ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING IN CORRUPT ACTIVITY DATA TO SUPPORT COMBATING CORRUPTION: THE CASE OF FEDERAL ETHICS AND ANTICORRUPTION COMMISSION OF ETHIOPIA.

BY

ELSABET WEDAJO

June, 2012

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING IN CORRUPT ACTIVITY DATA TO SUPPORT COMBATING CORRUPTION: THE CASE OF FEDERAL ETHICS AND ANTICORRUPTION COMMISSION OF ETHIOPIA.

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

BY

ELSABET WEDAJO

June, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

Application of Data mining in Corrupt Activity Data to Support Combating Corruption: The Case of Federal Ethics and Anticorruption Commission of Ethiopia.

BY

ELSABET WEDAJO

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
_____		_____	_____
_____		_____	_____
_____		_____	_____

Acknowledgement

First I would like to acknowledge and extend my heartfelt deep gratitude to my Advisor Gashaw Kebede (PHD) for his vital encouragement, support and constructive comments in the whole research progress. This Thesis wouldn't have been a success for me without his valuable comments and suggestions.

I would like to express my heartfelt gratitude to my friends I have in Information Science Department for their supportive ideas.

Most especially to my Family and to God, who made all things possible.

Last but not least, I would like to thank the FEACCE staffs for helping in the data collection, other supports where they are needed in the study and much needed motivation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

ARFF = Attribute Relation File Format

CRISP-DM = cross industry standard process for data mining

CSV = Comma Delimited File

FEACCE= the Federal Ethics and Anti Corruption Commission of Ethiopia

ICT= Information Communication Technology

IT = Information Technology

KDD = Knowledge Discovery process in Data Bases

OLAP = Online Analytical Process

WEKA = Waikato Environment for Knowledge Analysis

Abstract

The aim of this research is to extract hidden, novel and potentially useful knowledge concerning corruption from the data taken from existing database of the FEACCE with the help of data mining techniques and tools selected. Persons who are working in any activity can commit corruption knowingly or unknowingly but we don't know what persons with what personal characteristics are vulnerable to corruption. Currently the FEACCE staffs who are working in investigation of corruption attempt to understand few relationship between corruption category and the characteristics of offenders. This can be achieved by utilizing data mining techniques efficiently, effectively and accurately than those staffs who are working in statistics department using traditional and simple statistical methods to analyze corruption data.

To attain the objective of the study the researcher has used CRISP-DM methodology that consist six steps and the steps can be used iteratively until the required results are achieved. From the steps, data preprocessing needs to be given the higher priority because the data bases of FEACCE was inconsistent and incomplete and this needs to be cleaned before it is given as an input for the data mining tool. This step needs more time and effort than the remaining steps. Thus results from data mining techniques that were relied on the structured approach were useful for the FEACCE to attain its objectives.

This study applied three different data mining techniques and came up with different models along with evaluations. The models can be used by the FEACCE and other bodies that are combating corruption in the country to predict the future events and classify corruption offenders through both predictive and descriptive modeling techniques.

Some of rules generated from association rule mining were not that much interesting because some of the attribute values in the data base were insufficient and some are many. Approximately rules generated from association rule mining and rules developed in classification model were the same as they both are predictive models. The classification model that uses the output of clustering model as an input has best performance than the direct classification of the data set. Application of data mining in corrupt activity data is introduced in this research.

CHAPTER ONE

1.1 Introduction

The definitions of corrupt practices are the offering, giving, receiving, or soliciting, directly or indirectly anything of value to influence improperly the actions of party (ADBGIACD, 2010). Corruption can occur whenever a person supplies or has access to information or resources or has responsibilities for decision making. Any work place activities should be designed up on awareness of the corruption risks in the means of management can reduce the risk to an acceptable level (CPM, 2006).

The definition of corruption crime exceeds the usual definition, considering as tainted not only assets controlled by the accused directly, but also those which he or she controls indirectly through other individuals. This extension is a most useful addition to the cache of anti-corruption investigators and prosecutors, whose work is facilitated somewhat by the possibility of pursuing illicit enrichment charges also against public officials who seek to evade justice by transferring their ill-gotten property to loyal proxies (Tewodros, 2011).

Corruption distorts resource allocation and government performance. The consequences of its development are many and differ from one country to the other. Among the contributing factors are policies, programs and activities that are poorly envisioned and managed, failing institutions, poverty, income inequalities, inadequate civil servants' compensation, and lack of accountability and transparency (Langseth, 1999).

International aid institutions' help its client countries attack is not because it is immoral, wrong, or even illegal. Rather, decision is based on the negative effect corruption has on economic development, the emergence of an enabling environment for the private sector, and in deepening poverty in the developing world; a situation that demands a response from the international community. There is considerable evidence that a strong negative

relationship exists between the extent of corruption and economic performance of the given country (Mauro, 1997).

Pervasive corruption reduces the efficiency of government in general and the effectiveness of private investment and foreign aid in particular. The negative impacts of corruption have served as the force for international aid institutions to insist the establishment of good governance measures in recipient developing countries. These measures attempt to improve integrity, transparency, and accountability in government and private administrative transactions, to achieve sustainable growth and improved service delivery to the public (Jean, 1996).

Corruption prevention is an area where data mining technology is not widely used. However the potential to use data mining in corruption prevention is massive. When the corrupt behavior involves monitory transactions these are always recorded within the organization or external data base systems. An analysis of the situation can define the non-corrupt behavior and data mining software can be modeled to look for or mine transactions that are different from the non-corrupt behavior (Langseth, 1999).

A good description of data mining is to discover useful, previously unknown knowledge by analyzing large and complex data sets. Data mining is one step in a broader Knowledge Discovery in Data base (KDD) process. Data mining itself is relatively a process of using algorithms to discover predictive and potentially useful patterns in data set (Derosa, 2004).

Data mining is becoming increasingly common in both the private and public sectors. Industries such as Banks, Insurance, Medicine, and Retailing commonly use data mining to reduce costs, enhance research and increase sales. In the public sector data mining application initially were used as a means to detect fraud and waste, but have grown to also be used for the purposes of measuring and improving program performance (Seifert, 2004).

Data mining is one of the most important research fields that are due to the expansion of both Computer hard ware and Software technologies which have imposed organizations to depend heavily on those technologies (Agrawal et.al, 1993). Data mining is a powerful tool that enables crime investigators who may lack extensive training as data analysts to explore large databases quickly and efficiently (Fayyad and Uthurusamy, 2002).

Data is considered as the number one asset of any organization; it is obvious that this asset should be used to predict future decisions of the business (Fukuda et. al., 1996). Data mining is a means to help organizations to make full use of the data stored in their data base. And when it comes to decision making, this is true in all areas, and is also true in all different types of Organizations (Agrawal and Srikant, 1994).

According to Seidman (2000), data mining is the process of realizing meaningful patterns and relationships that recline hidden within very large database. Hand, et.al (2001) also define data mining as the analysis of observational data sets to find unexpected relationships and to summing-up data in novel ways that are both in forms understandable and useful to the data owner.

It has been reported that data mining has helped the Federal Government of U.S.A (United States of America) to recover millions of dollars in fraudulent Medicare payments. The justice department has been able to use data mining to assess crime patterns and justice resource allotments accordingly. Similarly the department of Veterans Affairs has used data mining to help predict demographic changes in the population it serves so that it can better estimate its budgetary needs. Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend defensive measures (Seifert, 2004).

The Federal Ethics and Anti Corruption Commission of Ethiopia (FEACCE) has a database that contains the records of corruption offences like abuse of power, acceptance of undue advantages, maladministration, unlawful disposal of objects in charge, misappropriation in discharge of duties, traffic in official influence, illegal collection (e.g.

taxes) or expenditures, undue (too much) delay of matters, taking things of value without or with inadequate consideration, granting and approving license improperly and possession of unexplained property. All this fall under the category of corruption crimes committed by public servants in violation of trust and good faith and these needs to be mined to create new knowledge.

This study was intended to show the means of the FEACCE is able to extract the hidden knowledge of its data base by using data mining application. Those of useful knowledge will be used for decision making in the circumstances of the FEACCE meet its objectives related to combating corruption with in Ethiopia.

1.2 Statement of the Problem

The UN (United Nations) points out that corruption can take many forms that vary in degree, from the minor use of influence to institutionalized bribery, and that "this can mean not only financial gain but also non-financial advantages" (UN, 2010).

Corruption is also defined by the World Bank and Transparency International (TI) as the misuse of public office for private gain. As such, it involves improper and unlawful behavior of public service officials, both politicians and civil servants, whose positions create opportunities for the diversion of money and assets from societies or government to them and their collaborators (Langseth, 1999).

Corruption is a function of both the opportunity to request and receive bribes and the risk of detection. Corruption exists in all sectors of society. It damages a country's development by undermining belief in public institutions, increase costs for firms and discourage both foreign and domestic investments and it is a growing challenge for the business sector both in the developing and industrialized countries. At the level of the individual firm, it raises transaction costs and introduces risks, as well as opens up for extortion. Regardless of sector and level of transactions, corruption hampers development.

Corruption is no longer a local matter, but a transnational phenomenon that affects all societies and economies, it is very essential to make cooperation to prevent and control. To promote, facilitate and enhance international cooperation and technical assistance in the prevention and fight against corruption, the community needs to adopt and put in place strong anticorruption prevention mechanisms.

The harmful effects of corruption on countries' economic development are widely acknowledged in the economics literature. Several authors show that corruption detracts investors, reduces the productivity of public expenditures, distorts the allocation of resources and thus lowers economic growth. These findings are reflected in the strategies of multinational organizations like the World Bank and International Monetary Fund (Axel and Thomas, 2005)

Corruption falls disproportionately on the poorer members of society and hinders them from accessing scarce services. Ethiopia is one of the Countries that face a great problem of corruption offences and to deal with the problem punishment is used highly as curative method. Every member of the society is required to give Information to the FEACCE whenever they know someone or groups of people committing corrupt activities by using free call line.

Rather than using highly curative method it is advantageous to make predictions about the future based up on the hidden knowledge constructed from historical database or repositories of the records of corruption crimes and this technique can help to take appropriate actions.

Crime investigation institutes should predict the future based up on the extracted knowledge from their data base. Hidden knowledge tells about the previous and current situations and these are useful to predict the future situations. Combating corrupt activities applies investigation techniques which are useful for prevention mechanisms rather than dealing after corruption has occurred. FEACCE do not utilize its corrupt

activity data in a way that enables it to extract new knowledge that is important to forecasting the problem of corrupt activities and take preventive actions.

To this end this study will attempt to discover critical and hidden knowledge by finding out associations between different attributes of the data base, creating groups or clusters for the data set and classifying the data set based on the class attributes of the data base by using data mining techniques. The output of this study benefits the FEACCE to have a good insight about the relationships of the attributes of corruption data base. This will help to make future prediction based up on the patterns that has generated by the data mining technique. For other researchers it will indicate research directions to be conducted on corruption fighting by using applications of Information Technology (IT).

1.3 Objective of the Study

This study has both General and Specific objectives:

1.3.1 General Objective

The General objective of this research is to develop models that identifies novel and hidden knowledge from the historical database of FEACCE by using data mining approach .The identified hidden knowledge can be used to predict the associated conditions with corruption activities, specifically taking undue advantages from the public committed by government officials.

1.3.2 Specific Objectives

To achieve the general objective of this research the following specific objectives are mandatory and they are:

- Conduct a thorough literature review to have a good insight about the problem domain and data mining techniques that are appropriate to select patterns and to describe the association of attributes that are helpful to predict about the future conditions of corrupt activities.
- Collect data to be analyzed that are in the domain of the research problem from the database of FEACCE.

- Prepare data for modeling by preprocessing in the way that is well suited to the data mining software which includes adjusting inconsistent data encoding, accounting for missing values, combining attributes which mean the same thing.
- Select best data mining algorithms that can help to come up with good models that would extract new patterns and useful to predict the future, naturally group the profile of corrupters based on their similarity and classify corruption offenders based on the class label of the data base.
- Build and train the data mining models on the prepared training datasets and select the best models in line with the idea of domain experts.
- Evaluate the model for the degree of suitability on predicting corrupt activities with related conditions and identify novel knowledge.
- Present the findings and recommend appropriate measures that could be taken by FEACCE to combat corruption.

1.4 Scope and Limitation of the Study

The scope of the research is limited to application of data mining technique for FEACCE to detect conditions associated to corrupt activities, create natural groups of corrupters profile and classifying the new data set of corrupters based on the class label of the data set. It will be limited to identifying patterns that are pertinent to corrupt activities specifically acceptance of undue (unnecessary) advantages from the public committed by the government officials in Ethiopia. Among the data mining techniques classification, clustering and association rules mining has been applied due to time and budget constraints.

1.5 Significance of the Study

This study is significant to FEACCE because the study introduces the application of the data mining technique that enables FEACCE to extract new, potentially useful and novel knowledge from the data repository and predict other associated situations that will be used to give acceptable decisions in the fighting corruption. Indirectly for the public it

gives prior knowledge about corrupt activities that are the stagnant factor for the overall development of the country. Ethiopia is now growing in the fastest rate and fighting against corruption is unquestionable to sustain the growth because the country's property should be allocated effectively when and where it is needed. For the researcher, the study provides the opportunity to engage in academic exercise.

1.6 Thesis Organization

This thesis consists six chapters. Chapter one introduces both corruption and Data mining, statement of the problem, objective of the study, scope and limitation of the study and significance of the study. Chapter two presents thorough literature review on both data mining and corruption. Chapter three presents the methodology of the study.

Chapter four is devoted to data understanding and preprocessing. Chapter five is model building and experimentation and chapter six presents conclusions and recommendations.

CHAPTER TWO

Literature Review

2.1 Overview of Data Mining

The amount of data continues to grow at an enormous rate even though data stores are already immense. The primary challenge is how to make the database a competitive business advantage by converting apparently meaningless data into useful information. How this challenge is met is critical because companies are increasingly relying on effective analysis of information simply to remain competitive in the market. A mixture of new techniques and technology is emerging to help sort through data and find competitive data.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both infer either sifting through a large amount of material or resourcefully probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for Gold in rocks is usually called “Gold Mining” and not “Rock Mining”, thus by analogy, Data mining should have been called “Knowledge Mining” (Zaïane, 1999).

Data mining refers to extracting or “mining” knowledge from large amount of data (Han et. al., 2001). Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large and quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 2000).

Data mining is a term that describes different techniques used in a domain of machine learning, statistical analysis, modeling techniques and data base technologies that can be used in different industries. With a combination of these techniques, it is possible to find different kinds of structures and relations in the data, as well as to derive rules and

models that enable prediction and decision making in new situations (Gamberger et.al. 2001).

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks and decision trees). Consequently, data mining consists of more than collecting and managing data; it also includes analysis and prediction (Adriaans and Zantinge 1996).

Data mining can make any institution to be able to extract hidden knowledge from their data repositories by applying different process of examining data from different perspectives and summarizing it into useful information. It is a process that allows users to understand constituent of relationships between data. It reveals patterns and trends that are hidden among data. It is often viewed as a process of extracting valid, previously unknown, non-trivial and useful information from large data bases. Data mining systems can be classified according to the kinds of database mined, the kinds of knowledge mined, the techniques used or the applications (Ravichandra, 2003).

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. Applications include association; classification and clustering of the given transactional data base (Seifert, 2004).

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased

ability to combine data from disparate sources into a single searchable source (Adriaans and Zantinge 1996).

2.2 Tasks of Data Mining

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute a fascinating outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers (Westphal and Blaxton, 1998).

As Berry and Linoff (1997) stated in practice, there primary goals of data mining tend to be Prediction and Description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data mining activities into one of the two categories:

1) **Predictive Data Mining**, which produces the model of the system described by the given data set, or

2) **Descriptive Data Mining**, which produces new, non-trivial information based on the available data set.

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for

particular data mining applications can vary considerably. The goals of prediction and description are achieved by using data mining techniques, explained later, for the following primary data mining tasks.

Data mining tasks refer to the essential procedures where intelligent methods are applied to extract useful patterns. There are different data mining tasks such as clustering, classification, associations and outlier detection. Each task can be thought as a particular kind of problems to be solved by a data mining algorithms. Meanwhile, some algorithms can be applied to different tasks (His et.al, 2003). The tasks discussed here are association rules, clustering and classification as follows:

2.2.1 Association Rule

The task of association rule mining is to find certain association relationships among a set of objects called items in a database. The association relationships are described in association rules. Each rule has two measurements, support and confidence. Confidence is a measure of the rule's strength, while support corresponds to statistical significance.

The task of discovering association rules was first introduced in 1993 (Agrawal et. al., 1993). Originally, association rule mining focused on market basket data which stores items purchased on a per-transaction basis. A typical example of an association rule on market data is that 70% of customers who purchase bread also purchase butter. Later, association rule mining was also extended to handle quantitative data.

In the literature of data mining, there are a lot of studies on designing scalable algorithms for mining association rules. Such studies are particularly useful with a large amount of data in order to understand the customer behavior in their stores. However, it is generally true that the rules of association rule mining are not directly useful for the business sector (Wong and Fu, 2004).

Association rules are used to generate recommendation actions. In their model, given a set of past transactions, pre-selected target items and it intends to build a model for

recommending target items and promotion strategies to new customers, with the goal of maximizing the net benefit (Wang et. al., 2002).

According to Ravichandra (2003) Association analysis is the discovery of association rules sharing attribute value conditions that occur frequently together in a given data set. It is widely used in the context of analysis of "transaction data." Association rules are of the form: X => Y, That is to say, A1 ^ A2 ^ … ^ Am => B1 ^ B2 ^ B3 … Bn, where Ai (for i º {1, 2, …m}) and BJ (for j º { 1, 2, 3, …n}) are attribute value pairs. This rule is interpreted, as "database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y." For example,

Age (x, "25 ….. 35") ^ income (x, 15k …. 25k) => buys (x, "Cell phones'): support 2%, confidence 60%.

The rule indicates that of the customers under study, 2% are 25 to 35 years of age with an income of 15k to 25k and have purchased a cell phone in a shop; there is a 60% confidence or certainty that a customer in the said age and income group will purchase a cell phone.

A rule is an implication a → C. The left part of the rule is called the antecedent and the right, the consequent. The sets A and C are disjointed as we cannot find the same item in both the antecedent and consequent. A rule makes sense thanks to its support s = sup (A → C) = P (An ∩ C) and its confidence is c = cone (A → C) = P (C /A).

An objective measure for association rules of the form x => y is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies; i.e. P (X È Y), where X È Y indicates that a transaction contains both X and Y – the union of items sets X and Y (in the context of set theory it is X Ç Y). Another objective measure for association rules is confidence, which assesses the degree of certainty of the identified association, i.e., P (Y|X) – probability that a transaction containing X also contains Y. Thus, support and confidence are defined as:

$$\text{Support } (X \Rightarrow Y) = (P \cup Y) = \frac{\text{No.of tuples containing both A\&}B}{\text{Total No.of tuples}}$$

$$\text{Confidence } (X \Rightarrow Y) = P(Y|X) = \frac{\text{No.of tuples containing both A\&}B}{\text{No.of tuples containing A}}$$

Association analysis studies the frequency of items occurring together in transactional data bases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules (Zaïane, 1999).

2.2.1.1 A-priori Algorithm

The first efficient algorithm to mine association rules is A-priori Algorithm. The first step of this algorithm is the research of frequent item sets. The user gives a minimum threshold for the support and the algorithm searches all item sets that appear with a support greater than this threshold. The second step is to build rules from item sets found in the first step. The algorithm computes confidence of each rule and keeps only those where confidence is greater than a threshold defined by the user (Agrawal and Srikant, 1994).

A Priori algorithm is divided in to three sections. After Initial frequent item sets are fed in to the system, first candidates are generated, pruned, and candidate support calculation is executed in rotate. The support information is fed back into the candidate generator and the cycle continues until the final candidate set is determined.

A frequent item is an item that often occurs in a database. A frequent item set, then, is a set of items that often occurs together in the same basket within the database if it consists of more than one item. The cutoff of how often a set must occur before it is included in the candidate set is the support.

Candidate generation is the process in which one generation of candidates is built into the next generation. This building process is from where the A-priori name derives. Each

new candidate is built from candidates that have been determined A-priori (in the previous generation) to have a high level of support. Thus, they can be confidently expanded into new potential frequent item sets. This is expressed formally as follows:

$\forall a1; a2 \in Cm$ do

With a1 = (i1; …, im 1; im)

and a2 = (i1; …, im 1; i*m)

and im < i*m

a = a1 [a2 = (i1; …, im 1; im; i*m)]

It should be noted that only ordered sets are utilized, that is, the item codes increase toward the last item in a set. Thus, when a is generated from a1 and a2, the sets remain ordered. Candidate generation pairs up any candidates that differ only in their final element to generate the candidate item sets for the next candidate generation.

The next step of candidate generation guarantees that each new candidate is not only formed from two candidates from the previous generation, but that all subsets that can be created by removing one element are also present in the previous generation, as follows:

$\forall a \in Cm$ do

$\forall i \in a: a-\{i\} \in Cm$

The initial candidate generation proves by design that if we remove either of the last two items (im, i*m) from the new candidate, we will get candidates from the previous generation, namely, a1 and a2. The second step verifies that if we remove any single item from the new candidate, we will find a candidate from the previous generation. This progressive build-up of candidates is the heart of the A-priori algorithm.

In the second phase candidates are pruned if they do not satisfy the value of minimum support within each transaction.

The third phase of the algorithm is the support calculation. It is by far the most time consuming and data intensive part of the application, as during this phase the data base is streamed into the system. Each potential candidate's support, or the number of occurrences over the data base set, is determined by comparing each candidate with each transaction in the data base. If the items appear in the transaction the support count for that candidate is incremented, as follows:

$\forall t \in T$ do

$\forall c \in C$ do

If c C t

Support (c) ++

The main problem with the A-priori algorithm is this data complexity. Each candidate must be compared against every transaction set. This gives a large running time for a single generation, O (|T || C|| t|), assuming the subset function can be implemented in time |t|.

2.2.2 Clustering

Clustering is similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity), so that objects sharing the same characteristics are confined in one group and minimizing the similarity between objects of different classes unlikely to the previous objects of the data sets are dispersed in different groups (inter-class similarity) (Frawley et.al., 1991).

According to Ravichandra (2003) clustering in data mining is the process of grouping data sets based on the similar characteristics they have and it have the following steps where all data sets are used to train and the process are:

A, Preprocessing and feature selection

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

B, Similarity measure

Similarity measure plays an important role in the process of clustering where a set of objects are grouped in to several clusters, so that similar objects will be in the same cluster and dissimilar ones in different clusters. In clustering, features represent an object and the similarity relationship between objects is measured by a similarity function. This is the function, which takes sets of data items as an input and returns out put a similarity measure between them. And the most commonly used is the Euclidean Metric (EM) which defines the distance between two data point's p = (p1, p2…) and q = (q1, q2…) d = $[\sum(Pi - Qi)^2]^{1/2}$.

C, Clustering algorithm

Clustering algorithms are general schemes, which use particular similarity measures as subroutines. The particular choise of clustering algorisms depends on the desired properties of the final clustering, e.g. what are the relative importance of compactness, parsimony, and inclusiveness. Other considerations include the usual time and space complexity. Clustering algorithm attemts to find natural groups of components (data) based on some similarity, it also finds the centroid of a group of data sets to determine cluster membership. Most algorithms evaluate the distance between a point and the

cluster centroids. And the output is a statistical description of the cluster centroids with the number of components in each cluster.

D, Result validation

Does the result make sense? If not, we may want to iterate back to some prior stages. It may also be useful to do a set of clustering tendency, to try to guess if clusters are present at all; note that any clustering algorithm will produce some clusters regardless of whether or not natural clusters exist.

E, Result interpretation and application

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

2.2.2.1 K-means algorithm

The k-means algorithm is a simple iterative method to partition a given dataset in to a user specified number of clusters, k. This algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of D-dimensional vectors, $D = \{xi \mid i = 1..., N\}$, where $xi \in R^d$ denotes the i^{th} data point. The algorithm is initialized by picking k points in R^d as the initial k cluster representatives or "centroids" (Xindong et.al, 2007). Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or alarming the global mean of the data k times. Then the algorithm iterates between two steps till convergence:

Step 1: Data Assignment. Each data point is assigned to its closest centroid, with ties broken at random. This results in a partitioning of the data.

Step 2: Relocation of Means. Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

2.2.3 Classification

Classification is a supervised learning technique that classify or identify class of the new data sets based on the given and known class attribute. Some of the data sets are used to be training and some are used as the test data sets.

Classification is a most important and frequently used technique in data mining. It is a process of finding a set of models that describe and distinguish data classes or concepts. The derived model may be represented in various forms such as classification (IF-THEN) rules, decision tree, neural networking, etc. A decision tree is a flowchart like tree structure when each node denotes a test on an attribute value where each branch represents an outcome of the test, and tree leaves represent classes. Decision trees can be easily converted to classification rules. A neural network when used for classification is typically a collection of neuron-like processing units with weighted connections between the units. While learning classification rules the system has to find the rules that predict the class from the prediction attributes. So firstly the user has to define conditions for each class, the data mine system then constructs descriptions for the classes (He et. al., 2003).

Classification problems aim to identify the characteristics that indicate the group to which each instance belongs. Classification can be used both to understand the existing data and to predict how new instances will behave. Classification is a well recognized data mining operation that has been studied extensively in the fields of Statistics, Pattern recognition, Decision theory, Machine learning and Neural network.

2.2.3.1 Decision tree

Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field (Ian and Eibe, 2005).

Decision trees are a simple, but powerful form of multiple variable analyses and they provide unique capabilities to supplement the following analytical techniques for data.

- Traditional statistical forms of analysis such as multiple linear regressions.
- Variety of data mining tools and techniques such as neural networks.
- Recently developed multidimensional forms of reporting and analysis found in the field of business intelligence.

Decision tree can reflect both a continuous and categorical object of analysis. The display of this node reflects all the data set records, fields, and field values that are found in the object of analysis. The discovery of the decision rule to form the branches or segments below the root node is based on a method that extracts the relationship between the object of analysis (that serves as the target field in the data) and one or more fields that serve as input fields to create the branches or segments. The values in the input field are used to estimate the likely value in the target field. The target field is also called an outcome, response, or dependent field or variable.

2.3 Data Mining Procedures

Data mining process is a step in Knowledge Discovery Process in Database (KDD) consisting of methods that produce useful patterns or models from the data (Hand and Kamber, 2001).

Data mining is a rather complicated process that has to be planned very carefully in order to be successful. It has to be organized within one of the proposed rigorous procedures. One of the stages describes the next stages of data mining to be performed (Rud, 2001).

Berger (2004) believes that to be effective in data mining, successful data analysts should generally be familiar with the following stages:

2.3.1 Identification of a Research Problem

Most data-based modeling studies are performed for a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies are likely to be focused on the data mining technique at the cost of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypothesis formulated for a single problem at this stage (Kantardzic, 2003).

This first step requires the combined expertise of an application domain and a data mining model. In successful data mining applications, this co-operation does not stop in the initial phase; it continues during the entire data mining process, the requirement to knowledge discovery is to understand data and business (TCC, 2005). Without this understanding, no algorithm, regardless of complexity, is able to provide result that can be confident.

2.3.2 Data Selection

This process is concerned with the collection of data from different sources and locations. The current methods used to collect data are:

- ***Internal data***

Data are usually collected from existing databases, data warehouses, and OLAP (Online Analytical Process). Actual transactions recorded by individuals are the richest source of information, and at the same time, the most challenging to be useful.

- ***External data***

Data items can be collected from Demographics, psychographics and web graphics or from external databases created by outsourcing if possible. In addition to data shared within a company.

2.3.3 Data Preparation

All raw data sets which are initially prepared for data mining are often large; many are related to humans and have the potential for being messy. Real world databases are subject to noise, missing, and inconsistent data due to their typically huge size, often several gigabytes or more (Kantardzic, 2003).

Data preprocessing is commonly used as a preliminary data mining practice. It transforms the data into a format that will be easily and effectively processed by the users. There are a numbers of data preprocessing techniques which include: Data cleaning; that can be applied to remove noise and correct inconsistencies, outliers and missing values. Data integration; merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations, such as normalization, may be applied; normalization improves the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction; can reduce the data size by aggregating, eliminating redundant features. The data processing techniques, when applied prior to mining, can significantly improve the overall data mining results (Hand and Kamber, 2001).

Since multiple data sets may be used in various transactional formats, extensive data preparation will be required. There are various commercial software products that are specifically designed for data preparation, which can facilitate the task of organizing the data prior to importing it into a data mining tool (Daniel, 2005).

2.3.4 Choice and Transformation of Variables

This is the final data preparation step before building models. There are four main parts to this step: (TCC, 2005).

- **Select Variables**

Ideally, you would take all the variables you have, feed them to the data mining tool and let it find those which are the best predictors. In practice, this doesn't work very well. One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models. A very common error, for example, is to use as a predictor variable data that can only be known if you know the value of the response variable. People have actually used date of birth to "predict" age without realizing it.

- **Select rows**

As in the case of selecting variables, you would like to use all the rows you have to build models. If you have a lot of data, however, this may take too long or require buying a bigger computer storage capacity than you would like. Consequently it is often a good idea to sample the data when the database is large. This yields no loss of information for most business problems, although sample selection must be done carefully to ensure the sample is truly random. Given a choice of either investigating a few models built on all the data or investigating more models built on a sample, the latter approach will usually help you to develop a more accurate and robust model.

- **Construct New Variables**

It is often necessary to construct new predictors derived from the raw data. For example, forecasting credit risk using a debt-to-income ratio rather than just debt and income as predictor variables may yield more accurate results that are also easier to understand. Certain variables that have little effect alone may need to be combined with others, using various arithmetic or algebraic operations (e.g., addition, ratios). Some variables that

extend over a wide range may be modified to construct a better predictor, such as using the log of income instead of income.

- **Transform Variables**

The tool you choose may dictate how you represent your data, for instance, the categorical explosion required by neural nets. Variables may also be scaled to fall within a limited range, such as 0 to 1. Many decision trees used for classification require continuous data such as income to be grouped in ranges (bins) such as High, Medium, and Low. The encoding you select can influence the result of your model. For example, the cut-off points for the bins may change the outcome of a model.

2.3.5 Modeling

According to TCC (2005) the most important thing to remember about model building is that it is an iterative process. You will need to explore alternative models to find the one that is most useful in solving your business problem. What you learn in searching for a good model may lead you to go back and make some changes to the data you are using or even modify your problem statement. This step is where the data mining algorithms sift through the data to find patterns and to build predictive models. Generally, a data analyst will build several models and change mining parameters in an attempt to build the best or most useful models (Mdzingwa, 2005).

Once you have decided on the type of prediction you want to make (e.g., classification or regression), you must choose a model type for making the prediction. This could be a decision tree, a neural net, a proprietary method, or that old standby, logistic regression. Your choice of model type will influence what data preparation you must do and how you go about it. For example, a neural net tool may require you to explode your categorical variables. Or the tool may require that the data be in a particular file format, thus requiring you to extract the data into that format. Once the data is ready, you can proceed with training your model (Berry et al, 2000).

As Nasereddin (2009) state in estimating and building the model process in data mining; there are four main parts: select data mining task (s), select data mining method (s), select the suitable algorithm that will not change if a new version of data is introduced (since they are concerned with the definition of the problem) and extract knowledge and they are described as follows:

A, Select Data Mining Task (s)

Selecting which task to use depends on the model whether it is Predictive or Descriptive (Hand and Kamber, 2001). Predictive models predict the values of data using known results and/or information found in large data sets, historical data, or using some variables or fields in the data set to predict unknown. Classification, regressions, time series analysis, prediction, or estimation are tasks for predictive model (Fayyad, 1996). A descriptive model identifies patterns or relationships in data and serves as a way to explore the properties of the data examined. Clustering, summarization and sequence discovery are usually viewed as descriptive (Hand et al., 2001).

The relative importance of prediction and description for particular data mining applications can vary considerably. That means selecting which task to use depends on the model whether it is predictive or descriptive.

B, Select Data Mining Method (s)

After selecting which task we can choose the method and assuming we have a predictive model and the task is classification while the method is Rule Induction, with Decision tree or Neural Network. In most research in this area; researchers estimates the relevant model this model to produce acceptable results. There are number of methods for model estimation includes these but not limited to neural networks, Decision trees, Association Rules, Genetic algorithms, Cluster Detection and Fuzzy Logic.

C, Select Suitable Algorithm

The next step is to construct a specific algorithm that implements the general methods. All data mining algorithm include three primary components these are: (1) Model representation, (2) model evaluation, and (3) search.

D, Extracting Knowledge

This is the last step in building the model which are results (or the answers for the problem solved in data mining) after making the simulation for the algorithm. This can be best explained by presenting an example of association rule mined from the transactional data base of supermarket. For example if (income=high and saving=true and employment=self then buying computer=yes). The rule indicates that if person have high income and have money saved in an account and self-employed then the person have habit of buying computer.

2.3.6 Validation of the Model

Validation of the data mining model is required in order to confirm the usability of the developed model. Validation can be conducted using a validation data set and assesses the quality of the model fit to the data as well as protecting the model from over or under fitting (Mdzingwa, 2005).

Modern data mining methods are expected to yield highly accurate results using high dimensional models, with new updated version of the data, it will surely change the results; say for example the number of rules in association rules could be changed; it may add a new rules, replace existing number of rules or change the percentage of the rules, and this change could affect the decision making process (Hand.et.al., 2001).

2.3.7 Model Deployment

Once a data mining model is built and validated, it can be used in one of two main ways. The first way is for an analyst to recommend actions based on simply viewing the model and its results. For example, the analyst may look at the clusters the model has identified,

the rules that define the model or the lift and ROI (Return on Investment) charts that depict the effect of the model. The second way is to apply the model to different data sets. The model could be used to flag records based on their classification, or assign a score such as the probability of an action (e.g., responding to a direct mail solicitation). Or the model can select some records from the database and subject these to further analyses with an OLAP (on-line Analytic processing) tool (TCC, 2005).

2.4 Data Mining and Knowledge Discovery Process in Data Base (KDD)

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision making.

While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge (Zaïane, 1999). The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract potentially useful patterns.

- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

By knowledge discovery process in data base, interesting knowledge, regularities, or high level information can be extracted from the relevant sets of data in data bases and be investigated from different angles, and large data bases thereby serve as rich and reliable sources for knowledge generation and verification.

Mining information and knowledge from large database has been recognized by many researchers as a key research topic and machine learning. Companies in many industries also take knowledge discovering as an important area with an opportunity of major revenue and the discovered knowledge can be applied to information management, query processing, decision making, process control and many other applications (Fayyad et al., 1996).

2.5 Data Mining and Online Analytical Process (OLAP)

On-Line Analytical Process (OLAP) operations allow the navigation of data at different levels of abstraction, such as drill-down, roll-up, slice, and dice. OLAP tools and methods have become very popular in recent years as they let users analyze data in a warehouse by providing multiple views of the data, supported by advanced graphical representations. In these views, different dimensions of data correspond to different business characteristics. OLAP tools make it very easy to look at dimensional data from any angle or to slice-and-dice it. Although OLAP tools, like data mining tools, provide answers that are derived from data, the similarity between them ends here (Zaïane, 1999).

OLAP tools do not learn from data, nor do they create new knowledge. They are usually special purpose visualization tools that can help end-users draw their own conclusions

and decisions, based on graphically condensed data. OLAP tools are very useful for the data-mining process; they can be a part of it but they are not a substitute. OLAP is part of the spectrum of decision support tools. Traditional query and report tools describe what is in a database. OLAP goes further; it's used to answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. For example, an analyst might want to determine the factors that lead to loan defaults. He or she might initially hypothesize that people with low incomes are bad credit risks and analyze the database with OLAP to verify (or disprove) this assumption. If that hypothesis were not borne out by the data, the analyst might then look at high debt as the determinant of risk. If the data did not support this guess either, he or she might then try debt and income together as the best predictor of bad credit risks (TCC, 2005).

2.6 Application of Data Mining in Different Domains

Data mining deals with the discovery of unexpected patterns and new rules that are "hidden" in large databases. It serves as an automated tool that uses multiple advanced computational techniques, including Artificial Intelligence (the use of computers to perform logical functions), to fully explore and characterize large data sets involving one or more data sources, identifying significant, recognizable patterns, trends, and relationships not easily detected through traditional analytical techniques alone. This information then may help with various purposes, such as the prediction of future events or behaviors (Reza and Thomas, 2001).

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial

databases, multimedia databases, time-series databases and textual databases, and even flat files (Piatetsky-Shapiro and Frawley, 1991).

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring) (Seifert, 2004).

Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine (Ibid, 2004).

Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together and Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor (Cahlink, 2000).

The Justice Department of America has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs. Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend precautionary measures (Ibid, 2004).

Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to

identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. Two initiatives that have attracted significant attention include the now discontinued Terrorism Information Awareness (TIA) project13 conducted by the Defense Advanced Research Projects Agency (DARPA), and the now canceled Computer Assisted Passenger Prescreening System II (CAPPS II) that was being developed by the Transportation Security Administration (TSA). CAPPS II is being replaced by a new program called Secure Flight (TIAP, 2003).

Telecommunications industry has undergone intensive growth and development during the last decade. Telecom operators and carriers are operating today in an extremely challenging business environment, due to the increasing customer dissatisfaction with existing services; market uncertainty; bandwidth commoditization; limited market capital; large, expensive, and inflexible IT infrastructures (Pareek , 2007).

According to Sivanandam and Sumathi (2006) and Mattison (1997) in order to survive and remain competitive, Telecommunications companies of all sizes are aggressively moving from a product-strategy-based to a customer-strategy-based business model, placing customer in the central focus of their activities. They have also realized that the enormous quantities of data they collect and possess, could be effectively used to support them in solving some of the important business problems and for getting competitive advantage, by turning it into useful information and knowledge the most important company assets in the knowledge-based society. The Telecommunications industry was one of the first to experience the benefits from the application of Business Intelligence (BI) solutions and the adoption of the Data Mining technologies (Sivanandam and Sumathi, 2006) and (Mattison, 1997).

Data mining can be used to improve quality control, for example Apte, et al. (1993) used computational techniques for quality control in manufacturing. They deployed it in a disk drive manufacturing line to reduce the number of expensive tests while meeting the

performance criteria. They applied rule induction, neural network, decision tree, and k - nearest neighbor in their experimentation (Apte et. al, 1993).

Data mining has been applied to data coming from different types of educational systems. On one hand, there are traditional face-to-face classroom environments such as special education, higher education, etc. On the other, there is computer-based education and web-based education such as well-known learning management systems, web-based adaptive hypermedia systems, intelligent tutoring systems, etc. The main difference between them is the data available in each system. Traditional classrooms only have information about student attendance, course information, curriculum goals, individualized plan data, etc. However, computer and web-based education have much more information available because these systems can record all the information about students' actions and interactions into log files, databases. Data mining builds analytic models that discover interesting patterns and tendencies from student's usage of information that can be used by the teacher to improve the student's learning and course maintenance (Tsantis and Castellani, 2001).

Learning management systems accumulate a great deal of log data about students' activities. They can record whatever student activities are involved, such as reading, writing, taking tests, performing various tasks, and even communicating with peers (Mostow et al., 2005). The application of data mining in e-learning systems is an iterative cycle in which the mined knowledge should enter the loop of the system and guide, facilitate and enhance learning as a whole, not only turning data into knowledge, but also filtering mined knowledge for decision making (Romero and Venchura, 2006).

Web usage mining is the other area in which data mining is applied. Web usage mining is automatic discovery of user access patterns from Web servers by using data mining approaches. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs which contain information about the referring pages for each page reference, and user registration or survey data gathered.

Analyzing such data can help organizations determine the life time value of customers, cross marketing strategies across products, and effectiveness of pro-motional campaigns, among other things. It can also provide information on how to restructure a Web site to create a more effective organizational presence, and shed light on more effective management of workgroup communication and organizational infrastructure. For selling advertisements on the World Wide Web (WWW), analyzing user access patterns helps in targeting ads to specific groups of users (Cooley and Srivastava, 1997).

Data mining is an essential tool for analyzing Internet and Web log data. Monitoring and characterizing "normal" activity can help to rapidly identify unusual or suspicious events in large datasets, providing actionable patterns for use in subsequent analysis and surveillance. Using data mining, the Richmond Police department identifies, characterizes, and analyzes unusual and suspicious activity in Web log data. This includes the identification and characterization of extremely rare events, anomalies, and patterns in relatively large datasets. The department also uses many data culling and descriptive features to analyze complex series of phone calls and linked conference calls (Colleen, 2006).

In general data mining is new area of Information Communication Technology where it is applied to any kind of business areas to have a competitive advantage in market areas.

2.7 Related Works

Application of Data Mining In Crime Prevention: The Case of Oromia Police Commission by Leul Woldu in 2003.

The purpose of the study was to explore the applicability of data mining technique in the efforts of crime prevention with particular emphasis to the Oromia Police Commission and to build a model that could help to extract crime patterns. With this objective decision trees and neural network were employed to classify crime records on the basis of the values of attributes crime label (Crime Label) and crime scene (Scene Label).

5,500 sample data were taken randomly from the criminal record database of the Oromia Police Commission to build and test the model.

Results of the experiments have shown that decision tree has classified crime records at an accuracy rate of 94 percent when the attribute Crime Label is used as a basis for classification. Whereas, in the same experiment, the accuracy rate of neural networks is 92.5 percent. On the other hand, in the case of classification of records on the values of the attribute Scene Label decision tree has shown an accuracy rate of 85 percent while neural network revealed 80 percent.

In both experiments the output indicated that decision tree performed better. Besides, decision tree generated understandable rules that could be easily presented in human language and thus police officers can make use of these rules for designing crime prevention strategies. The experiment has proved that data mining is valuable to support the crime prevention process and particularly, decision trees seem more appropriate for the domain problem.

The researcher identified that data mining techniques could contribute a lot in identifying areas that are crime prone and the characteristics of offenders who involve in different levels of crimes. Thus, it could be more important to use the data mining technique as a tool for the decision making process. In other words, the Oromia Police Commission could optimize its crime prevention efforts by employing data mining technology (Leul 2003).

Association Rule Mining for Community Crime Pattern Discovery by Anna and Christopher in 2010.

According to the Authors current manual inspection of crime data by analysts is limited, primarily due to the amount of data that can be processed concurrently and in a reasonable time frame. Further, complex relationships between various crime attributes can be overlooked by human analysts. Providing automated knowledge

discovery tools becomes attractive to accelerate the efforts of local law enforcement.

The intention of the paper was application of association rule mining for community crime pattern discovery.

A relative support metric was defined to extract rare, novel rules from thousands of discovered rules. Such an approach relieves the need of law enforcement personnel to sift through uninteresting, obvious rules in order to find interesting and meaningful crime patterns of importance to their community.

Association rule mining has proven useful for this crime application, and has utility for other crime-related data sets. To the knowledge of the authors, this is the first experimental study of applying association rule mining to a crime data set.

A novel relative support measure was proposed to prune the set of rules and to extract rare rules from the larger original set. The use of relative support achieves a 95.2% reduction in the final number of rules. These resulting 675 final rules represent a much more manageable number of rules for a crime analyst to investigate. This enables law enforcement personnel to more easily understand the discovered rules by removing the need to sift through uninteresting, obvious rules in order to find interesting and meaningful patterns (Anna and Christopher, 2010).

An Enhanced Algorithm to Predict a Future Crime using Data Mining by Malathi and Santhosh in 2011.

In this paper the Authors look at the use of missing value and clustering algorithm for a data mining approach to help predict the crimes patterns and fast up the process of solving crime.

They has been concentrate on missing value algorithm and A-priori algorithm with some enhancements to aid in the process of filling the missing value and identification of crime patterns. They applied these techniques to real crime data. They

also use semi-supervised learning technique in this paper for knowledge discovery from the crime records and to help increase the predictive accuracy.

The crime dataset used in the present research work was downloaded from the Integrated Network for Societal Conflict Research (INSCR) website.

The Authors indicated that a major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. As information science and technology progress, sophisticated data mining and artificial intelligence tools are increasingly accessible to the law enforcement community.

These techniques combined with state-of-the-art Computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personnel. Computers are also less prone to errors than human investigators, especially those who work long hours.

Experimental results prove that the tool is effective in terms of analysis speed, identifying common crime patterns and future prediction. The developed tool has promising value in the current changing crime scenario and can be used as an effective tool by Indian police and enforcement of law organizations for crime detection and prevention (Malathi and Santhosh, 2011).

1 Overview of Corruption

Corruption has been widely studied and its effect on numerous areas of public and private life are well documented. Most studies agree that corruption is bad. For example Shleifer and Vishny (1993) show that corruption leads to a misallocation of abilities which is very costly for economy.

First, the very definition of corruption is worth a brief discussion. There is great debate today about how to properly define corruption and of what use the various definitions play in our understanding of the phenomenon. The most common definition involves private gain via public authority: the abuse of public power for private benefit or profit. This is the working definition that many world organizations use in discussing corruption, including the World Bank, the International Monetary Fund, and Transparency International. It is also, we should note, an attempt to define corruption universally and without regard to a specific culture. To refine this, corruption is literally a transaction. Much of the academic research, particularly from the economic perspective, emphasizes this approach. Corrupt transactions occur at the intersection of the public and private sectors (Rose-Ackerman, 1978).

The classical definition of corruption is often attributed to Colin Nye as an action that deviates from the formal duties of a public role (elective or appointive) because of private regarding (personal, close family, private clique) wealth or status gains (Nye, 1967).

Khan (1996) offers that corruption is an activity that deviates from the formal rules of authority because of private-regarding motives such as wealth, power or status. Corruption then may exceed legal boundaries and become a question of morality, following this logic naturally.

Often, the definition of corruption is tied to a particular style or category of the occurrence as a whole. For example, the distinction between systemic or institutionalized corruption and private or petty payoffs is seen as defining the study and understanding of the impact of corruption on state and society (Rose-Ackerman, 2006).

Another naming convention often used as a substitute for corruption is "rent seeking." Rent seeking, however, is really a category of corruption: it defines an economic transaction based on a value. In other words, rent is in excess of all relevant costs. As such, rent seeking may not always be considered immoral or illegal but is often inefficient (Kunicova and Rose-Ackerman, 2005).

Favoritism and nepotism are both types of clientelism a system of patron client relationships through which exchanges take place. Since social traditions and culture are often inextricably linked with such practices, viewing clientelism as corruption may be challenging without providing sufficient context: According to case studies, in many Sub-Saharan African countries, long lasting patrimonial and clienteles practices have over the years established what has been called hegemonic elites, or ruling state-classes. These are composed of rather small elite of politically and economically dominating families (Amundsen, 1999).

Corruption may also be considered in large scale terms, often referred to as grand corruption in the literature versus petty, opportunistic ventures. Political corruption is often considered on the grand scale such as corrupt branches of government, corrupt electoral systems, or corrupt public-private projects (Rose-Ackerman, 2006).

An interesting way to understand how such corruption becomes pervasive in an organizational structure is to view it as collectivization. Such corruption takes on a conspiratorial quality with the costs of participating in the collective action lower than the costs of whistle blowing or refusal (Amundsen, 1999).

Finally, some literatures distinguish corruption on the basis of outcomes, specifically labeling it as redistributive versus extractive. This distinction is based on the idea that corruption is not necessarily mutually beneficial; in other words, not a quid pro quo deal. If the corruption is understood as immediately benefiting one party more than the other, then the direction or flow of that benefit may determine whether it's extractive or redistributive. The latter is most easily understood in the classic Robin Hood formula of

stealing from the rich to give to the poor, while the former is a more feudal style of authoritarian rule and benefit. It is important, therefore, to understand corruption as an evolving concept that may not always be simply a transactional issue (Blundo and Sardan, 2006).

According to Blundo and Sardan (2006) corruption may characterize many types of social, civil, and governmental interactions depending upon the view point. To fully capture the impact of corruption on smallholders, it is critical that these perspectives be considered.

1.1 Forms of Corruption

According to Blundo and Sardan (2006) to drill down further, typical manifestations of corruption may include:

• **Bribery**: An offer of money or favors to influence a public official.

• **Embezzlement**: Stealing money or other government property.

• **Fraud**: Cheating the government through dishonesty.

• **Extortion**: may take place when money is demanded to do something. It can also be demanded by corrupt officials who otherwise threaten to make illegitimate use of state force in order to inflict harm. This is similar to extortion by organized crime groups.

• **Patronage**: refers to favoring supporters, for example, with government employment. It can be seen as corruption if this means that incompetent persons, as a payment for supporting the regime, are selected for loyalty rather than ability.

• **Nepotism**: Favoritism shown by public officials to relatives or close friends.

• **Kickbacks**: are official's shares of misappropriate funds allocated from his or her organization to an organization involved in corrupt bidding.

1.2 Types of Corruption

According to Riley (1997) Corruption can be categorized under three headings.

1 Incidental Corruption: this is small scale. It involves junior public officials, such as policemen or customs officers; it produces profound public alienation; it has little macro-economic cost, but it is often hard to curb.

2 Systematic Corruption: this is corruption that affects, for example, a whole government department. It can have a substantial effect on government revenues; it may divert trade and/or development; it can only be dealt with by sustained reform.

3 Systemic Corruption: that is government by theft. In this Situation honesty becomes irrational, and there is a huge developmental impact (Riley, 1997).

2 Perspectives of Corruption in Different Academic Disciplines

2.1 Economics

Much of the historical research and classical methodology for gathering and analyzing empirical data about corruption stems from an economic perspective. In fact, the study of corruption is often contained to an econometric approach, utilizing models to better understand the transactions, costs, and general effects on production. This approach essentially characterizes corruption as inefficiency in the market which can be remedied by some imposition of incentive, structure, or regulation. While there is plenty of data to support this conclusion, there is also plenty of data to controvert it. Currently, much of the economic scholarship is recognizing this and addressing it by embracing aspects of other academic disciplines to provide more a comprehensive understanding of the issue (Transparency International, 2006).

2.2 Political Science

Almost as prevalent as the economic research into the effects of corruption on markets, development, and production is the political science theory on the state, civil society, and

collective organization. Typically, the research from the political science camp involves more analysis of the power structures inherent in corruption.

2.3 Anthropology

Perhaps one of the more interesting approaches to understanding corruption is from the field of anthropology. Relatively recently, the literature in this area explores the causes and effects of corruption on a more humanistic level, emphasizing the social conduct, history, and culture as factors. Where political science sees the intersection of state and society as the source of corruption, and economics understands corruption as a flaw in the system, anthropology looks at corruption as social exchange. And in this context, perception a theme that will reappear often is as important as reality (Haller and shore, 2005).

2.4 Public Policy

A quick search online yields numerous reports and recommendations from a variety of think tanks, NGOs (Non-Governmental Organizations), and public policy institutions. Most such reports emphasize the pragmatic: fighting or reducing corruption in developing countries. These reports tend to amalgamate some of the academic literature from economics, political science, and to a lesser extent, anthropology in characterizing the problem and proposing solutions.

3 Current Trends and Themes in Corruption Literature

Overall, the study of corruption in the various disciplines continues to struggle with a number of themes. First, the analysis of corruption is almost always tied to the state. This results in two principal issues: 1) a comparative study, most often referred to as cross-country analysis and 2) a distinction between government and civil society (Johnston, 2005).

The first of these, cross country analyses, often provide startling results. For example, in studies comparing corruption in Asia to Sub-Saharan Africa, the results demonstrate

similar issues with corruption yet very differing results in economic growth and productivity. Furthermore, much of the cross country analysis makes grand assumptions, representing an entire region or continent with the study of a few countries. The distinction between government and civil society also leads much of the public policy development to blame weak governance. But this distinction is currently being challenged by much of the research, especially from the anthropological point of view.

Failing to demonstrate weak governance as the sole cause for corruption, the next easiest culprit to assign blame is culture. This can lead to a fatalistic viewpoint and hardly delivers a more accurate understanding of the issues. A more refined view of weak governance or culture, in some regards, is the emphasis on institutional structure as the source of corruption. This theory suggests that much of corruption is built in either purposefully or unknowingly to institutional structures (Gebenga, 2007).

Transparency is the often cited goal in curing much of corruption's afflictions. One particularly interesting way of understanding transparency is to view patronage, rent-seeking or other acts that might be considered corruption as acceptable if enacted in a transparent environment. For example, the vast resources spent on political lobbying the United States could certainly be understood as corruption. Yet, the transparent or at least somewhat transparent environment in which that occurs somehow changes the dynamic to an acceptable practice (Jain, 2001).

Many studies have correlated higher wages to less corruption in government agencies, but there is a tipping point at which higher wages actually lead to increased corruption as government workers struggle to keep ahead of their peers. And the notion of incentive systems as either the cause or the solution can be challenged by research from academic disciplines. "Rewarding A while hoping for B," our best intentions to provide incentives for a particular behavior may ultimately reward an undesired behavior (Kerr, 1995).

4 Corruption and Development in Africa

Since the post-colonial Africa, corruption has been a cause for concern because it diverts already limited funds, undermines economic progress and impedes policy changes required for development. Africa presents a typical case of the countries in the world whose development has been undermined and retarded by the menace of corrupt practices. A series of reforms have been carried out in all the African countries so as to make the system (African states) efficient and result oriented. However, the anticipated gains of such efforts or reforms have not been visible due to series of factors which include that of corruption. Without doubt, corruption has permeated the African society and anyone who can say that corruption in Africa has not yet become alarming is either a fool, a crook or else does not live in this continent (Chinua, 1988).

The situation has gone so bad to the extent that whichever way one views corruption, it involves a violation of public duty or deviation from high moral standards in exchange for (or in anticipation of) personal financial gains. It is connected with moral and fraudulent acts (Bamidele, 1995).

The effects of corruption are felt in the political and social, as well as the economic, spheres. Although the direct costs of corruption may be high in terms of lost revenue or funds diverted from their intended use, the indirect costs in terms of the economic distortions; inefficiencies and waste resulting from corrupt practices are more problematic over the long-term and thus make it more difficult to address. Corruption increases the costs of doing business, wastes resources, hence radically reduce revenues accruing to the state. It also results in poor service delivery, "moonlighting" or multiple concurrent sources of employment and refusal to perform normal functions without additional payment. Moreover, corruption deepen poverty and make it difficult for ordinary people to get ahead as the result of their own efforts (Gbenga, 2007).

Here is increasing evidence that the social and economic cost of corruption disproportionately affects the poor, who not only suffer from the lack of services and efficient government, but who are also powerless to resist the demands of corrupt

officials. Different arguments have been put forward to explain the pervasiveness of corruption in Africa these include poverty, the personalization of public office, the political culture and the inability of leaders to overcome their colonial mentality in respect of their perception of public office (Lawal and Tobi, 2001).

Development is all about the capacity of members of the society to actualize them by participating actively in the social engineering of their life and destiny. It entails the ability of the individuals to influence and manipulate the forces of nature for their betterment and that of humanity" (Nnavozie, 1990).

Rodney (1972) sees beyond the individual or people's perception of development and conceived development whether economic, political or social to imply both increase in output and changes in the technical and institutional arrangement by which it is produced. In other words and more importantly, development is a multi-dimensional concept and in spite of the various conceptions, development is basically about the process of changes which lies around the spheres of societal life (Rodney, 1972).

It is clear to us that there is a linkage between corruption and development. In other words, there is a direct reaction of the devices of corruption on development. If, for instance, development is conceived to include the capacity of a government or system to manage resources efficiently to improve the well-being of the citizens and then corruption can be thus regarded as one of the main obstacle to good governance and development of countries.

4.1 Effects of Corruption in Africa

The effects of corruption in Africa can be analyzed from three main perspectives and they are Political, economic and socio-cultural aspects. From the political view point, corruption has the capacity of engendering political instability, breakdown of law and order, brain drain, inefficiency of the public service among others. Viewed from the economic perspective, corruption is no doubt an enemy of economic development in the international scene, as it gives the continent a poor image in the international scene and it

gives the continent poor image in interpersonal and business relationships. A nation that condones corruption is often besieged with a lot of economic and social vices. Economic and social infrastructural facilities are vandalized to create room for unnecessary replacement and purchases or conversion to personal use. Trade and commerce cannot thrive, as investors will be unwilling to invest much trade or business in this part of the world, the overall resulting effects of all these malpractices will increase in the rate of inflation, unemployment and decline in output, foreign reserves and deterioration in the standard of living of the people (Gebenga, 2007).

In the socio-cultural context, corruption apart from engendering poverty has the capacity of changing the social values of a good and progressive society dramatically to nothing else than the crazy pursuit of wealth affluences, power and society recognition. People no longer appreciate the virtues of good morale, conduct and practices. Without doubt corruption has eaten deep into the fabric of the African people and the African society and it continues with the people almost permanently. Africa presents a typical case whose development and the desired change have been undermined and retarded by the menace of corrupt practices (Gbenga, 2007).

5 Historical Development of Corruption in Ethiopia

As many scholars and experts would agree, corruption is not a social phenomenon that can be explained by a simple cause/effect model. It is a complicated issue, often the result of many contingent circumstances, which produce varied and wide-ranging effects. Without limiting the generality of this argument, however, one can cite numerous factors that are widely believed to be the major causes of corruption in Ethiopia. Family and ethnic loyalties and obligations, imprecise distinction between private and public interests, privatization, weak financial management, inadequate accounting and auditing, weak legal and judicial system, over regulated bureaucracy, deterioration of acceptable moral and ethical values, unsound policies and inefficient civil service system have also been cited by some scholars and researchers as the primary causes of corruption in Ethiopia (Tesfaye, 2007).

5.1 Consequences of Corruption in Ethiopia

According to AEO (2007) during the Imperial and the Derg Regimes, corruption is said to have resulted in undermining the legitimacy of the governments and weakening their structures, reducing productivity, hindering development, worsening poverty, marginalizing the poor, creating social unrest and finally speeding up their downfall. Unfortunately, it has continued to pose threats to the Country's development and democratization processes. Currently, corruption is believed to be one of the major factors that significantly contribute to the reduction of government revenue. It can also negatively affect the on-going poverty reduction program at the national level.

According to expert analysis by the Ethiopian civil service reform program, the major causes of corruption in Ethiopia are poor governance, lack of accountability and transparency, a low level of democratic culture and tradition, lack of citizen participation, lack of clear regulation and authorization, low institutional control, extreme poverty and inequality, harmful cultural practices, a command economy during the Derg regime, weak financial management, inadequate accounting and auditing, and a weak legal and judicial system (AEO, 2007).

5.2 Areas where Corruption is believed to be Rampant in Ethiopia

According to the outcome of the corruption survey conducted in 2001, the areas where corruption is believed to be rampant are those where financial resources are transferred from the private to the public sector and vice versa. Other agencies where corruption is believed to be flourishing include those engaged with the allocation of land and government housing, provision of telephone and electric services, granting of loans, licensing and issuance of permits, collection of taxes and procurement of consumable and fixed assets. Customs and excise offices are also believed to be highly affected by corrupt practices (Tesfaye, 2007).

6 Corruption and the Smallholder (micro level firms)

Corruption impacts the smallholder in a variety of distinct and overlapping manners, almost always impairing growth or benefits. The following represent a number of case studies, analyses, and research projects that demonstrate the impact of corruption on smallholders. They are separated by topic or theme in order to tie the examples back to the academic theory on corruption (TI, 2006).

Grand or Systemic

On the most macro level, grand corruption can impact smallholders directly and indirectly. As the perception indices indicate, the impact of the mere perception of corruption can be a financial charge in terms of lost potential foreign investment and business development that is supposed to be the backbone of the given Nation (Rose-Ackerman, 1978).

Embezzlement

To some extent, the effect of embezzlement on the smallholder may be characterized as magnificent or universal: it's an indirect loss of resources that should have otherwise "trickled down" to the lowest level. For example, in a well-known and often cited audit of the school system in Uganda, it turned out that only 13% of the funds allocated for non-salary items like textbooks and supplies reached the schools. All allocations were subsequently published in newspapers and on the radio, increasing the funding that reached the schools by 90% (Spector, 2005).

Services

There is growing empirical research indicating the provision of services is directly and indirectly impacted by corruption. Corruption is linked to reduce costs on operations, maintenance, medicine, schools, health care, and other social services. Specifically, a high level of corruption has difficult cost for a country's child and infant mortality rates,

percent of low-birth weight babies in total births, and dropout rates in primary schools (Gupta in Jain, 2001).

Bribery

The simplest and most easily understood form that corruption often takes, bribery is seemingly ubiquitous in developing countries and imposes a tremendous burden on smallholders. One study in the Dominican Republic demonstrates that smallholders there cannot participate in forestry as an economic activity due to the overwhelming bribery that exists in the government bureaucracy (Peter, 2004).

7 ICT (Information Communication Technology) Tools for Combating Corruption

As Ake et. al. (2007) stated there are numerous ICT tools that can be used during various phases of combating corruption, including prevention, detection, analysis, and corrective action. Anticorruption software is a label used for various tools designed specifically for detecting and taking action against fraud, including both "intelligent mining" of data sets and administrative procedures. The origin of the tools can be traced to methods used for intelligence and police work. For example, the "Pursuit" software from Distillery Software (http://www.distillerysotware.com/industries/anti_corruption.html) contains tools for intelligence "allowing investigators to capture rich entity and association data and build up a picture of relationships between persons of interest, assets, events and organizations", complaints management and investigation management. Complaints and investigations management are basically administrative tools facilitating various procedures involved, such as "Witness and Exhibit Management" (management of structured and unstructured data), "Brief of Evidence/Case File Production" (produces court documents automatically as output from the other tools), and "Asset Tracking" (linking to individuals and organizations, and tracking actions in relations to those assets).

These are tools focusing mainly on systematic and or large stake corruption and they operate only in the electronic world so things that take place exclusively in the physical world escape attention. However, small-stake corruption can also be traced this way. For example, one of the effects of even petty corruption is that civil servants own more expensive property than they can reasonable afford given their official salary, and asset tracking is one way of systematically finding this out. Another way to do this is, of course, by coordinating existing government records of people and properties, if such exist sufficiently (Ake et.al, 2010).

In India, the National Rural Employment Guarantee Act (NREGA) guarantees wage employment to every household whose adult members volunteer to work on labor intensive public works annually. The system is administered in a way that even its proponents admit holds ample opportunity for corruption and further exploitation of the rural poor it intends to serve. An investigation has found that officials and politicians inflate work bills, fake wages and pocket funds. To bypass the human agents involved in the administrative process, computers, not officials, now issue job cards, provide work estimates, and generate each worker's pay slip online (at www.nrega.ap.gov.in). Payments are made into individual postal accounts created for the purpose. Reportedly this technologically uncomplicated measure has so far recovered a substantial amount of misappropriated funds (Srivastava, 2008).

CHAPTER THREE

3.1 Methodology of the Study

To achieve the objective of the study the researcher has used appropriate data mining methodologies as stated below:

3.2 Research Design

3.2.1 Cross Industry Standard process for Data Mining (CRISP-DM) Methodology

CRISP-DM has utilized as a framework for the research. According to Azevedo (2008) CRISP-DM is the most popular data mining model that stands for Cross-Industry Standard Process for data mining. CRISP-DM provides the overview of the life cycle of data mining tasks comprising six stages, their respective task and the relationship of these tasks. The sequence of the phases is not rigid. Moving back and forth between different phases is always required until the required result is gained. It depends on the outcome of each phases to determine which phase is done next, or which particular task of a phase, should be performed next.

In terms of relationships between the data mining tasks, at this description level, it is not possible to identify all relationships. Essentially, relationships could exist between any data mining tasks depending on the goals, the background and interest of the user and most importantly on the data (Pete et.al, 2000).

The CRISP-DM organizes the process model into hierarchical process model. At the top level, the task is divided into different phases. Each phase consists of several second level generic tasks. These tasks are complete (covering the phase and all possible data mining applications) and stable (valid for yet unanticipated developments). These generic tasks are mapped to specialized tasks. Finally these specialized tasks contain several process

instances which are record of the actions, decisions and results of an actual data mining engagement process.

CRISP-DM model describes the cyclical nature of data mining itself. A data mining process is continuous even after a solution has been installed. The lessons learned during the process can generate new, often more focused on business questions. Subsequent data mining process will benefit from the experience of previous ones. A brief outline of the six phases is presented below:

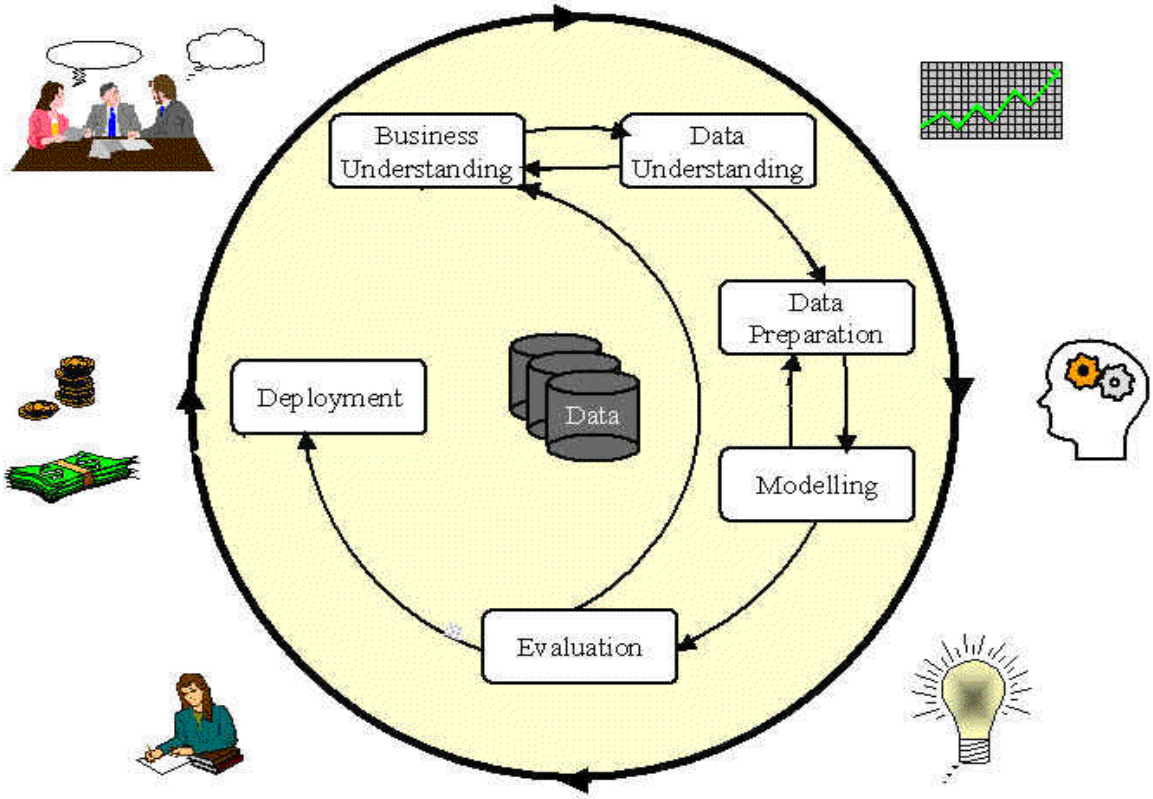

Figure 3.1 Life Cycle of CRISP-DM (Azevedo, 2008)

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.

In this phase the researcher has tried to focus on understanding the business objective of FEACCE by discussing with the employees and analyzing their documents or publications like pamphlets, journals, newspapers, research papers and magazines found in its library. The major business of FEACCE is combating corruption within Ethiopia by working with the society and this is the significant role for the economic, social, cultural and political development of the country. The research problem has constructed in the way that can be solved by the data mining application.

2 Data Understanding Phase

This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypothesis for hidden information.

For this research the data has been taken from FEACCE data base that contains information about the corrupters who are government officials, private organization workers, jobless and self-employed with their socio demographic characteristics (name, job, address, educational level, gender, benefit gotten from corruption, number of children corrupters has, marital status, decision given by courts, crime area and crime level) and these were the selected attributes from the data base to be mined to achieve the research objective by the researcher with the help of the domain experts and the data mining tool selected.

3 Data Preparation Phase

This phase covers all activities to construct the final data set from the initial raw data in a way the data will be given to the data mining software. The tasks are likely to be performed multiple times, and not in any prescribed order. In this process entry and attribute selection should be done because all the attributes and attribute values may not be critical for the data mining. Transformation of data in to one category with the same context is the other task. The quality of data should be ensured by cleaning missed values, noisy data, outliers, inconsistencies, duplicates and irrelevant. Finally data from different repositories should be integrated in to one format.

In this phase the researcher prepared the data set to feed to the selected data mining tool. From The data base of the FEACCE containing 22 attributes 11 attributes were selected and attribute values in ordinal form transformed in to nominal forms. To insure the quality of data MS-Excel (Microsoft Excel) built in function has used like filter, find and replace. And lastly data from different repositories were merged in to MS-Excel data base. The brief description of data preprocessing will be described in chapter four.

4 Modeling Phase

In this phase different modeling techniques were selected and applied and also their parameters are standardized 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, stepping back to the data preparation phase is often necessary. In this research, association rule model, clustering model and classification models has been applied on the prepared data set.

5 Evaluation Phase

At this stage the model or models obtained are more thoroughly evaluated and the step executed to construct the model or the models are reviewed to be certain that they properly achieve the objectives of the study. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results has been reached.

In this phase the researcher participated domain experts to select interesting rules generated from the association rule mining models. Clustering and Classification models were examined based on performance they registered. Rules generated from both predictive models (association rule mining and classification) were evaluated.

6 Deployment Phase

In general Creation of the model will not be the end of this research. Even if the intention of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the FEACCE can use it.

Depending on the requirements the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the Commission. In many cases it is the commission, not the researcher, who carries out the deployment steps. However, even if the researcher will not carry out the deployment effort it is important for the Commission to understand up front what actions need to be carried out in order to actually make use of the generated models.

To deploy the models generated the researcher documented the hidden knowledge generated in different models in the means of the commission can use it and further more ready for expected efforts to introduce application of data mining in corrupt activities.

In general CRISP-DM methodology is really complete and can be documented and all its steps are accordingly organized, structured and defined so that the research would be understood.

3.3 Tool selection

Waikato Environment for Knowledge Analysis (WEKA) is a collection of state-of- the-art machine learning algorithms and data preprocessing tools. It provides extensive support for the whole process of the experimental data mining, including preparing the input data, evaluating the learning schemes statistically, and visualizing the input data and the result of learning. This comprehensive toolkit is accessed through a common interface and users can compare different methods and identify those that are most appropriate for the problem at hand.

To create the Association, Clustering and Classification models WEKA has been used. The selection of this particular tool was because the researcher is familiar with the software and also because the software is one of the recommended software for the purpose of data mining researches. Nowadays WEKA is recognized as a land mark system in data mining and machine learning and it has achieved widespread acceptance within academia and businesses, and has been become a widely used tool for data mining researches.

CHAPTER FOUR

Data Understanding and Preprocessing

This chapter is about identification of source data to be used by the researcher in order to meet the objectives of the study and preprocessing the data (cleaning, integration, reduction, transformation and selection) in the way that is appropriate to the data mining tool and the models. Different kinds of efforts are undertaken on the data by the researcher from the data collection until it becomes suitable for the experiment and analysis.

4.1 Data Collection

The data for this study has taken from the FEACCE which was the collection of investigation of Corruption offences committed on both the country property and the citizens. The actual data collection was started by extracting MS-Excel (Microsoft Excel) data base and SPSS (Statistical Package for Social Science) data base that was in English language. Most of data was converted from the manual documents and files (the annual publications and investigation files) that were in Amharic language.

The data was sorted by file numbers and contains all corruption offences investigated by FEACCE in different kinds of governmental institutes, nongovernmental and private organizations from the year 1993 until 2004 E.C. The data was collected from the society by using the free call line of FEACCE. The computer data base has 22 attributes and 1000 entries and the remaining data sets are collected from the manual databases which are 3189 entries.

The data set was represented as p × q data matrix. The p rows represent situations on which measurements of the q attributes were taken. The p rows are the representatives of the records while the q columns represent the attributes, features, fields or variables of the data matrix. These different expressions are intended to the same meaning that represents columns (attributes).

Although this study has intended to uncover the hidden knowledge of situations of corruption offences committed by government officials, due to insufficient data, the study covers all individuals who have committed corruption and punished by different courts in the country. Such offences include corruption committed on justice, land, cash transfers by using computers, forged licenses and documents, bribes, embezzlement, maladministration, sexual harassment, inappropriate payments, tax, partiality on project bids etc. As it is not that much necessary in data mining to use the attributes such as name, phone numbers, and addresses the researcher has not used these attributes.

According to Seifert (2004), while technological capabilities are important, there are other accomplishment and oversight issues that can influence the success of a project's outcome. That issue is data quality assurance, which refers to the accuracy and completeness, believability and interpretability of the data being analyzed. Before feeding the data set to the data mining tool the quality of data should be assured in order to achieve best results. Lack of using uniform standards for data set and human typing errors are special areas that the researcher needs to focus on, because these problems possibly create misleading knowledge and needs to be standardized and cleaned.

To make the data free of the above problems the quality should be improved through data preprocessing which encompasses avoiding the data fields which are not required for the data mining problem, identifying outliers or errors (such as person age represented as age of 231), standardizing attributes (such as different attributes are used to mean the same thing), cleaning redundant or duplicate data and normalizing the attribute values used to represent information for the columns attributes.

To have a clear insight of the data used the researcher categorized the data set attributes into two groups and they are:

- Offender Information: age, sex, education level, marital status, number of children, employment, position salary.

- Offence Information: benefit, decision, crime area, punishment year and crime level.

4.2 Data Preprocessing

Data preprocessing is commonly used as a preliminary data mining practice and it is more time consuming when it is compared to the other data mining activities. It transforms the data in to a format that will be easily and effectively processed by the selected data mining tool. Data preprocessing steps should not be considered completely independent from other data mining phases. Iterations of the data mining processes, all activities, together could define new and improved data sets for subsequent iterations.

To get the required information from huge, incomplete, noisy and inconsistent set of data, it is compulsory to use data preprocessing techniques. Data cleaning is a procedure to clean the data by filling in missing values, smoothing noisy data and identifying or removing inconsistencies. Data integration merges data from different sources in to a coherent data store like from manual and computer databases, data marts and data warehouses.

Data transformation operations, such as normalization and aggregation, are additional data preprocessing procedures that would contribute toward the success of the mining processes. Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. All the data preprocessing techniques are done by using MS-Excel built in functions like search and replace, filtering and auto fill mechanisms.

4.2.1 Data Cleaning

Data in the real world are dirty and incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data) noisy (containing errors or outliers) and inconsistent (containing discrepancies in codes or names). Before integrating data from different data bases the researcher has tried to curve such data quality problems. Missing data, noisy data, and inconsistent data reduces the accuracy and efficiency of the data mining techniques therefore care must be taken to produce

correct, potentially useful and hidden knowledge for this study. The data the researcher has collected is not clean and contain typographical errors, missing values, noisy or inconsistent data. So the researcher needs to apply different techniques to get rid of such anomalies.

- **Outliers**: - Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors and sometimes they are natural and abnormal values. Such non representative samples can seriously affect the model produced later. For this study this problem has been presented in the offenders age attribute like the value ”281” that no one can have.
- **Missing data**: - Missing data, which occurs in reality, should be minimized. Missing data could reduce the quality of a global model. On the other hand, some Data mining techniques are robust enough to support analyses of data sets with missing values. All of the attribute values were susceptible for this problem; there were no attribute with missing name and it is known that this is common to all of the data bases. From out of 11 selected attributes for the data mining problem the attributes job and salary have the highest number of missing values constituting 15 % and they were replaced by”?” to mean not given or not known.
- **Noisy data**: - attribute values that might be incorrect or invalid. Like typographical errors. Such a typographical errors are like “bed” instead of “bid” in the values of crime area and values for crime area *“መሬት”* is miss spelled as *“መረት”* to mean “un low full land collection”. And these errors were corrected by asking the data base administrator**.**
- **Inconsistent data**: - attribute values containing discrepancies which are the problem of format or standardization. The values of attribute sex contain value like “fm” to mean female and “m” to mean male. And there is abbreviation for the value of attribute job “gov. civil service.e” to mean government civil service employee and “s.employee”, to mean self-employee and “g.e.employee” to mean government enterprise employee which is unique only for the data base

administrator. All of these problems are changed to common naming and formats by asking the data base administrator.

4.2.2 Data Integration

The data needed may reside in a single database or in multiple databases. The source databases may be transaction databases used by the operational systems of the organization. Other data may be in data warehouses or data marts built for specific purposes. Still other data may reside in a proprietary database belonging to another company such as a credit bureau.

Data integration and consolidation combines all data from different database sources into a single mining database and requires bringing together the differences in data values from the various sources. Improperly joined data are a major source of quality problems. For example, under the values of the attribute "benefit" there are some values putted in U.S (United States) Dollar ($) and Gold without converting them in to Ethiopian birr in the different databases used for this study. The researcher converts this by multiplying the given number by the current value of 1 U.S $ in Ethiopian birr that is (17.36).

4.2.3 Data Reduction

Data reduction is a way to reduce the size of the data set without affecting the data mining result; an approximate data set is created that can be made accessible. For example sample of the data set can be used instead of dealing with the full data Hand et al, (2001). Data reduction techniques like data cube aggregation, dimensionality reduction, data compression, numerosity reduction and discretization can be used to get a reduced representation of data without losing the contents of information from the original data set.

- **Dimensionality Reduction**

Dimensionality reduction is a technique in which irrelevant, weakly relevant and duplicate attributes (number of attributes intended to mean the same thing) have got solutions from the data base. In other words it is where the most relevant and needed

attributes which are helpful to meet the objective of the study are selected by the researcher. Dimensionality reduction helps the model to be free of over fitting and enhance the accuracy of the result, reduce time and space required in data mining and allows easier visualization of the result. Dimensionality problems are concerned with irrelevant attributes, redundant attributes, attributes with many different values and attribute with non-variant values.

- **Irrelevant data reduction: -** It is a phase in which noisy data and irrelevant data are removed from the collections which are not of interest to the data mining task being developed because of it may have the same value for the measure of the variable or it may have distinct values for each of the data set and also the attribute may have the same intention or meaning with the other attribute. Generally attributes that are not useful to generate knowledge are irrelevant.

  Attributes "tip codes", "decision1", "working institutes", "peers working institutes", "reported date", "decision date", "address", "file number", "job", "out of the commission" and "additional evidence" were attributes that were discarded because of their irrelevance with the interest of the data mining problem in this study.

  "Tip code" has the same intention with the attribute "crime area", "decision 1", has non-variant value for all data sets ("medebegna kereker"), "working institutes", "peers working institutes" and "job" contain many attribute values for the instances (more than 15 values), " reported date", "address", "file number" and "decision date" contains distinct values and "out of the commission" and "additional evidence" were containing extra detailed information. Thus these 11 attributes were discarded because they decrease the performance of the models.
- **Attributes with the same meaning: -** Attributes may have the same information content or they may have very close intentions at this time only one attribute should be selected for the best performance of the data mining. In this case the researcher has faced this problem. In the data base there are two attributes that have the same intention: "tip code" and "crime area". Since the study has the

objective of revealing the hidden knowledge in the corruption area the researcher preferred to select the attribute "crime area" rather than "tip code". It means the attribute "tip code" has been irrelevant and discarded.

- **Non variant-Attributes: -** attributes that holds true for all the data sets are expected to be ignored, for instance the attribute "decision 1" have the same values for all the datasets that is "medebegna kereker" to mean common court process and this attribute is believed to be irrelevant attribute and the researcher has discarded it from the data base because it creates a problem on models by decreasing the level of accuracy.
- **Attribute taking many different values:** - attributes that have very many values for the data set also need to be excluded from the data set. Attributes like "file number" and "tip codes" and "offenders working institutes" have over 75 attribute-values with very detailed information.

➢ **Numerosity Reduction**

Numerosity reduction is used when the huge data set needs to be represented by small amounts of data set (sample) and this could help to save the space and time complexity used by the data mining algorithm. For the purpose of this study the whole data set has been used by the researcher because it is believed to be not problematic on the space and time complexity needed by the data mining algorithm being developed.

4.2.4 Data Transformation

Data transformation is a function that maps the entire set of values of a given attributes to a new set of replacement values such that each old values can be identified and replaced by new values. The data set collected has not been in the way that is appropriate format to be accepted and processed by the selected data mining tool (WEKA) easily. For example, age of the offender is an ordinal attribute and this needs to be transformed in to a new category format like a nominal attribute (young, adult and old) or in a range form like 20-30 and 30-40. The attribute salary is also ordinal and needs to be transformed in to a new format like (very low, low, medium and high).

- **Concept Hierarchy**

Data can be abstracted at different conceptual levels. The raw data in a data base is called at its primitive level and the knowledge is said to be at a primitive level if it discovered by using raw data only. For example most of the statistical tools for data analysis are based on the raw data in a data base. Abstracting a raw data to a higher conceptual level, discovering knowledge and expressing knowledge at a higher abstraction levels have a great advantage over data mining at a primitive level to produce summarized and meaningful knowledge (Yijun, 1997).

For example if we discover a rule from the data set without representing it in a new format (conceptual hierarchy) the rule looks like this:

Rule 1: 80% of peoples who are titled as professor, senior engineer, doctor and lawyer have salary between 50,000 and 100,000.

After abstracting the data set in to certain higher levels, we may get this kind of rule:

Rule 2: well educated people get well paid.

Generally rule 2 is much more summarized than rule 1 and in a certain extent it is more preferable. What we mean here is people entitled with professors, doctors, engineers and lawyers have the new transformed title and to the higher level concept with educated people and we summarize salary between 50,000 and 100,000 to the higher level concept well paid.

For this study the researcher used concepts for the variables offenders' age, salary, benefit and number of children. According to the FDRE (Federal Democratic Republic of Ethiopia) criminal code number 53 and 56 age groups are given in three groups for corruption offenders. The first group is from age 9-15 and these age groups are indicated as free of corruption crime. The second age group is from 15-18 and they are said to be "ወጣት አጥፊዎች" to mean "Young Criminals" and they could prisoned in special places.

The third group is above age 18 and these age groups are called “adult” and they will take justice according to the crime they has committed.

The dataset the researcher collected has values for the “Age” attribute from 20-80, from the criminal code these age are grouped as above age 18 and the researcher has tried to conceptually represent these age ranges in to range of 10 that is from age 20-30, 30-40, 40-50, 50-60, 60-70 and 70-80.

The attribute “Benefit” transformed to ranges such as benefit that is less or equal to 50,000 Ethiopian birr changed to little, from 60,000 to 70,000 is transformed to medium, from 70,000 up to 80,000 is changed to high, from 80,000 up to 90,000 changed to very_heavy and grater or equal to 100,000 birr has changed to extreme benefit.

The attribute “salary” also conceptually converted in ranges. Salary less than 3000 Ethiopian birr has changed to very_low, from 300 up to 6000 transformed to low, from 6000 up to 9000 medium and greater than 9000 changed to high.

For the attribute “Children” ch0 has given to offenders with no children, ch1 for children 1-5, ch2 for children 5-10 and ch3 for children greater or equal to 10.

4.2.5 Final Selected Attributes

Among the steps in preprocessing, attribute selection has a special role. Attribute selection is a process in which a subset of M attributes out of N is chosen, complying with the constraint $M \leq N$, in such a way that characteristic space is reduced according to some criterion and also attribute selection guarantees that data getting to the mining phase are of good quality (Lui and Motoda in Helyane, 1998).

For this research the researcher identified 11 attributes out of 22 attributes with the help of the domain experts who are working in FEACCE. Next the selected attributes were ranked by information gain method using the tool selected (WEKA) (figure4.1). These attributes are presented with their descriptions in the table 4.1 and (see their values in data base in appendix A).

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 11 crimelevel):
        Information Gain Ranking Filter

Ranked attributes:
 0.481846   8 benefit
 0.205226  10 crimearea
 0.031539   1 age
 0.0199     7 salary
 0.017001   3 educ
 0.015558   6 employment
 0.011267   9 decision
 0.009107   5 marital
 0.004428   4 children
 0.000955   2 sex

Selected attributes: 8,10,1,7,3,6,9,5,4,2 : 10
```

Figure 4.1 Rank of attributes using WEKA information gain selection method

Rank	Attribute name	Data type	Description
1	Age	Nominal	Age of the offender
2	Sex	Nominal	Gender of the offender
3	Educ	Nominal	Education status of the offender
4	Children	Nominal	The number of children the offender have
5	Marital	Nominal	Marital status of the offender
6	Employment	Nominal	Employment or recruitment of the offender
7	Salary	Nominal	Salary of the offender
8	Benefit	Nominal	Benefit gained by the offender
9	Decision	Nominal	Decision passed to the offender from courts
10	Crime-area	Nominal	Crime areas offender committed
11	Crime-level	Nominal	Class variable (predictable state)

Table 4.1 Descriptions of the 11 selected attributes

4.3 Switching in to Final Dataset Format

The whole numeric values of attributes like age, salary, children, and benefit are converted in to nominal values because the algorism selected for association rule mining (A-priori) can't be active when the data set is in numeric values. First data set are prepared and saved in MS-Excel file. Secondly they were saved in .CSV (Comma-delimited) format to separate each of the data sets with the intention of replacing the tabs and spaces. Thirdly, the .CSV file format was opened in Notepad and the header information is added with the symbol @ placed at the front of each of key information: "relation", "attribute" and "data". This means "@relation" was put in front of relation name, "@attribute" was put in front of all attributes affirmation and "@data" was also placed at the beginning of the attribute values. Fourthly, nominal attributes are followed by the set of values they can take on, enclosed in curly braces. Finally, the file was saved

with .ARFF (Attribute Relation File Format) extension as the next figure indicates. @ Line signals the start of the instances in the data set. Instances are written one per line, with values for each attribute in turn and separated by commas.

The attribute specifications in .ARFF files allow the data set to be checked to ensure that it contains legal values for all attributes, and programs that read ARFF files do this checking automatically. Figure 4.2 shows how data set are represented in .ARFF extension.

```
@relation corruption
@attribute,age{20-30,30-40,40-50,50-60,60-70,70-80}
@attribute,sex{male,female}
@attribute,educ{elementary,secondary,certificate,deploma,degree,masters_and_above}
@attribute,children{ch0,ch1,ch2,ch3}
@attribute,marital{married,never_married,divors,widdowed}
@attribute,employment{gov_employee,jobless,self_employee,private_employee}
@attribute,salary{high,low,midium,very_low}
@attribute,benefit{little,high,medium,very_high,extream}
@attribute,decision{discontinued,guilty,not_guilty}
@attribute,crimearea{asset_reg,bid,?,cash_transfer,justice,land,license,loan,others,payment,purchase,tax}
@attribute,crimelevel{simple,heavy,very_heavy}

@data
40-50,male,secondary,ch1,married,gov_employee,very_low,high,guilty,others,simple
50-60,male,elementary,ch1,married,gov_employee,very_low,extream,not_guilty,others,very_heavy
30-40,male,secondary,ch1,married,gov_employee,very_low,extream,guilty,land,very_heavy
20-30,female,degree,ch0,never_married,gov_employee,very_low,little,guilty,tax,heavy
40-50,female,elementary,ch0,never_married,self_employee,very_low,medium,guilty,others,simple
30-40,male,deploma,ch1,married,self_employee,very_low,extream,guilty,others,heavy
30-40,male,elementary,ch1,married,self_employee,very_low,high,guilty,purchase,heavy
70-80,male,elementary,ch3,married,self_employee,very_low,extream,guilty,others,heavy
40-50,male,secondary,ch2,married,self_employee,very_low,extream,guilty,others,heavy
30-40,male,elementary,ch2,married,self_employee,very_low,extream,guilty,others,heavy
40-50,female,secondary,ch0,never_married,self_employee,very_low,extream,guilty,tax,very_heavy
30-40,female,secondary,ch2,married,self_employee,very_low,little,guilty,tax,simple
40-50,female,degree,ch1,never_married,gov_employee,low,little,guilty,others,simple
20-30,male,degree,ch1,married,gov_employee,very_low,little,guilty,justice,simple
40-50,male,deploma,ch1,married,gov_employee,very_low,little,guilty,land,simple
40-50,male,degree,ch1,married,gov_employee,very_low,little,not_guilty,justice,simple
20-30,male,degree,ch0,never_married,gov_employee,low,little,guilty,justice,simple
40-50,female,secondary,ch1,married,gov_employee,very_low,little,guilty,others,simple
40-50,female,masters_and_above,ch2,married,gov_employee,high,very_high,guilty,others,heavy
50-60,male,secondary,ch1,married,gov_employee,low,extream,not_guilty,others,simple
30-40,male,secondary,ch1,married,self_employee,low,extream,guilty,tax,very_heavy
20-30,male,deploma,ch1,never_married,gov_employee,low,extream,guilty,land,very_heavy
30-40,male,deploma,ch1,never_married,gov_employee,low,extream,guilty,payment,heavy
30-40,male,degree,ch1,never_married,gov_employee,low,extream,not_guilty,land,very_heavy
30-40,male,deploma,ch0,never_married,gov_employee,very_low,extream,guilty,purchase,very_heavy
30-40,male,secondary,ch1,married,gov_employee,very_low,very_high,guilty,purchase,heavy
30-40,male,secondary,ch0,never_married,gov_employee,very_low,extream,guilty,purchase,simple
```

Figure 4.2 Representation of the dataset in .ARFF extension

CHAPTER FIVE

Model Building and Model Evaluation

This chapter briefly discusses the data mining techniques used, interpretations and evaluations of experimental results. As the researcher tried to indicate earlier, the techniques used for this research (in chapter one) are Association rule (to discover the relationships of the data set attributes), Clustering (to create the natural grouping of the data sets based on their similarity measurement) by setting K values (number of clusters needed) and Classification techniques (to classify new instances of the data sets based on the training data sets) so that classification rules are created to reveal persons with what characteristics of the given attribute are committing corruption crimes. These techniques are used by different data mining algorithms through re-adjusting their parameters to get the best result (hidden knowledge in corruption database).

5.1 Model Building

As three of the data mining techniques were identified in scope of this study (Association rule, clustering and classification) the researcher has built three different models for these different data mining techniques. For association rule modeling A-priori algorithm has been used; for clustering simple k-means algorithm; and for classification model decision tree has been used. To create classification model, input has been taken from clustering model data set and implemented in decision tree (J48 classification tree).

The multiple variable analysis capability of decision trees enables the researcher to go beyond simple one-cause, one-effect relationships and to discover and describe things in the context of multiple influences. Multiple variable analyses are particularly important in current problem-solving because almost all critical outcomes that determine success are based on multiple factors. Decision trees demonstrate decision alternatives, possible outcomes and events schematically through the visual approach that is helpful to easily comprehend sequential decisions and outcome dependencies so that one can quickly express alternatives clearly. Because of the advantages of decision trees mentioned here

the researcher preferred to use it from other classification algorithms such as Neural-networks and Bayes for the classification model building for this study.

Simple K-means algorithm is the algorithm used in building clustering model because it is a widely used clustering technique that seeks to minimize the average squared distance between data points in the same cluster; it lets the user to specify K values and it is very understandable than the other clustering algorithms like EM (Euclidian Measure) and Divisive clustering. Although it offers no accuracy guarantees, its simplicity and speed are very appealing in practice. the accuracy level of clustering model has checked by the classification model since their intention is creating classes of the data sets and data sets used by clustering model has used as an input for classification model.

Association rule mining finds interesting associations or correlation relationship among large set of data items. Association rules shows attribute value conditions that occur frequently together in a given data set. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and unlike the if-then rules of logic; association rules are probabilistic in nature. The rules appear in the following form: if (Antecedent) then (Consequent).

In addition to the antecedent (the "if" part) and the consequent (the "then" part) an association rule has two numbers that express the degree of uncertainty about the rule. Before the association rule mining applied the antecedent and consequent are set of items called item sets that are disjoint (do not have any item in common). The following figure shows this in Venn diagram (Figure 5.1.):

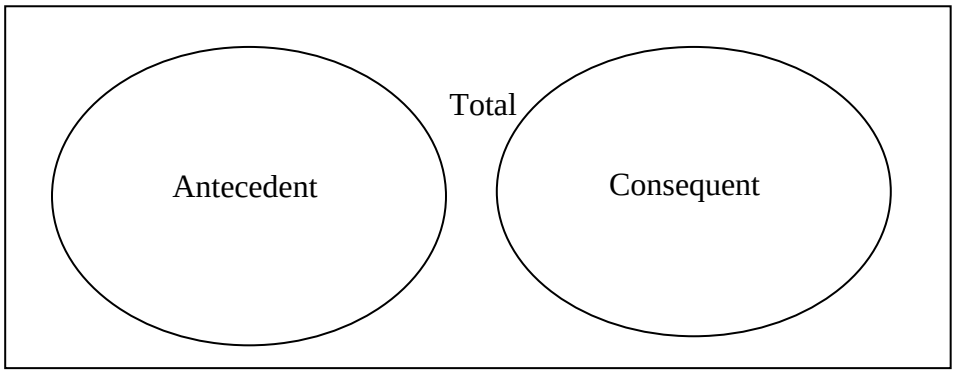

Figure 5.1 Representation of antecedent and consequent using Venn diagram

After the association rule mining the knowledge mined have two different numbers (figure 5.2.). The first number is called support for the rule, which is simply the number of transactions that include all items in the antecedent and consequent parts of the rule in other words it is the percentage of the total number of records in the database.

The second number is the Confidence of the association rule. Confidence is the ratio of the number of transactions that include all items in the Consequent as well as the Antecedent (namely the support) to the number of transactions that include all items in the Antecedent.

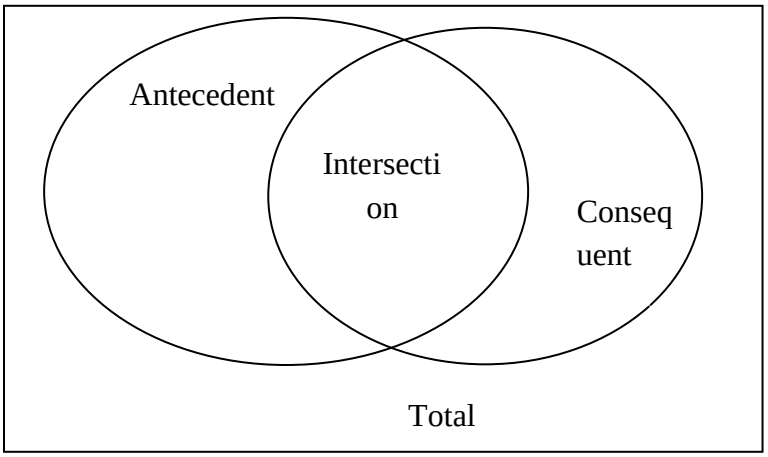

Figure 5.2 Representation of Antecedent, Consequent and their intersection using Venn diagram

5.1.1 Experiments and Analysis of Association Rule Model

Association rules really predict any attribute, not just predicting the class label of the incoming dataset like the classification rule, and this gives them the freedom to predict combinations of attributes too. Different association rules express different regularities that underlie the dataset and they can generally predict interesting and known facts of the dataset.

A-priori algorithm of WEKA for Association rule mining technique lets the users to set min-support and min-confidence. This algorithm reduces the factor delta support (Δs)

which is introduced by the user, until the minimum support is reached or the required numbers of rules are generated from the dataset. Information on running A-priori algorithm on the database is presented in the table below (Table 5.1.).

Method	Meaning
-N(required number of rules)	20, 10
-T (metric type to rank rules)	Confidence
-C (the minimum confidence score of the rule)	0.8, 0.9
-D(delta at which the minimum support is decreased at each iteration)	0.05
-U(upper bound for minimum support)	1.0
-M(the lower bound for the minimum support)	0.1
-S(significance of a rule at a given level)	-1.0
-Relation	Corruption
-Instances	4189

Table 5.1 List of parameters to run the association rule

For association rule algorithm four different models have been developed in this study and they are presented with their brief analysis as follows.

Experiment #1

In experiment 1, 11 attributes have been used to generate 10 best association rules (figure 5.3.).

As the model indicates, there are fourteen one-item frequent item-set, thirty-seven two-item frequent item-set, forty-four three-item frequent item-set, twenty-two four-item

frequent item-set and two five-item frequent item-set generated that satisfy 29% minimum support and 90% minimum confidence. Only ten numbers of rules are displayed as the parameter “number of rules” set to 10.

```
Apriori
=======

Minimum support: 0.29 (1214 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 9

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 37

Size of set of large itemsets L(3): 44

Size of set of large itemsets L(4): 22

Size of set of large itemsets L(5): 2

Best rules found:

 1. sex=male marital=married crimelevel=very_heavy 1305 ==> benefit=extream 1218    conf:(0.93)
 2. sex=male crimelevel=very_heavy 1489 ==> benefit=extream 1383    conf:(0.93)
 3. marital=married crimelevel=very_heavy 1602 ==> benefit=extream 1487    conf:(0.93)
 4. sex=male decision=guilty crimelevel=very_heavy 1328 ==> benefit=extream 1229    conf:(0.93)
 5. marital=married decision=guilty crimelevel=very_heavy 1427 ==> benefit=extream 1320    conf:(0.93)
 6. crimelevel=very_heavy 1845 ==> benefit=extream 1701    conf:(0.92)
 7. children=ch1 benefit=extream decision=guilty 1371 ==> marital=married 1260    conf:(0.92)
 8. decision=guilty crimelevel=very_heavy 1642 ==> benefit=extream 1509    conf:(0.92)
 9. children=ch1 benefit=extream 1580 ==> marital=married 1436    conf:(0.91)
10. salary=very_low crimelevel=very_heavy 1363 ==> benefit=extream 1237    conf:(0.91)
```

Figure 5.3 Association model generated in the first Experiment using all attributes

A-priori model produces a number of rules that satisfy the above minimum metrics of support and confidence set by the user. If a rule has a metrics value above the minimum thresh hold, then the rule is included in the large item-set.

There are two basic characteristics that have to be satisfied by large item-set. First the large item set property which states that, any subset of a large item set must be large. Second the contra-positive which says if an item set is not large, none of the superset is

large. Four of the ten rules generated by the first A-priori model are presented below (see fig. 5.3 for complete list of 10 rules generated).

Best rules found:

Rule #1 Sex=male marital=married crimelevel=very_heavy 1305 ==> benefit=extreme 1218 conf: (0.93)

If (Sex=male and marital=married and crimelevel=very_heavy then benefit=extreme). The knowledge is represented by the "if then" relation is to mean that first the database contains corruption offence records about corrupters; there exist relationships among the attributes sex, marital, crime level and benefit and their relation look like, if the corrupter is male, married and committed very heavy corruption crime then he has gotten extreme benefit (greater or equal to 100,000 Ethiopian Birr) from corruption with **confidence of 1218/1305 (93%) and 1218/4189 (29%) support**.

Rule #2 Sex=male crimelevel=very_heavy 1489 ==> benefit=extreme 1383 con: (0.93)

If (Sex=male and crimelevel=very_heavy then benefit=extreme). Rule number two is interpreted as if the offender is male and committed very heavy corruption crime, the benefit he has gotten is extreme benefit with **1383/1489 (93%) confidence and 1383/4189(33%) support**.

Rule #3 Marital=married crimelevel=very_heavy 1602 ==> benefit=extreme 1487 conf: (0.93)

If (marital=married and crimelevel=very_heavy then benefit= extreme).The third rule indicates that there is a relationship in the corruption database among the attributes marital, crime level and benefit. The rule is interpreted as, if the offender has marital status married and committed very heavy corruption crime then the offender has gotten extreme benefit from the crime committed with 1**487/1602 (93%) confidence and 1487/4189(35%) support**.

Rule #4Sex=male decision=guilty crimelevel=very_heavy 1328 ==> benefit=extreme 1229 con: (0.93)

If (Sex=male and decision=guilty and crimelevel=very_heavy then benefit=extreme). If the offender is male, in the wrong side of corruption suggested by the court and committed very heavy corruption then he has gotten the benefit of greater or equal to 100,000 Ethiopian Birr with **1229/1328 (93%) confidence and 1229/4189 (30%) support**.

The first experiment generated association rules specifically relayed on the two attributes (sex and marital). The reason is for their frequent values sex=male and marital=married. Here the researcher wanted to have more rules by discarding these two attributes.

Experiment #2

In experiment 2, 9 attributes have been used to generate 15 best association rules (figure 5.4.)

```
Apriori
=======

Minimum support: 0.17 (704 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 19

Size of set of large itemsets L(2): 58

Size of set of large itemsets L(3): 56

Size of set of large itemsets L(4): 18

Size of set of large itemsets L(5): 3

Best rules found:

 1. educ=secondary crimelevel=very_heavy 784 ==> benefit=extream 746    conf:(0.95)
 2. children=ch1 employment=gov_employee crimelevel=very_heavy 787 ==> benefit=extream 729    conf:(0.93)
 3. employment=gov_employee decision=guilty crimelevel=very_heavy 1009 ==> benefit=extream 933    conf:(0.92)
 4. employment=gov_employee crimelevel=very_heavy 1117 ==> benefit=extream 1030    conf:(0.92)
 5. crimelevel=very_heavy 1845 ==> benefit=extream 1701    conf:(0.92)
 6. employment=gov_employee salary=very_low decision=guilty crimelevel=very_heavy 793 ==> benefit=extream 731    conf:(0.92)
 7. children=ch1 crimelevel=very_heavy 1231 ==> benefit=extream 1134    conf:(0.92)
 8. employment=gov_employee salary=very_low crimelevel=very_heavy 869 ==> benefit=extream 800    conf:(0.92)
 9. decision=guilty crimelevel=very_heavy 1642 ==> benefit=extream 1509    conf:(0.92)
10. children=ch1 decision=guilty crimelevel=very_heavy 1072 ==> benefit=extream 983    conf:(0.92)
11. employment=gov_employee salary=very_low benefit=extream crimelevel=very_heavy 800 ==> decision=guilty 731    conf:(0.91)
12. employment=gov_employee salary=very_low crimelevel=very_heavy 869 ==> decision=guilty 793    conf:(0.91)
13. educ=secondary crimelevel=very_heavy 784 ==> decision=guilty 715    conf:(0.91)
14. salary=very_low crimelevel=very_heavy 1363 ==> benefit=extream 1237    conf:(0.91)
15. employment=gov_employee benefit=extream crimelevel=very_heavy 1030 ==> decision=guilty 933    conf:(0.91)
```

Figure 5.4 Association rule model generated in the second experiment

Second experiment has 9 attributes (“educ”, “age”, “children”, “benefit”, “employment”, “salary”, “decision”, “crime area” and “crime level”) of the corrupters. As the second association model above shows, there are 19 one-item frequent item-set, 58 two-item frequent item-set, 56 three-item frequent item-set, 18 four-item frequent item-set, 3 five-item frequent item-set and 1 six-item frequent item-set generated with 17% minimum support and 90% minimum confidence. Out of which only 15 rules are displayed. The generated 15 best rules are in ranges of **17%-40% support and 91%-95% confidence**. From 15 rues displayed 5 rules are analyzed by randomly selection (rule #6, #10, #14 and #15).

Rule #6 Employment=gov_employee salary=very_low decision=guilty crimelevel=very_heavy 793 ==> benefit=extreme 731 conf: (0.92)

If (Employment=gov_employee and salary=very_low and decision=guilty and crimelevel=very_heavy then benefit=extreme). This rule is the set of large item-set having five items and it is interpreted as, if the criminal is recruited by government, has very low salary (less than 3000 Ethiopian Birr), in wrong side of corruption as examined by court, committed very heavy corruption then s/he has got extreme benefit (greater or equal to 100,000 Ethiopian birr) from crime with **731/4189 (17%) support and 731/793(92%) confidence**.

Rule #10 Children=ch1 decision=guilty crimelevel=very_heavy 1072 ==> benefit=extreme 983 conf: (0.92)

If (Children=ch1 and decision=guilty and crimelevel=very_heavy then benefit=extreme). If the offender has children up to 5, in the wrong side of corruption and committed very heavy corruption then s/he has extreme advantage from the crime with **983/4189 (23%) support and 92% confidence**.

Rule #14 Salary=very_low crimelevel=very_heavy 1363 ==> benefit=extreme 1237 conf:(0.91)

If (Salary=very_low and crimelevel=very_heavy then benefit extreme). This rule indicates that there is a relationship between the attributes salary, crime level and benefit and the rule has interpreted as, if the offender's salary is very low (less than 3000 Ethiopian birr) and committed very heavy crime then s/he has extreme (greater or equal to 100,000 Ethiopian birr) advantage with the **support of 1237/4189 (30%) and 91% confidence**.

Rule #15 Employment=gov_employee benefit=extreme crimelevel=very_heavy 1030 ==> decision=guilty 933 conf:(0.91)

If (Employment=gov_employee benefit=extreme crimelevel=very_heavy then decision=guilty). This rule has four number of items and **support of 933/4189 (22%) and 933/1030 (91%) confidence.** The rule shows the relation to employment of the offender, benefit from the crime gotten, crime level and decision guilty that has been inspected and given by the court. Thus offenders reported to the FEACCE who is government employee, got greater or equal to 100,000 Ethiopian birr (extreme benefit), committed very heavy corruption then s/he is in the wrong side of corruption crime (guilty).

Experiment #3

In experiment 3, 8 attributes have been used to generate 15 best association rules (figure 5.5.).

```
Apriori
=======

Minimum support: 0.1 (419 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 21

Size of set of large itemsets L(2): 84

Size of set of large itemsets L(3): 92

Size of set of large itemsets L(4): 28

Size of set of large itemsets L(5): 1

Best rules found:

 1. age=30-40 crimelevel=very_heavy 541 ==> benefit=extream 518    conf:(0.96)
 2. children=ch1 crimearea=land crimelevel=very_heavy 458 ==> benefit=extream 438    conf:(0.96)
 3. educ=secondary crimelevel=very_heavy 784 ==> benefit=extream 746    conf:(0.95)
 4. educ=secondary children=ch1 crimelevel=very_heavy 496 ==> benefit=extream 470    conf:(0.95)
 5. crimearea=land crimelevel=very_heavy 725 ==> benefit=extream 686    conf:(0.95)
 6. educ=secondary salary=very_low crimelevel=very_heavy 588 ==> benefit=extream 554    conf:(0.94)
 7. salary=very_low crimearea=land crimelevel=very_heavy 599 ==> benefit=extream 563    conf:(0.94)
 8. educ=degree 456 ==> employment=gov_employee 425    conf:(0.93)
 9. children=ch1 employment=gov_employee crimelevel=very_heavy 787 ==> benefit=extream 729    conf:(0.93)
10. children=ch1 employment=gov_employee salary=very_low crimelevel=very_heavy 601 ==> benefit=extream 555    conf:(0.92)
11. employment=gov_employee crimelevel=very_heavy 1117 ==> benefit=extream 1030    conf:(0.92)
12. crimelevel=very_heavy 1845 ==> benefit=extream 1701    conf:(0.92)
13. children=ch1 crimelevel=very_heavy 1231 ==> benefit=extream 1134    conf:(0.92)
14. employment=gov_employee salary=very_low crimelevel=very_heavy 869 ==> benefit=extream 800    conf:(0.92)
15. age=40-50 crimelevel=very_heavy 582 ==> benefit=extream 529    conf:(0.91)
```

Figure 5.5 Association rule model generated in the third experiment

The third experiment has generated by using 8 attributes (age, educ, children, employment, salary, benefit, crime area and crime level) by ignoring 3 attributes (sex, marital and benefit). From 15 displayed rules with the minimum support of 10% and the minimum confidence of 90% here is the analysis of three rules (rule #1, #4 and #6).

Rule #1 Age=30-40 crimelevel=very_heavy 541 ==> benefit=extreme 518 conf:(0.96)

If (Age=30-40 and crimelevel=very_heavy then benefit=extreme). The relation between the attributes age and crime level in this rule is interpreted as, if the offender is between ages of 30 to 40 and committed very heavy corruption crime then s/he has accepted greater or equal to 100,000 Ethiopian birr with the confidence level of 96 %.

Rule #4 Educ=secondary children=ch1 crimelevel=very_heavy 496 ==> benefit=extreme 470 conf: (0.95)

If (Educ=secondary and children=ch1and crimelevel=very_heavy then benefit=extreme). The rule shows the relationships between the attributes education, children, crime level and benefit and it is interpreted as, if the offender has education level secondary, has children 1 up to 5 and very high corruption committed then benefit taken is extreme with 95% confidence level.

Rule #6 Educ=secondary salary=very_low crimelevel=very_heavy 588 ==> benefit=extreme 554 conf: (0.94)

If (education=secondary and salary=very_low and crimelevel=very heavy then benefit=extreme). If the offenders education level is secondary, have very low salary and commit very heavy corruption crime then the benefit gained is extreme (greater or equal to 100,000).

Experiment #4

The fourth association model shows fifteen rules by using 8 number of attributes (age, sex, children, salary, marital, employment, decision and crime level) and three of them were selected to be described below (figure 5.6.).

```
Apriori
=======

Minimum support: 0.19 (788 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 33

Generated sets of large itemsets:

Size of set of large itemsets L(1): 18

Size of set of large itemsets L(2): 64

Size of set of large itemsets L(3): 85

Size of set of large itemsets L(4): 50

Size of set of large itemsets L(5): 9

Size of set of large itemsets L(6): 1

Best rules found:

 1. sex=male children=ch1 decision=guilty crimelevel=very_heavy 838 ==> marital=married 791    conf:(0.94)
 2. children=ch2 907 ==> marital=married 849    conf:(0.94)
 3. children=ch1 salary=very_low crimelevel=very_heavy 885 ==> marital=married 827    conf:(0.93)
 4. sex=male children=ch1 crimelevel=very_heavy 969 ==> marital=married 904    conf:(0.93)
 5. age=40-50 sex=male 927 ==> marital=married 861    conf:(0.93)
 6. children=ch1 decision=guilty crimelevel=very_heavy 1072 ==> marital=married 995    conf:(0.93)
 7. age=40-50 children=ch1 893 ==> marital=married 822    conf:(0.92)
 8. children=ch1 crimelevel=very_heavy 1231 ==> marital=married 1132    conf:(0.92)
 9. employment=gov_employee salary=very_low crimelevel=very_heavy 869 ==> decision=guilty 793    conf:(0.91)
10. sex=male children=ch1 employment=gov_employee salary=very_low decision=guilty 1035 ==> marital=married 944    conf:(0.91)
11. age=40-50 decision=guilty 1060 ==> marital=married 965    conf:(0.91)
12. age=40-50 1218 ==> marital=married 1106    conf:(0.91)
13. sex=male employment=gov_employee crimelevel=very_heavy 902 ==> decision=guilty 819    conf:(0.91)
14. marital=married employment=gov_employee crimelevel=very_heavy 963 ==> decision=guilty 874    conf:(0.91)
15. age=40-50 salary=very_low 886 ==> marital=married 804    conf:(0.91)
```

Figure 5.6 Association rule model generated in the fourth experiment

Rule #1 Sex=male children=ch1 decision=guilty crimelevel=very_heavy 838 ==> marital=married 791 conf :(0.94)

If (Sex=male and children=ch1 and decision=guilty and crimelevel=very_heavy then marital=married). The rule is interpreted as: if the offender's sex is male, have children 1 up to 5, guilty of corruption, committed very heavy corruption then he is married.

Rule #2 Children=ch2 907 ==> marital=married 849 conf: (0.94)

If (Children=ch2 then marital=married). If the offender has children 5 up to 10 then s/he has married.

Rule #3 Children=ch1 salary=very_low crimelevel=very_heavy 885 ==> marital=married 827 conf: (0.93)

If (Children=ch1 and salary=very_low and crimelevel=very_heavy then marital=married). If the criminal has children 1 up to 5, paid very low salary (up to 3000 Ethiopian birr), consigned very heavy corruption then s/he is married.

Having the association models developed above the following clustering model has been developed to naturally (without having the class label of the dataset) group the corruption dataset.

5.1.2 Experiments and Analysis of Clustering Model

Clustering the data set is unsupervised learning. The clustering models do not need the class label (attribute) of the dataset. The whole dataset are used to be trained to place the new data sets in to the appropriate clusters based on the similarity measure, unlike the classification model (supervised learning) in which some of the dataset are used to be trained with the class label to classify the remaining datasets. Thus in this study the whole records of the data set (4189) were used to be trained and the class label (crimelevel) has been ignored. To develop the clustering model WEKA simple k-means algorithm has been used with the fine-tunings of the following different parameters:

- Display Standard Deviations -- Display standard deviations of numeric attributes and counts of nominal attributes. There are two options (true or false) true for display standard deviation and false to do not display.
- Don’t Replace Missing Values -- Replace missing values globally with mean for numeric attributes/mode for nominal attributes and there are true or false options.
- Max Iterations -- set maximum number of iterations (cycles) performed on the dataset while building the required clustering model and this parameter is fitting for numerical values only and have the default value 500 .
- Number of Clusters -- set number of clusters (groups) needed from the dataset and the default value is 2.
- Seed -- The random number seed to be used and 10 is the default value.

Data base Attributes can be selected in both forms of automatic selection using the data mining tool and by the domain experts working in the area. Thus for this research the researcher lets two domain experts to select the most needed attributes. 8 attributes (age, educ, children, employment, salary, benefit and crime area) were selected (Table 5.2) and the class label was ignored to build the clustering model as presented below.

Experiment #1

The first clustering model shows the run information of the selected attributes with k value 3 and the default value of parameter "seed".

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R
Relation:     corruption-weka.filters.unsupervised.attribute.Remove-R2,5,9
Instances:    4189
Attributes:   8
              age
              educ
              children
              employment
              salary
              benefit
              crimearea
Ignored:
              crimelevel
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 11616.0
Missing values globally replaced with mean/mode
```

Figure 5.7 Run Information of first clustering model with Value K=3 and Seed=10

Cluster index	Clustered Instances	Age of the offender	Education level	Childr en	Employment	Salary	Benefit	Crime area
1	3103	30-40	Secondary	Ch1	Gov_employee	Very_low	Extreme	Others
2	795	40-50	Secondary	Ch2	Self_employee	Very_low	Extreme	Land
3	291	30-40	Elementary	Ch1	Gov_employee	Very_low	Medium	Others

Table 5.2 The Summarized result of the first clustering model

As the above table shows (Table 5.2.) the selected attributes are used as independent and the entire data sets were used as the training data. The clustering model has assigned cluster indexes for each instance of the data set and descriptive approach has also been utilized for the identification of the characteristics of each clusters.

Out of the three clusters the first cluster (cluster #1) has the highest number of records that was 74% of the data set. Cluster #2 and cluster #3 have the smallest number of records 19% and 7%. In the three selected number of clusters cluster #1 and cluster #3 has the highest number of attributes shared (six attributes). There was the highest percentage assigned for age group 30-40, more has children 1 up to 5 and more of government employs with very low salary. Other types of crime area include (embezzlement, bribery, delay of matters, abuse of power, acceptance of undue advantages and maladministration).

Cluster #1 and cluster #2 are similar in three attribute values; they were education level secondary, very low salary and extreme benefit. Cluster #2 and #3 share only one attribute value, namely, very_low salary. And in the second experiment the researcher

tried to merge cluster #1 and #3 by setting the parameter seed value to 100 because they share the highest number of attribute values.

Experiment #2

The second clustering model shows run information of the clustering model by using k values set to 3 and the parameter "seed" value 100.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R
Relation:     corruption-weka.filters.unsupervised.attribute.Remove-R2,5,9
Instances:    4189
Attributes:   8
              age
              educ
              children
              employment
              salary
              benefit
              crimearea
Ignored:
              crimelevel
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 11215.0
Missing values globally replaced with mean/mode
```

Figure 5.8 Run Information of Second clustering model with Value K=3 and Seed=100

Cluster index	Clustered Instances	Age of the offender	Education level	Childr en	Employment	Salary	Benefit	Crime area
1	1963	30-40	Diploma	Ch1	Gov_employee	Very_low	Extreme	Land
2	1519	20-30	Secondary	Ch1	Gov_employee	Very_low	Extreme	Others
3	707	40-50	Secondary	Ch1	Self_employee	Very_low	Little	Land

Table 5.3 Summarized result of the second clustering experiment

Here in the second model cluster #1 and #3 share only three attribute values (Table 5.3) unlike with the first experiment that was six attribute values. For more understanding and to see the whole values of attributes in each cluster the research adjusted the parameter of "display standard deviation" altered to true (Table 5.4.).

Cluster index	Frequency of records	Age of the offender	Education level	Children	Employment	Salary	Benefit	Crime area
1	1963	6%20-30 54%30-40 21%40-50 11%50-60 5%60-70 0%70-80	10%elementary 18%secondary 21%certificate 34%deploma 13%degree 1%master and above	7%Ch0 73%ch1 18%ch2 0%ch3	79%gov 0%jobless 14%self 5%private	3%high 17%low 1%midium 77%verylow	19%little 1%high 5%midim 1%veryhigh 71%extream	0%asset 1%bid 3%cash 7%justice 36%land 0%licence 1%loan 24%others 0%payment 8%purchase 13%tax

2	1519	40%20-30 11%30-40 28%40-50 13%50-60 4%60-70 0%70-80	7%elementary 57%secondary 18%certificate 3%deploma 10%degree 1%masters and above	11%ch0 65%ch1 21%ch2 1%ch3	72%gov 1%jobless 18%self 6%private	7%high 12%low 3%midium 76%veryl ow	33%little 6%high 12%midium 1%veryhigh 46%extream	0%asset 1%bid 1%cash 5%justice 5%land 0%licence 0%loan 74%others 0%payment 3%purchase 5%tax
3	707	6%20-30 21%30-40 50%40-50 14%50-60 6%60-70 0%70-80	9%elementary 71%secondary 10%certificate 3%deploma 4%degree 0%masters and above	7%ch0 62%ch1 29%ch2 1%ch3	10%gov 0%jobless 84%self 4%private	5%high 13%low 8%midiu m 72%veryl ow	56%little 2%high 9%medium 0%veryhigh 30%extream	1%asset 0%bid 0%cash 7%justice 42%land 11%license 0%loan 14%others 0%payment 11%purchas e 8%tax

Table 5.4 Detailed result of the second clustering experiment

In this cluster it is possible to group the data set in to their perspective groups by the level of similarity they have on corruption crime characteristics. The following table presents the features of the clusters based on the above experiment (Table 5.5).

Cluster index	Cluster description	Remark
1	This cluster has the highest number of records in the data set (1963). From this age group 30-40 consists of the highest percentage (54%), age 40-50 are 21%, 50-60 are 11 % and age group 20-30 and 60-70 are accounting 6% and 5 % of the data set. The education level of this cluster is dominated by diploma having 34% preceded by certificate which is 21%, secondary, degree and elementary have almost the same instances which are 18%, 13% and 10% and masters and above accounts only 1% from the data set. The next column is number of children the offenders have and shows that most of the offenders have children 1 up to 5 covering 73%, criminals having5-10 and no children accounts 18% and 7 % of the data set and no offenders have greater or equal to 10 children. Almost all of the cluster instances are government employees having 79% of the dataset, self-employed and private workers account 14% and 5 % of the dataset and no jobless are presented. The salary of offenders in this cluster is dominated by Very low accounting for 77% of instances, low and high accounting for 17% and 3% and 1% is for medium salary. Most of the offenders in this clusters got extreme benefit from the corruption they committed having 71%, little benefit19%, medium benefit accounts 5%, high and very high benefit covers 1% of the instances. For the attribute crime area this cluster is characterized by 36% crime committed on land this is the highest proportion from the other areas, 24% for others (crimes committed like embezzlement, bribes and maladministration), tax, purchase and justice are the next crime types committed by offenders in this clusters.	Good
	This cluster contains 1519 of the data set. Age group 20-30 covers 40% of the data set, 40-50 covers 28%, 30-40 and 60-70 accounts 11% and 4% of	

2	the instances. The education level of this cluster is dominated by secondary level accounting 57% preceded by certificate, degree and elementary which cover 18, 10 and 7%. While masters and above accounts 1% of the data set. In this cluster, 65% offenders have 1 up to 5 children, 21% have 5 up to 10 children, 11% of the offenders haven't children and 1% of the data set have children greater or equal to 10. Government employees dominate employment of the offenders by 72% preceded by self-employed with 18%. The salary of the cluster is characterized by very low (less than 3000 Ethiopian Birr) and low salary (between 3000 and 6000 Ethiopian Birr) accounting 12% of the data set. Offenders in this data set have gotten extreme benefit from the crime accounting 46% of the instances and 33% of them have little benefit. The crime area other dominated others crime area proceeded by justice, land and tax with 5% for each.	Good
3	The third cluster contains 707 instances of the data set with offenders having age 40-50 accounting for 50%, 30-40 for 21%, 50-60 for 14% and 71% of them have secondary education level preceded by certificate and elementary. Offenders having 1 up to 5 children constitute 62% and 5 up to 10 constitute 29%. Most of the offenders here are self-employed constituting 84% of the instances, preceded by government and private employees with very low salary and little benefit of corruption. 42% of the crime was committed on land.	Good

Table 5.5 Description of the three clusters

Experiment #3

The third clustering model shows run information by using k value 3 and seed 200.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance
Relation:     corruption-weka.filters.unsupervised.attribute.Remove-R2,5,9
Instances:    4189
Attributes:   8
              age
              educ
              children
              employment
              salary
              benefit
              crimearea
              crimelevel
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 11929.0
Missing values globally replaced with mean/mode
```

Figure 5.9 Run Information of Second clustering model with Value K=3 and Seed=200

The features of clusters in the third experiment are presented in the following table (table 5.6).

Cluster index	Clustered Instances	Age of the offender	Education level	Children	Employment	Salary	Benefit	Crime area
1	1487	20-30	Secondary	Ch1	Self-employed	Very_low	Extreme	Land
2	1371	40-50	Diploma	Ch1	Gov_employed	Very_low	Extreme	Others
3	1333	30-40	Secondary	Ch1	Gov_employed	Very_low	Little	Others

Table 5.6 The summarized result of the third clustering experiment

As the table above (Table 5.6.) shows all cluster instances have approximately the same instance of records. The first cluster has 35% of the records; the second cluster contains 33% instances and the third cluster has 32 % objects of the data set. In the first cluster age of the offenders is from age 20 up to 30, in the second cluster age group is from age 40 up to 50 and in the third cluster it is from 30 up to 40.

In the first cluster the education level of the criminals is secondary level (from grade 9 up to grade 10), in the first cluster secondary is education level, in the second cluster diploma is the education level and in the third cluster the level is secondary. For the case of attribute children (number of children the corrupter has) is 1 up to 5.

In the first cluster corrupters are self-employed whereas clusters #2 and #3 have the same attribute value that was government employee. For the attribute value salary of the offender all the whole clusters have the same attribute value very low (less than 3000 Ethiopian birr). For the attribute benefit cluster #1 and #2 have the same value, i.e., extreme benefit received from the corruption they have committed. For the attribute value crime area cluster #2 and #3 have the same attribute value others that includes (embezzlement, bribery, delay of matters, abuse of power, acceptance of undue advantages and maladministration) whereas cluster #1 has the attribute value land.

5.1.2.1 Choosing the Best Clustering Model

Three clusters were developed to come up with the appropriate clustering model and from the three a cluster model that satisfies the criteria of best cluster was selected. Criteria to select appropriate clustering model are high intra-class similarity, low inter-class similarity, minimum number of iteration and minimum sum of squared error.

In the first experiment, cluster #1 and #3 share the greatest number of attribute values and the sum of the squared error was 11616. When the model is compared to the second experiment it is not good to select because there is another model that said to be good. Because of the above mentioned reasons the researcher tried to create the difference

between the two clusters (cluster #1 and #3) by altering the value of the parameter “seed” to 100 and clusters has more different patterns than before.

The second experiment has k value 3 and seed 100 and the problem was solved. Thus the three clusters almost have different patterns than before and the sum of squared error has reduced to 11215 from 11616. Keeping this the researcher has tried to get another model because the attribute value of employment in the second experiment has the same value in the three clusters.

The third clustering model was developed by altering the parameter “seed” to 200. In this experiment there has another problem occurred, in addition to having the same attribute value for “government employee”, attribute “children” has the same value (ch1) for the three clusters. Because of these reason the third model doesn’t chosen as the best clustering model.

Finally the researcher has selected the clustering model of the second experiment. Relaying on good clustering selection methods such as, the attribute values in different clusters are different as compared to the others (intra-class similarity), registered high similarity in one cluster (inter-class similarity) and registered smaller squared error. Thus it groups the data set of corruption crime and helps the crime investigators of corruption in the FEACCE.

The output of clustering model has been feed to the classification model to help the description of the clustering model to be predicted by the classification model.

5.1.3 Model Building and Analysis of Classification Model

The output of clustering model has been used as an input for classification model generation. The algorithm used was decision tree J48 to classify the data set. Unlike to the clustering algorithm that uses unsupervised learning to cluster the data instances, classification algorithm uses supervised learning method. This means it needs class label of the data set to classify the new incoming data instances.

It is easy to read a set of rules directly off a decision tree. One rule is generated for each node. The antecedent of the rule includes a condition for every node, one path from the root to that leaf, and the consequent of the rule is the class assigned by the leaf.

Experiment #1

Percentage split test mode has used by setting the split value to 45% for train and the remaining 55% for test. Default values of the parameter J48 has used such as "number of objects" with value 2 and "seed" 1. The tree generated has 306 numbers of leaves and 356 tree sizes, the time taken to build the model was 0.05 seconds and the accuracy of the classification model was 97.5%.

If (antecedent) then (consequent) relationship was used to analyze the generated classification rule. Antecedent part of the rule can be any attribute values that has been used as a precondition to label the data sets in to three different clusters (cluster #1 #2 and #3). In this case label or the class is consequent that is found at the end of the leaf.

The portion of the J48 pruned tree model by using the default value of the parameter "number of objects" equal to 2 is presented as follows (figure 5.10): (see Appendix B for more information).

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     corruption-weka.filters.unsupervised.attribute.Remove-R2,5,9_clustered-weka.
Instances:    4189
Attributes:   9
              age
              educ
              children
              employment
              salary
              benefit
              crimearea
              crimelevel
              Cluster
Test mode:    split 45.0% train, remainder test

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

J48 pruned tree
------------------

employment = gov_employee
|   benefit = little
|   |   educ = elementary
|   |   |   crimearea = asset_reg: cluster2 (0.0)
|   |   |   crimearea = bid: cluster2 (0.0)
|   |   |   crimearea = ?: cluster2 (0.0)
|   |   |   crimearea = cash_transfer: cluster2 (0.0)
|   |   |   crimearea = justice: cluster1 (6.0)
|   |   |   crimearea = land: cluster2 (12.0)
|   |   |   crimearea = license: cluster2 (0.0)
|   |   |   crimearea = loan: cluster2 (0.0)
|   |   |   crimearea = others: cluster2 (56.0)
|   |   |   crimearea = payment: cluster2 (0.0)
|   |   |   crimearea = purchase: cluster2 (0.0)
|   |   |   crimearea = tax: cluster2 (0.0)
|   |   educ = secondary
|   |   |   children = ch0: cluster0 (16.0)

|   |   |   children = ch2
|   |   |   |   crimearea = asset_reg: cluster0 (0.0)
|   |   |   |   crimearea = bid: cluster0 (0.0)
|   |   |   |   crimearea = ?: cluster0 (0.0)
|   |   |   |   crimearea = cash_transfer: cluster0 (0.0)
|   |   |   |   crimearea = justice: cluster0 (4.0)
|   |   |   |   crimearea = land
|   |   |   |   |   age = 20-30: cluster1 (4.0)
|   |   |   |   |   age = 30-40: cluster0 (6.0)
|   |   |   |   |   age = 40-50: cluster1 (6.0)
|   |   |   |   |   age = 50-60: cluster1 (0.0)
|   |   |   |   |   age = 60-70: cluster1 (0.0)
|   |   |   |   |   age = 70-80: cluster1 (0.0)
|   |   |   |   crimearea = license: cluster0 (0.0)
|   |   |   |   crimearea = loan: cluster0 (0.0)
|   |   |   |   crimearea = others: cluster0 (50.0)
|   |   |   |   crimearea = payment: cluster0 (0.0)
|   |   |   |   crimearea = purchase: cluster0 (0.0)
|   |   |   |   crimearea = tax: cluster0 (0.0)
|   |   |   children = ch3: cluster0 (0.0)
```

Figure 5.10 Part of decision tree generated in the first experiment

Cluster0 means very _heavy, cluster1 is heavy and cluster2 is simple. Some of the generated classification rules in the first experiment are presented as follows:

- If (employment=gov_employee and benefit=little and education elementary and crime area=asset _reg then cluster2 (0.0)).

This rule is interpreted as offenders employed by government, have got little benefit (less or equal to 50,000 Ethiopian birr) from the crime, have elementary education level and commit corruption on asset registration then classified as cluster2. No test data set satisfy this condition.

- If (employment=gov_employee and benefit=little and education elementary and crime area=bid then cluster2 (0.0)).

This rule indicates that offenders who are government employed, have little benefit from corruption committed, have elementary education level and with unknown crime area then classified in cluster2. In the test data set there are no instances with these characteristics.

- If (employment=gov_employee and benefit=little and education elementary and crime area=justice then cluster1 (6.0)).

The rule indicates that offenders who are employed by the government, have got little benefit, have elementary education level and commit corruption offence on justice then classified as cluster1. There are 6 instances in test data set and all are classified correctly.

- If (employment=gov_employee and benefit=little and education elementary and crime area= land then cluster 2 (12.0)).

This rule shows that offenders employed by the government, have got little benefit, have elementary education level and commit corruption on land then classified as cluster2. 12 instances are with these characteristics and all are correctly classified.

- If (employment=gov_employee and benefit=little and education secondary and children=ch0 then cluster0 (16.0)).

The rule is interpreted as offenders, who are employed by the government, have got little benefit from the crime committed, have secondary education level and have not children then classified as cluster0. All instances (16) are correctly classified.

- If (employment=gov_employee and benefit=little and education secondary and children=ch2 and Crimearea=asset_reg then cluster0 (0.0)).

Offenders who are government employees, have got little benefit from corruption offence, have secondary education level, have children 5 up to 10, and commit corruption on asset registration then classified as cluster 0. From the test data set no instances satisfy these preconditions.

The next table (Table 5.7) presents the output of confusion metrics generated from the first classification experiment. The table shows that the test data set was 2304, leaving 1885 dataset for the training. The reason for setting the split value to 45% was that the researcher wants to see the accuracy of the classification model to predict or classify the incoming data set based on the small number of training data set.

Cluster index	Predicted			Total	Score rate
	Cluster #1	Cluster #2	Cluster #3		
Cluster #1	1685	12	2	1699	97.5%
Cluster #2	8	428	0	436	97.3%
Cluster #3	35	0	134	169	98.5%
Total	1728	440	136	2304	97.5%

Table 5.7 Output from decision tree J48 using the default value "Number of objects" and "seed"

The confusion matrix in the table above shows that the first cluster was predicted as the first cluster 1685 times, predicted as the second cluster 8 times and predicted as third cluster 35 times. Second cluster was classified as the first cluster 12 times, predicted as the second cluster 428 times and it hasn't been predicted as the third cluster. Third cluster has been predicted as first cluster 2 times, hasn't been predicted as second cluster and predicted as its own cluster 134 times.

The score rate of the classification model to predict the firs cluster as the first cluster was 97.5%, the model scores the accuracy rate of predicting the second cluster as the second cluster was 97.3%, also labeling the third cluster as the third cluster has 98.5% and the

total accuracy rate of the first classification model gotten in the first experiment has registered 97.5%.

Experiment #2

In the second classification model the split value has remained the same as the first experiment but the parameter of J48 number of objects has increased to 10 (figure 5.11). And the tree has decreased to 119 numbers of leaves and 138 size of tree. The time taken to build the classification model was 0.03 seconds and the accuracy level of the model fall to 91.14% from 97.5%. This is because when number of objects is increased number of the tree size decreased so the classification tree loses the required numbers of antecedents (preconditions to be classified as correctly).

```
employment = gov_employee
|   benefit = little
|   |   educ = elementary: cluster2 (74.0/6.0)
|   |   educ = secondary: cluster0 (292.0/10.0)
|   |   educ = certificate: cluster0 (176.0/3.0)
|   |   educ = deploma: cluster0 (180.0/3.0)
|   |   educ = degree: cluster0 (153.0)
|   |   educ = masters_and_above: cluster0 (19.0)
|   benefit = high: cluster0 (113.0)
|   benefit = medium
|   |   educ = elementary: cluster2 (10.0)
|   |   educ = secondary: cluster0 (58.0/9.0)
|   |   educ = certificate: cluster2 (59.0)
|   |   educ = deploma: cluster2 (41.0/5.0)
|   |   educ = degree: cluster2 (15.0)
|   |   educ = masters_and_above: cluster2 (8.0)
|   benefit = very_high
|   |   age = 20-30: cluster2 (10.0)
|   |   age = 30-40: cluster0 (21.0/4.0)
|   |   age = 40-50: cluster0 (19.0/6.0)
|   |   age = 50-60: cluster0 (0.0)
|   |   age = 60-70: cluster0 (0.0)
|   |   age = 70-80: cluster0 (4.0)
|   benefit = extream
|   |   children = ch0: cluster0 (117.0)
|   |   children = ch1: cluster0 (1081.0)
|   |   children = ch2
```

Figure 5.11 Part of decision tree generated in the second experiment with "Number of Objects" = 10

Some of the rules generated in the J48 pruned tree model are analyzed as follows.

- *If (employment = gov_employee and benefit = little and education = elementary then cluster2 (74.0/6.0)).*

The rule indicates those government employees who have got little benefit from the crime committed and have elementary education level are classified as cluster #2. There are 74 objects of the data set having this property and from which 68 records are correctly classified and the remaining 6 records are miss classified.

- If (*employment = gov_employee and benefit = little and education = secondary then cluster0 (299.0/10.0)).*

The rule shows that those government employees who have got little benefit from the crime and have secondary education level are classified as cluster 0. There were 299 objects sharing this property and out of these, 289 objects were correctly classified and the rest 10 objects were incorrectly classified.

- If (*employment = gov_employee and benefit = little and education = certificate then cluster 0(176.0/3.0)).*

The rule was interpreted as those government employees who has got little benefit from the crime and has certificate education level are labeled as cluster 0 which was the first cluster. There were 176 instances has shared this property and from these objects 173 were correctly labeled and the remaining 3 were mislabeled.

- If (*employment = gov_employee and benefit = little and education = diploma then cluster0 (180.0/3.0)).*

This rule shows that government employees who have got little benefit from the crime and have diploma education level are labeled as cluster 0. There were 180 instances sharing this property and from these instances 177 were correctly classified and three of them were misclassified.

- *If* (*employment = gov_employee and benefit = little and education = degree then cluster0 (153.0)).*

If the offender is recruited by government, have got little benefit from the crime and have a degree then cluster 0. 153 instances satisfy this condition and all of them are classified correctly.

- *If* (*employment = gov_employee and benefit = little and education =masters and above then cluster0 (19.0)).*

The rule indicates that government employees who have got little benefit from the crime and have masters and above education level then in cluster 0. All instances that satisfy these conditions (19 instances) are classified correctly.

- *If* (*employment = gov_employee and benefit =high then cluster0 (113.0)).*

The rule interpreted as if the offender is employed by government and have got high benefit then cluster0. All instances that satisfy these conditions (113 instances) are classified correctly.

The following table presents the confusion matrix for the test dataset which was 55% of the data set displayed by the second classification mode1 (Table 5.8).

Cluster index	Predicted			Total	Score rate
	Cluster #1	Cluster #2	Cluster #3		
Cluster #1	1648	36	15	1699	91.8%
Cluster #2	81	355	0	436	89.6%
Cluster #3	67	5	97	169	86.8%
Total	1796	396	112	2304	91.14%

Table 5.8 Output from decision tree J48 using the value of the parameter "number of objects"= 10

The above table (Table 5.8) shows that first cluster is predicted as first cluster 1648 times, predicted as the second cluster 81 times and predicted as the third cluster 67 times.

The second cluster is predicted as the first cluster 36 times, predicted as the second cluster 355 times and predicted as the third cluster 5 times. The third cluster is predicted as the first cluster 15 times, not predicted as the second cluster and predicted as the third cluster 112 times.

The accuracy rate for predicting the first cluster as a first cluster is 91.8%, for the second cluster 89%, for the third cluster 86.8% and 91.14% is the total accuracy rate to classify all test data set.

Lastly based on the performance of the two classification models developed, the most performing model has been chosen to predict the class label of the new incoming data. Although the first experiment has much leaf, size of tree and greater time compared to the second model for developing the model, it has better accuracy in classifying the data set. Thus the first classification model developed with the default value of the parameter "number of objects" (2) is better than the model developed with "number of objects" altered to 20.

5.2 Evaluation

The generated models during the experiment have been assessed if they meet the aim of this study. The objectives of the study were to come up with the best models that can extract hidden knowledge from the data base of the FEACCE, group the profile of corrupters according to the similarities they have by assigning cluster index and predict the class of new incoming data set relaying on training datasets. To evaluate the developed models domain experts and performance of the models has been used.

Association Rule Model

Domain experts were participated to choose the best association rule model and they were interested by the whole developed models. From each developed models they choose two interesting rules (presented in Interpretation and discussion part) and all models are best for them.

The association rule models were also evaluated by their performance. Maximum support and maximum confidence was used as a measurement. When the rule support might be of interest during the association discovery process the rule confidence is what finally determines the rule validity.

In the first association rule mining model the maximum support was 36% and 93% maximum confidence, In the second model registered the maximum support of 40% and 95% confidence, the third model 24% maximum support and 96% confidence level was registered and in the fourth model 27% support and 94% confidence has been registered. From the evaluation of the measurements the researcher chooses the second association rule model as the best model.

Clustering Model

To evaluate the clustering models number of iterations and sum of squared errors has been utilized. In the first clustering model number of iteration 3 and sum of squared error 11616 has registered, in the second model number of iteration 3 and 11215 sum of squared error has registered and in the third model number of iterations 6 and 11929 squared error has been registered.

As the best clustering model is the one that registered minimum number of iteration and minimum number of squared error, the researcher selected the second clustering model as the best model.

Classification Model

In this study classification of corrupters under the three classes of the data base (very heavy, heavy and simple) is based up on the attribute values they have in the data set. For the susceptibility of the criminals their behavior and the environmental conditions can be an effect for the crime level they are committing. But the FEACCE classify offenders only based up on the two attribute values of “benefit” and “decision”.

Most of the time to evaluate the classification models accuracy level and time taken to build the model are used as the measurement. In this study the accuracy level of the model has been used as a measurement. In the first classification model the accuracy of 97.5% has registered and in the second model 91.14% accuracy has registered. Thus the researcher selects the first classification model as a best model.

5.3 Interpretation and discussion of predictive rule findings

This part will discuss the rules generated in predictive models (association rule mining and classification models). In descriptive model (clustering model) the whole parts were discussed in chapter five.

Association rule

In the experiments and analysis of association rule mining, different models were generated with different parameters and set of attributes. Numbers of different rules with satisfactory measures (confidence and support) and most importantly meeting the subjective judgment of the domain experts on their interestingness were evaluated. Summary of the association rule mining done on the data base of FEACCE with 4189 instances, the subsequent interpretations and discussions of the discovered interesting rules are presented as follows:

Experiment #1

In this experiment all attributes were used to generate the first association rule model.

Attributes: “age”, “sex”, “education”, “children”, “marital”, “employment”, “salary”, “benefit”, “decision”, “crime area” and “crime level” of the offenders.

Rule

Salary=very_low and crime level=very_heavy ==> benefit=extreme

The rule indicates that if the offenders salary is very low and have committed very heavy crime level then the benefit s/he have gotten is above 100,000 Ethiopian birr or extreme benefit.

Children=ch1and benefit=extreme and decision=guilty ==> marital married

The rule indicates the characteristics of the criminals. If offender has children one up to five and has extreme benefit and in the wrong side of corruption then s/he is married.

Domain experts were interested with these rules. Because, the antecedent parts of the rules are the characteristics for the reason why the offenders have got extreme benefit from the crime they committed.

Experiment #2

The second experiment has generated association rules using 9 attributes: namely, "educ", "age", "children", " benefit", "employment", "salary", "decision", "crime area" and "crime level") of the corrupters.

Rule

Educ=secondary and crime level=very heavy ==> benefit extreme

Whenever the corrupter has secondary education level s/he is committing very heavy corruption level that have extreme benefit.

Decision=guilty and crime level=very heavy ==> benefit=extreme

When the offender is guilty of corruption s/he has committed very heavy corruption that has extreme benefit.

As in the first experiment the domain experts were interested in these association rules and the rules represent interesting regularity in the data base.

Experiment #3

The third experiment has generated association rules using 8 attributes (“age”, “educ”, “children”, “employment”, “salary”, “benefit”, “crime area” and “crime level”) buy ignoring 3 attributes (“sex”, “marital” and “benefit”).

Rules

Age=30-40 and 40-50 and crime level=very heavy ==> benefit=extreme

The rule predicts that if the offenders are between the ages of 30 up to 50 and commit very heavy crime level then s/he have got extreme benefit.

Children=ch1 and crime area=land and crime level=very heavy ==> benefit=extreme

If the offender have children one up to five and commit crime on land and the crime is very heavy then s/he have got extreme benefit from the crime.

Experiment #4

In the last association rule model, 8 attributes were used (“age”, “sex”, “children”, “salary”, “marital”, “employment”, “decision” and “crime level”).

Sex=male and children=ch1 and employment=gov_employee and salary=very low and decision=guilty ==> marital=married

If the offender is male and has children one up to five and government employee and have very low salary and guilty of the crime then he is married.

Age=40-50 and salary=very_low ==>marital=married. If the offender is between age 40 up to 50 and have very low salary then s/he is married. The two rules are very interesting for experts.

Classification rule

In classification rule development there were two models developed. From the selected classification model (first classification model) some of rules that hasn't discussed in chapter five will be presented as follows:

- If (employment=gov_employee and benefit=little and education secondary and children=ch2 and crime area=asset_reg then cluster0 (0.0)).

If the offender is employed by government, have got little benefit, have secondary education level, have children 5 up to 10 and commit corruption on asset registration then predicted as cluster0. The test data set did not have instances with these characteristics.

- If (employment=gov_employee and benefit=little and education secondary and children=ch2 and crime area=land and age=20-30 then cluster1 (4.0)).

If the offender is employed by government, have got little benefit, have secondary education level, have children 5 up to 10, commit corruption on land and age from 20 up to 30 then classified as cluster1. Instances that satisfy these conditions are correctly classified.

- If (employment=gov_employee and benefit=little and education secondary and children=ch2 and crime area=land and age=30-40 then cluster0 (6.0)).

Offenders who are employed by the government, have got little benefit, have secondary level education, have children 5 up to10, commit corruption on land and have age of 30-40 then labeled as cluster0.

- If (employment=gov_employee and benefit=little and education secondary and children=ch3 then cluster0 (0.0)).

Criminals employed by the government, have got little benefit, have secondary level education, and have children greater than 10 then predicted as cluster 0.

5.4 Comparison of Rules Generated from both Association Rule and Classification Models

This part will present the similarity of the rules generated and the difference between classification and association rule mining models. Most of rules appeared in the association rule model also appeared in the list of classification rule model.

For example If (employment=government employee and benefit=extreme then cluster 0) has generated in the first classification rule model. If (benefit= extreme and crime level=very_heavy then employment= government employee) generated in the third association rule experiment (rule number 10). Crime level very_heavy (class label of the data set) was replaced by cluster0.

The difference between the two models (association and classification models) is there are no rules generated in the association rule models but appeared in the classification rule model. The reason is the attribute instances are few and these instances are failed to generate the association rule.

For example attribute “benefit” value little and medium, attribute “age” value from age 50-60, 60-70 and 70-80 and attribute “crime area” value license, tax, bid, payment and loan. But in the classification rule model all the testing records were classified in to the predefined label.

CHAPTER SIX

Conclusions and Recommendations

6.1 Conclusions

FEACCE collects corruption evidences given by the people to be investigated in front of the country law. The computer data base is not in a way that is suitable for the application of data mining. The commission lack awareness on application of Information Communication Technology (ICT) specifically data mining to combat corruption.

Data mining can be applicable in the area of corruption crime to analyze data that might be difficult to humans and create hidden, important and novel knowledge. Thus the extracted knowledge enables the commission to come up with the way its objectives comes to be true. So the FEACCE can utilize the advantages of data mining by firstly work on the computer data base.

To have more rules in association rule development it is better to change attributes to be mined in different association model generation. And it is helpful to alter different parameter values.

To come up with a good clustering model it is good to change the parameter seed to randomize instances. It is not good to change the parameter k value if the class label of the data is known. Moreover it is better to choose the appropriate attributes to be described.

The accuracy level of the classification model is better if instances used to develop the clustering model are an input for classification rule model development.

This research has used Association rule mining, Clustering and Classification data mining techniques. Nine models have been developed, four models under the association rule mining, three models under the clustering, and two models under the classification

model. For these three different techniques three different types of data mining algorithms have been used. For association rule mining, a-priori, for clustering simple k-means and for classification decision tree algorithms have been used.

Ethiopia is now the country that is substantially growing in the Industrial, Agricultural, Service sectors and others. Corruption stagnate development, create economic inflation and instability like the European countries face this problem these days, so the country really needs to work hardly on corruption to sustain development.

This research would introduce the FEACCE to the application of data mining in corrupt activity data to combat corruption.

6.2 Recommendations

Even though this research has been done to fulfill an academic requirement; the outputs can be used by the FEACCE to work on corruption by identifying the profile of corruption offenders. Identification of offenders could be very useful for preventive approach of corruption and this could protect the country's property before offenders commit corruption.

If (employment=gov_employee and benefit=little and education secondary and children=ch2 and crime area=land and age=30-40 then cluster0 (6.0)).

Offenders who are employed by the government, have got little benefit, have secondary level education, have children 5 up to10, commit corruption on land and have age of 30-40 then labeled as cluster0 (commit very heavy corruption).Such kinds of rules generated in this research help the commission to predict the future events and take appropriate actions on corruption. Thus focus on the characteristics of offenders should be given to create awareness.

To attain the above mentioned ideas the FEACCE should create the means to effectively record the reported crimes, alter the data base and familiar with data mining. Especially the data base design has to be changed to discover special and interesting features in the data mining process and this could lead the commission to its objectives. In the existed computer data base the attributes has not been as good as the manual data base, the entries were insufficient ,half of the attributes were not in the interest of the data mining activity that has been applied, some have the same value for the whole attributes and some of them has more than fifteen values.

The improvement of the contents of computer data base can enables better results. Further researches can test other mining techniques like artificial neural networks, time series and document summarization. To improve the performance of the model

continuous data can be used in place of categorical data, which is the application of text mining on the unstructured data of corruption records.

References

Adriaans, Pieter and Zantinge, Dolf. (1996). Data Mining. New York: Addison Wesley, pp.5-6.

African Development Bank Group Integrity and Anti Corruption Department. (2010). Integrity and Anti corruption: progress report.

African Economic Outlook. (2007). Ethiopia: A survey of Development of Ethiopia.

Agrawal, R. Imielinksi, T. and Swami, A. (1993). Mining association rules between sets of items in large database. The ACM SIGMOD Conference, pp. 207-216, USA: Washington DC.

Agrawal, R., Srikant, R., (1994). Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile.

Åke, G., Rebekah H., & David Sasaki. (2010). Increasing transparency and fighting corruption through ICT. ISBN:978-91-85991-02-0. Universitetsservice US-AB, Stockholm.

Amundsen, I. (1999). Political corruption: An introduction to the issues, Working Paper 99:7, Bergen: Chr. Michelsen Institute.

Anna L. Buczak and Christopher M. Gifford . (2010). Fuzzy Association Rule Mining for Community Crime Pattern Discovery. ISBN 978-1-4503-0223-4/10/07 $10.00. Washington, D.C., USA.

Apte, C., Weiss, S., and Grout, G. (1993). Predicting Defects in Disk Drive Manufacturing: A Case Study in High Dimensional Classification" IEEE Annual

Computer Science Conference on Artificial Intelligence in Application, Los Alamitos, pp. 212–218.

Axel, Dreher and Thomas, Herzfeld. (2005) . The Economic Costs of Corruption: A Survey and New Evidence.

Azevedo, Ana. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference data mining. ISBN: 978-972-8924-63-8.

Bamidele, Ayo. (1995). The Evolution of the Nigerian Local Government System. ISBN:58-587-065.

Berger, C. (2004). Oracle Data Mining, Know More, Do More, Spend Less.

Berry, J.A. and Linoff, Gordon S. (2000). Mastering Data Mining: The Art and Science of Customer Relationship Management, Michael, USA, Wiley Computer Publishing. Rhodes University.

Berry, MJ and Linoff, G. (1997). Data mining techniques: for marketing, sales and customer support. USA: Wi-ley.

Blundo, G. and Sardan, O. de. (2006). Everyday Corruption and the State Citizens & Public Officials. London: Zed Books.

Cahlink, George. (2000). Data Mining Taps the Trends, Government Executive Magazine.

Chinua, Achebe. (1988). The Trouble with Nigeria. Enugu, Fourth Dimension Publishers.

Colleen, McCue. (2006). Data Mining and Crime Analysis in the Richmond Police Department.

Cooley, R. Mobasher, B. and Srivastava, J. (1997). Grouping web page references into transactions for mining World Wide Web browsing patterns. University of Minnesota, Department of Computer Science, Minneapolis.

Corruption Prevention Mechanism. (2006). State of New south Wales Department of Education and Training.

Daniel, T. (2005). "Discovering knowledge in data: an introduction to data mining", Larose, p. cm, ISBN:0-471-66657-2.

Derosa, Mary. (2004). Data mining and Data Analysis for Counterterrorism. ISBN:0-89206-443 9. Washington, D.C.

Fayyad, U. M., G. P. Shapiro, P. Smyth. (1996). From Data Mining to Knowledge Discovery in Databases, ISSN: 0738-4602-1996, AI Magazine, pp 37–53.

Fayyad, U.M. and Uthurusamy, R. (2002). Evolving Data Mining into Solutions for Insights. ACM. pp. 28-31.

Fukuda, T. Morimoto, Y.Morishanti, S. and Tokuyama, T. (1996). Mining optimized association rules from numeric attributes: Symposium on principles of Data base systems. ACM Pp.182-191.

Gamberger, D. Šmuc, T. Marić, I. (2001). DMS - poslužitelj zaanalizu podataka, Institut Ruđer Boškovic.

Gbenga, Lawal. (2007). Corruption and Development in Africa: Challenges for Political and Economic Change. ISSN: 1818-4960. Humanity & Social Sciences Journal 2.

Haller, D. and Shore, C. (2005). Corruption Anthropological Perspectives. London: Pluto Press.

Hand, David. Mannila, Heikki. Padhraic, Smyth. (2001). Principles of Data Mining. ISBN: 026208290 MIT Press, Cambridge.

Hand, Jiawei and Kamber, Micheline. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

He, Zengyou. Xu, Xiaofei. Deng, Shengchun. (2003). Data Mining for Actionable Knowledge: A Survey. Department of Computer Science and Engineering, Harbin Institute of technology, Harbin P.R China.

Ian, H. Witten. Eibe, Frank. (2005). Data mining: practical machine learning tools and techniques. ISBN: 0-12-088407-0. Morgan Kaufmann publishers.

Ibid. (2004). Data Mining: Federal Efforts Cover a Wide Range of Uses. GAO-04-548 Washington.

Jain, A. (2001). The Political Economy of Corruption. London: Routledge.

Jean, Goerzen. (1996). Service Delivery Surveys: Applying the Sentinel Community Surveillance Methodology, Country Overviews.

Johnston, M. (2005). Civil Society and Corruption Mobilizing for Reform. Lanham, MD: University Press of America.

Kantardzic, Mehmed J. (2003). Data Mining: Concepts, Models, Methods, and Algorithms, ISBN: 0471228524, IEEE Computer society, Wiley-Interscience.

Kerr, S. (1995). On the folly of rewarding A, while hoping for B, Academy of Management Executive, 9(1): pp.7-14.

Khan, M. (1996). A typology of corrupt transactions in developing countries. IDS Bulletin, vol. 8, no. 5.

Kunicova, J. and Rose-Ackerman S. (2005). Electoral rules and constitutional structures as constraints on corruption: British Journal of Political Science, 35(4) 573-606.

Langseth, Peter. (1991). Prevention: An effective tool to reduce corruption. Global program against corruption: center for international crime prevention. Office of drug control and crime prevention. United Nations office at Vienna.

Lawal, O.O. and Tobi, A.A. (2001). Corruption.

Leul Woldu. (2003). Application of Data Mining In Crime Prevention: The Case of Oromia Police Commission. Addis Ababa University.

Liu, H., Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers.

Malathi. A. and Santhosh Baboo. (2011). An Enhanced Algorithm to Predict a Future Crime using Data Mining. International Journal of Computer Applications (0975 – 8887) Volume 21– No.1.

Mattison, R. (1997). Data Warehousing and Data Mining for Telecommunications. House Inc.

Mauro, P. (1997). Why Worry about Corruption? Washington D.C.: International Monetary Fund.

Mdzingwa, Nhamo. (2005). Data Mining with Oracle 10g using Clustering and Classification Algorithms: Computer science honors project Literature Review.

Mostow, J., Beck, J., Cen, H., Cuneo, A. (2005). An educational data mining tool to browse tutor Student interactions: Time will tell In Proceedings of the Workshop on Educational Data Mining, Pittsburgh, USA, pp. 15–22.

Nasereddin, Hebah H. O. (2009). Stream Data Mining: Department of computer Information System Faculty of Information Technology Amman Arab University for Graduate Studies.

Nnavozie, O.U. (1990). The Bureaucracy and National Department: The case of Nigeria.

Nye, J.S. (1967). Corruption and political development: American Political Science Review, vol. 61, no. 2, pp. 417-427.

Pareek, D. (2007). Business Intelligence for Telecommunications: Auerbach Publications, Taylor & Francis Group LLC.

Pete, Chapman Julian, Clinton Randy, Kerber (NCR). (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Peter, H. (2004). Disillusion and hope of smallholders in Cordillera Central, Dominican republic: Is procaryn the right way to change the trend of forest deterioration? In: Baumgartner, David M.; (ED).

Piatetsky-Shapiro, G. Frawley, W. J. (1991). Knowledge Discovery in Databases. AAAI/MIT Press.

Ravichandra, K. Rao. (2003). Data mining and clustering technique: Workshop on Semantic Web. Indian Statistical Institute, Bangalore.

Reza, Fadaei-Tehrani, Thomas, M. Green. (2002). Crime and society International Journal of Social Economics Volume 29 Number 10 pp. 781-795.

Riley, M. (1997). In Corruption and Development, p.141.

Rodney, Walter. (1972). How Europe underdeveloped Africa. London. Bogle L'ouverture.

Romero, C., & Ventura, S. (2006). Data mining in e-learning, Southampton, UK: Wit Press.

Rose-Ackerman, S. (1978). Corruption: A Study in Political Economy. London/New York: Academic Press.

Rose-Ackerman, S. (2006). International Handbook on the Economics of Corruption. Northampton, MA: Edward Elgar.

Rud, Parr O. (2001). Data Mining Cookbook: modeling data for marketing, risk, and customer relation-ship management. USA: Wiley.

Seidman, Claude. (2000). Data mining with Microsoft SQL Server Technical reference. ISBN:07356-1271-4.Amazon.com/mining-Microsoft-Server-Technical-Reference/dp/0735612714.

Seifert, Jeffery W. (2004). Data mining: an overview. Congress Research Report for congress.

Shleifer, Andrei and Vishny, Robert W. (1993). Corruption: Quarterly Journal of Economics, 108: 599-617.

Sivanandam, S. Sumathi, S. (2006). Introduction to Data Mining and its Applications. Springer Verlag Berlin Heidelberg.

Spector, B. (2005). Fighting Corruption in Developing Countries: Strategies and Analysis. Bloomfield, CT: Kumarian Press, Inc., p. 7

Srivastava, P. (2008). A New software that keeps politicians as honest as possible: Think Change India.

Terrorism Information Awareness Project. (2003).

Tesfaye Shamebo. (2007). Anti Corruption Efforts in Ethiopia. Federal Ethics & Anti Corruption Commission of Corruption Prevention & Research Department.

Tewodros Mezmur. (2011). The Ethiopian Federal Ethics and Anti-Corruption Commission: A Critical Assessment. Law, Democracy & Development. VOL 15.

Transparency International. (2006). Corruption perception index.

Tsantis, L. Castellani, J. (2001). Enhancing learning environments through solution-based knowledge discovery tools. Journal of Special Education Technology, 16 (4), 1-35.

Two Crows Corporation. (1996). Introduction to Data Mining and Knowledge Discovery, Third (Ed.) (Potomac, MD: Two Crows Corporation, 1999); Pieter, Adriaans and Dolf, Zantinge, Data Mining (New York: Addison Wesley).

Two Crows Corporation. (2005). Introduction to Data Mining and knowledge Discovery. ISBN: 1-892095-02-5.

UN. (2010). United Nations Global Compact. Principle 10.

Wang, K. Zhou, S. Han, J. (2002). Profit Mining: From Patterns to Actions. In: Proc. Of EDBT'02, pp.70-78.

Westphal, C. and Blaxton, T. (1998). Data Mining Solutions: Methods and Tools for Solving Real-World Problems, John Wiley, New York.

Wong, R.C-W. and Fu, A.W-C. (2004). ISM: Item selection for Marketing with Cross-Selling Considerations. In: Proc .of PAKDD'04, pp.431-440.

Xindong, Wu. Vipin, Kumar. Ross, Quinlan. Joydeep, Ghosh. (2007). Top 10 algorithms in Data Mining. Pp: 6-7.Survey paper.

Yijun, Lu. (1997). Concept hierarchy in Data mining: Specification, Generation and Implementation. Simon Fraser University.

Zaïane, Osmar R. (1999). Introduction to Data Mining: CMPUT690 Principles of Knowledge Discovery in Databases. University of Alberta. Department of Computing Science.

Appendix A: Attributes and their values in the data set

Age

- a, 20-30
- b, 30-40
- c, 40-50
- d, 50-60
- e, 60-70
- f, 70-80

Sex

- a, male
- b, female

Education

- a, elementary
- b, secondary
- c, certificate
- d, diploma
- e, degree
- f, masters and above

Children

- a, ch0
- b, ch1
- c, ch2
- d, ch3

Marital

- a, married
- b, never married
- c, divorced
- d, widowed

Employment

a, government employee

b, jobless

c, self employee

d, private employee

Salary

a, high

b, low

c, medium

d, very low

Benefit

a, little

b, medium

c, high

d, very high

e, extreme

Decision

a, discontinued

b, guilty

c, not guilty

Crime area

a, asset registration b, bid

c, cash transfer d, justice

e, land f, license

g, others h, payment

i, purchase j, tax k, loan

Crime level

a, simple

b, heavy

c, very heavy

Appendix B: Partial view of the pruned decision tree

```
J48 pruned tree
------------------

employment = gov_employee
|   benefit = little
|   |   educ = elementary
|   |   |   crimearea = asset_reg: cluster2 (0.0)
|   |   |   crimearea = bid: cluster2 (0.0)
|   |   |   crimearea = ?: cluster2 (0.0)
|   |   |   crimearea = cash_transfer: cluster2 (0.0)
|   |   |   crimearea = justice: cluster1 (6.0)
|   |   |   crimearea = land: cluster2 (12.0)
|   |   |   crimearea = license: cluster2 (0.0)
|   |   |   crimearea = loan: cluster2 (0.0)
|   |   |   crimearea = others: cluster2 (56.0)
|   |   |   crimearea = payment: cluster2 (0.0)
|   |   |   crimearea = purchase: cluster2 (0.0)
|   |   |   crimearea = tax: cluster2 (0.0)
|   |   educ = secondary
|   |   |   children = ch0: cluster0 (16.0)
|   |   |   children = ch1: cluster0 (206.0)
|   |   |   children = ch2
|   |   |   |   crimearea = asset_reg: cluster0 (0.0)
|   |   |   |   crimearea = bid: cluster0 (0.0)
|   |   |   |   crimearea = ?: cluster0 (0.0)
|   |   |   |   crimearea = cash_transfer: cluster0 (0.0)
|   |   |   |   crimearea = justice: cluster0 (4.0)
|   |   |   |   crimearea = land
|   |   |   |   |   age = 20-30: cluster1 (4.0)
|   |   |   |   |   age = 30-40: cluster0 (6.0)
|   |   |   |   |   age = 40-50: cluster1 (6.0)
|   |   |   |   |   age = 50-60: cluster1 (0.0)
|   |   |   |   |   age = 60-70: cluster1 (0.0)
|   |   |   |   |   age = 70-80: cluster1 (0.0)
|   |   |   |   crimearea = license: cluster0 (0.0)
|   |   |   |   crimearea = loan: cluster0 (0.0)
|   |   |   |   crimearea = others: cluster0 (50.0)
|   |   |   |   crimearea = payment: cluster0 (0.0)
|   |   |   |   crimearea = purchase: cluster0 (0.0)
```

```
|   |   |   |   crimearea = tax: cluster0 (0.0)
|   |   |   children = ch3: cluster0 (0.0)
|   |   educ = certificate: cluster0 (176.0/3.0)
|   |   educ = deploma: cluster0 (180.0/3.0)
|   |   educ = degree: cluster0 (153.0)
|   |   educ = masters_and_above: cluster0 (19.0)
|   benefit = high: cluster0 (113.0)
|   benefit = medium
|   |   educ = elementary: cluster2 (10.0)
|   |   educ = secondary
|   |   |   crimearea = asset_reg: cluster0 (0.0)
|   |   |   crimearea = bid: cluster0 (0.0)
|   |   |   crimearea = ?: cluster0 (0.0)
|   |   |   crimearea = cash_transfer: cluster0 (0.0)
|   |   |   crimearea = justice: cluster0 (0.0)
|   |   |   crimearea = land: cluster1 (9.0)
|   |   |   crimearea = license: cluster0 (0.0)
|   |   |   crimearea = loan: cluster0 (0.0)
|   |   |   crimearea = others: cluster0 (39.0)
|   |   |   crimearea = payment: cluster0 (0.0)
|   |   |   crimearea = purchase: cluster0 (10.0)
|   |   |   crimearea = tax: cluster0 (0.0)
|   |   educ = certificate: cluster2 (59.0)
|   |   educ = deploma
|   |   |   crimearea = asset_reg: cluster2 (0.0)
|   |   |   crimearea = bid: cluster2 (0.0)
|   |   |   crimearea = ?: cluster2 (0.0)
|   |   |   crimearea = cash_transfer: cluster2 (0.0)
|   |   |   crimearea = justice: cluster2 (4.0)
|   |   |   crimearea = land: cluster1 (5.0)
|   |   |   crimearea = license: cluster2 (0.0)
|   |   |   crimearea = loan: cluster2 (0.0)
|   |   |   crimearea = others: cluster2 (27.0)
|   |   |   crimearea = payment: cluster2 (0.0)
|   |   |   crimearea = purchase: cluster2 (0.0)
|   |   |   crimearea = tax: cluster2 (5.0)
|   |   educ = degree: cluster2 (15.0)
|   |   educ = masters_and_above: cluster2 (8.0)
|   benefit = very_high
|   |   educ = elementary: cluster2 (14.0)
```

```
|   |   |   crimearea = license: cluster2 (0.0)
|   |   |   crimearea = loan: cluster2 (0.0)
|   |   |   crimearea = others: cluster2 (27.0)
|   |   |   crimearea = payment: cluster2 (0.0)
|   |   |   crimearea = purchase: cluster2 (0.0)
|   |   |   crimearea = tax: cluster2 (5.0)
|   |   educ = degree: cluster2 (15.0)
|   |   educ = masters_and_above: cluster2 (8.0)
|   benefit = very_high
|   |   educ = elementary: cluster2 (14.0)
|   |   educ = secondary
|   |   |   children = ch0: cluster0 (4.0)
|   |   |   children = ch1: cluster0 (11.0)
|   |   |   children = ch2: cluster1 (6.0)
|   |   |   children = ch3: cluster0 (0.0)
|   |   educ = certificate: cluster0 (5.0)
|   |   educ = deploma: cluster0 (0.0)
|   |   educ = degree: cluster0 (11.0)
|   |   educ = masters_and_above: cluster0 (3.0)
|   benefit = extream
|   |   children = ch0: cluster0 (117.0)
|   |   children = ch1: cluster0 (1081.0)
|   |   children = ch2
|   |   |   crimearea = asset_reg: cluster0 (0.0)
|   |   |   crimearea = bid: cluster0 (4.0)
|   |   |   crimearea = ?: cluster0 (0.0)
|   |   |   crimearea = cash_transfer: cluster0 (3.0)
|   |   |   crimearea = justice
|   |   |   |   age = 20-30: cluster0 (8.0)
|   |   |   |   age = 30-40: cluster0 (0.0)
|   |   |   |   age = 40-50: cluster1 (4.0)
|   |   |   |   age = 50-60: cluster0 (0.0)
|   |   |   |   age = 60-70: cluster0 (0.0)
|   |   |   |   age = 70-80: cluster0 (0.0)
|   |   |   crimearea = land
|   |   |   |   age = 20-30: cluster1 (0.0)
|   |   |   |   age = 30-40: cluster0 (17.0)
|   |   |   |   age = 40-50: cluster1 (20.0)
|   |   |   |   age = 50-60: cluster1 (22.0)
|   |   |   |   age = 60-70: cluster1 (11.0)
```

```
|   |   educ = secondary: cluster0 (80.0)
|   |   educ = certificate: cluster0 (43.0)
|   |   educ = deploma: cluster0 (18.0)
|   |   educ = degree
|   |   |   age = 20-30: cluster0 (0.0)
|   |   |   age = 30-40: cluster0 (0.0)
|   |   |   age = 40-50: cluster2 (2.0)
|   |   |   age = 50-60: cluster0 (3.0)
|   |   |   age = 60-70: cluster0 (0.0)
|   |   |   age = 70-80: cluster0 (0.0)
|   |   educ = masters_and_above: cluster0 (0.0)
|   children = ch2
|   |   crimearea = asset_reg: cluster0 (0.0)
|   |   crimearea = bid: cluster0 (0.0)
|   |   crimearea = ?: cluster0 (0.0)
|   |   crimearea = cash_transfer: cluster0 (0.0
|   |   crimearea = justice
|   |   |   age = 20-30: cluster0 (0.0)
|   |   |   age = 30-40: cluster0 (8.0)
|   |   |   age = 40-50: cluster1 (4.0)
|   |   |   age = 50-60: cluster0 (0.0)
|   |   |   age = 60-70: cluster0 (0.0)
|   |   |   age = 70-80: cluster0 (0.0)
|   |   crimearea = land: cluster0 (0.0)
|   |   crimearea = license: cluster0 (0.0)
|   |   crimearea = loan: cluster0 (0.0)
|   |   crimearea = others
|   |   |   benefit = little: cluster0 (0.0)
|   |   |   benefit = high: cluster0 (0.0)
|   |   |   benefit = medium: cluster2 (3.0)
|   |   |   benefit = very_high: cluster0 (0.0)
|   |   |   benefit = extream: cluster0 (12.0)
|   |   crimearea = payment: cluster0 (0.0)
|   |   crimearea = purchase: cluster0 (3.0)
|   |   crimearea = tax: cluster1 (10.0)
|   children = ch3: cluster0 (0.0)

Number of Leaves  :     306

Size of the tree :      356
```

Declaration

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

_____

Elsabet Wedajo

June, 2012

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

_____

Gashaw Kbede (Ph.D)

June, 2012