed a prototype of Customer Classification System. The prototype enables to classify a new customer into one of the customer clusters, generate cluster results, search for a customer and find the cluster where the customer belongs, and also provides with the description of each customer clusters. The researcher recommended that the study is encouraging and extra more studies by using a possible large amount of customer records with demographic information and employing other clustering and classification algorithms could yield better results.

Hailemariam [23] conducted a study entitled as "Application of Data Mining Techniques to Customer Profile Analysis in the Ethiopian Electric Power Corporation". The objective of the study was to support Customer Relationship Management (CRM) activities at Ethiopian Electric Power Corporation (EEPCo) by employing appropriate data mining techniques on the customers profile database to clusters and classify customers.

The methodology used by the researcher was CRISP-DM methodology. In this study, the K-means clustering algorithm was used to segment customer records into clusters with similar behaviors.

Finally, the researcher conducted classification technique. In the classification sub-phase, J48 decision tree and Naive Bayes algorithms were employed. Among the different experiments, a J48 decision tree model that showed a classification accuracy of 99.89% was selected. The researcher recommended that the results of this study were encouraging and confirmed the belief that applying data mining techniques could indeed support CRM activities at EEPCo.

Henok [28] conducted a study entitled as "The application of Data Mining to support customer relationship management at Ethiopian Airlines". The objective of the study was to study the application of data mining techniques on the frequent flyer customers' database, in order to discover strategic customer segments that could support CRM activities at ETHIOPIAN AIRLINES.

The methodology used by the researcher was CRISP-DM. The researcher used clustering techniques to segment customer information stored in Frequent Flyer Program (FFP's) database. K-means clustering algorithm was selected to segment customers data into five clusters depending on the similarity behavior that they exhibit. And also decision tree classification is used to assign new customers data to the segments created. The tool selected for model building was Knowledge studio.

Finally, the result gained from five clusters, three of the clusters contained 21% of customers record generated the highest revenue. The cluster containing medium and low revenue generated customers contained 27% and 52% of the customers record respectively. The researcher concluded that applying data mining techniques could definitely support CRM activities at Ethiopian Airlines. Then he recommended that, further study is needed to do on customer

segmentation using demographic information and utilizing other clustering algorithms and association rule algorithms could give better result.

Belete [6], mentioned the gaps and recommendations on his study. There was a gap of not including domestic tax ERCA's customer database, attribute selection constraint, algorithm selection constraint and less data complexity. So this study aims to address the gaps identified and recommendations pointed out by Belete [6].

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Methodology is the steps or procedures that the researcher follows to achieve the stated objectives of the research. It is a road map that shows the direction how the researcher is going to conduct the research to reach at its target. Accordingly, this study followed different methods in order to develop good customers segmentation models for designing and implementing successful customer relationship management. In this research specifically, descriptive models have been implemented. Through descriptive, the behaviors of individual customers have been studied and categorized into similar clusters [23].

3.2 Research Design

In order to achieve the stated objectives of this study, the researcher adopted the **CR**oss-Industry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) methodology. This is because CRISP-DM has been widely used and applied in data mining studies. CRISP-DM model is flexible and can be customized easily for different data mining studies. In addition, it is flexible to account for differences i.e. for different business problems and different data. And also it is an open-source and industry standard data mining process model. It allows the researcher to create a data mining model that fits the required particular needs [23][25]. The CRISP-DM model is divided into six phases. These are Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment [4][23][25].

3.3 The CRISP-DM Process

CRISP-DM is a standard process model in industries which are consisting of a sequence of steps that are usually involved in a data mining study [48]. CRISP-DM stands for CRoss Industry Standard Process for Data Mining developed by Consortium of Data Mining industrial Companies [4]. The CRISP-DM process model provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks and the relationships between these tasks [48].

The CRISP-DM process organizes the data mining processes into six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. The sequence of the phases is not rigid rather it is flexible and iterative. Moving back and forth between different phases is always required and allowed in the process. The outcome of each phase determines which phase, or particular task of a phase has to be performed next. The arrows indicate the most important and frequent dependencies between phases [48]. And the phases help organizations to understand the data mining process and provide a road map to follow while planning and carrying out a data mining project.

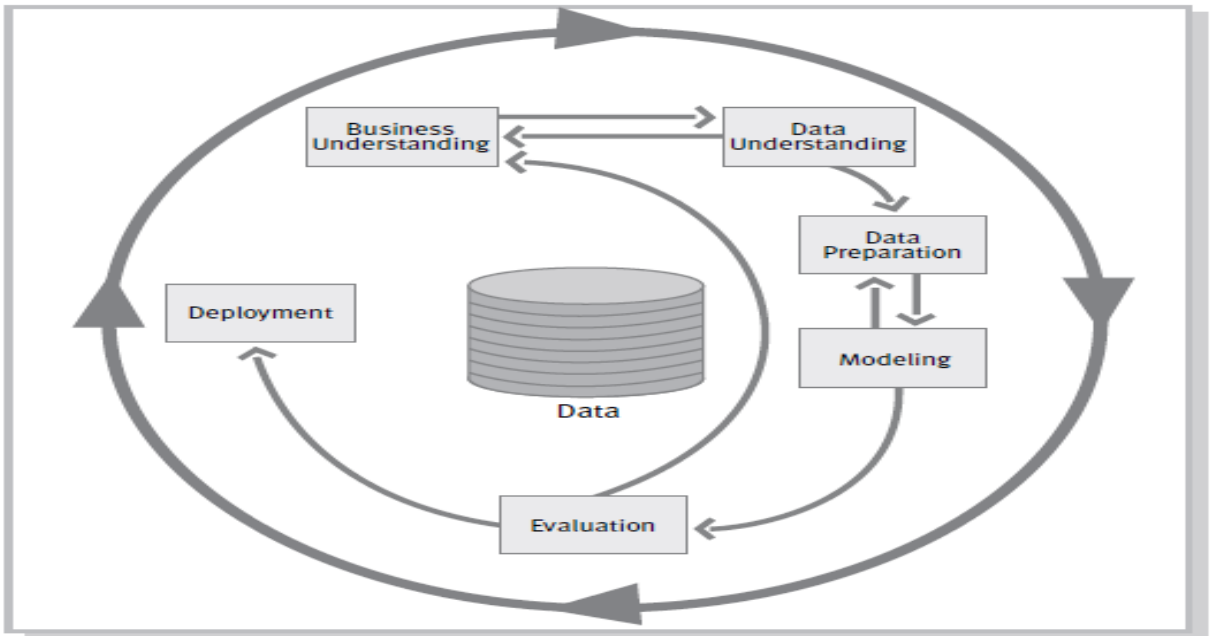

Fig. 3.1 Phases of the CRISP-DM process model [48].

As it has been indicated in the above Fig 3.3, the six steps of CRISP-DM processes are cyclical that can be implemented iteratively. And the subsequent explanations are the descriptions of the methodology within the study content.

3.3.1 Business Understanding Phase

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives [48].

During the business understanding phase a literature review was conducted in order to gain insight with data mining studies which are related to customer segmentation and customer relationship management studies which have been solved by the application of data mining techniques and methods in previous research projects.

The researcher made an effort to study the objective of Ethiopian Revenues and Customs Authority and the situation of the business process through document analysis and interviews. To understand the business, formal interview, informal interview and direct observation have been conducted at the Information Technology Management Directorate: business analyst team, database and application administration team and Customer Service Directorate employees who are engaged in operational tasks on existing system of the authority. This was done in order to get insight with the specific customer related problems at Ethiopian Revenues and Customs Authority which have not been solved yet.

Based on the domain experts explanation and evaluation made, the researcher found different kinds of customers in the Authority those need to be segmented according to their revenue generating capability. So, in this study, the researcher is attempted to apply data mining techniques to differentiate customers who engaged in import and export business activities, whether they are high, medium and low value customers based on the revenue they generate and the transaction of the data.

3.3.1.1 Data Mining Objective

The data mining objective for this study was to apply data mining technique for segmenting or clustering ERCA's customers into different groups based on their characteristics to treat or handle them accordingly. To accomplish this task effectively, the customer data was collected and the important variables had been identified. The identified variables were in turn used to build a model by using clustering techniques.

3.3.2 Data Understanding phase

The data understanding phase starts with initial data collection and proceeds with activities that enable researcher to be familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information [48].

To achieve the objectives of this study, the researcher attempted to understand the data that resides with ERCA's customers database. So, the data source was the database of ERCA which contains the data of the customers those engaged in import and export business activities of different goods. The dataset contains customer types and data related to business operations with respect to their attributes and selection of suitable data has been made according to the authority's need and problem.

At this stage, the data used in this research has been described briefly. The description includes listing out attributes, data types and their respective values, evaluation of their importance, Careful analysis of the data and its structure have been done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular data mining tasks to be performed.

3.3.3 Data preparation phase

The data preparation phase covers all activities needed to construct the final dataset that will be fed into the modeling tool(s) from the initial raw dataset. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include data cleaning, table, record, and attribute selection, data reduction, as well as transformation and cleaning of data for modeling tools [48].

After the data are identified and collected, different data preprocessing techniques are applied for processing and cleaning the data. WEKA data mining tool is used for attribute selection preprocessing techniques (to reduce dimensionality) and by derivation of new attributes. Data preparation and preprocessing comprises selection of attribute, handling noisy data, handling missing data fields, integrating, descritization, normalization, transforming and preparing the processed data in a file format suitable for the data mining tool or software that has been implemented.

In this research, MS-Excel 2007 tool is also used for data preparation, preprocessing and analysis by using its functions such as sort, formula, filter, find and replace so on. The datasets which are prepared in MS-Excel format should be converted to CSV (Comma Separator Value) and ARFF (Attribute Relational File Format for data processing). The input of the data mining applications proved to be simple with the conversion of Excel spread sheet datasets into a CSV file format and then an ARFF file format for modeling process.

Finally, with the help of domain experts most important attributes, which help to segment ERCA's customers based on their value have been selected as a final datasets for the model preparation purpose.

3.3.4 Modeling phase

The choice of the data mining technique has to be applied depending on the nature of the problem. Actually, various techniques can always be applied on the same type of problem, but there is always a technique that yields the best results for a specific problem. It is sometimes necessary to model several techniques and algorithms, and then select the one yielding the best results. In other words, several models are built in a single iteration of the phase, and the best one is selected [7].

In this data mining phase, various modeling techniques have been selected and applied, and their parameters have been calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary [48]. So in this study, the clustering models are applied to build the customer segmentation model.

Clustering technique is selected with its capability of finding groups that have similarity in the data. It has been applied to explore homogeneity of records of customers with respect to clustering fields and assign them to the revealed clusters. Using clustering technique can enable to discover the groups with internal homogeneity and interclass heterogeneity [13]. Clustering technique is unsupervised learning functions that finds the relationship of data from the given dataset. In doing so, for this study different algorithms have been tested and the one that performs the best is selected. So, k-means clustering algorithm is selected and applied in order to group or segment customers with their homogeneous characteristics. The reason why k-means

algorithm selected is, it was identified as the most fastest and efficient clustering algorithm to handle many records and wider datasets [25].

In this research, in order to build a data mining models so as to solve the identified problem, WEKA (Waikato Environment for Knowledge Analysis) was selected as a data mining tool. WEKA is open source software which can support most of the technical aspects of the CRISP-DM standard data mining methodology [11]. Besides, it was selected because of its accessibility and familiarity of the researcher with this tool.

3.3.5 Evaluation phase

Before proceeding to final deployment of the model and to be certain the model properly achieves the business objectives or not, it is important to thoroughly evaluate the model and review the steps executed to create it. So this phase assesses the final model how much it meets the stated goals in the first phase of the process. If the information gained at this point affects the quality of the entire project, this means returning to the first phase and re-initiating the cyclical process has to be taken into account. If the model meets all the business goals and is considered satisfactory for application, this step is followed by the review of the entire data mining process, in order to secure the quality of the entire process, i.e. check whether there have been any oversights in its execution. Basically, it is a general quality management and provision, within the domain of the project management discipline [7] [48].

In this study, the researcher examined different models and algorithms which yields a better result. The model has been evaluated together with the domain experts regarding to its interestingness and novelty. The different clustering models those developed in this research has been evaluated based on cluster sum of squared error (SSE) values, number of iteration the algorithm takes to converge, the attributes average values that satisfy the threshold, and experts' judgment.

3.3.6 Deployment phase

In general, model creation is 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 such a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as to generate a report or as complex as to implement a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models [48].

Through the knowledge discovered in the earlier phases of CRISP-DM process, sound models can be obtained that may be applied to business operations for many purposes including prediction or classification and clustering depending on the key situations of the authority. These models need to be monitored for change in operating conditions, because what might be true today may not be true for the future starting from now.

CHAPTER FOUR

EXPERIMENTATION

4.1 Introduction

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

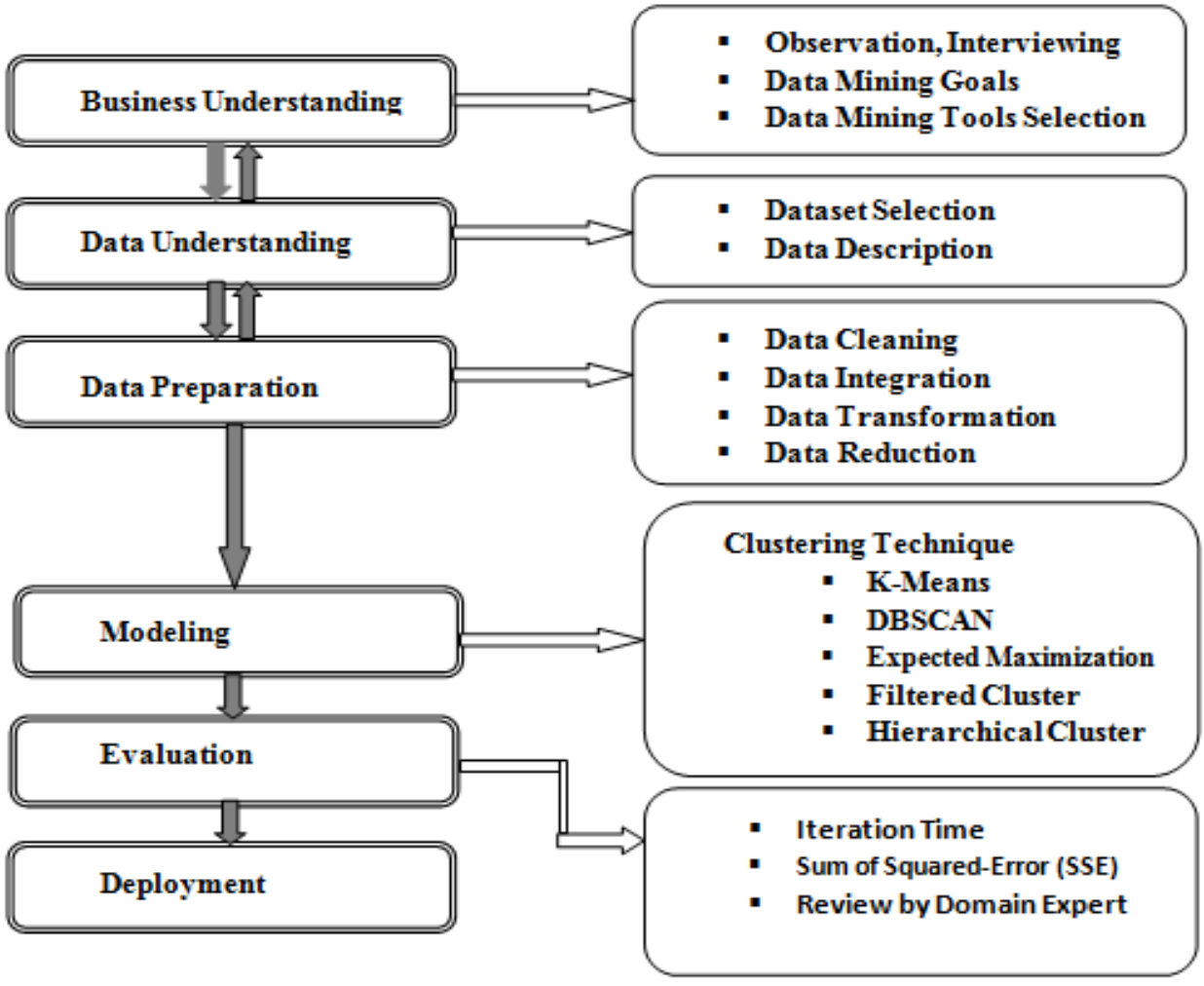

Fig 4. 1 Experimentation process flow of the study

4.2 Business Understanding of ERCA

The Ethiopian Revenues and Customs Authority (ERCA) came into existence on 14 July 2008, by the merging of the Ministry of Revenue, Ethiopian Customs Authority and the Federal Inland Revenue Authority. ERCA is the responsible body for collecting customs duties and domestic taxes from all business sectors; Importers and Exporters to support the government plan into practice. In addition to raising revenue, ERCA is also responsible to protect the society from adverse effects of smuggling and contraband. It seizes and takes legal action on the people involved in the act of smuggling while it facilitates the legitimate movement of goods and people across the border [19].

According to article 3 of the proclamation No. 587/2008, the Authority is looked upon as "an autonomous federal agency having its own legal personality". Reasons for the merging of the foregoing administrations into a single autonomous Authority are varied and complex.
Some of those reasons include:

- To provide the basis for modern domestic tax and customs administrations
- To cut through the red tape or avoid unnecessary and redundant procedures that results delay and considered cost-inefficient etc
- To be much more effective and efficient in keeping and utilizing information, promoting law and order, resource utilization and service delivery
- To transform the efficiency of the revenue sector to a high level

ERCA has its headquarters in Addis Ababa. It is led by a Director General who reports to the Prime Minister and is assisted by four Deputy Director Generals: namely:- D/Director General of Inland Tax Division, D/Director General of Customs Division, D/Director General of Enforcement Division and D/Director General of Capacity Building and Resource Administration Division. Each deputy director general oversees at least four directorates. Both the Director General and the Deputies are appointed by the Prime Minister.

As the ERCA's experts stated, the newly designed tax and customs procedures have brought about a positive result and it has produced beneficial effects to both the Authority and its customers or stakeholders. However, to further improve the satisfaction of customers and the beneficial effects of the authority, there is a need to implement modern CRM in the organization.

Objective of the Ethiopian Revenues and Customs Authority (ERCA)

ERCA has the following objectives:

- To establish or set up modern revenue assessment and collection system; and provide customers with fair, efficient and quality service
- To encourage taxpayers voluntarily discharge their tax obligations
- To enforce tax and customs laws by preventing and controlling contraband as well as tax fraud and evasion
- To collect timely and effectively tax revenues generated by the economy
- To provide the necessary support to regions with a view to harmonizing federal and regional tax administration systems.

Powers and Duties of the ERCA

According to the proclamation No. 587/2008 of the *Ethiopia Federal Negarit Gazeta*, ERCA has the powers and duties to establish and implement modern revenue assessment and collection system; ERCA has also the powers and duties of provide efficient, equitable and quality service within the sector; properly enforce incentives of tax exemptions given to investors and ensure that such incentives are used for the intended purposes; conduct study and research activities with greater emphasis to improve the enforcement of customs and tax laws, regulations and directives and the collection of other revenues; and based on the result of the study and research initiate laws, policies and implement the same upon an approval.

According to the experts explanation, ERCA has additional responsibility to collect and analyze information which is necessary for the control of import goods, export goods, the assessment and determination of taxes; compile statistical data on criminal offenses relating to the sector, and disseminate the information to others respected sectors. Furthermore, the Authority has been providing information and appropriate support to the Federal Police in the control of illegal trafficking of goods and struggled contraband; and cause appropriate measures be taken in accordance with the law. The powers of deciding the place where import and export goods are to be deposited and controlled are also given to ERCA.

To understand ERCA's business processes and the problem domain regarding to the customers; as it was stated in the methodology section, the researcher used different techniques or methods. The researcher has communicated with domain business process owner experts, customer service experts, technical experts and direct on site observation. As the researcher's observation indicates, there is a need to implement effective customer segmentation in the Authority so as to increase the satisfaction of the customers and maximize the revenue collection volume of the Authority.

Based on the experts' explanation and on site observation, the authority has been giving services for different kinds of customers. According to the authority's current practice, some customers are high value customers; these types of customers are those customers who generate large amount of revenue for the organization by importing smaller amount of items, spending low invoice and higher frequency of items. On the other hand, there are also customers who generate lower revenue for the Authority. These types of customers are called low value customers. Low value customers import large amount of items, spend higher invoice and lower frequency of items. These types of customers generate smaller amount of revenue for the Authority. The intermediate between high and low value customers are medium value customers. In addition, the goods those are imported for manufacturing Industries (Industry goods) have been given higher priory during the importation process. So, the main aim of this research is to differentiate customers whether they are high, medium or low value customers based on the revenue they generate and the transaction volume of the data.

4.2.1 Data Mining Goals

The data mining goal for this study was to apply clustering techniques to segment customers into different groups based on their characteristics to treat or handle them accordingly. To carry out this task effectively, it was necessary to identify important variables from the total customers collected data. The identified variables were used to build a model by using clustering techniques. In this research, different clustering algorithms like Expected Maximization, Filtered cluster, Hierarchical Cluster, DBSCAN and k-means algorithms have been selected for testing.

4.3 Data Understanding

After understanding of the problem domain, defining of data mining goal and selecting an appropriate tool; the next step was to collect, analyze and understand the available customers' data. Therefore, to fulfill the data requirement, customers' data was initially collected from ERCA's ASYCUDA++ customer database. Then, analysis of the data and its structure was done together with the technical experts and other domain experts by evaluating the relationship of the data with the problem at hand. In addition, the business processes were dealt with the functional domain experts in the Authority.

4.3.1 Initial Data Collection

For the initial data collection process, the major source of data for this research was the ERCA's ASYCUDA++ customer database. ASYCUDA++ database is very huge system and rich data source because every declaration of import/export transactions and other detail information of the customers are found in the ERCA's ASYCUDA++ customer database. For instance, the customer/company name which is populated from ERCA's domestic tax SIGTAS database, the item type, total revenue collected, the price of the item, total number of items imported, total amount of invoice spent to import the items of every customer are found in this database. The researcher collected a half-year (six months) import declaration of customers' data which was around 131072 records. The responsibility of administering, following up and controlling the ERCA's customs and domestic tax databases is given to the Information Technology Management Directorate (ITMD).

4.3.2 Description of the Collected Data

In this study, the researcher collected a half-year (six months) import transaction data. Since every import/export transaction of goods within the country is registered in the ASYCUDA++ database, the amount of data found is very huge but the researcher took only selected data by consulting domain experts. To select the data, the researcher used the WEKA preprocessing facilities which is stratified remove folds by setting number of folds with two; and only around 65535 records. However, in the ERCA's ASYCUDA++ database there are a lot of tables with many attributes, the researcher selected 11 highly relevant attributes by consulting the domain experts . A sample data which is found from the initial data collection is attached into this research as **Appendix 1.**

Among from different tables of the ASYCUDA++ database the researcher selected the customer related tables those had been required for this study are mentioned below:

SAD_GEN TABLE (SAD General Segment Table): The SAD_GEN table contains all information related to the SAD general segments. SAD means an abbreviation for Single Administration Document.

Attribute/Field	Description	Data Type
SAD_CONSIGNEE	Company/Importer code	Char
SAD_REG_DATE	Declaration Registration date	Date
SAD_REG_SERIAL	Registration Serial Number	Char
SAD_ITEM_TOTAL	Total number of items imported	Small Int
SAD_PACK_TOTAL	Total number of packet	Char
SAD_TOT_INVOICED	Total amt of invoiced for entire declaration	Char
SAD_REG_NBER	Registration Number	Integer
SAD_RCPT_NBER	Receipt Number	Integer
SAD_ASMT_NBER	Assessment Number	Integer
SAD_ASMT_DATE	Assessment Date	Date
SAD_NUM	Declaration Version Number	Integer
SAD_TYP_DEC	Type of declaration (import or export)	Char
SAD_CTY_IDLP	If the SAD is an import SAD this field indicates the country of last consignment. If the SAD is for an export declaration, then this is the field showing the country of first destination.	Char
SAD_CTY_EXTCOD	Country of Export Code	Char
SAD_CTY_DESTCOD	Country of Destination code	Char
SAD_CUR_COD	Currency code used on the Invoice	Char
SAD_MOT_BORD	Mode of transport at border	Char
SAD_WHS_COD	Warehouse Code	Char
SAD_TOTAL_TAXES	The total amount of all duties, taxes and charges levied on the general segment and all of the items	Decimal
SAD_TYP_PROC	The type of customs procedure for which SAD is being lodged	Char

Table 4.1 Attributes and description of the SAD general segment table

SAD_ITM TABLE: - The SAD_ITM TABLE contains all Item level data about the declaration of items.

Attribute/Field	Description	Data Type
ITM_NBER	Item number	Small int
SADITM_HS_COD	Commodity code	Char
SADITM_CTY_ORIGCOD	Country of Origin code	Char
SADITM_GOODS_DESC	Description of the goods	Char
SADITM_GROSS_MASS	Gross Mass	Decimal
SADITM_GROSS_NET_MASS	Net Mass	Decimal
SADITM_SUPP_UNITS	Supplementary Unit	Decimal
SADITM_PREV_DOC	Previous document reference (for temporary import for instance)	Char
SADITM_ITM_PRICE	Item Price	Decimal
SADITM_ITM_TOTAMT	Total amount of duties and taxes for the item	Decimal
SADITEM_NUM	Declaration Internal Version Number	Integer
SADITM_PREV_WHSCOD	Previous Warehouse code	Char

Table 4.2 Attributes and description of the SAD_ITEM table

UNCMPTAB (Companies Table): - This table records all companies/customers related data. As this table is not historized, ASYCUDA doesn't record all modifications done on a company code but only keep the last updated value

Attribute/Field	Description	Data Type
CMP_COD	Company code	Char
CMP_NAM	Company name	Char
CMP_ADR	Company address	Char
CMP_TEL	Telephone Number	Char
CMP_FAX	Fax number	Char
CMP_TLX	Telex number	Char
LST_OPE	Last operation	Char
SER_STA	Server status number	Integer

Table 4.3 Attributes and description of the company table

UNCTYTAB (country table): - This table contains information about the country where the item is imported.

Attribute/Field	Description	Data Type
CTY_COD	Country code	Char
CTY_DSC	Country description	Char
LST_OPE	Last operation	Char
SER_STA	Server status number	Integer

Table 4.4 Attributes and description of the country table

TAX_PAYER Table: SIGTAS domestic tax table where Customers' TIN is initially captured.

Attribute/Field	Description	Data Type
TIN_NO	Customer Tax Identification Number	Decimal
TAX_PAYER_NO	Tax Payer Number	Decimal
REGIST_NAME	Company/Customer Name	Char
CITY_NO	CUSTOMER/Company address	Char
TAX_CENTER_NO	Tax Center Number	Integer
INTER_DATE	Inter Date	Date
UPDATE_DATE	Update Date	Date

Table 4.5 Attributes and description of the Tax Payer table

4.4 Data Preparation

The purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a form that is suitable for further analysis and processing. Data preprocessing helps to fill some missing values; to detect some outliers that may jeopardize the result of data mining; and to detect and remove/correct some noisy data. In addition data normalization, discretization, data reduction and data transformation have been performed at this phase. Moreover, to conduct the experimentation the dataset has been prepared in the appropriate format [25].

In the data mining process, data are usually collected from the existing databases, data warehouses, and other data sources. However, preprocessing is a challenging and time taking task, especially in large databases. It is not only time consuming to specify a preprocessing operations but also to apply it on a large databases. The data in the real world is dirty and it may contain incomplete, noisy (outliers), inconsistence and irrelevant data. "No quality data, no quality mining results". So, this stage involves a serious of steps to provide the final dataset from raw data for modeling purpose. According to Han and Kamber [25], data preprocessing phase of data mining process includes: data cleaning, data integration, transformation, data selection and data formatting.

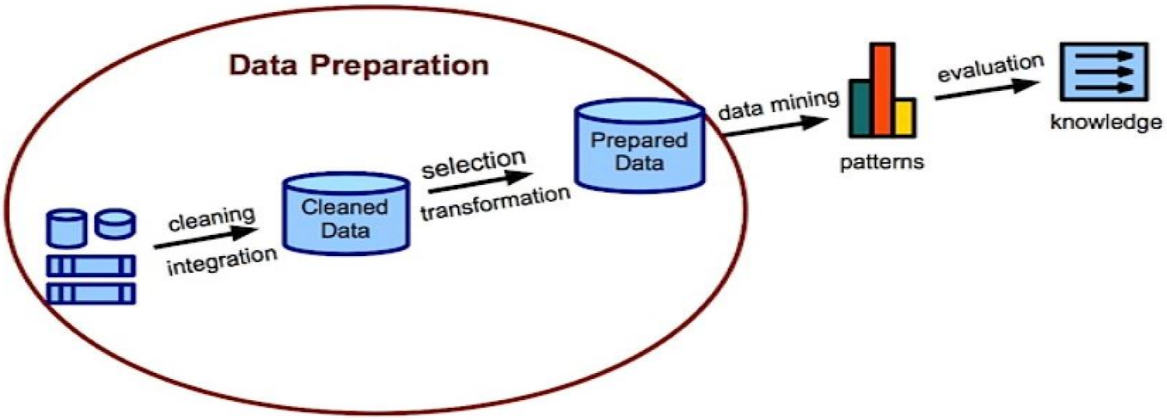

Fig 4.2 Data Preparation steps [23].

4.4.1 Data Cleaning

Data Cleaning is the process of filling missing values, removing noise (outliers) and correcting data inconsistency. Usually, real world databases and data warehouses contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process [23]. Therefore, data cleaning is one of the key data mining processes which is applied to realize the quality of the data so as to improve the performance of the accuracy and efficiency of the data mining techniques.

In the data cleaning phase of this study, the researcher has identified missing values of the attribute and remove some outliers and incomplete data which does not affect the entire dataset. From the total dataset, TOTAL_REVENUE attribute has had outliers and missing values. When the total revenue is missed or if it is zero value for a record, for such types of data, the

researcher learnt from the domain experts these types of items are tax exempted (tax free) goods, as a result the researcher discarded or removed these instances from the dataset; because, these types of instances were not important for the data mining process.

4.4.2 Data Integration

In this research, the data stored in different tables of customers' databases are raw data, which were not in the form of appropriate for the business goal to be addressed and the corresponding data mining task to be done. So, the data integration process was done before deriving the attributes. Data integration is the activity of merging or combining data from multiple sources like databases, data cubes, or flat files into coherent data store, as in data warehousing [25].

Data integration method for retrieving important fields or attributes from different files was done in the effort to prepare the data ready for the data mining techniques process. In this research, the customs ASYCUDA++ and domestic tax SIGTAS databases have been used to carry out the data integration process. Even if the main source of data for this study is ASYCUDA++, the primary source of company/customer Tax Identification Number (TIN) information was domestic tax SIGTAS database. Once the customers' TIN information is captured in domestic tax SIGTAS database, then it has been populated to customs ASYCUDA++ database. Company/customers information is one of the main attribute for this research. So, after selecting different fields from different tables together with the domain experts, the integration part which has been brought different attributes from different tables into a single dataset was performed by using SQL commands in the original ASYCUDA++ database. The attributes were retrieved from the selected tables which were stated in the description of the data collected parts.

4.4.3 Data Reduction: Data Dimensionality

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Data dimensionality helps to eliminate irrelevant attributes and reduce noise that contains no information useful for the data mining task for the specific study. In this study, the researcher removed 11 attributes out of 22 since these attributes have less importance for the study. Before removing the attributes, the researcher fed all the attribute to weka and the

performance was less. So based on the weka GainRationAttributeEval and ranker together with the domain experts advice, the reseacher tested until the performance is maintained. Finally, 11 attributes are employed for the experimentation process.

4.4.4 Data Transformation

The other important step in data preprocessing is deriving other fields from the existing ones. According to Berry and Linoff (2004), adding fields that represent the relationships in the data are likely to be important in increasing the chance of the knowledge discovery process yield useful result. Data creation involves the creation of new variables by combining existing variables to form ratios, difference and so forth. REVENUE_ITEM, REVENUE_PRICE and INVOICED_REVENNUE are attributes which were derived from the other original attributes. Accordingly, the derived attributes are defined using the existing attributes as follows:

$$\text{REVENUE\_ITEM} = \frac{\text{TOTAL\_REVENUE}}{\text{TOTAL\_ITEM}}$$

$$\text{INVOICED\_REVENUE} = \frac{\text{TOTAL\_INVOICED}}{\text{TOTAL\_REVENUE}}$$

$$\text{REVENUE\_PRICE} = \frac{\text{TOTAL\_REVENUE}}{\text{ITEM\_PRICE}}$$

Moreover, the researcher has discretized the attribute SADITM_CTY_ORIGCOD. This attribute had around 121 values which was the name of the country where the item is imported or originated. However, the researcher discretized this value into 6 which is in their continent of origin. The attributes business location and business activity are derived from SAD_CONSIGNEE customer attribute. MS-Excel was selected and used for cleaning the data. According the authority' business procedure, the values of business location like Federal (ET), AA and other cities are treated separately.

4.4.5 Data Formatting

At this step once the data integration is finalized, the researcher has changed the data into a format which is suitable for the data mining tool. Since the preprocessing of the data was performed in Ms Excel, the final dataset was also in Ms Excel format. However, the selected tool WEKA 3.6.9 doesn't accept the data in Excel format. Like other Software, WEKA needs data to be prepared in some formats and file types. The dataset provided to this software was prepared in a format that is acceptable for it. So, first the researcher converted the data in comma delimited (CSV) format which is acceptable for WEKA and ARFF (Attribute Relation File Format) format. Comma delimited is applied for a list of records where the items are separated by commas, whereas ARFF is an extension of a file format that the WEKA data mining software can read.

4.4.6 Attribute Selection

After the preparation of the final dataset the next step was to identify the best attributes which are used for cluster modeling. The researcher evaluated the information content of the attributes using the select attributes technique of WEKA 3.6.9 version with GainRationAttributeEval attribute evaluator and Ranker search method. Moreover, the researcher consulted domain experts for the attribute selection. Because, experts are more familiar with the attributes those have better importance to identify the customers in the authority from business perspective. So, there were variables to be discarded. However, it does not mean that these attributes have no importance rather these variables were believed to provide very little useful information for the problem at hand and since clustering is unsupervised learning they may reduce the accuracy of the clustering algorithm.

The following table shows attributes with their description after the final preprocessed dataset. The dataset containing these attributes is fed into the data mining software.

Attribute/Field	**Description**	**Data Type**
CMP_NAM	Company name	Char
CMP_ADR/LOCATION	Company address	Char
SAD_REG_DATE	Declaration registration number	Date
SAD_ITEM_TOTAL	Total item	Decimal
SAD_TOT_INVOICED	Total amt of invoiced for entire declaration	Decimal
SAD_REG_NBER	Registration Number	Integer
SAD_ASMT_NBER	Assessment Number	Integer
SAD_CUR_COD	Currency Code	Char
SAD_RCPT_NBER	Receipt Number	Char
SADITM_ITM_PRICE	Item price	Decimal
SADITM_HS_COD	Commodity code	Char
SADITM_GOODS_DESC	Description of the goods	Char
SADITM_CTY_ORIGCOD	Country of Origin code	Char
SADITM_GROSS_MASS	Gross Mass	Decimal
SADITM_NET_MASS	Net Mass	Decimal
SADITM_SUPP_UNITS	Supply Units	Decimal
SADITM_ASMT_NBER	Assessment Number	Integer
SAD_PACK_TOTAL	Total number package	Char
SAD_CONSIGNEE	Company code	Char
SADITM_ITM_TOTAMT	Total amt of duties and taxes for the item	Decimal
BUSINESS ACTIVITY	Sector that the importer/exporter engaged	Char
ITEM_FEQUIENCY	Repetitiveness of item for importation	Decimal

Table 4.6 Selected attributes with their description

4.5 Modeling

In the data mining steps, modeling phase is the process of providing the processed data to the selected clustering or classification algorithm and selecting the model that shows better performance and output. In this part of data mining process, there are a number of tasks that needs to be performed. Some of the tasks involved in this phase are: selection of modeling technique, generating experiment design, building model and selection of the best model.

4.5.1 Selection of Modeling Techniques

Before selecting a modeling technique, it was better to know the nature of the dataset. A good segmentation model can divide customers into homogeneous group on the basis of shared common attributes. Clustering technique is a direct data mining technique (i.e. no predefined or labeled classes) and the instances are being divided into natural groups. To select the best clustering algorithm, the researcher used WEKA tool to measure the model building performance different clustering algorithms; DBSCAN, Expected Maximization, Filtered Cluster, Hierarchical cluster and k-means algorithms. The WEKA run performance measurement output is illustrated in table 4.7.

	Time taken in Seconds				
Cluster No.	**DBSCAN**	**Expected Maximization**	**Filtered Cluster**	**Hierarchical Cluster**	**K-means**
2	0.04	0.09	0.06	0.32	0.01
3	0.04	0.24	0.06	0.47	0.02
4	0.03	0.21	0.02	0.23	0.01
5	0.04	0.16	0.02	0.22	0.01
6	0.04	0.19	0.01	0.20	0.01
7	0.04	0.22	0.02	0.19	0.02
8	0.04	0.15	0.01	0.18	0.02
9	0.04	0.15	0.02	0.19	0.02

Table 4.7 Time taken to build the model

As it has been indicated in the table 4.7, k-means clustering algorithm elapsed the least amount of time to build the model and performed the best. Filtered Cluster clustering algorithm

performed the second. And, DBSCAN, Expected Maximization and Hierarchical Cluster algorithms performed third to fifth respectively. Finally, k-means clustering algorithm was selected as the most appropriate algorithm for researcher's data. In addition, the reason why k-means algorithm had been selected for this study was that, it automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The Simple k-means algorithm uses Euclidean distance measure to compute distances between instances and clusters [23].

4.5.2 Experiment Design

Before going to the actual model building, it was better to set a guide for the training, testing and evaluation process. For the case of clustering technique the total instance of the dataset has to be employed to build a model. So, in order to perform cluster model building, 11 attributes and 65535 instances were used in the experimentation process. The following variables are selected and used for model building, while the rest of the attributes are not used because they are less important for the model building.

- The location of the customer where the business is established (BUSINESS LOCATION)
- The type of the business activity the customer engaged (BUSINESS ACTIVITY)
- Total number of items imported (SAD_ITEM_TOTAL)
- Total amount of invoice spent to import the item (SAD_TOTAL_INVOICED)
- Total amount of revenue generated (TOTAL_REVENUE)
- The price of the item (ITEM_PRICE)
- The amount of item how much frequent (ITEM_FREQUENCY)
- Origin (Continent) of the item where it is imported (CTY_DSC)
- Ratio of total revenue to total number of items (REVENUE_ITEM)
- Ratio of total revenue to item price (REVENUE_PRICE)
- Ratio of total invoiced to total amount of revenue generated (INVOICED_REVENUE)

4.5.3 Cluster Modeling

Once the final dataset has been selected and ready for use, the next step was to build the clustering models by using the selected data mining tool. As it was mentioned in section 4.5.1,

the k-means clustering algorithm was selected for this study. One of the challenging tasks in the k-means clustering algorithm was to determine the value of k, which finds out the optimal clustering model that creates dissimilar segments of customers according to organizations need. Thus, the researcher used the WEKA run information; based on number of iterations and sum of squared error results to determine the k value. SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster and if all cases within a cluster are identical the SSE would be equal to 0 and the minimum value will be taken as 0 [53]. The clusters that have generated a model with a minimum amount iteration time and sum of square errors within clusters have been taken as the bench mark to assign k value. WEKA run information is depicted in the table 4.2.

Cluster No.	No. of Iterations	Sum of Squared Errors (SSE) within clusters
2	4	2383.95486279155
3	14	2368.45860509924
4	**6**	**2242.82494935422**
5	**15**	**2231.57106806998**
6	**13**	**2180.56370605793**
7	22	2269.55078155402
8	46	2283.85911159456
9	50	2269.66304530107

Table 4.8 WEKA run output information to set k value.

As it has been indicated in the table 4.8, Cluster numbers 4-6 have resulted with minimum amount of iteration time and sum of squared errors within clusters. So, by taking into consideration the organization's capacity to manage various clusters properly with the consultation of domain experts and WEKA output information, the researcher assigned k values (four, five and six).

Hence, different clustering models have been built for three different values of k (four, five and six) as well as (10, 100, 1000) seed size using WEKA software version 3.6.9. Finally, the best clustering model was selected based on specified criteria; number of iterations, inter-class similarity error and the domain experts' judgment that match with the business objective of the authority.

According to the domain experts explanation, the authority has been using a different procedures to treat both inland tax administration customers and customs administration customers. Regarding to inland tax administration customers, the authority has put in place a threshold based on the revenue generation capacity of the customers as <100,000 = C, <=200,000 = B and >= 500,000 = A. Regarding to customs administration customers who engaged in import and export business activities, the authority has put in place a threshold as diamond, gold, silver and bronze customers based on the authority's criteria for authorized economic operators (AEO) who are privileged customers. So, the threshold value has been determined together with the domain experts by taking into consideration the current practices of the authority. The threshold value for each attribute has been mentioned as follows:

TOTAL_ITEM	Very High	High	Average	Low	Very Low
TOTAL_INVOICE	Very High	High	Average	Low	Very Low
ITEM_PRICE	Very High	High	Average	Low	Very Low
ITEM_FREQUENCY	Very High	High	Average	Low	Very Low
TOTAL_REVENUE	Very High	High	Average	Low	Very Low
INVOICED_REVENUE	Very High	High	Average	Low	Very Low
REVENUE_ITEM	Very High	High	Average	Low	Very Low
REVENUE_PRICE	Very High	High	Average	Low	Very Low

List of Attributes	**Threshold values**				
	Very Low	**Low**	**Medium**	**High**	**Very High**
TOTAL_ITEM (TOIT)	<10	10-21.9	22-47.9	48-70	>70
ITEM_PRICE (ITPR)	<500	500-9999	10000-49999	50000-150000	>150000
TOTAL_INVOICE (TOIN)	<50000	50000-119999	120000-159999	160000-220000	>220000
ITEM_FREQUENCY (ITFR)	<6	6-14.99	15-39.99	40-70	>70
TOTAL_REVENUE (TORV)	<350000	350000-649999	650000-999999	1000000-1999999	>2000000
INVOICED_REVENUE (INRV)	<1	1-4.99	5-7.99	8-10	>10
REVENUE_ITEM (RVIT)	<100000	100000-299999	300000-599999	600000-999999	>1000000
REVENUE_PRICE (RVPR)	<1000	1000-2999	3000-6999	7000-9999	>10000

Table 4.9 List of range of conditions by which a cluster result is assessed

The following table depicts abbreviations of variables or attributes used for experimentation analysis.

Abbreviated Attributes	Description	Abbreviated words	Description	Abbreviated words	Description
BL	The location of the customer's business where it is established	AFR	Africa	AA	Addis Ababa
BA	The type of the business activity that the customer engaged	ASI	Asia	ET	Federal
TOIT	Total number of imported items	AUS	Australia	CG	Commercial Goods
ITPR	The price of the item	NAM	North America	IG	Industry Goods
TOIN	Total amount of invoice spent to import the item	SAM	South America	FI	Federal Investment
ITFR	The frequency of item how much recurrent	VL	Very Low	RI	Regional Investment
TORV	Total amount of revenue generated	L	Low		
INRV	Ratio of total invoiced to total amount of revenue generated	M	Medium		
RVIT	Ratio of total revenue to total no. of items	H	High		
RVPR	Ratio of total revenue to item price	VH	Very High		
CTY	Country of origin where the good is produced or manufactured				

Table 4.10 List of abbreviated words and attributes of the dataset along with their description

Determining Clusters Rank

To assign a proper rank for each cluster, the researcher has determined the required value for each discrete ranges mentioned in the threshold. The value assignment process is performed based the authority's business objective.

High value customers

Attributes	Very Low	Low	Medium	High	Very High
TOIT	5	4	3	2	1
TOIN	5	4	3	2	1
ITFR	1	2	3	4	5
TORV	1	2	3	4	5
ITPR	5	4	3	2	1
INRV	5	4	3	2	1
RVIN	1	2	3	4	5
RVPR	1	2	3	4	5

Table 4.11 value assignment for cluster ranks

High value customers need to be ranked by adding the discrete values mentioned in the table 4.11 based the authority's business objective.

Low value customers

Attributes	Very Low	Low	Medium	High	Very High
TOIT	1	2	3	4	5
TOIN	1	2	3	4	5
ITFR	5	4	3	2	1
TORV	5	4	3	2	1
ITPR	1	2	3	4	5
INRV	1	2	3	4	5
RVIN	5	4	3	2	1
RVPR	5	4	3	2	1

Table 4.12 value assignment for cluster ranks.

Low value customers need to be ranked by adding the discrete values mentioned in the table 4.12 based the authority's business objective.

Experiment I

The first experiment was conducted by using k-means algorithm to build cluster model. In this experiment, 65535 instances/records and 11 attributes were employed. Meanwhile, the number of iterations, inter-class similarity error and the objective of the authority were examined during the experiment. The output of the experiment one is depicted in the following Table 4.13.

Segmentation (K=4 and seed size 10)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	9839(15)%	10.24	15.00	117138	5.71	336856	0.67	511467.26	4771.58	ASI	ET	CG
2	14611(22%)	9.28	5234	37211	6.20	593399	0.14	260025.05	2253.11	ASI	AA	CG
3	24236(37%)	17.72	1400	86285	52.42	5544642	2.92	355507.93	21919.31	ASI	ET	CG
4	16849(26%)	15.63	2.00	107664	15.33	5725715	0.19	290500.37	13613.74	EUR	ET	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	9839(15)%	L	VL	L	VL	VL	VL	M	M			
2	14611(22%)	VL	L	VL	L	L	VL	L	L			
3	24236(37%)	L	L	L	H	VH	L	M	VH			
4	16849(26%)	L	VL	L	M	VH	VL	L	VH			
Total	**65535 (100%)**											

Table 4.13 Cluster description based on values of attributes for K=4 and seed size 10

As it has been indicated in **Table 4.13**, the upper part of the table shows the attributes' average values in each segment (1-4) and bottom part of the table shows the corresponding mapping of the values into five discrete values or ranges. After the average value of each attribute in each

cluster has been replaced with the corresponding discrete values, a summarized description has been done for each cluster in **Table 4.14**. And, the ranking of cluster is determined based on the basic attributes and the revenue generation of customers to the authority.

In experiment one, four clusters were formed but these four clusters didn't contain equal number of customers. The majority of the customers are grouped under Cluster 3 that is 24236 (37%) customers out of 65535. Cluster 4 consists of the second larger group 16849 (26%) customers out of 65535, cluster 2 contains 14611 (22%) customers, while Cluster 1 contains the least number of customers (9839 (15%).

Cluster	Description	Possible Rank
1	Low amount of items imported, very low price paid, low amount of invoice spent, very low item frequency, very low revenue generated, very low invoice/revenue, average revenue/total item, average revenue/item price, imported from ASI, location ET and Business Activity CG.	4th
2	Very low amount of items imported, low price paid, very low amount of invoice spent, low item frequency, low revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from ASI, location AA and Business Activity CG.	3rd
3	Low amount of items imported, low price paid, low amount of invoice spent, high item frequency, very high revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	1st
4	Low amount of items imported, very low price paid, low amount of invoice spent, average item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, very high revenue/item price, imported from EUR, location ET and Business Activity CG.	2nd

Table 4.14 Cluster summary and corresponding ranks based on basic attributes for K=4 and seed size 10

In the first cluster experimentation, where the value of K is four and seed size is 10, Cluster 3 is ranked first that contains customers who generated very high revenue, high item frequency, imported low amount of items and spent low amount of invoice. So Cluster 3 contains high value customers to the authority. Cluster 4 is ranked second that contains customers who generated very high revenue, average item frequency, imported low amount of items and spent low amount of invoice. Cluster 2 is ranked third that contains customers who generated low revenue, low item frequency, imported very low amount of items and spent very low amount of invoice. And, Cluster 1 is ranked fourth that contains customers who generated low revenue, by importing low amount of items and spending low amount of invoice. Finally, Cluster 3 contains high value customers, Cluster 4 contains medium level customers and Cluster 2 and 1 contain low value customers. In different literatures it is found that to make fair the distribution of instances or records in all clusters or segments, testing with different seed size is very important.

Experiment II

In the second experiment, the value of K was assigned 4 and size of seed to 100. The output of the experiment depicted in the Table 4.15 below.

Segmentation (K=4 and seed size 100)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	4964(8%)	19.08	2.00	110904	72.30	1424478	0.11	627997.79	19385.90	AFR	ET	CG
2	19681(30%)	17.05	30.00	386285	18.28	744642	0.19	258519.67	7869.90	EUR	ET	CG
3	25900(40%)	12.84	1400	71400	37.44	4182038	2.94	396515.57	21198.33	ASI	ET	CG
4	14990(23%)	9.98	5234	301324	5.71	564764	0.15	257986.88	2690.21	ASI	AA	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	4964(8%)	L	VL	L	VH	H	VL	H	VH			
2	19681(30%)	L	VL	VH	L	VH	VL	L	H			
3	25900(40%)	L	L	L	M	VH	L	M	VH			
4	14990(23%)	VL	L	VH	VL	H	VL	L	L			
Total	**65535 (100%)**											

Table 4.15 Cluster description based on values of attributes for K=4 and seed size 100

Cluster	Description	Possible Rank
1	Low amount of items imported, very low price paid, low amount of invoice spent, very high item frequency, high revenue generated, very low invoice/revenue, high revenue/total item, very high revenue/item price, imported from AFR, location ET and Business Activity CG.	1st
2	Low amount of items imported, very low price paid, very high amount of invoice spent, average item frequency, average revenue generated, very low invoice/revenue, low revenue/total item, high revenue/item price, imported from EUR, location ET and Business Activity CG.	3rd
3	Low amount of items imported, low price paid, low amount of invoice spent, average item frequency, very high revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	2nd
4	Very low amount of items imported, low price paid, very high amount of invoice spent, very low item frequency, low revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from EUR, location AA and Business Activity CG.	4th

Table 4.16 Cluster summary and corresponding ranks based on basic attributes for K=4 and seed size 100

In the second cluster experimentation, where the value of K is four and seed size is 100, Cluster 1 is ranked first that contains customers who generated very high revenue, very high item frequency, imported low amount of items and spent low amount of invoice. Cluster 3 is ranked second that contains customers who generated very high revenue, average item frequency, imported low amount of items and spent low amount of invoice. So Cluster 1 and 3 contain high value customers to the authority. Cluster 2 is ranked third that contains customers who generated average revenue, average item frequency, imported low amount of items and spent very high amount of invoice. And, Cluster 1 is ranked fourth that contains customers who generated low revenue, by importing low amount of items and spending very high amount of invoice. Cluster 3

and 4 contain medium and low level customers respectively. This cluster experimentation shows the proper cluster dissimilarity of high, medium and low customers.

As we can see from the table 4.14 experimentation two, the clusters are ranked based on the authority's business objective. Cluster 1 which contains 4964(8%) weka run result indicates the least number customers but ranked first. It is noticed that customers grouped under cluster 1 generated larger amount revenue than customers grouped under cluster 3 which contains 25900(40%) weka run result of customers. So the authority should prepare a special treating mechanism for the customers grouped under cluster 1.

4.5.3.3 Experiment III

The third experiment was conducted by setting the value of K to 4 and seed size to 1000 and it produces the following output as depicted in Table 4.17.

Segmentation (K=4 and seed size 1000)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	6469(10%)	14.14	1400	71400	126.84	5588246	4.89	153536.91	4005.78	ASI	ET	CG
2	20879(32%)	14.29	2.00	407664	13.79	5725715	0.20	290544.30	8844.23	EUR	ET	CG
3	8122(12%)	48.23	20.00	386285	41.88	5544642	4.04	118889.14	12006.81	ASI	ET	CG
4	30065(46%)	8.67	131300	131300	7.55	338415	0.47	476206.92	17707.99	ASI	ET	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	6469(10%)	L	L	L	VH	VH	L	L	M			
2	20879(32%)	L	VL	VH	L	VH	VL	L	H			
3	8122(12%)	H	VL	VH	H	VH	L	L	VH			
4	30065(46%)	VL	H	M	L	VL	VL	M	VH			
Total	**65535 (100%)**											

Table 4.17 Cluster description based on values of attributes for K=4 and seed size 1000

Cluster	Description	Possible Rank
1	Low amount of items imported, low price paid, low amount of invoice spent, very high item frequency, very high revenue generated, low invoice/revenue, low revenue/total item, average revenue/item price, imported from ASI, location ET and Business Activity CG.	1st
2	Low amount of items imported, very low price paid, very high amount of invoice spent, low item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, high revenue/item price, imported from EUR, location ET and Business Activity CG.	3rd
3	High amount of items imported, very low price paid, very high amount of invoice spent, high item frequency, very high revenue generated, low invoice/revenue, low revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	2nd
4	Very low amount of items imported, high price paid, average amount of invoice spent, low item frequency, very low revenue generated, very low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	4th

Table 4.18 Cluster summary and corresponding ranks based on basic attributes for K=4 and seed size 1000

In the third cluster experimentation, where the value of K is four and seed size 1000, Cluster 1 is ranked first that contains customers who generated very high revenue, very high item frequency, imported low amount of items and spent low amount of invoice. So Cluster 1 contains high value customers to the authority. Cluster 3 is ranked second that contains customers who generated very high revenue, high item frequency, imported high amount of items and spent high amount of invoice. Cluster 2 is ranked third that contains customers who generated very high revenue, low item frequency, imported low amount of items and spent very high amount of invoice. And, Cluster 4 is ranked fourth that contains customers who generated very low revenue, low item frequency, by importing low amount of items and spending average amount of invoice. Finally, Cluster 1 and 2 contains high value customers, Cluster 3 and 4 contain medium and low level low value customers.

Experiment IV

The fourth experiment was conducted by setting the value of K to 5 and seed size to 10 and it produces the following output illustrated in Table 4.19.

Segmentation (K=5 and seed size 10)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	9134(14%)	8.09	15.00	117138	5.62	336856	0.69	467919.63	4197.68	ASI	ET	IG
2	17702(22%)	10.15	5234	310438	6.07	3076875	0.15	252124.60	2504.94	ASI	AA	CG
3	21299(33%)	28.06	71400	86285	44.26	5544642	3.31	375245.66	22837.76	ASI	ET	CG
4	4346(7%)	20.55	131300	111300	80.17	1424478	0.12	624877.29	14858.29	AFR	ET	CG
5	16054(24%)	16.29	2.00	407664	15.19	5725715	0.20	227575.41	13349.76	EUR	ET	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	9134(14%)	VL	VL	L	VL	VL	VL	M	M			
2	17702(22%)	L	L	VH	L	VH	VL	L	L			
3	21299(33%)	M	H	L	H	VH	L	M	VH			
4	4346(7%)	L	H	L	VH	VH	VL	L	VH			
5	16054(24%)	L	VL	VH	L	VH	VL	L	VH			
Total	**65535 (100%)**											

Table 4.19 Cluster description based on values of attributes for K=5 and seed size 10

Cluster	Description	Possible Rank
1	Very low amount of items imported, very low price paid, low amount of invoice spent, very low item frequency, very low revenue generated, very low invoice/revenue, average revenue/total item, average revenue/item price, imported from ASI, location ET and Business Activity IG.	5th
2	Low amount of items imported, low price paid, very high amount of invoice spent, low item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from ASI, location AA and Business Activity CG.	4th
3	Average amount of items imported, high price paid, low amount of invoice spent, high item frequency, very high revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	2nd
4	Low amount of items imported, high price paid, low amount of invoice spent, very high item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, very high revenue/item price, imported from AFR, location ET and Business Activity CG.	1st
5	Low amount of items imported, very low price paid, very high amount of invoice spent, low item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, very high revenue/item price, imported from EUR, location ET and Business Activity CG.	3rd

Table 4.20 Cluster summary and corresponding ranks based on basic attributes for K=5 and seed size 10

In the fourth cluster experimentation, where the value of K is five and seed size 10, Cluster 4 is ranked first that contains customers who generated very high revenue, very high item frequency, imported low amount of items and spent low amount of invoice. So Cluster 4 contains high value customers to the authority. Cluster 3 is ranked second that contains customers who generated very high revenue, high item frequency, imported average amount of items and spent low amount of invoice. Cluster 5 is ranked third that contains customers who generated very high revenue, low item frequency, imported low amount of items and spent very high amount of invoice. . Cluster 2 is ranked fourth that contains customers who generated very high revenue, low item frequency, imported low amount of items and spent very high amount of invoice. And,

Cluster 1 is ranked fourth that contains customers who generated very low revenue, very low item frequency, by importing low amount of items and spending low amount of invoice.

Experiment V

The fifth experiment was conducted by setting the value of K to 5 and seed size to 100 and it produces the following output illustrated in Table 4.21.

Segmentation (K=5 and seed size 100)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	5262(8%)	19.04	2.00	110904	70.01	1424478	0.11	602843.59	18297.64	AFR	ET	CG
2	19420(30%)	16.91	30.00	386285	13.20	5544642	0.19	253592.95	7686.42	EUR	ET	CG
3	20914(32%)	17.98	71400	71400	15.93	336856	1.60	457333.45	26634.21	ASI	ET	CG
4	14197(22%)	8.93	5234	301324	5.59	1064764	0.15	267824.18	2366.23	ASI	AA	CG
5	5742(9%)	41.82	107100	107100	111.17	4182038	7.49	152973.47	750.35	ASI	ET	CG
Total	**65635 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	5262(8%)	L	VL	L	VH	H	VL	H	VH			
2	19420(30%)	L	VL	VH	L	VH	VL	L	H			
3	20914(32%)	L	H	L	M	VL	L	M	VH			
4	14197(22%)	VL	L	VH	VL	H	VL	L	L			
5	5742(9%)	M	H	L	VH	VH	M	L	VL			
Total	**65635 (100%)**											

Table 4.21 Cluster description based on values of attributes for K=5 and seed size 100

Cluster	Description	Possible Rank
1	low amount of items imported, very low price paid, low amount of invoice spent, very high item frequency, high revenue generated, very low invoice/revenue, high revenue/total item, very high revenue/item price, imported from AFR, location ET and Business Activity CG.	1st
2	Low amount of items imported, very low price paid, very high amount of invoice spent, low item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, high revenue/item price, imported from EUR, location ET and Business Activity CG.	3rd
3	Low amount of items imported, high price paid, low amount of invoice spent, average item frequency, very low revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	5th
4	Very low amount of items imported, low price paid, very high amount of invoice spent, very low item frequency, high revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from ASI, location AA and Business Activity CG.	4th
5	Average amount of items imported, high price paid, low amount of invoice spent, very high item frequency, very high revenue generated, average invoice/revenue, low revenue/total item, very low revenue/item price, imported from ASI, location ET and Business Activity CG.	2nd

Table 4.22 Cluster summary and corresponding ranks based on basic attributes for K=5 and seed size 100

In the fifth cluster experimentation, where the value of K is five and seed size 100, Cluster 1 is ranked first that contains customers who generated very high revenue, very high item frequency, imported low amount of items and spent low amount of invoice. So Cluster 1 contains high value customers to the authority. Cluster 5 is ranked second that contains customers who generated very high revenue, very high item frequency, imported average amount of items and spent low amount of invoice. Cluster 2 is ranked third that contains customers who generated very high revenue, low item frequency, imported low amount of items and spent very high amount of invoice. . Cluster 4 is ranked fourth that contains customers who generated high revenue, very low item frequency, imported very low amount of items and spent very high amount of invoice. And, Cluster 3 is ranked fifth that contains customers who generated very low revenue, average item frequency, by importing low amount of items and spending low amount of invoice.

Experiment VI

The sixth experiment was conducted by setting the value of K 5 and size of seed to 1000 and it produces the following output illustrated in Table 4.23.

Segmentation (K=5 and seed size 1000)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	3421(5%)	33.57	71400	71400	150.39	5588246	4.93	136447.07	4426.60	ASI	ET	CG
2	20969(32%)	13.97	2.00	407664	14.27	5725715	0.20	283527.08	8573.80	EUR	ET	CG
3	5789(9%)	48.11	20.00	57635	18.31	622483	3.78	39348.77	11526.55	ASI	ET	CG
4	28708(44%)	7.26	131300	131300	7.87	166912	0.35	498536.87	18110.02	ASI	ET	CG
5	6648(10%)	41.64	61380	386285	79.71	5544642	4.48	209150.42	8863.57	ASI	ET	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	3421(5%)	M	H	L	VH	VH	M	L	M			
2	20969(32%)	L	VL	VH	L	VH	VL	L	H			
3	5789(9%)	H	VL	L	M	L	L	VL	VH			
4	28708(44%)	VL	H	M	L	VL	L	M	VH			
5	6648(10%)	M	H	VH	VH	VH	M	L	H			
Total	**65535 (100%)**											

Table 4.23 Cluster description based on values of attributes for K=5 and seed size 1000

Cluster	Description	Possible Rank
1	Average amount of items imported, high price paid, low amount of invoice spent, very high item frequency, very high revenue generated, average invoice/revenue, low revenue/total item, average revenue/item price, imported from ASI, location ET and Business Activity CG.	1st
2	Low amount of items imported, very low price paid, very high amount of invoice spent, low item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, high revenue/item price, imported from EUR, ET and Business Activity CG.	3rd
3	High amount of items imported, very low price paid, low amount of invoice spent, average item frequency, low revenue generated, low invoice/revenue, very low revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	4th
4	Very low amount of items imported, high price paid, average amount of invoice spent, low item frequency, very low revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	5th
5	Average amount of items imported, high price paid, very high amount of invoice spent, very high item frequency, very high revenue generated, average invoice/revenue, low revenue/total item, high revenue/item price, imported from ASI, location ET and Business Activity CG.	2nd

Table 4.24 Cluster summary and corresponding ranks based on basic attributes for K=5 and seed size 1000

In the sixth cluster experimentation, where the value of K is five and seed size 1000, Cluster 1 is ranked first that contains customers who generated very high revenue, very high item frequency, imported average amount of items and spent low amount of invoice. So Cluster 1 contains high value customers to the authority. Cluster 5 is ranked second that contains customers who generated very high revenue, very high item frequency, imported average amount of items and spent very high amount of invoice. Cluster 2 is ranked third that contains customers who generated very high revenue, low item frequency, imported low amount of items and spent very high amount of invoice. . Cluster 3 is ranked fourth that contains customers who generated low revenue, average item frequency, imported high amount of items and spent very high amount of invoice. And, Cluster 4 is ranked fifth that contains customers who generated very low revenue, low item frequency, by importing very low amount of items and spending average amount of invoice.

Experiment VII

The seventh experiment was conducted by setting the value of K to 6 and seed size to 10 and it produces the following output as depicted in Table 4.25.

Segmentation (K=6 and seed size 10)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	7424(11%)	14.58	15.00	117138	5.57	336856	0.83	381495.19	7743.51	ASI	ET	IG
2	14235(22%)	9.03	5234	37211	6.32	593399	0.14	264621.32	2178.92	ASI	AA	CG
3	4447(7%)	31.40	2.00	86285	78.56	5544642	0.11	606302.24	19947.57	AFR	ET	CG
4	15571(24%)	12.78	30.00	3875	6.80	117970	0.20	289692.27	5810.21	EUR	ET	CG
5	2792(4%)	19.03	5.00	107664	56.56	5725715	0.13	192852.29	7336.22	EUR	ET	CG
6	21066(32%)	25.77	2.00	71400	44.20	4182038	3.35	379649.81	26218.41	ASI	ET	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	7424(11%)	L	VL	L	VL	VL	VL	M	H			
2	14235(22%)	VL	L	VL	L	L	VL	L	L			
3	4447(7%)	M	VL	L	VH	VH	VL	H	VH			
4	15571(24%)	L	VL	VL	L	VL	VL	L	M			
5	2792(4%)	L	VL	L	H	VH	VL	L	H			
6	21066(32%)	M	VL	L	H	VH	L	M	VH			
Total	**65535 (100%)**											

Table 4.25 Cluster description based on values of attributes for K=6 and seed size 10

Cluster	Description	Possible Rank
1	Low amount of items imported, very low price paid, low amount of invoice spent, very low item frequency, very low revenue generated, very low invoice/revenue, average revenue/total item, high revenue/item price, imported from ASI, location ET and Business Activity IG.	6th
2	Very low amount of items imported, low price paid, very low amount of invoice spent, low item frequency, low revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from ASI, location AA and Business Activity CG.	4th
3	Average amount of items imported, very low price paid, low amount of invoice spent, very high item frequency, very high revenue generated, very low invoice/revenue, high revenue/total item, very high revenue/item price, imported from AFR, location ET and Business Activity CG.	1st
4	Low amount of items imported, very low price paid, very low amount of invoice spent, low item frequency, very low revenue generated, very low invoice/revenue, low revenue/total item, average revenue/item price, imported from EUR, location ET and Business Activity CG.	5th
5	Low amount of items imported, very low price paid, low amount of invoice spent, high item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, high revenue/item price, imported from EUR, location ET and Business Activity CG.	3rd
6	Average amount of items imported, very low price paid, low amount of invoice spent, high item frequency, very high revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	2nd

Table 4.26 Cluster summary and corresponding ranks based on basic attributes for K=6 and seed size 10

In the seventh cluster experimentation, where the value of K is six and seed size 10, Cluster 3 is ranked first that contains customers who generated very high revenue, very high item frequency, imported average amount of items and spent low amount of invoice. So Cluster 3 contains high value customers to the authority. Cluster 6 is ranked second that contains customers who generated very high revenue, high item frequency, imported average amount of items and spent very high amount of invoice. Cluster 3 is ranked third that contains customers who generated very high revenue, high item frequency, imported low amount of items and spent low amount of invoice. . Cluster 2 is ranked fourth that contains customers who generated low revenue, low item frequency, imported very low amount of items and spent very low amount of invoice.

Cluster 4 is ranked fifth that contains customers who generated very low revenue, low item frequency, imported low amount of items and spent very low amount of invoice. And, Cluster 1 is ranked sixth that contains customers who generated very low revenue, very low item frequency, by importing low amount of items and spending low amount of invoice.

Experiment VIII

In the eighth experiment, value of K set to be 6 and size of seed to 100 and it produces the following output illustrated in Table 4.27.

Segmentation (K=6 and seed size 100)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	4475(7%)	20.83	2.00	110904	78.46	1424478	0.11	608241.65	21359.01	AFR	ET	CG
2	11640(18%)	21.87	30.00	386285	18.31	5544642	0.13	209195.23	9707.59	EUR	ET	CG
3	19041(29%)	19.25	71400	71400	16.73	336856	1.73	455384.63	28329.20	ASI	ET	CG
4	14332(22%)	9.06	5234	301324	5.60	1064764	0.13	263950.57	2378.11	ASI	AA	CG
5	5701(9%)	41.83	107100	107100	111.51	4182038	7.52	153818.91	749.06	ASI	ET	CG
6	10346(16%)	8.60	5.00	569876	7.42	5328625	0.29	372562.29	5232.40	EUR	ET	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	4475(7%)	L	VL	L	VH	H	VL	H	VH			
2	11640(18%)	L	VL	VH	M	VH	VL	L	H			
3	19041(29%)	L	H	L	M	VL	L	M	VH			
4	14332(22%)	VL	L	VH	VL	H	VL	L	L			
5	5701(9%)	M	H	L	VH	VH	M	L	VL			
6	10346(16%)	VL	VL	VH	L	VH	VL	M	M			
Total	**65535 (100%)**											

Table 4.27 Cluster description based on values of attributes for K=6 and seed size 100

Cluster	Description	Possible Rank
1	Low amount of items imported, very low price paid, low amount of invoice spent, very high item frequency, high revenue generated, very low invoice/revenue, high revenue/total item, very high revenue/item price, imported from AFR, location ET and Business Activity CG.	2nd
2	Low amount of items imported, very low price paid, very high amount of invoice spent, average item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, high revenue/item price, imported from EUR, location ET and Business Activity CG.	3rd
3	Low amount of items imported, high price paid, low amount of invoice spent, average item frequency, very low revenue generated, low invoice/revenue, average revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	6th
4	Very low amount of items imported, low price paid, very high amount of invoice spent, very low item frequency, high revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from ASI, location AA and Business Activity CG.	5th
5	Average amount of items imported, high price paid, low amount of invoice spent, very high item frequency, very high revenue generated, average invoice/revenue, low revenue/total item, very low revenue/item price, imported from ASI, location ET and Business Activity CG.	1st
6	Very low amount of items imported, very low price paid, very high amount of invoice spent, low item frequency, very high revenue generated, very low invoice/revenue, average revenue/total item, average revenue/item price, imported from EUR, location ET and Business Activity CG.	4th

Table 4.28 Cluster summary and corresponding ranks based on basic attributes for K=6 and seed size 100

In the eighth cluster experimentation, where the value of K is six and seed size 100, Cluster 5 is ranked first that contains customers who generated very high revenue, very high item frequency, imported average amount o f items and spent low amount of invoice. So Cluster 5 contains high value customers to the authority. Cluster 1 is ranked second that contains customers who generated very high revenue, very high item frequency, imported low amount of items and spent very high amount of invoice. Cluster 2 is ranked third that contains customers who generated very high revenue, average item frequency, imported low amount of items and spent very high amount of invoice. Cluster 6 is ranked fourth that contains customers who generated very high revenue, low item frequency, imported very low amount of items and spent very high amount of

invoice. Cluster 4 is ranked fifth that contains customers who generated very high revenue, very low item frequency, imported very low amount of items and spent very high amount of invoice. And, Cluster 3 is ranked sixth that contains customers who generated very low revenue, average item frequency, by importing low amount of items and spending low amount of invoice.

Experiment VIIII

In the eighth experiment, value of K set to be 6 and size of seed to 100 and it produces the following output illustrated in Table 4.29.

Segmentation (K=6 and seed size 1000)												
Cluster #	Freq. of records (dist. in %)	TOIT	ITPR	TOIN	ITFR	TORV	INRV	RVIT	RVPR	CTY	BL	BA
1	3419(5%)	33.57	71400	71400	150.14	5588246	4.92	136507.33	4425.10	ASI	ET	CG
2	19198(29%)	14.23	2.00	107664	15.26	5725715	0.19	278834.88	9031.17	EUR	ET	CG
3	5265(8%)	48.99	20.00	57635	19.90	622483	4.10	41286.38	12177.96	ASI	ET	CG
4	18197(28%)	6.46	131300	131380	9.27	215896	0.50	653612.38	27265.63	ASI	ET	CG
5	6600(10%)	41.72	61380	86285	79.93	5544642	4.51	209987.42	8910.18	ASI	ET	CG
6	12856(20%)	10.31	5234	37211	5.23	593399	0.15	235331.24	2558.76	ASI	AA	CG
Total	**65535 (100%)**											
Cluster #	Freq. of records (dist. in %)											
1	3419(5%)	M	H	L	VH	VH	L	L	M			
2	19198(29%)	L	VL	L	M	VH	VL	L	H			
3	5265(8%)	H	VL	L	M	L	L	VL	VH			
4	18197(28%)	VL	H	M	L	VL	VL	H	VH			
5	6600(10%)	M	H	L	VH	L	L	L	H			
6	12856(20%)	L	L	VL	VL	L	VL	L	L			
Total	**65535 (100%)**											

Table 4.29 Cluster description based on values of attributes for K=6 and seed size 1000

Cluster	Description	Possible Rank
1	Average amount of items imported, high price paid, low amount of invoice spent, very high item frequency, very high revenue generated, low invoice/revenue, low revenue/total item, average revenue/item price, imported from ASI, location ET and Business Activity CG.	1st
2	Low amount of items imported, very low price paid, low amount of invoice spent, average item frequency, very high revenue generated, very low invoice/revenue, low revenue/total item, very high revenue/item price, imported from EUR, location ET and Business Activity CG.	2nd
3	high amount of items imported, very low price paid, low amount of invoice spent, average item frequency, low revenue generated, low invoice/revenue, very low revenue/total item, very high revenue/item price, imported from EUR, location ET and Business Activity CG.	4th
4	Very low amount of items imported, high price paid, average amount of invoice spent, low item frequency, very low revenue generated, very low invoice/revenue, high revenue/total item, very high revenue/item price, imported from ASI, location ET and Business Activity CG.	6th
5	Average amount of items imported, high price paid, low amount of invoice spent, very high item frequency, low revenue generated, low invoice/revenue, low revenue/total item, high revenue/item price, imported from ASI, location ET and Business Activity CG.	3rd
6	Low amount of items imported, low price paid, very low amount of invoice spent, very low item frequency, low revenue generated, very low invoice/revenue, low revenue/total item, low revenue/item price, imported from ASI, location AA and Business Activity CG.	5th

Table 4.30 Cluster summary and corresponding ranks based on basic attributes for K=6 and seed size 1000

In the eighth cluster experimentation, where the value of K is six and seed size 100, Cluster 1 is ranked first that contains customers who generated very high revenue, very high item frequency, imported average amount of items and spent low amount of invoice. So Cluster 1 contains high value customers to the authority. Cluster 2 is ranked second that contains customers who generated very high revenue, average item frequency, imported low amount of items and spent low amount of invoice. Cluster 5 is ranked third that contains customers who generated low revenue, very high item frequency, imported average amount of items and spent low amount of invoice. . Cluster 3 is ranked fourth that contains customers who generated low revenue, average

item frequency, imported high amount of items and spent very low amount of invoice. Cluster 6 is ranked fifth that contains customers who generated low revenue, very low item frequency, imported very low amount of items and spent very low amount of invoice. And, Cluster 4 is ranked sixth that contains customers who generated very low revenue, low item frequency, by importing very low amount of items and spending average amount of invoice.

In general, most of the clustering experiments conducted in this research are good in segmenting customers according to the business objectives of the authority (identification of customers as high, medium, and low values to the authority). Thus, to select the best clustering model from the delivered models, the researcher used different clustering model comparison methods. So, the comparison methods applied for this cluster model are: domain experts decision about number of clusters, number iteration that the algorithms run to read the dataset and sum of square-error (SSE) between instances within a cluster.

4.5.4 Choosing the best Clustering Model

In finding out of the best clustering model or segmentation of customers, nine cluster modeling experiments were conducted and come up with the appropriate clustering model. In this regard, the domain experts also played a great role, because the experts can understand each segmentation needs and business know-how in the authority. And finally, the best clustering model that satisfy the criteria was selected.

As it has been indicated in the above experiments, to improve the distribution of instances in different segments, different values of K (4, 5 and 6) with different seed values (10, 100 and 1000) have been conducted and after a number of experiments, better cluster model has been selected. The value of K=4 with seed value 100 gives better distribution of instances in the segments. Moreover, based on additional advices from the domain experts, about the number of clusters where the value of K=4 and seed size 100, is good. Moreover, in this experiment, dissimilar clusters were formed and the number of iterations that the algorithm runs to read the dataset and the sum of squared- errors between inter-class is minimum than in other clustering experiments. The summary of experimentation is illustrated at table **4.26**.

Experiment No.	Number of K	Seed Size	Number of iteration	Within cluster sum of squared errors
1	4	10	10	2249.78646217118
2	**4**	**100**	**6**	**2142.82494935422**
3	4	1000	11	2363.88501674263
4	5	10	15	2208.936490582
5	5	100	15	2231.57106806998
6	5	1000	11	2360.6019080062
7	6	10	24	2205.05289388125
8	6	100	13	2180.56370605793
9	6	1000	17	2271.6725495066

Table 4.31 Comparison of clustering model experiment

4.6 Evaluation

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

To understand business process of the authority, it was important to communicate the authority's domain experts. As it has been indicated in the data understanding section, the researcher conducted informal and formal interviews with ERCA's domain experts. So, the researcher prepared interview questions and five experts were selected for interview. The selection method was conducted randomly based on their knowhow on the area. After the selected five experts are interviewed, the result of the interview was promising for the study. Sample interview questions are annexed as annex 3.

Segmenting customers into meaningful groups according to the organizations' need is very important to treat them accordingly and to apply appropriate CRM strategies. The clustering process was focused on customer's value to the business. This customer value was defined by the selected basic attributes: revenue generation, item frequency, total item, and total invoice to the authority. Thus, the clustering result selected (where k=4 and seed size 100) was good since it

clusters customers having similar revenue generation patterns in the same group and the groups formed were different from each other. It clusters customers as high, medium and low value customer according to the organization's need. Besides, the selected clustering model segmented customers based on minimum number of iterations and sum of squared error (SSE) within a cluster. Thus, as it has been clearly indicated in the experiment, the selected clustering model brought customers into different clusters according to their individual behaviors.

4.7 Deployment of the Result

In the data mining process, once the data mining model is built and validated, it can be used in one of the two main ways: recommending actions by viewing the model with its results and applying the model to different data sets [54]. Deployment is the last step of data mining process and the result of data mining for this research was building a clustering model. The knowledge gained or mined from data needs to be organized and presented in a way that the organization can understand and use it for proper customer handling and effective customer relationship management.

The result of this study will require modification on the existing policies and procedures of the authority. So, before the implementation of deployment process, the changes those resulted from the research's findings has to be accepted, approved and owned by the business domain experts.

To make the result which is obtained from the data mining process applicable; integration of resources like people, business processes and technology are needed. Moreover, the integration of resources is based on the information or result obtained from the segmentation model. After integrating the necessary adjustments of domain experts the result of this study can be deployed for identifying high, medium, and low value customers and treating them accordingly and also for decision making and successful CRM purpose in the authority and also the authority will be beneficiary from the research result.

As it has been indicated in the experimentation process, customers are identified as high value customers and to be provided special service for them if they are generating high revenue by importing small amount of goods, spending low amount of invoice and high item frequency. On the other hand customers are identified as low value customers who are generating low revenue by importing large amount of items, spending high amount invoice and low item frequency.

CHAPTER FIVE

5. Conclusion and Recommendations

5.1 Conclusion

The essence of the information technology revolution and in particular, the World Wide Web is an emerging opportunity to build better relationships with customers than has been previously possible in the offline world. Currently, the organizations have been focusing on customer attraction, retention and shifting from product centric to the customer centric approach. In this research an attempt has been made to reveal the big potential of data mining applications for customer segmentation, referring to the optimal usage of data mining techniques to thoroughly analyze the collected historical customers' data of the Ethiopian Revenues and Customs Authority (ERCA) and to segment customers based on their value.

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

In the data mining process of this research, a half year (six months) 131072 records of data was collected and from this 65535 records of data was employed. At the same time, 11 attributes were selected out of 22 attributes for the final experimentation process. In this study, extensive time has been spent on the data preparation phase and consulting the domain experts in the authority on the interpretation and selection of the appropriate model developed, which were built for clustering model of data mining.

The goal of the business to be achieved in this research was to group ERCA's customers into similar groups based on their revenue generation and transaction volume pattern. The criterion to segment customers was based on customers' value to the business, which is measured with basic attributes. These basic attributes were revenue generated, item frequency, invoice spent, total number of items imported, the price of items imported, the items originated behaviors, business

location and Business area. With these basic attributes the descriptive data mining task (clustering) was employed. Based on the preprocessed dataset, different clustering experiments have been performed with the K-means clustering algorithm by assigning values for K(4, 5 and 6) seed values (10,100,1000). The model where **K=4** and **seed size 100** had shown better clustering performance. It enables to cluster the authority's customers into dissimilar clusters of high, medium and low value customer groups with minimum iteration of algorithm and minimum sum of square error (SSE) between records in a cluster, **6** and **2142.82**, respectively. So, this experiment or cluster model has been chosen for cluster index.

5.2 Recommendations

The researcher believes that the contribution of this research work could be a good knowledge for a competitive study in customer-oriented business organizations in the future. The findings of the research would encourage business-oriented organizations to work on the application of data mining techniques for successful CRM implementation so as to gain a competitive advantage in the global economy.

The application of data mining techniques has been increasingly becoming very important technology for different sectors that have a huge data, such as retail trade, airline industry, banking, telecommunication, electric power industry, healthcare sectors and service giving governmental organizations.

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

- ERCA should develop an integrated data warehouse for big data analysis application and generate valuable knowledge from its experimental data.
- To increase the satisfaction of customers so as to generate higher revenue, ERCA should plan to develop awareness creation to its employees about the importance of CRM. Moreover, data mining techniques for customer segmentation should enable the authority to identify high value and low customers.
- The result of this study should be implemented by the authority on the ground to improve the authority's customer support services.

- Now a days, due to the dynamic technological changes, business problems have been becoming complex and diverse. So, the researcher believes that, application of other data mining techniques will improve applicability of customer relationship management in the authority.
- The result of this study is promising, but the researcher recommends that this result will be used as input for further research in the authority by applying different data mining technologies like classification technique. In addition, the organization has many potential areas for the future research.

Reference

[1] Agrawal D. (2013). *A Comprehensive Study of Data Mining and Application.* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 1.

[2] Almotairi M. A. T. (2009*). Evaluation of the Implementation of CRM in Developing Countries.* Brunel Business School, Brunel University, A thesis submitted for the degree of Doctor of Philosophy

[3] Ayre L.B. (2006). *Data Mining: What It Is and How It Is Related To Information Retrieval And Text Mining.*

[4] Babita C., Vivek B. and Balram K. (2011). *Implementation of Data Mining Techniques for Strategic CRM Issues.* Department of Computer Sciences, Vivek Bhambri et al, Int. J. Comp. Tech. Appl., Vol 2 (4), 879-883.

[5] Belachew R. (2013). *Application of Data Mining Techniques for Customers Segmentation and Prediction:* The Case of Buusaa Gonofa Microfinance Institution. (*Master's Thesis*), Addis Ababa University.

[6] Belete B. (2011). *Knowledge Discovery for effective customer segmentation: the case of ERCA.* (Master's Thesis), Addis Ababa, Ethiopia: Department of Information Science, Addis Ababa University.

[7] Benjamin A. K. and Andrews K. D. (2010). *Towards a successful customer relationship management: A conceptual framework.* Faculty of Engineering, University of Mines and Technology, Tarkwa, Ghana, African Journal of Marketing Management Vol. 2(3) pp. 037-043

[8] Berry M. J.A. and Linoff G. S. (2004), *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*, 2nd edition. Wiley Publishing, Inc., Indianapolis, Indiana

[9] Boris B. (2012). *Knowledge Acquisition in Databases.* Management Information Systems, Vol. 7 (2012), No. 1, pp. 032-041.

[10] Buchanan B. G. (2006). *A (Very) Brief History of Artificial Intelligence*. AI Magazine Volume 26 Number 4.

[11] Dandi M. (2016). *Application of Data Mining Techniques for Customers Segmentation in Insurance Business:* The Case of Ethiopian Insurance Corporation. (*Master's Thesis*), Addis Ababa University.

[12] Daniel M. (2013). *Application of Data Mining Technology to support fraud protection: the case of ERCA.* (Master's Thesis), Addis Ababa, Ethiopia: Department of Information Science, Addis Ababa University.

[13] Denekew A. (2003). *The application of Data Mining to support customer relationship management at Ethiopian Airlines*. (*Master's Thesis*), Department of Information Science Addis Ababa University.

[14] Dilmuradjon Z. and Nodar M. (2015). *Application of Data Mining in the Banking Sector*. Journal of Technical Science and Technologies; ISSN 2298-0032; Volume 4, Issue 1.

[15] Dunham M.H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

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

[17] Emilia M. (2008*). Behavioral Segmentation of Telecommunication Customers*. (*Master's Thesis*), Royal Institute of Technology School of Computer Science and Communication KTH CSC SE-100 44 Stockholm, Sweden.

[18] Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996). *From Dat Mining to Knowledge Discovery in Databases.* American Association for Artificial Intelligence.

[19] Federal Democratic Republic of Ethiopia (2008). *Customs Proclamation No. 587/2008, Federal Negarit Gazeta*, Addis Ababa, Ethiopia.

[20] Gray P. and Byun J. (2001). *Customer Relationship Management.* Center for Research on Information Technology and Organization UC Irvine, Claremont Graduate School, University of California.

[21] Gupta G. and Aggarwal H. (2012). *Improving Customer Relationship Management Using Data Mining.* International Journal of Machine Learning and Computing, Vol. 2, No. 6.

[22] Hajizadeh E., Ardakani H. D. and Shahrabi J. (2010*). Application of data mining techniques in stock markets: A survey.* Industrial Engineering Department, Amirkabir University of Technology, Tehran, Iran, Journal of Economics and International Finance Vol. 2(7), pp. 109-118.

[23] Hailemariam A. (2011). *Application of Data Mining Techniques to Customer Profile Analysis in the Ethiopian Electric Power Corporation*. (*Master's Thesis*), Addis Ababa University.

[24] Hand, D.(2006), Principles of data mining. Prentice-Hall of India, New Delhi, India.

[25] Han J. and Kamber M. (2006), Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann.

[26] Han J. and Kamber M., Pei J. (2012), Data Mining: Concepts and Techniques, 3nd edition, Morgan Kaufmann.

[27] Harry Z. (2004). *The Optimality of Naive Bayes*. Faculty of Computer Science University of New Brunswick Fredericton, New Brunswick, Canada E3B 5A3

[28] Henok W. (2002). *The application of Data Mining to support customer relationship management at Ethiopian Airlines*. (*Master's Thesis*), Department of Information Science Addis Ababa University.

[29] Hisham S. S. (2011). *Customer Relationship Management and Its Relationship to the Marketing Performance.* Dep't of Business Management Faculty of Commerce Cairo University, International Journal of Business and Social Science Vol. 2, No . 10

[30] Jansen S.M.H. (2007). *Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior*. A Vodafone Case Study.

[31] Joyce J. (2002). *Data Mining: A Conceptual Overview*. Management Science Department University of South Carolina, Communications of the Association for Information Systems, Volume 8, 267-296.

[32] Khalid R., Haslina M., and Huda I. (2011*). Customer Relationship Management (CRM) Processes from Theory to Practice: The Pre-implementation Plan of CRM System.* International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 1, No. 1

[33] Kishana R. K., *Member, IACSIT*, and Velu C. M. (2013). *Customer Segmentation Using Clustering and Data Mining Techniques.* International Journal of Computer Theory and Engineering, Vol. 5, No. 6.

[34] Konstantinos T. and Antonios C. (2009). Data Mining Techniques in CRM: Inside Customer Segmentation, 1st Edition. John Wiley and Sons, Ltd., Publication, United Kingdom

[35] Krzysztof J. Cios, Witold P., Roman W. S. and Lukasz A. K. (2007). *Data Mining: A Knowledge Discovery Approach.* Springer Science+Business Media, LLC

[36] Mike M. (2007). *Customer Segmentation and Predictive Modeling*. Behavioral Science.

[37] Milad F. and Tannaz R. A. (2014). *The Role of Information Technology Infrastructure for Customer Relationship Management Implementation of Manufacturing Companies.* Department of Management, Islamic Azad University, Babol, Iran, A Journal of Economics and Management Vol. 3 Issue 8 August 2014, ISSN 2278-0629 , 159-179

[38] Monika G. and Rajan V. (2012). *Applications of Data Mining in Higher Education*. CSE Department, BahraUniversity, Waknaghat, H.P 173234, India, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1

[39] Ngai E.W.T., Xiu L. and Chau D.C.K. (2009). *Application of data mining techniques in customer relationship management: A literature review and classification*. E.W.T. Ngai et al. / Expert Systems with Applications 36 (2009) 2592–2602.

[40] Patil T. R. and Sherekar M. S. S. (2013). *Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*. Sant Gadgebaba Amravati University, Amravati, International Journal Of Computer Science And Applications Vol. 6, No.2

[41] Priya T. U. (Dr) (2015*). Customer Relationship Management And Organizational Competitiveness Of Commercial Banks In Chennai.* International Journal of Management Academy, 3 (4): 15-24

[42] Rahul B. Diwate , Amit S. (2014). *Data Mining Techniques in Association Rule : A Review.* Department of computer Science & Engineering, G.H.R.C.E.M. Amravati, Amravati University, India, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 227 -229

[43] Rodpysh K. V., Aghai A. and Majdi M. (2012). *Applying Data Mining in Customer Relationship Management*. Dep't of IT, Gilan University of Applied Sciences Crescent, Rasht, Iran, Industrial Management, University Ramseur, Iran and Department of Industrial Engineering, University Ayandegan Tonkabon, Iran International Journal of Information Tech, Control & Automation (IJITCA) Vol.2, No.3.

[44] Rygielski C., Jyun-C.W. and Yen D. C. (2002). *Data mining techniques for customer relationship management.* Technology in Society, 24, Pp. 483–502.

[45] Siraj F. and Abdoulha M. A. *Mining Enrolment Data Using Predictive and Descriptive Approaches.* Applied Sciences, College of Arts & Sciences, Universiti Utara Malaysia Malaysia.

[46] Tariku A. (2011*). Mining Insurance Data for fraud Detection: The Case of Africa Insurance Share Company*. (*Master's Thesis*), Department of Information Science Addis Ababa University.

[47] Tenkir S. (2009). *Customs Management System in Ethiopia.* (Master's Thesis), Addis Ababa, Ethiopia: Department of Accounting and Finance, Addis Ababa University.

[48] The CRISP-DM consortium (2000). *Step-by-step data mining guide*, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA),and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands).

[49] Two Crows Corporation (1999), *Introduction to data mining and knowledge discovery, 3rd*. Potomac, MD 20854, USA.

[50] Umair S. and Haseeb Q. (2014). *A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA).* International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 12 No. 1 Nov. 2014, pp. 217-222.

[51] Upadhyay T., Vidhani A.and Dadhich V. (2016). *Customer Profiling and Segmentation using Data Mining Techniques*. Nirma University, Volume 7, Number 2 March 2016 - Sept 2016, pp. 65-67.

[52] VLADIMIR B. and JOHN Z. (1999). *Knowledge discovery and data mining in biological databases.* The Knowledge Engineering Review, Vol. 14:3, 1999, 257-277.

[53] Yu Xia and Jiming Peng. *A Cutting Algorithm for the Minimum Sum-of-Squared Error Clustering*.

[54] Zaki M. J. and Meira W. J. (2014). *Data Mining and Analysis Fundamental Concepts and Algorithms*. Cambridge University Press is part of the University of Cambridge, 32 Avenue of the Americas, New York, NY 10013-2473, USA

[55] Ziafat H. and Shakeri M. (2014). *Using Data Mining Techniques in Customer Segmentation.* Hasan Ziafat Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.70-79.

[56] Zhiyuan Y. (2013). *Visual Customer Segmentation and Behavior Analysis - A SOM-Based Approach*. Abo Akademi University, Department of Information Technologies Joukahaisenkatu 3-5, FIN-20520 Turku, Finland.

Appendix 1: Sample initial data

1	SAD_CONSIGNEE	CMP_NAM	SAD_REG_DATE	SAD_ITM_TOTAL	SAD_PACK_TOTAL	SADITM_ITM_PRICE	SAD_TOT_INVOICED	SADITM_HS_COD	CTY_DSC	CDI_AMOUNT_TOTAL
2	0039788043AARI	SAMSON ASRAT HAILEGIORGIS	01-01-2016	31.00	589.00	691.25	13,018.20	85166000	China	80,569.90
3	0002447647AA02	SINKENESH WORKU JIRRU	01-01-2016	1.00	10.00	15,310.00	15,310.00	62044900	Thailand	324,578.44
4	0000517721AA03	ORION SHEBABAW HAILE	01-01-2016	8.00	1.00	220.90	7,485.94	85299090	China	139,754.73
5	0003999592ETFI	SOFOMAR TOUR & TRAVEL PRIVATE LIMIT	01-01-2016	6.00	3.00	11,600.00	11,915.00	82071900	India	146,679.85
6	0000022765ET02	EQUATORIAL BUSINESS GROUP PRIVATE L	01-01-2016	10.00	6.00	22,308.51	30,310.13	87089900	Sweden	398,453.09
7	0000699290AA02	ZEMEDHUN HAILEMESKEL GEBREHANA	01-01-2016	5.00	12.00	1,800.00	14,249.00	62046300	Thailand	300,172.90
8	0005987750AA02	NATNAEL TESSEMA MOUNI	01-01-2016	1.00	47.00	42,680.00	42,680.00	62044900	Thailand	944,322.92
9	0026319333AA02	YILMA MEKONNEN TADESSE	01-01-2016	2.00	30.00	8,244.60	15,324.60	61083200	India	320,886.32
10	0005105655TG02	ASMELASH ASGEDOM WELDESAMUALE	01-01-2016	3.00	179.00	22,543.75	47,293.75	39269090	China	624,700.32
11	0012414260AA02	TIGIST TSEGAYE GUALEMARIAM	01-01-2016	10.00	10.00	3,355.02	10,328.44	40169300	India	151,397.96
12	0016295696ET07	THE SUGAR CORPORATION	01-01-2016	1.00	144.00	40,426.23	40,426.23	39140000	China	318,449.78
13	0000253982AA03	YEWBDAR GIZAW BARUDA	01-01-2016	17.00	15.00	7,896.50	17,052.21	87089900	Germany	201,213.21
14	0000022063ET02	ETHIO CERAMICS PRIVATE LIMITED COMP	01-01-2016	2.00	440.00	26,814.40	26,936.00	94060000	China	447,347.83
15	0003292319ET03	GMT INDUSTRIAL PRIVATE LIMITED COMP	01-01-2016	1.00	30.00	280,210.23	280,210.23	11071000	Netherlands	3,321,121.21
16	0002456668ET02	OMAN GLOBAL PRIVATE LIMITED COMPANY	01-01-2016	1.00	8,928.00	205,344.00	205,344.00	34011110	Indonesia	1,702,476.03
17	0000051195AA02	ADALBERTO FREZZA ALBERTO	01-01-2016	2.00	201.00	27,320.50	40,030.50	34021120	Kenya	502,432.60
18	0005943148ET03	NANODAS TRADE & INDUSTRY P.L.C	01-01-2016	8.00	1,058.00	90.00	200,000.01	48191000	China	1,213,661.91
19	0016853288AA02	SEMIRA YESUF AHMED	01-01-2016	3.00	2,293.00	33,648.00	34,809.17	96180000	China	657,594.06
20	0001560100ET07	TM FOOD COMPLEX PLC	01-01-2016	1.00	1,733.00	35,518.43	35,518.43	39232990	United Arab Emirates	209,686.61
21	0005677540AA02	ELIZABETH NEGIA WOLDE	01-01-2016	7.00	476.00	36,905.00	41,791.75	42022900	China	734,774.04
22	0020553023DR02	AREAYASELASIE TSEGA G/MICHAEL	01-01-2016	1.00	706.00	211,600.00	211,600.00	64029900	China	3,388,445.14
23	0040373612ET02	S M E D TREADING PLC	01-01-2016	2.00	1,860.00	8,840.00	12,984.00	07095100	China	217,449.73
24	0000005426AA02	AKALU AMSALU WOLDEMICHAEL	01-01-2016	1.00	483.00	23,912.64	23,912.64	54076900	China	531,479.22
25	0001875897AA02	AEMERO ALEM BERHANE	01-01-2016	4.00	954.00	6,800.00	23,818.00	96031010	China	375,457.10
26	0000148813DR02	MENSUR IBRAHIM YUSUF	01-01-2016	1.00	2,880.00	105,480.00	105,480.00	39012000	Thailand	601,398.36
27	0000148826AA02	JEWAHER IBRAHIM YOUSUF	01-01-2016	1.00	1,200.00	21,600.00	21,600.00	84238100	India	302,399.85
28	0016672108AA03	LULIT DABA KASSA	01-01-2016	5.00	5.00	8,471.96	38,751.89	87032210	Japan	1,129,873.76
29	0004363629AA07	EYOB NIGUSE WOLDESELASSIE	01-01-2016	1.00	480.00	130,032.00	130,032.00	39072000	United Arab Emirates	617,653.20

Appendix 2: WEKA run output information

kMeans

======

Number of iterations: 10

Within cluster sum of squared errors: 2249.78646217118

Missing values globally replaced with mean/mode

Cluster centroids:	Cluster#				
Attribute	Full Data	0	1	2	3
	(65535)	(9839)	(14611)	(24236)	(16849)
LOCATION	FEDERAL	FEDERAL	ADDIS ABABA	FEDERAL	FEDERAL
BUSINESS ACTIVITY	COM_ GOODS	INDS_GOODS	COM_GOODS	COM_GOODS	COM_GOODS
ITEM_TOTAL	17.8761	10.2414	9.2785	17.7179	15.6332
ITEM_PRICE	2.0	15.0	5,234	1,400	2.0
TOTAL_INVOICED	386,285	117,138	37,211	86,285	107,664
ITEM_FREQUENCY	25.5673	5.7053	6.2039	52.4244	15.3253
COUNTRY_OF_ORIGIN	ASI	ASI	ASI	ASI	EUR
REVENUE_ITEM	340921.3899	511467.2584	260025.0547	355507.9275	290500.3694
REVENUE_PRICE	12824.934	4771.5766	2253.1113	21919.3092	13613.7408
INVOICED_REVENUE	1.2615	0.6695	0.1408	2.9226	0.1894
TOTAL_REVENUE	5,544,642	336,856	593,399	5,544,642	5,725,715

Output of WEKA run with K=4 and seed=10

kMeans
======
Number of iterations: 15
Within cluster sum of squared errors: 2231.57106806998
Missing values globally replaced with mean/mode

Cluster Centroids:		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(65535)	(5262)	(19420)	(20914)	(14197)	(5742)
LOCATION	FEDERAL	FEDERAL	FEDERAL	FEDERAL	ADDIS ABABA	FEDERAL
BUSINESS ACTIVITY	CG	CG	CG	CG	CG	CG
ITEM_TOTAL	17.8761	19.0378	16.9052	17.9845	8.9303	41.8184
ITEM_PRICE	2.0	2.0	30.0	71,400	5,234	107,100
TOTAL_INVOICED	386,285	110,904	386,285	71,400	301,324	107,100
ITEM_FREQUENCY	25.5673	70.0106	13.2025	15.9298	5.5851	111.1663
COUNTRY_OF_ORIGIN	ASI	AFR	EUR	ASI	ASI	ASI
REVENUE_ITEM	340921.39	602843.59	253592.95	457333.45	267824.18	152973.47
REVENUE_PRICE	12824.934	18297.6441	7686.4199	26634.2047	2366.2256	750.3497
INVOICED_REVENUE	1.2615	0.1086	0.1867	1.5964	0.1468	7.4886
TOTAL_REVENUE	5,544,642	1,424,478	5,544,642	336,856	1,064,764	4,182,038

Output of WEKA run with K=5 and seed=100

kMeans
======
Number of iterations: 17
Within cluster sum of squared errors: 2271.6725495066
Missing values globally replaced with mean/mode

Cluster Centroids:		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(65535)	(3419)	(19198)	(5265)	(18197)	(6600)	(12856)
LOCATION	FEDERAL	FEDERAL	FEDERAL	FEDERAL	FEDERAL	FEDERAL	ADDIS ABABA
BUSINESS ACTIVITY	CG	CG	CG	CG	CG	CG	CG
ITEM_TOTAL	17.8761	33.568	14.2327	48.9922	6.4615	41.723	10.3146
ITEM_PRICE	2.0	71,400	2.0	20.0	131,300	61,380	5,234
TOTAL_INVOICED	386,285	71,400	107,664	57,635	131,300	86,285	37,211
ITEM_FREQUENCY	25.5673	150.4144	15.2609	19.9026	9.2733	79.9379	5.2257
COUNTRY_OF_ORIGIN	ASI	ASI	EUR	ASI	ASI	ASI	ASI
REVENUE_ITEM	340921.39	136507.33	278834.88	41286.38	653612.38	209987.42	235331.24
REVENUE_PRICE	12824.93	4425.10	9031.17	12177.96	27265.63	8910.18	2558.76
INVOICED_REVENUE	1.2615	4.9263	0.186	4.1017	0.495	4.5065	0.1484
TOTAL_REVENUE	5,544,642	5,588,246	5,725,715	622,483	215,896	5,544,642	593,399

Output of WEKA run with K=6 and seed=1000

Appendix 3: Interview questions check list

No	**Issues/points of analysis**	**Yes**	**No**
	Issues related to domain experts		
1	Is there any document that states about the authority?		
2	Are there customer related documentations in the authority?		
3	Does the authority have clearly stated business rule?		
4	Is there any prescribed procedure regarding to customers relationship management?		
5	Is there any business process that shows the customers interaction with the authority?		
	Issues concerning researches in the authority		
1	Are there researches concerning data mining in the authority's library?		
2	Are there researches related to application of data mining in the customs administration domain in the library?		
3	Are there researches related to application of data mining in the tax administration domain in the library?		
	Issues related to ERCA's database		
1	Does your organization have a database system?		
2	Does the database capture customers information those engaged on import/export business activities?		