ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

MINING CRIME DATA FOR EFFECTIVE RESOURCE ALLOCATION AND CRIME PREVENTION:

THE CASE OF ADDIS ABABA POLICE COMMISSION

By

LETEZGI HAGOS

JUNE, 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

MINING CRIME DATA FOR EFFECTIVE RESOURCE ALLOCATION AND CRIME PREVENTION:
THE CASE OF ADDIS ABABA POLICE COMMISSION

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
LETEZGI HAGOS

JUNE, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

MINING CRIME DATA FOR EFFECTIVE RESOURCE ALLOCATION AND CRIME PREVENTION:

THE CASE OF ADDIS ABABA POLICE COMMISSION

By

LETEZGI HAGOS

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
Dereje Teferi (PhD)	Advisor(s),	_____	_____
Million Meshesha (PhD)	Examiner,	_____	_____

Declaration

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

_____

Letezgi Hagos

June 28, 2011

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

_____

Dereje Teferi (Ph.D)

Acknowledgment

First and foremost I would like to thank to my advisor, Dereje Terferi (PhD) for his invaluable assistance and constructive comments on the research document.

I would like to express my appreciation to the wonderful friends I have had outside the department and those that are at the Department of Information Science.

Last but not least, I would like to thank all the FSCE staffs and Addis Ababa Police commission staffs, without whose good will to supply me with the data, I would not have accomplished this research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

AAPCO ----------	Addis Ababa police commission office
ARFF ----------	Attribute-Relation File Format (ARFF)
CSV----------	Comma Separated Value
FSCE----------	Forum on Street Children Ethiopia
GUI ----------	Graphical User Interface
KDD----------	Knowledge Discovery in Databases
NGO----------	Non-governmental Organization
CRISP-DM----------	Cross industry standard process for data mining

Abstract

The aim of this research is to extract meaningful crime trends regarding offences against children from the data in existing police records with the help of data mining techniques. We know children are exposed to different offences but we do not know which children are exposed to what type of offence (crime category). The output from this research helps to identify which children are exposed to which crime categories. Currently the police officers try to understand the relation between any two attributes but they do not know the relation among more than two attributes and the relationship between other variables and a class variable. This is why this can be achieved by using data mining techniques in an efficient and accurate manner than those achieved by trained personnel and traditional simple statistics to analyze crime data.

The researcher used the six-phased CRISP-DM for data mining process and each of the steps in this model starting from business understanding up to evaluation and deployment phases are performed step wise and iteratively when needed. Even though all the phases are equally important the data pre-processing part has got due emphasis since police records are inconsistent and frequently incomplete making task of formal analysis inaccurate and time consuming. These analytical processes would benefit from using data mining techniques in a structured approach. Both unsupervised and supervised learning are used within the structured methodology to mine the police data.

This research will serve as a reference material for researchers, crime investigators, planners and NGOs that work on prevention and control of offences against children. Based on this, it can also help to implement different crime preventive programs like through awareness creation programs**.**

The research demonstrates that data mining techniques can be successfully used in proactive policing to prevent crimes. This is more applicable for high volume crimes such as theft, violence and sexual assaults that have been committed most commonly. These crimes can often be segmented and classified and the generated models can be used to

predict potential victims of a specified crime category through predictive models as well as to attribute the profile of victims with the help of descriptive techniques.

Some of the rules in association rule are not interesting due to few values that are unable to generate patterns. From all the crime categories in the crime records sexual assault has the highest number and best rules are generated related with sexual assault. Almost all interesting rules generated in association rule are included in the rules generated by the classification model. Generally, promising results are registered that encourage further researches in the area.

CHAPTER ONE

INTRODUCTION

1.1. Background

Like all human being, children require living and growing up in place where it is safe for their lives. Their parents and care givers should make sure that there is a concerned body that protects them from different harmful conditions and labor exploitations. Police forces across the developed world have attempted to apply advanced computing technologies for extracting the trends and tackling such crimes (Adderley and Musgrove, 2001). However, comparatively little use has been made of data mining techniques in analyzing and modeling the behavioral patterns which occur in the commission of a particular crime.

Data mining has attracted a great deal of attention in the information industry in recent years due to the wide availability of huge amount of data and the imminent need for converting data into useful information and knowledge. Data mining can be contended as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of functionalities such as data collection and creation, data management and data analysis and understanding (involving data warehousing and data mining) (Han and Chamber, 2001).

Defining a scientific discipline is always a controversial task; researchers often disagree about the precise range and limits of their field of study. Bearing this in mind, and accepting that others might disagree about the details, the researcher shall adopt a working definition of data mining.

Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand and Mannila, 2001).

Data mining is a process that uses various techniques to discover hidden relevant information (knowledge or useful patterns) from heterogeneous and distributed historical

data stored in large databases, warehouses and other massive information repositories (Han and Kamber, 2001).

In this regard one of the areas where huge amount of data found is in police databases where different crime reports are recorded and stored for a long period of time. Therefore, application of data mining tools is needed to convert such data to useful information and knowledge.

One of the most important activities of law-enforcement agencies is the investigation of suspicions and clues about persons and events. Investigation can result in the prevention of crime in two distinct ways. First, the suspicions and clues are correct and the law-enforcement agency responded sufficiently fast to stop the crime before it has been committed. Second, the investigation did not lead to an early termination of the specific crime, but it resulted in new knowledge that is used to manage activities more efficiently and effectively than was previously possible. The later is called proactive investigation and assumes that sufficient data is available to be analyzed (Thibault *et al*, 2006).

Analysis has to be done automatically due to the large amounts of data. The knowledge obtained is necessary to establish and maintain good understanding of the domain area. This is the motivation behind intelligence led policing: good information enables law enforcement to prevent crimes and reduce risks of potential dangers (Cope, 2004; Tilley, 2005).

The technological advancement resulted in the management of huge computerized data acquisition and storage of databases which contain hidden knowledge that can be very important and useful for decision making. It is impossible and also time consuming to unravel this knowledge. Moreover, improper conclusions ultimately affects decision making. Consequently, a need to use more efficient techniques and to have or provide knowledge in a comprehensible form as well as to arrive at better results has developed from both the owner and users of the databases. This has lead to the exploration of a new field of research known as data mining.

Data mining is the process of analyzing large amount of data from different perspectives and summarizing it into useful information. It allows users to analyze data from different dimensions, categorize and summarize the relationships identified. Technically, it is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining techniques are the result of a long process of research and a product of development. Data mining tools predict future trends and behaviors, allow it to make proactive, knowledge-driven decision. It has wide application and supported by massive data collection and data mining algorithms.

Data mining encompasses the process of discovering hidden patterns and relationships in large amounts of information and data. This enables human beings to make accurate and reliable predictions of future events based on the identification and characterization of patterns and trends in historical crime data. Data mining helps solve a common problem: the more information, someone has the more difficult and time-consuming it will be to effectively analyze and draw meaning from it. By using a clear process and powerful analytic technologies, data mining quickly and thoroughly explores mountains of data, helping us identify the valuable or actionable nuggets of information. In the case of Addis Ababa police commission; this study tries to identify the profile of victims with respect to offender age, decision given, and the sub-city where the crime is committed.

Data mining is one of the sophisticated tools used in law enforcement to discover new patterns or confirm suspected patterns or trends. One of the strengths of data mining, as opposed to more traditional statistical methods, is that it is not necessary to know exactly what we are looking for before we start. Data mining uses powerful analytic tools to quickly and thoroughly explore mountains of data and pull out valuable and usable information. The primary use of data mining is to find something new in the data to discover a new piece of information that no one knew previously. This is bottom-up or data-driven approach because we start with the data and then build theories based on discovered patterns or trends.

Along with the prevention and investigation of crime, police makes use of previous crime data as an input for the formulation of crime prevention policies and strategic plans (Wilson, 1963). It is obvious from the outset that to make use of data and records, relevant data have to be kept and managed properly. For this reason the Addis Ababa Police Commission in collaboration with the FSCE have been collecting criminal records regarding children since 1997 EC and have maintained numerous criminal records.

As expressed by the experts in the police commission and reported in recent years unlike developed countries, developing countries, including Ethiopia, did not use modern data mining tools and techniques to facilitate the processing of crime data. It is only in the last few years Ethiopian police commission has began to develop databases for departments that possess bulky data. The databases developed so far includes criminal database, traffic database, and personnel database.

The database in FSCE is part of the criminal database which is used for this study. It has 5355 records with 27 attributes. Criminal records from each sub-city are sent manually to the commission by filling the criminals profile form and then the FSCE collects manual documents from the commission and the database administrator in the FSCE is responsible to enter all the records in to the database.

Even though there are improvements in developing databases with moderate amount of data in some of its departments (Traffic, FSCE), the Addis Ababa Police Commission has not yet exploited the knowledge embedded in those data. The Addis Ababa Police commissioner noted that the current system is manual except the fact that the some aforementioned databases are being used as a data repository. In line with this most of the time crime prevention measures are being taken based on crime incidents although it would have been based on crime trends. This indicates that, currently, the crime prevention approach is based on the crime reports incoming to the police, which is a reactive approach although there are some awareness creation programs based on simple statistics.

There are several factors that contribute to the under-utilization of the FSCE database system. According to the database administrator of FSCE, one reason is lack of knowledge on what could be done using these databases or deficiency of appropriate tools that could make use of these databases and the other problem is lack of skilled manpower.

Data mining is one of the powerful tools that have evolved to play a role as an instrument to discover patterns buried in large databases. Data mining is a technology for the exploration and analysis of large quantities of data by automatic or semiautomatic means in order to discover meaningful patterns and rules (Berry and Linoff, 1997).

1.2. Statement of the problem

This experimental research is undertaken to discover crime trends, which are pertinent in the prevention and control of offences committed against children in Addis Ababa. The research also attempts to determine the contribution of victim's profile in relation to offender age and decisions given as well as the sub-city where the crime was committed.

Nowadays crime is becoming a complex social phenomenon and its cost is increasing due to a number of societal and technological changes. Hence, law enforcement organizations like that of police need to learn the factors that constitute higher crime trends (Wilson, 1963). To control this social evil there is always a need for systematic crime prevention strategies and policies. Thus, achieving this understanding and processing of crime records are one method to learn about crime, victims and criminals.

Data mining helps to learn from the previous and current conditions to predict the developing conditions. Otherwise it is impossible to predict or control the occurrence of offences and ensure the safety of the group targeted as victims and prevent individuals from committing crimes.

Generally, law enforcement agencies need to work in identifying, predicting, responding to and preventing offences committed against children. This is achieved by analyzing past criminal behavior patterns, mapping their anticipated future occurrence through identifying offenders' and victims' profiles. This helps to stop or minimize crime by

avoiding opportunistic committing of crimes, easily react to up on call for help, deliver critical real-time information to the field and improve deployment of law enforcement assets to protect the victims and improve response times for crimes. There are new emerging police roles like, protect life and property to reduce the opportunity to commit crimes, to maintain social order and to protect individual freedom and privacy.

Data mining serves as a tool to solve the above mentioned problems and extract hidden useful patterns on massive amounts of heterogeneous and historical crime data collected and stored in an electronic format. Like all other databases used for data mining, Ethiopian police department has different sub-departments and hierarchic where a particular criminal case stays for unspecified amount of time. Different sub-departments collect data about different crimes (traffic violations, sex crime, theft crime, violent crime etc). Similarly different hierarchies (city, sub-city, kebele) store data at different levels of detail. This necessitates the use of data mining technology since the complexity of the data make harder to extract accurate and reliable information using human beings and simple statistical tools.

1.3. General objective of the study

The main objective of this study is to apply data mining technologies on developing a model that pinpoint the events associated with offences against children and predict related conditions with some of the common offences to extract noble patterns regarding the patterns of crimes committed against children.

1.3.1. Specific objective

To achieve the above general objective this research has the following specific objectives.

- To develop an understanding of the application domain through reviewing related documents and communicating with domain experts from Addis Ababa police commission.
- To understand the possible opportunities of data mining applications on selecting and prioritizing future crime occurring conditions.

- To asses and choose data mining techniques which are appropriate to predict and describe unsafe situations.
- To collect and analyze working data which is relevant to the data mining problem.
- To pre-process and prepare the raw data into suitable dataset for the data mining software.
- To train the data mining model on the training data.
- To evaluate the model for its appropriateness on describing crimes or predicting crime occurrences
- To present the result using different visualization techniques.
- To recommend on what should be done from the finding of the experiments.

1.4. Scope and Limitation of the study

The scope of this research is limited to assessing the possible application of data mining technology for Addis Ababa police commission crime detection. It was limited to identify events associated with offences against children and associated conditions. Different data mining techniques can be applied to the problem domain. Due to time and budget constraint only association rule, clustering and classification mining techniques are applied. There are also data related problems such as:

- Criminal records are found in report having unstructured format and need special data miming software and highly skilled miner to deal with.
- Criminal data has inconsistency regarding the data of suspects often giving false names, birth dates, or addresses to police officers and thus have multiple database entries, making it difficult for officers to determine a suspect's true identity and relate past incidents involving that person.
- Some criminal records are found in hard copy formats

Due to the above mentioned limitations the researcher forced to limit her study to offences committed against children, only 5355 records are used for analysis.

1.5. *Methodology*

The following data mining methodologies were employed.

1.5.1. Research design

1.5.1.1. CRISP-DM model

This process model for data mining provides an overview of the life cycle of a data mining tasks. It contains the corresponding phases and their respective tasks, and relationships between these tasks. At this description level, it is not possible to identify all relationships. There possibly exist relationships between all data mining tasks depending on goals, background and interest of the user, and most importantly depending on the data. This contains step-by-step directions, tasks and objectives for each phase of the Data Mining Process.

Generally, in CRISP-DM the life cycle of a data mining process consists of six phases. The sequence of the phases is not strict. Moving back and forth between different phases is always required. It depends on the outcome of each phase to determine which phase, or which particular task of a phase, has to be performed next.

CRISP-model describes the cyclic nature of data mining itself. A data mining process continues even after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones. A brief outline of the six phases is given below:

Business Understanding

At this stage the researcher tries to understand the objective of Addis Ababa Police commission and FSCE by reading their annual reports and discussing with the domain experts. The duties of AAPC are divided into two broad categories. The first is operational policing and the second strategic policing. Operational policing is the day to day activities performed by crime investigators and police officers. Whereas strategic policing is planning for future crime investigation and crime prevention having knowledge from those previously conducted crime investigations' information and data. This is known as proactive policing which requires analysis methods and tools to understand hidden patterns as the data increases in size and complexity. The current research is mainly

focused on using data mining to facilitate strategic policing to protect children from offences.

Data Understanding

The data for this study is taken from crime reports registered in Addis Ababa police commission and FSCE. One of the prominent child-focused local NGOs in the country, Forum on Street Children Ethiopia (FSCE), has been involved in implementing the aforementioned roles since its establishment in 1989. As a child-oriented organization, the mission of FSCE is to work for the respect of the rights of street children, sexually abused and exploited children, physically abused children, and children in conflict with the law. The data contains information about offenders or suspects (name, age, gender, address, educational status, job, marital status), victim (name, age, gender, address, job, marital status, educational level) and crime incident (type, time, location (sub-city)), unique location mane, instruments used, special events) are registered.

Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include record, and attribute selection as well as cleaning and transformation of data for modeling tools. At this phase the researcher started with translating the data from local Amharic language to foreign language English. This task is perfumed using MS-Excel built-in functions auto filter and auto fills. Next the cleaning task was performed. Following this the integration task was performed. Finally the reduction and transformation tasks are performed.

Data mining tools require appropriate data that fulfils the required quality to be supported by the selected algorithm. Some algorithms accept only numeric input while others only categorical data or both but not missing attribute value etc.

This is why the availably of data by itself does not fulfill everything for developing the model or performing the data mining task rather it needs further processing. These can be

summarized as preprocessing, namely, data cleaning, data integration, data reduction, and data transformation.

As mentioned above the data are collected from forum for street children in Ethiopia (FSCE). Preprocessing tasks such as cleansing, integration and reduction are imposed in order to improve the data quality. This comprises attribute selection, handling noisy data, accounting for missing data fields, coding text valued attributes and preparing the preprocessed data in a file format acceptable to the WEKA software.

In the data cleaning preprocessing task the data content related problems (issues) should be solved. These are filling missing values, avoiding irrelevant attributes and removing noisy. In data integration combining different MS-Excel tables is performed. Solving problems of integration is crucial here. Even though data reduction is related to both dimensionality and numeriocity reductions, which are better attributes selection and (sampling & clustering) respectively, only dimensionality reduction is performed on the current data. In data transformation hierarchy generation to see data at different levels of abstraction and fit the data to the data mining system that is going to be developed like converting the age values to appropriate ranges and giving manes to each range is performed.

Modeling

In this phase, three modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem. The association rule modeling used measures of interestingness (support and confidence in line with domain experts judgments) to extract the victims profile, the clustering model used the simple K-means algorithm requires to set the K-value, seed, display-standard-deviation, distance-function by the researcher. Finally the classification is applied using the clustered data as input and the parameter number-of-objects is set by the researcher to compare it with the default value. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. After generating some association rules the researcher tried to return back to the data and merge some attribute values like the values for habit were no-habit, smoker,

chewing chat and drug addicted. The four values have been merged to two values (“No” for no-habit and “Yes” for the other three values).

Models with their algorithms used in this data mining research are; first association to identify which events and conditions are associated with particular crime by using the apriori algorithm. Second clustering to group the crime records and serve as an input for the decision tree. Third classification using J48 decision tree algorithm to predict what type of offences are expected to be crimes committed against particular children.

Evaluation

At this stage in the data mining task we have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. 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 should be reached. This is performed based on the domain experts’ advice and the parameters set and the researchers’ personal judgment.

Deployment

In this part all the generated rules should be converted to more specific rules which are simple to understand and apply. Additionally, there should be funding appropriate conditions necessary for the proper deployment. It demands the right integration and availability of qualified officers, technology and resources. In line with this, the commission together with the local NGO should have to keep important information about its victims, offenders, and the crime situation in a clear and suitable manner.

1.5.2. Tool selection

There are varieties of tools available for data mining such as the knowledge studio, WEKA, Xlminer, SA, SPSS, STATA, and other. Among those tools, WEKA is selected.

WEKA is selected for its familiarity with the researcher and its being open source and availability. WEKA consists of a collection of machine learning algorithms for solving real-world data mining problems. The package has three different interfaces: a command line interface, an Explorer GUI interface (which allows one to try out different preparation, transformation and modeling algorithms on a dataset), and an Experimenter GUI interface (which allows to run different algorithms in batch and to compare the results) (Witten & Frank, 2000). WEKA version 3.4.6 is used in this research.

The Waikato Environment for Knowledge Analysis (WEKA) was utilized to perform the association rule mining, clustering model and classification on the data. WEKA is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and WEKA has been tested under Linux, Windows, and Macintosh operating systems. Java provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset (Rogers, 2001).

The WEKA system uses a common file format to store its datasets and thus presents the user with a consistent view of the data regardless of what machine learning scheme may be used. This file format, the Attribute-Relation File Format (ARFF), defines a dataset in terms of a relation or table made up of attributes or columns of data. Information about the names of the relation, and the data types of the attributes are stored in the ARFF header, with the examples or instances of data being represented as rows of data in the body of the ARFF file. Though only nominal attributes are used in this research, attributes are currently allowed to take on three different data types, namely integers, real or floating point numbers and enumerations. With the numeric attributes an optional range may be specified for range checking and Boolean attributes are treated as an enumeration with two values (Garner, nd).

1.6. Significance of the study

This research will be used as an important material for the application of data mining in the domain to address the different crime and criminal behavior such as offence against

children. Therefore, victim education, offender counseling, and protective orders might be implemented successfully. Similarly, this research also served as academic practice for the researcher.

1.7. Thesis Organization

The research report in this study has six chapters. The first chapter deals with the basic overview including background, statement of the problem, objective, scope and limitation, methodology, literature review and this part which is thesis organization.

In the second chapter the review of data mining literatures are presented. In this chapter how and why data mining was evolved as a field of study and how it is applied in different domain areas are discussed.

In the third chapters critical literature review which includes review of crime and crime reports together with application of data mining crime on crime data are presented.

The fourth Chapter deals with data preprocessing tasks. In this chapter how the four data preprocessing tasks are applied to the current data are shown.

The fifth chapter deals with experimentations and result interpretations. In this chapter the nine different experiments like association (four experiments), clustering (three experiments) and classification (two experiments) with their interpretation and evaluation are done.

In the sixth chapter conclusions and recommendations are presented.

CHAPTER TWO

Data Mining Technologies

People have been gathering and analyzing data to get information and have knowledge about their environment and explain natural phenomenon. After investigating this data they have developed different theories, observations, and approaches that could help them identify and interpret phenomena of the natural world. People had been analyzing data and looking for patterns even without using machines and analysis tools.

However, gradually, new technologies have begun to play a vital role in storage, facilitating analysis and processing of data. Specially, the advent of computer technology in both hardware and software has revolutionized the way in which data are saved, interpreted and managed. These new methods of looking into data as well as the eagerness to learn from data have brought the chances to evolve disciplines like that of data mining (Saygin and Ulusoy, 2002).

2.1. Overview of Data Mining

Almost all organizations have developed the habit of collecting data that they believe directly or indirectly benefit their institutions. In line with the capabilities of collecting there is also generating large amount of data inside and outside of the organizational system. There are several contributing factors to the proliferation of bulky data. This includes, the use of bar code, the automation of many businesses and other transactions, advances in data collection tools ranging from scanned image documents and image platforms to satellite remote sensing systems.

As the size of data grows in organizations, there exists a need for new automated methods that can enable them to process and convert the accumulated data into useful information and knowledge. The tremendous amount of data collected and stored in large databases is beyond human capability for comprehension of patterns inside them without using powerful analysis tools. To get benefits from integrated and historical data, there should be a way to identify relevant and useful information (Hand et al, 2001).

As mentioned above, technological developments that aid to collect and store vast quantities of data have enabled organizations to capture and accumulate huge amount of data in their databases, within which, large amount of valuable information is buried (Han and Kamber, 2001). As the volume of data increases, the proportion of information in which people could understand decreases substantially. This reveals that the level of understanding of people about knowledge in the data at hand could not keep pace with the rate of generation of data in various forms, which results in increasing information gap. Consequently, people begin to realize this bottleneck and to look into possible remedies like data mining.

Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data. Data mining techniques are derived from three different sources: artificial intelligence, statistics and machine learning (Han and Kamber, 2001).

Much of the tools and techniques of statistics are adopted in the field of data mining. However, although statistics is very useful technique, it is not capable to address all of the general and domain specific data mining problems (Berry and Linoff, 1997). For instance, some problems may demand learning from experience and statistical methods could not address such problems. Moreover, statistics usually employs sample data (portion of the population data thought to be relatively appropriate representative) to build statistical models and this method can miss much information about the population (Thearling, 2003).

The other discipline from which data mining is adopted is artificial intelligence. This field of study is developed on the basis of replication of human mind in contrast to statistics. It is an attempt to apply "human-thought-like" approach to statistical problems. The application of artificial intelligence has become pervasive with the introduction of useful power and multipurpose computers at affordable prices (Ibid).

Similarly, machine learning is also a field of study that contributed a lot to data mining. Machine learning is more properly described as the hybrid of statistics and artificial

intelligence (Han and Kamber, 2001). Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced artificial intelligence heuristics and algorithms to achieve its goals (Larose, 2005). This depicts that the application of machine learning in the study of large volume of data is a radical shift not only from statistics but also from artificial intelligence via merging both fields.

From the above arguments it seems plausible to conclude that data mining, in many ways, is basically the adaptation of machine learning techniques to scientific and business problems. This is why data mining is considered as the union of historical and recent developments in statistics, artificial intelligence, and machine learning. The tools and techniques borrowed from these fields of studies are used together to extract previously unknown patterns buried in large database. As a result, data mining is becoming popular in both science and business areas where there is large amount of data that require special tools to extract patterns.

2.2. Data Mining Life Cycle

The data mining modeling cycle involves a number of stages. Initially, it is important to have a clear understanding of the business domain in order to understand the operational analytical processes (Thomsen 1998), the problems that are to be emphasized, the opportunities that may be realized and to assess the availability of data. Exploring and preparing the data, although time consuming (Sherman 2005), is a crucial stage in the cycle. In preparing data new fields may be derived from one or more existing fields, missing and boundary values identified and processed. Additionally, relationships between fields in each column and records in the rows of the entire data is identified to some of the pre-processing tasks that assist in cleaning and made appropriate the data prior to the mining process. Once data has been prepared for mining, the modeling stage can begin. Choosing and developing models involve domain knowledge (Chen et al, 2004), the results of which are validated against expected results.

Data mining is an iterative process as the results produced by the techniques need to be integrated with other results produced by other techniques to provide the desired business advantage. The Cross Industry Standard Process for Data Mining (CRISP-DM) was the data mining cyclic methodology used within this study. This methodology was designed by a group of businesses to be used with any data mining tool and within all business areas (Chapman et al, 2000). As reported by different authors CRISP-DM is the most widely used methodology. (Mena, 2003) recommended for its use in application of data mining crime prevention and detection.

Though the sequence of the phases is not rigid, the life cycle of a data mining research consists of six phases. Moving back and forth between different phases is always required depending on the outcome of each phase. These phases are:

2.2.1. Business Understanding

This initial phase focuses on understanding the objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. At this stage the researcher tried to identify the objective of Addis Ababa Police commission which prevention or reduction and detection of crimes as well as avoiding opportunistic situations that criminals could get. One such objective could be to reduce offences by 5% over the next year. Assess whether these problems need data mining and are solvable with the help of data mining.

Within a data mining modeling environment understanding the core business and its associated requirements is essential.

2.2.2. Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. Prior to any analysis the purpose must be defined and for any such analytical process to take place it is more cost effective to utilize currently stored data.

However, to reach the goal, it may require supplemental data to be captured separately by manual processes and subsequently stored in electronic format. It may also be possible to pay costs for data sets to compliment that which is currently used (such as census/demographic data, weather data etc). While dealing with the data mining modeling environment like understanding the core business, understanding the data is critically important. There are various issues when trying to understand the underlying data: -

- What information is required for the business to be known?
- What variables can be used to transform the data into information?
- Are some variables only used in special instances of the data?
- If so, what are those instances?
- Can the significance of certain variables be enhanced by combining or deriving other variables?

2.2.3. Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include merging or splitting table data, sampling record, and attribute selection as well as transformation of the data in to common format and cleaning, such as removing or inserting missing values, removing or smoothing outliers etc, of data to be supportable by the modeling tools. The potential of using data mining is only as good as the quality of the data (Dragoon, 2003). Data preparation is the most important feature of the CRISP-DM process and also the most time consuming (Sherman, 2005), taking up to 70% of the total research scheduling time. Another author has stated that this stage can take as much as 80% to 90% of the total time (Pyle, 1999).

The consistency of the data and its subsequent encoding is directly proportional to the results of the mining process (Weiss and Indurkya, 1998). It is not uncommon for data sets to have fields that contain unknown or incorrectly entered information and missing values,

how should they be treated? Are those fields essential to the mining process? There are five basic processes for treating records that contain missing values:

1. Omit the field(s) columns containing missing values: - this is to remove the entire column containing the missing values.
2. Omit the entire record that contains the missing values: - this is to discard the entire row containing missing values. This is advisable when it has more missing values and the attribute with missing value is class variable when the technique used is classification.
3. Automatically correct the missing data with default values e.g. select the mean from the range,
4. Derive a model to enter/correct the data: - this is to calculate mean for numerical values and mode for categorical values and replace with the one which is appropriate for replacing the missing.
5. Tag the value as incorrect:- just leave it as it is.

However, it is important to note that the absence of data may itself be valid, for example, when completing a questionnaire some people may refuse to state their age group thereby leaving this field as a null value but this in itself may be an identifying factor in any analysis.

During this stage in the cycle a variety of encoding techniques may be utilized to provide additional fields for analysis and enable fuzzy concepts. An example of missing data could be that some offender description fields may not contain the offender's age. This could either mean that the victim could not remember, the Officer did not ask the question or the information was not entered onto the crime report. Experience would suggest that the later option would be correct as it is common practice to take a full written statement at the time of the offence being reported and to transfer brief information onto the crime report. Within policing terms, text based memo fields are rich in information, however, the text is notably more noisy than other text sources such as news reports etc., containing many spelling errors, typos and grammatical errors (Chau et al 2002). However, in the absence of fully structured data, as in the Addis Ababa Police crime reporting system

(hand written memos), the text need to be written and parsed to extract information that can be employed in the analysis of crime.

There are a number of fields within the Police databases that do not contain data and are stored in the database as "-" or as an empty string like unknown. These mainly relate to location information. If an address is incorrectly entered into the system an 'unconfirmed location' is registered which permits the crime to be recorded but an operator has to manually enter the correct information when time permits. These are often subsequently left blank.

When using enterprise wide volumes of data it would be unusual to use the entire data set to build the models for the mining process, therefore, a relevant subset must be extracted. Care must be taken to ensure that the problem space is fully represented by the extracted data and that there are sufficient examples to be modeled.

In this thesis, data could be a number on a continuous scale (such as age), binary (such as gender), and nominal (such as habit). Data can be transformed from continuous to discrete using the decartelization process. This is known as transformation. Valuable information can be lost in transforming from one form to another (Gordon 1981). It is common to standardize variables, but this can in itself cause problems due to the discriminating effect of the variable being lost. For example in smoothing age value which might range from 15 to 65 in to the range 0 to 1 would lead to a 20 year old being scaled as 0.1 and a 30 year old 0.3. Thus a difference of ten years in age (a value of 0.2) would be ten times less important compared to a difference in fuzzily defined attributes such as a person's age (i.e. young, old) which is coded as a strict 0 or 1.

2.2.3.1. Variable Selection

Which variables should be used in the modeling process? Dependent upon the systems being mined there could be many hundred available variables but not all are relevant and will affect the outcome. It is within this area that the domain expert must provide a guide (Wang et al, 2004). Various statistical methods like information gain, gain ratio, chi-square, etc can be used for feature selection (Weiss and Indurkhaya, 1998). In many

researches, Genetic Algorithms (GAs) have been successfully employed in this area and some commercial neural network packages incorporate this feature.

2.2.4. Data Mining Modeling Techniques

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. Modeling algorithms assist in the extraction of complex interrelationships between data variables and identify the decision making rules. The models are used in prediction, estimation and classification thereby providing 'expert' decision support.

2.2.5. Evaluation

At this stage in the data mining task we have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. The resultant model may be validated in terms of its clarity, parsimony, generality and testability to assess the degree to which it meets the required objectives. A number of techniques may be used, for example: -

1. *K*-fold cross validation (Bishop, 1995).
2. Use a domain specialist to examine the results (Montgomery, 1998; Chen et al, 2004).
3. Cluster evaluation.
4. Statistical analysis.

In each of the works later described, the validation was conducted by one or more of the above methods. In case of domain experts evaluation they are crucial to evaluate because they are close to the business problem and they do have expected pattern from which their evaluation is based. This is better to be supplemented with the different evaluation tools.

2.2.6. Deployment

Building of the model is generally not the end goal. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. 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. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps.

2.3. Data Mining and Knowledge Discovery in Databases (KDD)

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields. In this research both data mining and knowledge discovery are used interchangeably.

2.4. Tasks of Data Mining

The two primary goals of data mining tend to be *prediction* and *description. Prediction* involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description,* on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of 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, nontrivial 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.

Practically, data mining can accomplish common tasks like; classification, estimation, prediction, association, clustering, and description. However, many of the practical problems of scientific, economic and business interests can be mapped into one of these common tasks (Berry and Linoff, 1997).

All data mining techniques are classified as either of the two models and are not equally applicable to all the above-mentioned tasks. Based on the nature of the problem under consideration and its proximity to the main divisions of data mining tasks, researchers need to choose the appropriate techniques among the numerous data mining techniques.

2.4.1. Classification

Classification is one of the most common data mining tasks, which is also pervasive in human life. Classification finds common properties among different entities in order to organize them into predefined classes. Human beings usually classify or categorize in order to understand and communicate about the world. For any object or instance, classes are predefined according to the value of a specific field (Berry and Linoff, 1997).

Data mining classification is a two-step process. These include, learning (training) and the classification (testing) steps. In the learning step, a classifier is built describing a predefined set of data classes or concepts. The classification algorithm builds the classifier model through learning from the training dataset and their associated class labels in the learning step.

Classification comprises examining the features of unseen instances and assigning it into one of the predefined classes. In fact, in the case of data mining the objects that are going to be examined comes from a database. The task of classification incorporates updating each record by filling in a field with a class code of some kind. This task starts with instances often called a training set which consists of predefined classes and are used to train as well as build a data mining model so that the model can be applied to classify unseen objects.

2.4.2. Clustering

Clustering is another task of data mining in which groups of instances or objects that are more similar belong to the same clusters where as dissimilar to different clusters (Han and Kamber, 2001). The only difference between classification and clustering is in the case of clustering unlike classification there are no predefined classes. As there are no predefined classes and examples in clustering, records are grouped together on the basis of similarity among the instances (Beaza-yates and Ribeiro-Neto, 1999).

2.4.3. Association rule

Discovers frequently occurring item sets in a database and presents the patterns as rules. This technique has been applied in problems like network intrusion detection to derive association rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks (Lee et al, 1999).

Association learning is a data mining scheme in which any affinity grouping between features is sought, not just one that predict a particular class value (Agarwal and Srikant, 1994). Association is suitable if the problem is to extract any structure that exists in the data at hand. The aim of association is to examine which instances are most likely to be grouped together. Market basket analysis is a typical practical application of association, which is concerned with what items are consumed with a particular item by individual buyer. Association rules can be developed in order to determine arrangements of items on

store shelves in a given supermarket so that items often bought together will be found arranged closer in location (Berry and Linoff, 1997).

Association rule is a powerful tool for discovering correlations among massive databases and the concept was introduced by (Hipp and Ulrich, 2000) for analyzing market basket data to mine customer shopping pattern. The inputs used in the process of generating association rule are taken from a table where corresponding values in each data items have correlations to one another.

Let a data set I = $\{I_1, I_2, I_3 \ldots\}$ and a database of transaction D= $\{t_1, t_2, t_3 \ldots\}$ where t_i = $\{I_{i1}, I_{i2}, I_{i3}, I_{i4} \ldots\ldots\ldots..\}$ and I_{ij} € I.

The idea is to identify relations among the items purchased so that the customers could be targeted for marketing specific products.

An association rule typically consists of three parts (i) an antecedent (X) (ii) a consequent(Y), (where X,Y contained in I are items of data set called as item sets and $X \cap Y = \Phi$), and (iii) a measure of the interestingness of the rule (support%, confidence%), as represented as

X => Y (support%, confidence %)

The antecedent and the consequent are a set of one or more predicates.

The **support** of a rule measures the frequency of collective occurrence of all the antecedent predicates of a rule in the dataset. The support for an association rule $X \rightarrow Y$ is the percentage of transaction in the database that consists of $X \cap Y$.

$$\textbf{\textit{Support}} = \frac{X \text{ n } Y}{\text{Total number of cases}}$$

The **confidence** measures the frequency of the occurrence of the consequent given the occurrence of the antecedent. The confidence for an association rule $X \rightarrow Y$ is the ratio of

the number of transactions that contains X∩Y to the number of transactions tha t contains X.

$$\textit{Confidence} = \frac{\mathbf{X} \text{ n } \mathbf{Y}}{\mathbf{Y}}$$

Criteria value filters the data based on the confidence value.

Thus, association rule mining algorithms aim to extract association rules with support and confidence greater than user-specified threshold. The process of Association rule has the following sequences (Kwasnicka and Switalski, 2005):

(i) Understanding of the problem domain and identification of the final goal of the process
(ii) Data collection
(iii) Preprocessing of data is carried out to refine the data
(iv) Encoding of data
(v) Data mining - Selection of association rule, selection of algorithm for data exploration, running the algorithm to generate patterns and
(vi) Interpretation, presentation and explanation of mined knowledge.

According to (Han and Kamber, 2000) a data mining system has the potential to generate thousands of patterns or rules. Not all of the patterns are useful or interesting. Hence we need to define what is an interesting pattern and how can we generate all the interesting patterns and only the interesting patterns.
A pattern is *interesting* if:
The pattern is easily understood by humans; the pattern is valid (with some degree of certainty) on new or test data; the pattern is potentially useful; or, the pattern is novel. A pattern is also interesting if it validates a hypothesis that the user wished to validate, or resemble a user's hunch. An interesting pattern represents knowledge.

Several objective measures of pattern interestingness exist, based on the structure of discovered patterns and of the statistics underlying them. The concepts of support and confidence are examples of objective measures of pattern interestingness. In general, each

interestingness measure is associated with a threshold, which may be controlled by the user.

Although objective measures help identify interesting patterns, they are insufficient unless combined with subjective measures that reflect the needs and interests of a particular measure. Subjective interestingness measures are based on user beliefs in the data. These measures find patterns interesting if they are unexpected (contradicting a user's belief) or offer strategic information on which the user can act.

It is often unrealistic and inefficient for data mining systems to generate all of the possible patterns. Instead, user-provided constraints and interestingness measures should be used to focus the search. Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining.

Association, unlike classification, can predict any field; not only the class and it can forecast more than one attribute's value at a time. For this reason we can find a number of association rules than classification rules.

2.4.3.1. Apriori Algorithm

Apriori is perhaps the earliest and the most known association rule mining algorithm is (Clare, 2003). The algorithm is used for mining association rules. The algorithm is an efficient association rule mining algorithm which explores the level-wise mining property. That is, given a database consisting of tuples, it finds association rules that frequently and reliably predict which items occur together. Since the association algorithm results in several rules, it is imperative that mining be limited by using certain parameters so that only interesting association rules with high coverage will be found (Agrawal et al, 1993).

Thus, two notions characterize the association rule, namely support and confidence. Support of an association rule refers to the number of instances they predict correctly While Confidence of an association rule refers to the same number expressed as a proportion of the number of instances that the rule applies to.

There are relationships between particular association rules. That is, some rules imply others. To minimize the number of rules that are generated, it makes sense to present to the user only the strongest one in cases where several rules are generated.

The Apriori algorithm works in the following manner.

Step 1: Finding frequent item sets

In this stage the algorithm focuses on generating item sets (any pair- can be one- of the attributes values) that satisfy minimum coverage. That is, each support value of these frequent item sets will be at least equal to a pre-determined minimum support.

Apriori iteratively searches frequent itemsets: at each iteration k, F_k, it identifies the set of all the itemsets of k items (k-itemsets) that are frequent. In order to generate F_k, a candidate set C_k of potentially frequent itemsets is first built. By construction, C_k is a superset of F_k. Thus, to discover frequent k-itemsets, the support of all candidate sets is computed by scanning the entire transaction database D. All the candidates with minimum support are then included in F_k, and the next iteration is started. The algorithm terminates when F_k becomes empty, i.e. when no frequent set of k or more items is present in the database.

It is worth considering that the computational cost of the k-th iteration of Apriori strictly depends on both the cardinality of C_k and the size of D. In fact, the number of possible candidates is, in principle, exponential in the number m of items appearing in the various transactions of D. Apriori considerably reduces the number of candidate sets on the basis of a simple but very effective observation: a k-itemset can be frequent only if all its subsets of k-1 items are frequent.

C_k is thus built at each iteration as the set of all k-itemsets whose subsets of k-1 items are all included in F_{k-1}. Conversely, k-itemsets that contain at least one infrequent (k - 1)-itemset are not included in C_k.

Each operation involves a pass through the dataset to count the items in each set, and after the pass the surviving itemsets are stored in a hash table- a standard data structure that allows elements stored in it to be retrieved very quickly.

Step 2: Generating strong association rules from the frequent itemsets

The second step of the algorithm focuses on producing rules that meet minimum accuracy. These rules must be the frequent itemsets and must satisfy minimum support and minimum confidence.

This phase of the procedure takes each itemset and generates rules from it, checking that they have the specified minimum accuracy. If only rules with a single test on the right-hand side were sought, it would be simply a matter of considering each condition in turn as the consequent of the rule, deleting it from the item set, and dividing the coverage of the entire itemset by the coverage of the resulting subset- obtained from the hash table- to yield the accuracy of the corresponding rule.

2.5. Scope of Data Mining Applications

The applications of data mining can be generic or domain specific. The generic application is required to be an intelligent system that by its own can take certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take on its own these decisions but are guided by users for selection of data, selection of data mining method and for the interpretation of the results. The multi agent based data mining application (Baazaoui et al, 2005) has capability of automatic selection of data mining technique to be applied. The Multi Agent System used at different levels: First, at the level of concept hierarchy definition then at the result level to present the best adapted decision to the user. This decision is stored in knowledge Base to be used in a later decision-making. Multi Agent System Tool used for generic data mining system development (Botia et al, 1998) uses different agents to perform different tasks. Generic systems are required to integrate as many learning algorithms as possible and decide the most appropriate algorithm to use.

The domain specific applications are focused to use the domain specific data and data mining algorithm that are targeted for specific objective. The applications studied in this context are aimed to generate the specific knowledge. In the different domains the data generating sources generate different type of data. Data can be from a simple text, numbers to more complex audio-video data. To mine the patterns and thus generate knowledge from this data, different types of data mining algorithms are used. The collection and selection of context specific data and applying the data mining algorithm to generate the context specific knowledge is thus a skillful job. In many of the domain specific data mining applications the domain experts plays vital role to specify interesting patterns and mine useful knowledge.

2.5.1. Application of data mining in different domains

Medical science is one of the areas for application of data mining. Diagnosis of disease, health care, patient profiling and history generation etc. are the few examples. Mammography is the method used in breast cancer detection. Radiologists face lot of difficulties in detection of tumors. Computer-aided methods could assist medical staff and improve the accuracy of detection (Antonie et al, 2001). Neural networks with back-propagation and association rule mining is used for tumor classification in mammograms. The data mining effectively used in the diagnosis of lung abnormality that may be cancerous or benign (Kusiak et al, 2000). The data mining algorithms significantly reduce patient's risks and diagnosis costs. The use of data mining in health care is the widely used application of data mining even though medical data is complex and difficult to analyze.

Data mining methods are also used to provide learners with real-time adaptive feedback on the nature and patterns of their on-line communication while learning collaboratively (*Anjewierden et al, 2007*). This makes it possible to increase the awareness of learners. The application of data mining methods to educational chats is both feasible and can bring the improvement in learning environments.

Data mining helps software maintenance engineers to comprehend the structure of a software system and assess its maintainability. The clustering algorithm effectively used

to produce overviews of systems by creating mutually exclusive groups of classes, member data or methods, according to their similarities and hence reduces the time required to understand the overall system. This method also helps in discovering programming patterns and “unusual” or outlier cases which may require attention.

Anomaly detection in the Network is very difficult and needs a very close watch on the data traffic. Intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic as normal or abnormal traffic. (Cai and Li, 2004). If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

A malicious executable is threat to system’s security, it damage a system or obtain sensitive information without the user’s permission. Data mining methods can be used to accurately detect malicious executables before they run (Schultz et al 2001). Classification algorithms RIPPER, Naïve Bayes, and a Multi-Classifier system are used to detect new malicious executables.

Sports are ideal for application of data mining tools and techniques. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning (Chapman et at, 2000). Data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport in a season, tour or game.

Data mining makes it easy, convenient and practical to explore very large databases for organizations. The different data mining techniques are used in crime data mining. (chen et at 2003, chen et al, 2004) Entity extraction used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports. Clustering techniques are used to automatically associate different objects such as persons, organizations, vehicles etc. in crime records. Deviation detection is applied in fraud

detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. Classification is used to detect email spamming and find authors who send out unsolicited emails. String comparator is used to detect deceptive information in criminal record. Social network analysis is used to analyze criminals' roles and associations among entities in a criminal network.

Bankruptcy is the major threat to the banking sector (*Foster and Stine, 2004)*. It increases the cost of lending. Data mining algorithms are effectively used for prediction of personal bankruptcy. Predicting bankruptcy has become the province of computer science rather than statistics. Data mining method least squares regression; neural nets and decision trees are proved to be the suitable for prediction of bankruptcy. Therefore, data mining can be applied in almost all areas.

CHAPTER THREE

Crimes and Application of Data Mining in Crime Records

This chapter tried to review some of the researches conducted on violence against children (sexual offence, rape, sexual assault, sexual violence, and other) or data mining that concentrate on offence (violence) against children and related crimes in general. This helps easily realize the focus the study and made the researcher familiar with the application of data mining for the problem domain.

3.1. Crime

Crime is defined as an act or omission of an act, which is punishable by criminal law. Criminal law, on the other hand, refers to a body of specific rules regarding human conduct, which have been explicitly stated by political authority. However, an act that is considered as a crime in one place and time may not hold true in another place or time.

According to (Andargachew 1988), a criminal is an individual person who has violated the legally forbidden act. In fact, there are some factors that have to be taken into account to convict whether a person should be considered as a criminal or not. Among these, an individual should be of competent age in light with the law of the land; and there must be a well-predefined punishment for the particular act committed.

Other authors defined Crime as a comprehensive concept that is defined in both legal and non-legal sense (Akpinar and Usul, 2004). From the legal point of view crime is the breaking or breaching of the criminal law (penal code) that governs a particular geographical area (jurisdiction) aimed at protecting the lives, property and the rights of citizens of belonging to that jurisdiction. Crime is an offence against a person (for example murder, and sexual assault), or his/her property (for example, theft and property damage) or the State regulation (for example traffic violations) (Akpinar and Usul 2004). In non-legal terms crime is a set of acts that violate socially accepted rules of human ethical or moral behavior (Akpinar and Usul 2004); for example acting against a custom in some society.

There are several causes for the growing rate of crime in a specific place or country. These include unemployment, economic backwardness, over population, illiteracy and inadequate equipment of the police force.

3.2. Crime Type

Crime occurs in a variety of forms which police informally categorizes as being either major or volume. Major crime consists of the high profile crimes such as murder, armed robbery and rape. These crimes can either be one-offs or serial. In the case of serial crimes it is relatively easy to link crimes together due to clear similarities in terms of modus operandi or descriptions of offenders. This linking is possible due to the comparatively low volume of such crimes. Major crimes usually have a team of detectives allocated to conduct the investigation. In contrast volume crimes such as burglary and shoplifting are far more prevalent. They are usually serial in nature as offenders go on to commit many such crimes. Property crimes, such as domestic burglary offences, committed by different individuals are highly similar and it is rare to have a description of the offenders (Adderley and Musgrove, 2001). Table 3.1 shows the classification of crime (Chen et al., 2003).

3.2.1. Violence against Children

Working definition of violence: for the UN study, the independent expert Professor Pinheiro has indicated that he used a broad definition of violence. Accordingly, violence against children is defined to include all forms of physical or mental violence, injury, abuse, neglect and negligent treatment, maltreatment, deprivation and exploitation, including sexual abuse. In this research, this definition of violence is adopted.

Sexual violence is any sexual act, attempt to obtain a sexual act, unwanted sexual comments and advances or acts to traffic or otherwise directed against a person's sexuality using coercion, by any person regardless of their relationship to the victim.

Sexual violence occurs throughout the world. Although in most countries there has been little research conducted on the problem, available data suggest that in some countries

nearly one in four women may experience sexual violence by an intimate partner, and up to one-third of adolescent girls report their first sexual experience as being forced .

Crime Type	Description	
Traffic Violations	Driving under the influence of alcohol, fatal/personal injury/property damage traffic accident, road rage	--
Sex crime	Sexual offences, Sexual abuse, rape, sexual assault, child molestation, Child pornography, prostitution	Trafficking in women and children for sexual exploitation, including prostitution and pornography
Fraud	Forgery and counterfeiting, frauds, embezzlement, identity deception Money laundering, counterfeiting, insurance fraud, fraud, and corruption; corruption and bribery, misappropriation of assets	Transnational money laundering, fraud, and corruption; trafficking in stolen software, music, movies, and other intellectual property
Arson	Arson on buildings	
Gang / drug offences	Narcotic drug offences (sales or possession)	Transnational drug trafficking, organized racketeering and extortion, people smuggling
Violent crime	Criminal Homicide, armed robbery, aggravated assault, other assaults	Terrorism, air and maritime piracy, bombings
Cyber crime	Internet frauds, illegal trading, network intrusion /hacking, virus spreading, hate crimes, cyber piracy, cyber pornography, cyber-terrorism, theft of confidential information.	

Table 3. 1 Crime types at different levels. Source: (Chen et al, 2003)

Sexual offenders often target children with particular characteristics (Mitchell et al, 2001). These may be children in the care of the state; children who have experienced prior

maltreatment; emotionally immature children with learning or social difficulties and problems with peer friendships; love or attention deprived children; children with strong respect for adult status; children from single parent families; children who will co-operate for a desired reward (such as money, computer games); and, children with low self esteem. A study of children aged 10-17 years in developed countries like US found that children over 14 years who were "troubled" (defined as being exposed to negative life events, maltreated and/or depressed) were more likely to be solicited (Mitchell et al. 2001).

Various terms have been used to describe the different offences recognized by law. They include: rape, sexual assault, sexual violence, and defilement. There appears to be no consensus on the exact elements constituting these specific sexual offences across countries. "Rape" is the most commonly recognized offence, which has been outlawed in most sub-Saharan African countries. Common definitions of rape consider it to include penetration of bodily orifices, without consent, utilizing a sexual organ or other object.

Offences against children can be directed against both male and female, however, under age children are main focuses of this research. The research tries to address the pattern in the various forms of sexual violence against both, as well as other offences directed against children by other people and parents (caregivers).

Sexual violence is defined as any sexual act, attempt to obtain a sexual act, unwanted sexual comments or advances, or acts to traffic, or otherwise directed, against a person's sexuality using coercion, by any person regardless of their relationship to the victim, in any setting, including but not limited to home and work.

Coercion can cover a whole spectrum of degrees of force. Apart from physical force, it may involve psychological intimidation, blackmail or other threats for instance, the threat of physical harm, of being dismissed from a job or of not obtaining a job that is sought. It may also occur when the person aggressed is unable to give consent for instance, while drunk, drugged, asleep or mentally incapable of understanding the situation.

The attempt to do so is known as attempted rape. Rape of a person by two or more perpetrators is known as gang rape. Especially in developed countries, Sexual violence

can include other forms of assault involving a sexual organ, including coerced contact between the mouth and penis, vulva or anus.

A wide range of sexually violent acts can take place in different circumstances and settings. These include, rape within marriage or dating relationships; rape by strangers; systematic rape during armed conflict; unwanted sexual advances or sexual harassment, including demanding sex in return for favors; sexual abuse of mentally or physically disabled people; sexual abuse of children; forced marriage or cohabitation, including the marriage of children; denial of the right to use contraception or to adopt other measures to protect against sexually transmitted diseases; forced abortion; violent acts against the sexual integrity of women, including female genital mutilation and obligatory inspections for virginity; Forced prostitution and trafficking of people for the purpose of sexual exploitation.

There is no universally accepted definition of trafficking for sexual exploitation. The term encompasses the organized movement of people, usually women, between countries and within countries for sex work. Such trafficking also includes coercing a migrant into a sexual act as a condition of allowing or arranging the migration.
Sexual trafficking uses physical coercion, deception and bondage incurred through forced debt. Trafficked women and children, for instance, are often promised work in the domestic or service industry, but instead are usually taken to brothels where their passports and other identification papers are confiscated. They may be beaten or locked up and promised their freedom only after earning − through prostitution − their purchase price, as well as their travel and visa costs.

Data on sexual violence typically come from police, clinical settings, nongovernmental organizations and survey research. The relationship between these sources and the global magnitude of the problem of sexual violence may be viewed as corresponding to an iceberg floating in water World Health Organization report in 1996. The small visible tip represents cases reported to police. A larger section may be elucidated through survey research and the work of nongovernmental organizations. But beneath the surface remains a substantial although un-quantified component of the problem.

In general, sexual violence has been a neglected area of research. The available data are scanty and fragmented. Police data, for instance, are often incomplete and limited. Many women do not report sexual violence to police because they are ashamed, or fear being blamed, not believed or otherwise mistreated. Data from medico-legal clinics, on the other hand, may be biased towards the more violent incidents of sexual abuse. The proportion of women who seek medical services for immediate problems related to sexual violence is also relatively small.

Although there have been considerable advances over the past decade in measuring the phenomenon through survey research, the definitions used have varied considerably across studies. There are also significant differences across cultures in the willingness to disclose sexual violence to researchers. Caution is therefore needed when making global comparisons of the prevalence of sexual violence.

3.3. Crime recoding systems (police records)

Whenever a crime is committed, a police officer visits the crime scene or the report is taken by telephone or person may come to the police office and report, which is known as the crime report. All police forces record their crime reports in a similar way. They also ask necessary information to serve as evidence for investigation when the report is conducted by other individual(s).

In both computerized and manual record of crime data variables stored may be known in a variety of ways but comprise the following: -

- Time, day and date of the crime
- Offence type
- Location of crime.
- Victim information
- Offender information
- Modus operandi (MO) identifies how the crime has been committed.

Depending upon the crime recording system used by each individual force the data fields will be a mixture of structured data fields and free text fields. The free text may not even contain key words or phrases and will contain non standard abbreviations, miss-spellings and, on occasion, contradictory information.

The aim of collecting the data by law reinforcement agencies is to solve the crime and to provide required performance information rather than to create a research database. Thus, the quantity and quality of information recorded varies considerably from case to case. It is often imprecise, and is almost certainly at times inaccurate. Crime data is very noisy (random error or variance in a measured variable) and contains lots of missing values. Unstructured and inconsistent data formats make it very complicated to automate the analytical processes.

3.4. Theories of Environmental Criminology

Understanding the behavior of offenders plays a significant role in understanding and predicting crime and criminality (Adderley and Musgrove 2003). It would, therefore, be helpful to be familiar with the theories of environmental criminology.

3.4.1. Routine Activity Theory

According to (Cohen and Felson 1979), the union of three elements in time and space are required for a crime to occur: a likely offender, a suitable target and the absence of a capable guardian against crime. This theory is summarized in Figure 3.1 below. Policing traditionally focuses upon the offender part of the triangle but crimes could be prevented or reduced by interacting with any aspect of the triangle.

Figure 3. 1: Diagrammatic representation of routine activity theory

3.4.2. Rational Choice Perspective Theory

This theory focuses upon the offender's decision making processes. Its main hypothesis is that offending is purposive behavior which helps the offender in some way. It believes that an offender has an objective to commit a crime even if these goals are immediate and consider only a few benefits and risk at a time (Clarke and Felson, 1993). For example, if mobile telephone or wallet is visible in a car and no one is around it might tempt an offender to grasp the opportunity.

3.4.3. Awareness Theory

(Brantingham and Brantingham, 1991) has suggested that crime has four dimensions. These include victim, offender, geo-temporal and legal. The spatial (environment) element of crime is significant to understand the behavior of offenders. A crime's space can be chosen either on purpose or accidentally by either the victim or the offender according to their life styles. Several things have an effect on the crime rate of an area. For example, what type of people live in particular space and what type of security is available.

The first two theories which are used for forecasting purposes are the rational choice perspective and routine activities theory. Both assume that crime is purposive and that individuals are self-determining: when people commit crime, they are seeking to benefit themselves, and certain calculations are involved in determining whether the criminal act will yield positive results (Clarke, 1997). Thus, offenders are influenced by situational and environmental features that provide desirable or undesirable offending opportunities. These theories are based upon the belief that criminals engage in rational (if bounded) decision-making and that characteristics of the environment offer clues to the offender that promising opportunities for crime exist (Cohen and Felson, 1979).

The real implications of these theories are that even motivated criminals may nonetheless be deterred from committing crime if they perceive a potential target to be too risky, to involve too much effort, to yield too meager a profit, or induce too much guilt or shame to make the venture worthwhile. From a predictive modeling perspective, then, these theories have the potential to guide the selection of independent variables with a focus on

those that characterize desirable targets and in turn, desirable locations of crime. Further, theory-based modeling enables us to identify which factors influence crime target selection, and thus inform crime prevention efforts. The models described below include an assessment of whether they are supported by theory, and the extent to which they inform prevention efforts.

3.5. Crime and Data Mining

Most, if not all, current systems both manual and computerized revolve around the investigation of crimes already committed. They are, therefore, reactive. In developed countries a majority of crime prevention forces use different types of relational database management systems (RDBMS) for recording and subsequent analysis of crime. Standard or interactive queries are written to produce patterns of crime, offending and various statistics (Adderley and Musgrove, 2001) but it is a common phenomenon in the developing countries to find mainly manual criminal record books used alongside for crime incidence location.

Intelligence Agencies collect and analyze information to investigate criminal's activities. One challenge to law enforcement and intelligent agencies is the difficulty of analyzing large volume of data involve in criminal activities. Data mining makes it easy, convenient and practical to explore very large databases for organizations. Like all other data, police data are classified as structured and unstructured types. In developed countries data mining tools are applied for both structured and unstructured.

Newer techniques identify patterns from both structured and unstructured data (Chau et al, 2002). Entity extraction used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports. This is a technique used to extract objects and their attributes with the help of special extraction tools from police reports or narratives. Thus, officers or analysts can create structured data from free texts. Other data mining techniques such as association analysis, classification and prediction, cluster analysis, and outlier analysis identify patterns in structured data (Han and Kamber, 2001). Clustering techniques used to automatically associate different objects such as persons, organizations, vehicles etc. in crime records. Deviation detection and outlier

analysis are applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. Classification is used to detect email spamming and find authors who send out unsolicited emails. String comparator is used to detect deceptive information in criminal record. Social network analysis used to analyze criminals' roles and associations among entities in a criminal network.

According to (Gupta et al. 2008) crime analysis tools can be integrated with latest visualization techniques such as Geographical Information System for enhancing the understanding of the results and patterns. The tool has very promising use in the current changing scenario and provides an effective tool to law enforcement agencies for crime detection and crime prevention.

Further (Deshpande and Thakare, 2010) stated that, different methods of data mining are used to extract patterns and knowledge from variety of databases. Selection of data and methods for data mining is an important task in the process and needs the knowledge of the domain. Several attempts have been made to design and develop a generic data mining system. However no system is found to be completely generic. Thus, for every domain, the domain expert's assistance is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generated.

When we come to our country few researches are made on crime data mining. These includes classification to predict which crimes are common in rural areas and which are common in urban areas, to classify which crimes are classified as serious, medium, and low crimes and association rule is used to extract the profile of victims.

According to (Woldekidan, 2003) with the help of association rule the learning algorithm was able to generate a number of rules over a series of experiments. On account of subjective (opinions of domain experts) and objective (support and confidence) measures

of interestingness, a number of rules having practical relevance or that can increase to the current knowledge in the problem domain were identified.

Another author in this area (Leul, 2003) using the classification technique concluded that from results of his experiment could be employed as an input for the decision making process on resource deployment, designing training programs, and crime prevention and investigation methods accordingly. In addition to this, detectors could use this model to scrutinize suspected individuals before a depth investigation is commenced. This means that it can be used to supplement operational policing. The researcher has got familiarity with the domain area and its data mining applicability, and concluded that there are problem in law enforcement agencies that can be solved with the help of data mining tools and techniques.

CHAPTER FOUR

Data understanding and Data Preparation

This chapter deals with the identification of source of data and preprocessing (cleaning, integration, reduction, and transformation) of the data to be appropriate for the tool and data mining models and achieve the objective of the study. Different efforts are exerted and techniques used in the process of data preparation.

4.1. Data Collection

The data for the study was drawn from a department in AAPC (Addis Ababa police commission) which is concerned with the investigation of offences committed against children, handling offences committed by children and the prevention of crimes through awareness creation. Before the actual preprocessing task was started, the necessary data were extracted from the department's MS-Excel database. The data on MS-Excel was in Amharic language.

The data are classified by the year at which the offence was committed or reported. The data includes all offence crimes against children and crime committed by children in the Addis Ababa ten sub-cities for the years between 1997 and 2002 E.C. In our society, some offences are not considered as crimes and never reported to police. This data does not include unreported ones. The data stored in the FSCE database has 27 attributes and 5355 entries.

Generally the data set is represented as n × p data matrix. The n rows represent the crime events on which measurements of the *p* attributes were taken. The rows of data matrix represent the records whereas the columns represent the *variables*, *features*, *attributes*, or *fields* of the data matrix. In all of the four different naming for columns the idea is the same: these names refer to the measurement that is represented by each column.

Though this research initiated to explore offences against both women and children but due to lack of data regarding women, it concentrates only on crimes committed against children such crimes include rape, violence, corporal punishment, buggery, sexual

harassment, abduction, serious bodily injury, etc. As we know privacy and confidentiality are issues to be resolved while we are dealing with police records. Both the victim's and the offender's personal identities require being off the record that contains their personal profile data while exposed to individuals other than law enforcement members like police and lawyers. Due to this there are some attributes which are not allowed to be used by the researcher. These include attributes such as name, telephone number, house number, and, even the Kebele (names for special locations) where the victim (s) and the offender lives.

According to (Han and Kamber, 2000) data quality can be measured in terms of accuracy, completeness, consistency, timeliness, believability and interpretability. Especially police data suffer from incompleteness and inconsistency problems. Since data quality can also be affected by the structure of the data being analyzed, the researcher gave due emphasis to this aspect of the data too. The lack of data standards in using abbreviations and human error are significantly available and need to be cleaned and standardized before applying data mining techniques. To improve data quality, it is necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database (e.g., ensuring that "no attribute" is represented as a "_" or any equivalent value to "empty" throughout the database), accounting for missing data points, removing unneeded data fields, identifying anomalous data points (e.g., an individual whose age is shown as 142 years or child greater than 18 years old etc), and standardizing data formats (e.g., changing dates so they all include MM/DD/YYYY).

To have clear understanding of the attributes used, the attributes are categorized in to three groups:

- Crime related: such as name of crime as crime type, sub-city, decision and year.
- Victim profile: sex, age, education, habit, religion, marital status, job.
- Offender profile: age.

4.2. Data Preprocessing

There are a number of data preprocessing techniques. Data cleaning for example can be applied to solve problems related with the data itself through removing noise and

correcting inconsistencies. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization and concept hierarchy development are applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction is another data preprocessing technique that reduces the data size by aggregating, eliminating redundant features, clustering, or sampling. These data preprocessing techniques are preferable to be applied prior to mining in order to substantially improve the overall data mining results. Almost all of the data preprocessing techniques are done using MS-Excel built-in functions like search and replace, filtering, and auto fill mechanisms.

4.2.1. Data Cleaning

The researcher tried to assess such problems which require due emphasis while making the data ready for applying data integration and other different data reduction techniques. Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2001). Thus, data cleaning has become a must in order to improve the quality of data so as to improve the accuracy and efficiency of the data mining techniques.

- **Outliers:-** are misleading data that do not fit to most of the data/facts in the entire data. Since most of the data have nominal value this is not a problem for the data used in this research. The attribute age for both the victim and the offender can have outlier values but both are written in interval format like, 0-9, 9-16, 16-18, for the victims and 9-15, 16-18, 16-20, 20-30, 21-30, 31-45 and 45 and above. The age for the victim starts from 0 to mean offences against infants and children below a year like abortion. The upper value excludes 18, since individuals with the age of 18 or greater are adults.
- **Missing data: -** attribute values might be absent. One can use different strategies such as removing, replacing with estimated values or treating as they are. Removing can be either the entire row that contains missing value (e.g. if the technique to be used is classification and the missing is on the class attribute, or

the row contains too much missing etc), or the entire column when it contains more than 30% missing values. Accordingly, all the 13 attributes satisfy this criterion. Although all the attributes except, crime type and sub-city have missing values represented as "ያልተገለፀ" to mean "unknown" or "not mentioned" or "no information" is available regarding this value. The fields that do not contain data are stored in the database as "unknown". It is not uncommon for data sets to have fields that contain unknown or incorrectly entered information and missing values.

However, it is important to note that the absence of data may itself be valid. Missing value of field entries within the data set were classified as "Unknown". Within this study "Unknown" fields were set to the value "?" when used to encode the variables. Out of the 13 attribute selected as relevant for the mining problem at hand, the attribute education has highest number of missing values. From the six distinct values the missing constitutes 10%.

- **Noisy dat**a: attribute values that might be invalid or incorrect. E.g. typographical errors. The typographic errors like spelling error such as ሌብት instead of ሌብነት which is to mean theft and ቤተሰብ instead of ከቤተሰብ to mean with-family while dealing at word level and they are corrected through consulting the database administrator.
- **Inconsistent data:-** containing discrepancies in codes or names, which is also the problem of lack of standardization or naming conventions. In this database the data are classified by year of report. The values of a particular attribute in the different years are different regarding naming and also abbreviation. The values of the attribute job contains words such as "የሌለው" and "ሰራአጥ" similar like that of jobless and unemployed which talks about the same idea but expressed using two different words even with in the same year of report. There are abbreviations like "አ/መድፈር to mean "አስገድዶ መድፈር" (forced rape) and "ግ/ሰዶም" to represent "ግብረ ሰዶም" (buggery) which are known only by the database administrator. All of these naming and abbreviations are converted into a common naming and format.

Similar with this the attribute education contains words like "ያልተማረ" and "ያልጀመረ" to say illiterate and never-joined-school. Special names for the areas- this information is also highly exposed to the problem of inconsistency, as it is common for a place to be referred using different names depending on the perception of the person responding to it("Sidst Kilo", "Menen School", "Yekatite 12 hospital", "Anbesa Gibi","Dibab", "Kenya Embassy", etc). While some might refer a place using the broad or more inclusive name for the area that includes several other specific places, others might choose to use the specific name of the place.

4.2.2. Data Integration

Data integration is a preprocessing task of combining data from multiple sources, databases, data warehouses, or different files structures. As a result of this combination process data that is fine on its own can become problematic because of different formats and structures, conflicting and redundant data and data at different levels of detail. In the data used in the current study there exit difference in structure. In some of the years the attribute age of victim comes before the attribute education, and there also exists a difference in age ranges of the offender. In some of the years it ranges from 9-15, 16-18, 16-20, 20-30, 31-45, and >45 and in some other years it ranges from 9-15, 16-20, 21-30, 31-45, and >45. These differences in format are solved by communicating with domain experts and the database administrator. The ranges are represented as 9-18, 18-30, 30-45 and above 45. After integrating the data for the year 1997-2002 E.C, a new important attribute year with six distinct values was derived. As shown in section 4.3.1 year is an attribute that holds true for all records when our intension is to extract patters for a single year. But integration allows it to have six different values "1997", "1998", "1999", "2000", "2001" and "2002".

4.2.3. Data Reduction

Data reduction is finding a way to reduce the size of the data set without affecting the data mining results. An approximate smaller data set is created that can then be made accessible for example in main memory by the data mining algorithm instead of dealing with the full data residing on disk (Hand et al, 2001). Data reduction techniques such as

data cube aggregation, dimension reduction, data compression, numerosity reduction, and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of information content. This general approach can, of course, only approximate the results we would have obtained had the algorithm been run on the full data. However, if the approximate data set is constructed in a clever enough manner, we can often get almost the same results on only a fraction of the data. No data compression technique was applied in this research.

4.2.3.1. Dimensionality Reduction

Dimensionality reduction is a data preprocessing techniques where irrelevant, weakly relevant or redundant attributes or dimensions are detected and removed. The selection of the most interesting attributes (or conversely, the removal of noisy variables or irrelevant attributes) among a pool of possible candidates may improve the quality of the output generated, reducing the chance of classifier over fitting and increasing its overall accuracy and visualization. These include irrelevant attributes, duplicate attribute, attributes with many different values and attributes with non-variant values.

- **Irrelevant data** are attributes in the database that might not be of interest to the data mining task being developed. This may be either it has the same value for all records or has distinct values for all records. Additionally, it can also have the same information content with other attribute(s) which is known as either duplicate records or duplicate attributes. In all cases the data is of no or low interest to generate rules and pattern. For example entry number is the order of reporting of the crime to police or recording to the database and since it serves as a primary key to identify the sequence of the crimes, no two entities have the same entry number.
- **Attributes with same information content: -** when fields with the same information content were encountered, only one of the fields is considered. For instance, Code of crime and type of crime holds the same information. In most cases, the police department uses code of crime for the purpose of investigation, while for reporting they used type of crime. Since this research deals with understanding crime patterns based on the reports on crimes committed against

children, the researcher preferred to use crime type than crime code. This means the attribute crime code is irrelevant (duplicate) column and was discarded.

- **Non-variant fields: -** Attributes that take values that holds true for all the records in the database are expected to be dropped, for instance, the attributes habit, marital status and year, have non-variant fields. As the researcher ties to introduce in the introduction part of this chapter the data are classified based on the year when the crime is reported. In some of the years the attribute values for these three variables are the same. Even though this does not hold true for the entire database, it creates a problem when we process data of a single years. For example if we process the data for the year 1999 E.C, the attribute habit has 940 of the records has the same value "no-habit" and 140 with value "ያለተገለፀ" which is to mean not-mentioned out of the 1080 records registered in that year. Accordingly if we are forced to replace the undefined with estimated value, according to the assumption by the domain experts, is that they assume that the criminals at that year were free from habits i.e. better to replace it with "no-habit". Even though this attribute was considered as being irrelevant and was excluded from the year 1999 EC, it is relevant when considered in the year 2000 EC and has values "drug-addicted", "chewing-chat", "smoking" as these values are associated with being street child and exposed to offences like rape and buggery.
- **Attributes taking many different values: -** The following attributes were also excluded since they take many different values. This is done with the intention to improve the speed, accuracy of analysis and training. Examples include:
- Name of the victim and name of the offender – even though this has the problem of taking many different values, in order to keep confidentiality and privacy of the data the names of the victim children and their offenders were excluded from the data exploration process.
- There are over 300 Kebeles and ten sub-cities in Addis Ababa. The values for the kebeles is too detail data especially for association rule mining, which is computationally expensive. Having more attribute values will make the exploration task difficult. The purpose of this research is not only to discover general rules that depict regularities or summarize the records in the database in

relation to the victim profile and age of the offender, but also in relation to year when the offence was committed and address where the crime was committed related to the attribute sub-city.

- The time when the offence was committed- time of the day at which crime event was committed on the victim child.

4.2.3.2. Numerosity Reduction

Numerosity reduction is where the data are replaced or estimated by alternative, smaller data representations. In this research the whole dataset in the database is used since the dataset size does not affect scalability of the algorithm. What is more, as long as the data can be handled by the learning algorithm, it is better to use the whole dataset, as it increases the performance of the algorithm since it has more examples to learn from.

4.2.4. Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for the mining model. Data has different formats. Fore example, age is written as ranges and need to be converted to some concepts; crimes names are written as phrases and are encoded as six offence categories (offe1, off12, offe3, offe4, offe5, offe6). Firstly the data was in Amharic language using MS-Excel application. As we know MS-Excel can accept phrasal words, words with any character and never specify data types. Due to this the data require specifying their data type converting to values which can be easily processed and accepted by WEKA tool.

4.2.4.1. Concept Hierarchy

Concept hierarchy can be defined as a partial order set. Given two concepts a and b belonging to a partial order relation R, i.e., (a, b) $\in$ R (described by a $\leq$ b, we say that concept a is more specific than concept b, or that b is more general than a. Usually partial order relation in a concept hierarchy represents the special-general relationship between concepts, also called subset-superset relation.

A tree is a special type of concept hierarchy, where a concept precedes only one concept and the notion of greatest concept exists, i.e., a concept that does not precede anyone. The tree root will be the most general concept, called ANY, and the leaves will be the attribute values in the database, that is, the lowest abstraction level of the hierarchy. In this work, we will use concept hierarchies that can be represented as a tree.

In this research the researcher used concepts for ages of both the victim and the offender. Especially the age of the victim is between 0-18 years. 0 is inclusive to mean infants subjected to abortion and children below a year. The value 18 is exclusive because 18 years aged individuals are not included in the FSCE database. This data was divided in to two intervals having 0-9 and 9-18 values and encoded to ag1 and ag2 respectively.

While the minimum and maximum of victims' ages are known, the boundaries of offenders' ages especially the upper limit is not known since it simply says above 45 years. The value of the lower limit is 9. Age values are registered as ranges 9-15, 16-18, 18-20, 20-30, 21-30, 30-45, 31-45 and above 45. Taking this data into account the researcher has divided these into three categories. Under age 9-18 as children, 18-45 as middle and above 45 as aged offenders and encoded as oag1, oag2 oag3 respectively. Some difficulty with this grouping was in the first intervals like 16-20 which can be either in children category or middle age category. This was solved using some probability. Since the range has 5 value (16, 17, 18, 19, 20) in it, 2/5 0f the data was selected as below 18 and 3/5 of them was selected as middle age. The approach adopted with interval data age was to partition each range in to a number of intervals (Connell & Brady 1985).

3.1. Final Selected Attributes

Attributes can be ranked either manually with the help of domain experts or automatically by using the application tools like WEKA attribute selection. From the WEKA attribute selection criteria the most common ones which are used for decision tree building are information gain attribute evaluation and gain ratio attribute evaluation. The researcher prefers to use information gain. Information gain works by calculating gain for each

attribute in relation to the entropy of class variable. The rank of the 13 selected attributes based on their information gain is presented in (figure 4.1) and their data type with the description of each attributes is presented in (table 4.1) below.

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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 cluster):
        Information Gain Ranking Filter

Ranked attributes:
 0.346374    3 Sex
 0.311359    1 typeofcrime
 0.162417    4 Education
 0.153624    5 Year
 0.142562    2 ageofvictim
 0.135496   11 Offenderage
 0.079775   13 subcity
 0.040359   12 decision
 0.004957   10 Living
 0.001938    8 MaritalStatus
 0.001395    9 Religion
 0.000604    7 Job
 0.000263    6 Habit

Selected attributes: 3,1,4,5,2,11,13,12,10,8,9,7,6 : 13
```

Figure 4. 1 Rank of attributes using information gain attributes selection method.

Rank	Attribute name	Data type	Description
1	Sex	Nominal	Gender of the victim
2	Martal_Status	Nominal	Marital status of the victim
3	Decision	Nominal	Decision given to the criminal act
4	Job	Nominal	Job of the victim
5	Sub-City	Nominal	The sub-city where the crime was committed
6	Living	Nominal	Living situation of the victim
7	Habit	Nominal	Special habits with the victims
8	Year	Number	There year when the crime was committed
9	Religion	Nominal	Religion of the victim
10	Education	Nominal	Educational status of thee victim
11	Age-of-victim	Nominal	Age of the victim
12	Offender-age	Nominal	Age of the offender
13	Crime-type	Nominal	Class variable (predictable state)

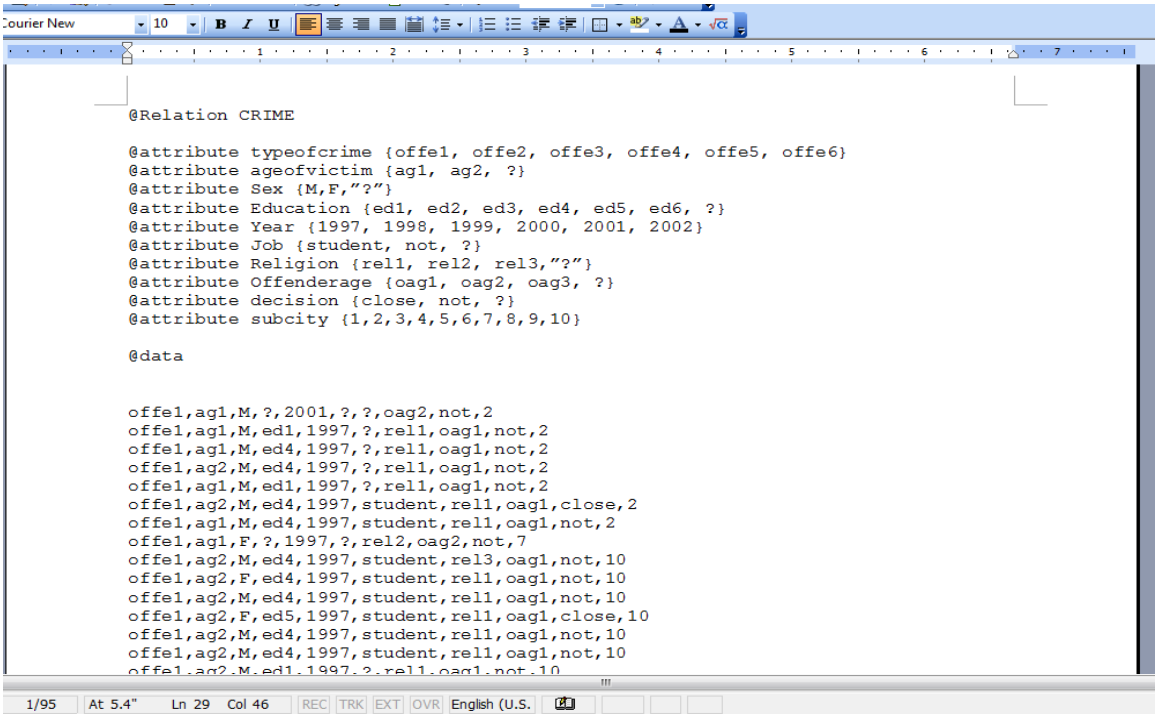
Table 4. 1 Description of the 13 selected attributes.

4.3. Converting into the Final Dataset Format

The fields in the database were tab separated. The database was opened in Excel. Since WEKA data mining techniques association rule, which is one of the selected data mining algorithm, is often used in situations where attributes are nominal, the numeric values (age) was converted to nominal, and certain similar categories were merged. There were missing data, that is, fields that were left unfilled or contain values to mean missing like “Unknown”. Missing values were replaced with the most frequent value or modal value except for some not nominal attributes, as it is one of the proper ways of representing such values in WEKA. As mentioned above the values “unknown”, and “not mentioned” were also considered as missing values.

The data was then saved in a .CSV format which is a format where commas are placed between values in adjacent columns or commas are replacing empty spaces and tabs. The database was then opened in Word, header information added. That is, the symbol @ was

placed in front of the key words “Relation”, “Attribute” and “Data” which means “@Relation” has been placed in front of the relation name, @Attribute has been placed in front of each attribute declarations and “@Data” has been placed in front of dataset values. Finally, during saving the file extension was changed to .arff. The following figure indicates the representation of the data in ARFF format.



```
@Relation CRIME

@attribute typeofcrime {offe1, offe2, offe3, offe4, offe5, offe6}
@attribute ageofvictim {ag1, ag2, ?}
@attribute Sex {M,F,"?"}
@attribute Education {ed1, ed2, ed3, ed4, ed5, ed6, ?}
@attribute Year {1997, 1998, 1999, 2000, 2001, 2002}
@attribute Job {student, not, ?}
@attribute Religion {rel1, rel2, rel3,"?"}
@attribute Offenderage {oag1, oag2, oag3, ?}
@attribute decision {close, not, ?}
@attribute subcity {1,2,3,4,5,6,7,8,9,10}

@data


offe1,ag1,M,?,2001,?,?,oag2,not,2
offe1,ag1,M,ed1,1997,?,rel1,oag1,not,2
offe1,ag1,M,ed4,1997,?,rel1,oag1,not,2
offe1,ag2,M,ed4,1997,?,rel1,oag1,not,2
offe1,ag1,M,ed1,1997,?,rel1,oag1,not,2
offe1,ag2,M,ed4,1997,student,rel1,oag1,close,2
offe1,ag1,M,ed4,1997,student,rel1,oag1,not,2
offe1,ag1,F,?,1997,?,rel2,oag2,not,7
offe1,ag2,M,ed4,1997,student,rel3,oag1,not,10
offe1,ag2,F,ed4,1997,student,rel1,oag1,not,10
offe1,ag2,M,ed4,1997,student,rel1,oag1,not,10
offe1,ag2,F,ed5,1997,student,rel1,oag1,close,10
offe1,ag2,M,ed4,1997,student,rel1,oag1,not,10
offe1,ag2,M,ed4,1997,student,rel1,oag1,not,10
```

Figure 4. 2 Dataset representations in .ARFF format

CHAPTER FIVE

Model Building and Model Evaluation

The goal of this chapter is briefly discussing the techniques used, interpreting and evaluating experimental results. As discussed in chapter one the researcher has used the Association rule mining techniques to describe the relationship among the selected attributes which shows the profiles of the victim children. In the clustering model the whole dataset was divided into clusters by setting K values from 3 to 5 and then deal with the classification model. Classification model is used in order to generating rules and show which children are exposed to particular crime categories. The tree models are used to support one another and should not be conflicting.

5.1. Model building

Model Implementation and Tool Selection are tasks that should be discussed in short in this chapter. In this part of the research the researcher is interested to explain the models selected to achieve the mining goals. As mentioned in the methodology section in chapter one, the problems or models addressed are association rule, clustering and classification. This includes data mining tool selection and the algorithms used for modeling technique. The classification modeling technique has used the clustered dataset as an input and implemented using decision tree (J48 classification tree). Decision trees are easy for interpretation and conversion into rules.

Generally decision tree has the following advantages: decision tree methods tend to produce models that are easy to interpret, its methods have a built-in feature selection method that makes them immune to the presence of useless variables, and are very adept at revealing complex interactions between variables. Each branch of a tree can contain different combinations of variables and the same variable can appear more than once in different parts of the tree. This can reveal how a variable can depend on another and in what specific context this dependency exists and tree models offer several ways of dealing with missing values that can often minimize or eliminate the effect of such values on

model performance. Due to this, decision tree is selected over the other classification methods like Bayes and artificial neural network.

For the purpose of association rule modeling technique, the algorithm used was WEKA's apriori where as the measures of interestingness of rules are the objectives ones support and confidence threshold values set by the algorithm itself. Unlike classification which searches the relationship between the predictable state or class variable and the other attributes, association rule is to find interesting patterns (relationship) among all the attributes used to build the association model. This algorithm minimizes the problem of setting threshold values by the user themselves.

In association rule mining, the rules appear in the following form:

If Antecedent Then Consequent.

The researcher started by defining four numerical values which can be determined for any rule through counting.

- $N_{Antecedent}$ Number of instances matching the antecedent
- $N_{consequent}$ Number of instances matching the consequent
- $N_{Intersection}$ Number of instances matching both the antecedent and consequent.
- N_{Total} The total number of records in the dataset

These numerical values can be represented using Venn diagram representation (see figure 5.1), a set to visualize the concept clearly.

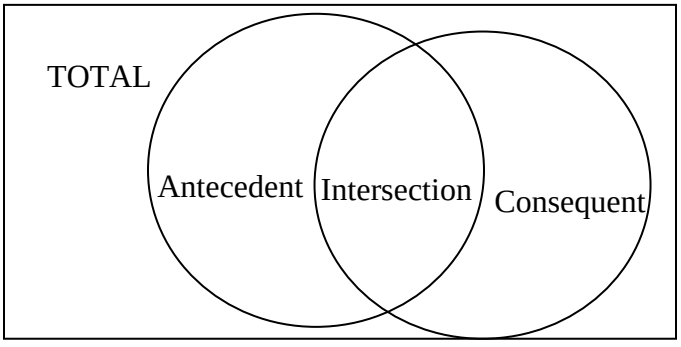


Figure 5. 1 Venn diagram showing Instances matching Antecedent, Consequent and Their intersection

WEKA tool consists of a collection of machine learning algorithms for solving real-world data mining problems. The package has three different interfaces: a command line interface, an Explorer GUI interface (which allows one to try out different preparation, transformation and modeling algorithms on a dataset), and an Experimenter GUI interface (which allows to run different algorithms in batch and to compare the results) (Witten & Frank, 2000).

The knowledge generated by different data mining techniques can be represented using many different ways from which classification and association rule are commonly mentioned for predictive and descriptive data mining models respectively.

5.1.1. Experiments and analysis of association rule

As expressed in chapter two association rules mining is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute values. Here association was applied in crime database to extract the profile of victims in relation to the place where crime are committed and age of offenders.

The WEKA tool implements an Apriori-type algorithm that solves the problem of setting min-support and min-confidence by the user partially. This algorithm reduces iteratively the minimum support, by a factor delta support (Δs) introduced by the user, until a minimum support is reached or a required number of rules has been generated.

The information on running the algorithm on the database is presented in the table below.

Scheme	Meaning
-N(required number of rules output)	20, 10
-T(metric type by which to rank rules)	Confidence
-C (the minimum confidence of a rule)	0.9, 0.8
-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	Crime
-Instances	5355

Table 5. 1 List of parameters used to run the association rule

In this research the 13 attributes and the following best rules were generated.

Experiment #1

```
Apriori
=======

Minimum support: 0.75 (4016 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 5

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 11

Size of set of large itemsets L(3): 7

Size of set of large itemsets L(4): 1

Best rules found:

 1. Job=student Living=famyes 4168 ==> MaritalStatus=single 4145    conf:(0.99)
 2. Job=student 4389 ==> MaritalStatus=single 4364    conf:(0.99)
 3. Habit=no Job=student 4176 ==> MaritalStatus=single 4151    conf:(0.99)
 4. Religion=rel1 Living=famyes 4223 ==> MaritalStatus=single 4186    conf:(0.99)
 5. Habit=no Religion=rel1 Living=famyes 4109 ==> MaritalStatus=single 4072    conf:(0.99)
 6. Habit=no Religion=rel1 4470 ==> MaritalStatus=single 4425    conf:(0.99)
 7. Religion=rel1 4600 ==> MaritalStatus=single 4552    conf:(0.99)
 8. Living=famyes 4876 ==> MaritalStatus=single 4807    conf:(0.99)
 9. Habit=no Living=famyes 4622 ==> MaritalStatus=single 4555    conf:(0.99)
10. Habit=no 5065 ==> MaritalStatus=single 4981    conf:(0.98)
11. decision=not 4173 ==> Habit=no 4103    conf:(0.98)
12. MaritalStatus=single decision=not 4102 ==> Habit=no 4033    conf:(0.98)
13. decision=not 4173 ==> MaritalStatus=single 4102    conf:(0.98)
14. Habit=no decision=not 4103 ==> MaritalStatus=single 4033    conf:(0.98)
15. Religion=rel1 Living=famyes 4223 ==> Habit=no 4109    conf:(0.97)
```

Figure 5. 2 ***Association rule model with best association rules found of the first experiment***

As can be seen in the above association rule model, there are six one-item frequent item-set, eleven two-item frequent item-sets, seven three-item frequent item-sets and one four-item frequent item-sets generated that satisfies the criteria 75% min-support, 90% min-

confidence and only 15 of them are displayed since the number of item-sets to be generated is set to 15.

Apriori generates a number of rules that satisfy the above set minimum metrics of support and confidence. If a rule has a metrics value above the minimum thresh hold, then the rule is included in the large item-set.

A large item-set has to satisfy two basic characteristics. First the large item-set property which states that, any subset of a large item-set must be large. Whereas the second the Contra-positive which says that if an item-set is not large, none of its supersets are large.

Four of the twenty rules generated by apriori are presented below (see fig. 5.2. for the complete list of rules generated).

Best rules found:

1. Habit=no Religion=rel1 Living=famyes 4109 ==> MaritalStatus=single 4072 conf:(0.99)

If (Habit=no and Religion=rel1 and Living=famyes then MaritalStatus=single).

The knowledge represented using the “If then” representation is to mean that, since this database contains crime records about victim children, there exists a relationship among the attributes marital-status, religion, living-situation and Habit. Their relation looks like, if the victim has no special habit, following the orthodox religion and lives with her/his family then s/he is single with confidence of **4072 /4109 (99%) and 4072/5355 (76%)** support**.**

2. MaritalStatus=single Religion=rel1 Living=famyes 4186 ==> Habit=no 4072 conf:(0.97)

If (marital status = single and religion=rel1 and living=famyes then habit= no).

This is to mean that if the victim is single, orthodox, and lives with her/his family then s/he has no special habit. Since this database contains crime records about victim children, there exists a relationship among the attributes marital-status, religion, living-situation and Habit. Their relation looks like, the victim is single, following the orthodox religion and lives with her/his family is the antecedent and s/he has no special habit is the consequent relation with confidence of **4072/4186 (97%).** This rule is a large item with highest number of attributes (i.e. four) in the rules found. Being a superset, the subsets of this rule should also hold true.

Even though there are other combinations like, MaritalStatus=single ==> religion = rel1, MaritalStatus = single ==> living = famyes, Religion=rel1 Living=famyes ==> MaritalStatus=single, Habit= no MaritalStatus = single living = famyes ==> Religion = rel1, Habit = no, MaritalStatus = single Religion = rel1 ==> living = famyes etc. some of these are included in the twenty best rules while others are not.

3. Job=student Living=famyes 4168 ==> MaritalStatus=single 4145 conf:(0.99)

If (Job = student and living = famyes then MaritalStatus = single). It is true for a victim that, if s/he is a student and lives with her/his family, s/he is single with 99% certainty and 77% support. This large item-set has three items (Job, Living, MaritalStatus) and is a superset for rule #4 below.

4. Job=student 4389 ==> MaritalStatus=single 4364 conf:(0.99)

If (Job=student then MaritalStatus=single). Because of similar reason, if the victim is a student is single with **81%** support and **4364/4389 (99%)** confidence. This large item-set having two items and is a subset of rule #3 above.

The association rule generated especially using the three attributes (Living, MaritalStatus and Habit) is because of their repeated values habit=no, living=family yes and marital status=singe. To give chance to the other attributes it better to remove the attributes and run using the other 10 attributes.

Experiment #2

Using the 10 attributes by ignoring the three most frequent attributes MaritalStatus, Habit, and living following association model is generated.

```
Apriori
=======

Minimum support: 0.25 (1339 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 40

Size of set of large itemsets L(3): 46

Size of set of large itemsets L(4): 17

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

Best rules found:

 1. Sex=F Education=ed4 1541 ==> Job=student 1417    conf:(0.92)
 2. typeofcrime=offe2 Sex=F 1567 ==> decision=not 1439    conf:(0.92)
 3. typeofcrime=offe2 Religion=rel1 1530 ==> decision=not 1402    conf:(0.92)
 4. typeofcrime=offe2 Job=student 1469 ==> decision=not 1339    conf:(0.91)
 5. ageofvictim=ag2 Sex=F Job=student decision=not 1511 ==> Religion=rel1 1376    conf:(0.91)
 6. typeofcrime=offe2 1744 ==> decision=not 1586    conf:(0.91)
 7. Sex=F Religion=rel1 Offenderage=oag2 1520 ==> decision=not 1380    conf:(0.91)
 8. typeofcrime=offe2 decision=not 1586 ==> Sex=F 1439    conf:(0.91)
 9. typeofcrime=offe1 Education=ed4 1574 ==> Job=student 1427    conf:(0.91)
10. typeofcrime=offe2 Religion=rel1 1530 ==> Sex=F 1386    conf:(0.91)
```

Figure 5. 3 ***The association rule model developed by the second experiment.***

As we can see from the above association model developed, there are 13 one-item frequent item-set, 40 two-item frequent item-set, 46 three-item frequent item-set, 17 four-item frequent item-set and 3 five-item frequent item-set generated with 25% min-support and 90% min-confidence. Out of which only 10 best rules are displayed. These 10 best rules generated ranges from 25% to 30% support and 91% to 92% confidence.

To see four of the rules, rules with two-item, three-item, four-item and five-item and members of the 10 best rules generated (Rule #6, Rule #1 Rule #7, and Rule #5) are selected for analysis..

1. Rule #6 (Type-of-Crime=offe2 1744 ==> decision=not 1586 conf:(0.91)

   If (Type-of-Crime=offe2 then decision=not).

This rule shows the relationship between two attributes, the attributes type of crime and decision. Those victim who are affected by crime category offe2 which represents sexual offences do not get quick decision.

2. Rule #7 (Sex=F Religion=rel1 Offenderage=oag2 1520 ==> decision=not 1380 conf:(0.91))

   If (Sex=F and Religion=rel1and OFFENDER_AGE = oag2 then Decision = not).

The relationship between the ten attributes was that if the victim is female, orthodox in religion and the offender is 18-45 then the decision is not closed(on progress) with 26% support and 91% confidence.

3. Rule #1 (Sex=F Education=ed4 1541 ==> Job=student 1417 conf:(0.92))

   If (Sex = F and Education = Ed4 then Job = Student).

Whenever the victim is female, educational status is 1-6 implies that the job of the victim is student with 26% of support and 92% confidence.

4. Rule #5 (Ageofvictim=ag2 sex=F Job=student decision=not 1511 ==> religion=rel1 1376 con: (0.91))

   If (Age_Of_victim=ag2 and Sex=F and Job=student and Decision=not then Religion=rel1)

This rule has the highest number of items i.e. five. It has support of 26% and confidence of 91%. This rule shows the profile of a victim in relation to age of offender, the type of crime, the decisions given and the sub-city where the crime was committed. Thus if the victim is between 9 and 18 years old, female in gender, student in occupation and decision is on progress, then she is orthodox religion follower.

Experiment #3

The association model developed using 9 attributes after removing the attributes four frequent ones. These includes habit, living, marital status and religion.

```
Apriori
=======

Minimum support: 0.2 (1071 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 39

Size of set of large itemsets L(3): 45

Size of set of large itemsets L(4): 13

Best rules found:

 1. typeofcrime=offe2 ageofvictim=ag2 1234 ==> decision=not 1140    conf:(0.92)
 2. typeofcrime=offe2 Sex=F Job=student 1323 ==> decision=not 1218    conf:(0.92)
 3. Sex=F Education=ed4 decision=not 1315 ==> Job=student 1210    conf:(0.92)
 4. Sex=F Education=ed4 1541 ==> Job=student 1417    conf:(0.92)
 5. typeofcrime=offe2 Sex=F 1567 ==> decision=not 1439    conf:(0.92)
 6. typeofcrime=offe2 Job=student 1469 ==> decision=not 1339    conf:(0.91)
 7. typeofcrime=offe2 ageofvictim=ag2 1234 ==> Sex=F 1123    conf:(0.91)
 8. typeofcrime=offe2 Job=student decision=not 1339 ==> Sex=F 1218    conf:(0.91)
 9. typeofcrime=offe2 1744 ==> decision=not 1586    conf:(0.91)
10. typeofcrime=offe2 decision=not 1586 ==> Sex=F 1439    conf:(0.91)
```

Figure 5. 4 ***The association rule model generated by the third experiment.***

As displayed in the association model the following 111(one hundred eleven) large item sets of which 14 are one-item frequent item-set, 39 are two-item frequent item-set, 45 are three-item frequent item-set and 13 four-item frequent item-set. These 111 large item-sets are generated with 20% minimum support and 90% minimum confidence and 10 best rules are displayed as shown in figure 5.3.

From the above experiments removing the attributes habit, Marital Status, and living give chance for the attributes sex, type of crime and education to appear in the association rule model. Similarly removing the attribute religion gives chance for the attribute age of victim. This is because the attributes living, habit, marital status and religion and decision have frequently appearing values.

Some of the association rule generated using the nine attributes are presented below.

1. (typeofcrime=offe2 ageofvictim=ag2 1234 ==> decision=not 1140 conf:(0.92))

   (If the type of crime=offe2 and ageofvictim=ag2 then decision= not).

Offence2 is a crime that includes rape, harassment, Buggery, and all other types of sexual assault. This is to mean the victims of offence2 at the age of 9-18 years old the decision to these crimes is delayed with 21% support and 92% confidence.

2. (typeofcrime=offe2 Sex=F Job=student 1323 ==> decision=not 1218 conf:(0.92))

   If (type of offence=sexual assault and sex=F and job=student then decision = not (on progress)).

   Female children who are students and exposed to sexual assault do not get quick decision with confidence of 92%.

3. (typeofcrime=offe2 ageofvictim=ag2 1234 ==> Sex=F 1123 conf:(0.91))

   If (typeofcrime=sexual assault and ageofvictim=9-18 then sex=F).

Victim children at the age of 9-18 years and exposed to sexual assault are female.

Experiment #4

The association model developed using 7 attributes after removing the attributes six frequently appearing ones. These includes habit, living, marital status, religion, job and decision.

```
Apriori
=======

Minimum support: 0.1 (536 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 18

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 56

Size of set of large itemsets L(3): 33

Size of set of large itemsets L(4): 8

Best rules found:

 1. typeofcrime=offe2 ageofvictim=ag2 Offenderage=oag2 791 ==> Sex=F 727    conf:(0.92)
 2. typeofcrime=offe2 Offenderage=oag2 1037 ==> Sex=F 944    conf:(0.91)
 3. typeofcrime=offe2 ageofvictim=ag2 1234 ==> Sex=F 1123    conf:(0.91)
 4. typeofcrime=offe2 ageofvictim=ag2 Education=ed4 682 ==> Sex=F 617    conf:(0.9)
 5. typeofcrime=offe2 1744 ==> Sex=F 1567    conf:(0.9)
 6. typeofcrime=offe2 Education=ed4 930 ==> Sex=F 825    conf:(0.89)
 7. typeofcrime=offe1 Education=ed5 741 ==> ageofvictim=ag2 638    conf:(0.86)
 8. Sex=F Education=ed5 803 ==> ageofvictim=ag2 687    conf:(0.86)
 9. Education=ed5 Offenderage=oag2 858 ==> ageofvictim=ag2 728    conf:(0.85)
10. Education=ed5 1490 ==> ageofvictim=ag2 1262    conf:(0.85)
```

Figure 5. 5 ***The association Rule model with 10 best rules generated with the firth experiment.***

The following are two of the rules generated in the above experiment.

1. typeofcrime=offe2 ageofvictim=ag2 Offenderage=oag2 791 ==> Sex=F 727 conf:(0.92)

If (typeoscrime=sexual assault and ageofvictim=9-18 and offenderage=18-45 then sex=F).

A victim at the age of 9-18 years, exposed to sexual assault and offended by age group 18-45 years are female.

2. typeofcrime=offe2 ageofvictim=ag2 Education=ed4 682 ==> Sex=F 617 conf:(0.9)

If (typeofcrime=sexual assault and ageofvictim=9-18 and Education=1-6 then Sex=F)

Victims at the age of 9-18 years, at education level of 1-6 and exposed to sexual assault are female.

Based on the rules generated from the association model the following …

5.2. Clustering model

During clustering the whole dataset have been used for the training purpose since clustering is unsupervised learning. In unsupervised learning, algorithms do not need class labels (dependent variable) rather learn from the dataset and place objects into their category (clusters) based on the underlying data distribution. In this study WEKA K-means algorithm was used to implement the clustering model. This algorithm includes the following configuration parameters:

- Display standard deviation: displays the standard deviation of numeric and counts of nominal attributes. The available choices are true to display and false not to display.
- Do not replace missing values: to replace or not missing values globally with mean for numeric attribute or mode for nominal attributes. The available choices are true to replace and no not to replace.
- Number of clusters: the number of clusters (K in K-mean) that is to be created. This value has to be manually input into the system. Most of the time the value of K ranges from 2-20, but it has to be determined by the number of segments that the business can successfully handle or manage.
- Seed: The random number of seeds to be used.

Attributes can be selected either using automatic tools like WEKA attribute selection evaluation algorithms or with the help of domain experts' experience. The researcher preferred to do both and compare their results. According to the domain experts (crime

investigators and predictors) of the AAPC members and the FSCE database administrator all the recorded attributes are important. But the following attribute are selected as important attributes based on the domain experts Sex, religion, habit, living, education, age-of-victim, offender-age, victim's occupation and year. Whereas WEKA information gain attribute selection evaluation ranked the 13 attributes as Sex, Type-of-Crime, Education, Year, Age-of-Victim, Offender-Age, Sub-city, decision, Living, Marital-Status, Religion, Job and Habit. The researcher used this rank and took eight attributes which have highest information gain and religion from domain experts' experience.

Value of the Variables	**Short Form**	**Value of the Variables**	**Short Form**
Offence 1	O1	Oag1(offender age group one) 9-18	G1
Offence 2	O2	Oag2 (offender age group two) 18-45	G2
Offence 3	O3	Oag3(offender age group three) >45	G3
Offence 4	O4	Addis Ketema	1
Offence 5	O5	Akaki	2
Offence 6	O6	Arada	3
Ag1 Age group 1(0-9)	A1	Bole	4
Ag2 Age group 2 (9-18)	A2	Gulele	5
Female	F	Kirkos	6
Male	M	Kolfe	7
Education 1(illiterate)	E1	Lideta	8
Education 2 (kg)	E2	NifasSilkLafto	9
Education 3(basic education)	E3	Yeka	10
Education 4(1-6)	E4	Orthodox	R1
Education 5(7-12)	E5	Muslim	R2
Education 6(diploma)	E6	Protestant	R3

Table 5. 2 ***The abbreviated values of the attributes used in clustering model.***

Experiment #1

This is the first clustering experiment using eight selected attributes. This is conducted to group the data into five clusters with the assumption from domain experts to divide the

crime records into five dissimilar groups.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     CRIME-weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R8
Instances:    5355
Attributes:   8
              typeofcrime
              ageofvictim
              Sex
              Education
              Year
              Religion
              Offenderage
              subcity
Test mode:    evaluate on training data

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 14538.0
Missing values globally replaced with mean/mode
```

Figure 5. 6 ***Run information of the first experiment with value of K=5***

As can be seen from the above experiment this cluster has seed 10 and K value set to five (5). Only 8 of the 13 attributes are selected based on the importance of the attributes from the domain experts and the WEKA attribute evaluation in combination.

Cluster index	Frequency Records	Type of crime	Age of victim	Sex	Education	Year	Religion	Offender age	Sub city
1	2356	O2	A2	F	E4	200	R1	G2	5
2	1377	O1	A2	M	E5	1999	R1	G1	3
3	391	O3	A2	F	E4	1998	R1	G2	9
4	862	O1	A2	M	E4	1999	R1	G2	3
5	369	O2	A1	F	E5	1999	R1	G2	10

Table 5. 3 ***Summarized result of the first experiment***

As shown in the above table, on its summary reports all the variables are taken as independent values plus the whole dataset 5355 entries are used as training data. The algorithm assigns cluster index for each record in the dataset. Such visual output provides a descriptive classification model which identifies the characteristics of each cluster.

Out of the five clusters cluster 3 (cluster #2 in WEKA) and cluster 5 (cluster #4 in WEKA) has the smallest number of records that they have only 7% of the objects in each. The two clusters are similar in three of the eight attributes they have highest percentage of female in gender, more orthodox in religion and are victims by 18-45 years old offenders. Due to this they are 37.5% similar in character. Cluster_3 and cluster_4 contain similar victims who have more or less similar patterns. These two clusters have 75% similar characteristics. They share highest percentage of the six attribute values. Thus the researcher believes that they are not good enough to describe two different groups. It is better to merge them and set K value to four.

Experiment #2

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     CRIME-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance
Instances:    5355
Attributes:   8
              typeofcrime
              ageofvictim
              Sex
              Education
              Year
              Religion
              Offenderage
              subcity
Test mode:    evaluate on training data

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


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 14613.0
Missing values globally replaced with mean/mode
```

Figure 5. 7 *Run information of the second experiment with value of K set to 4*

Cluster index	Frequency of Records	Type of crime	Age of victim	Sex	Education	Year	Religion	Offender age	Sub city
1	2289	O2	A2	F	Ed4	200	Rel1	G2	10
2	1481	O1	A2	M	E5	1999	R1	G1	3
3	789	O1	A2	F	E4	1998	R1	G2	9
4	796	O1	A2	M	E4	1999	R1	G2	3

Table 5. 4 ***Summarized result of the second experiment***

Rather than describing clusters using only the mode attribute values, it is better to see using values of the entire attributes in each cluster.

Cluster index	Frequ-ency of records	Type of crime	Age of victim	Sex	Education	Year	Religion	Offender age	Sub city
1	2289	13%O1 72%O2 6% O3 2%O4 1%O5 3%O6	29%A1 70%A2	11%M 88%F	10%E1 3%E2 0%E3 65%E4 20%E5 0%E6	7%1997 10%1998 11%1999 37%2000 20%2001 13%2002	87%R1 10%R2 2%R3	27%G1 66%G2 6%G3	10% 1 8% 2 3% 3 9% 4 13% 5 7% 6 11% 7 7% 8 3% 9 23% 10
2	1481	74%O1 4%O2 12%O3 3%O4 3%O5 1%O6	18%A1 81%A2	78%M 21%F	3%E1 1%E2 0%E3 33%E4 61%E5 0%E6	6%1997 21%1998 36%1999 13%2000 11%2001 11%2002	86%R1 12%R2 0%R3	68%G1 28%G2 3%G3	6% 1 7% 2 27% 3 7% 4 21% 5 4% 6 6% 7 3% 8 10% 9 4% 10
3	789	83%O1 1%O2 4%O3 3%O4 1%O5 5%O6	26%A1 73%A2	21%M 78%F	6%E1 1%E2 0%E3 77%E4 14%E5 0%E6	6%1997 38%1998 7%1999 4%2000 28%2001 14%2002	80%R1 16%R2 1%R3	18%G1 71%G2 10%G3	7% 1 5% 2 5% 3 6% 4 23% 5 9% 6 8% 7 4% 8 29% 9 0% 10
4	796	77%O1 6%O2 12%O3 4%O4 4%O5 1%O6	26%A1 73%A2	98%M 1%F	3%E1 2%E2 0%E3 94%E4 0%E5 0%E6	7%1997 7%1998 28%1999 19%2000 19%2001 17%2002	90%R1 8%R2 1%R3	0%G1 94%G2 5%G3	8% 1 12% 2 30% 3 6% 4 14% 5 6% 6 8% 7 5% 8 4% 9 3% 10

Table 5. 5 ***Detailed result of the second experiment***

Similarly as shown in the above table cluster #3 and cluster #4 have the same pattern. The characteristics of these clusters with respect to different attributes are summarized as follows:

Cluster index	Cluster description	Remark
1	This cluster has greater number of victims than any other clusters (43%) of the total records. Majority of them are victims of offence2 which is a category that includes sexual assaults, 70% of the victims are 0-9 years old, 88% of them are female, 65% of them are at 1-6 educational level, 87% of them are followers of orthodox religion, 66% of their offenders are at the age of 18-45 years and 23% of these crimes are committed in Yeka sub-city. On the other lower side they are 1% of the victims are affected by offence5, only 29% are male, people at the education levels basic education and above 12 are not included in this cluster, only 2% of the victims are protestant, 6% are at the age of above 45 and 6% of the crimes are committed in both Arada and Nifas Silk sub-cities.	Good
2	Almost 2/3 of the victims are exposed to offence1 which is a category that includes all types of injuries and simple and corporal punishments, 81% of the victims are at the age of 9-18 years old, 78% of the them are male, almost all 94% of them are at education levels of Ed4 and Ed5 which represent 1-6 and 7-12 respectively, above 1/3 of the crimes was committed in the year 1999, 86% of the victims are orthodox, 68% of their offenders are at the age of G1i.e. 9-18 years old, and from the ten sub-cities the Arada sub-city is the place where 27% 0f the crimes are committed.	Moderate
3	This cluster has similar number of records with cluster #4. Both constitute only 30% of the total records which is less than the percentage of cluster #1. most of the victims (83%) are affected by offence1, 73% of the victims are 9-18 years old, 78% of them are female, almost all (91%) of them are at the educational level of 1-6 and 7-12 grades. 80% the victims are orthodox, 70% their offenders are at the age of 18-45 years old, 52 (29% Nifas-Silk-Lafto and 23% Gulele) of the entire crimes are committed in both Nifas-Silk-	Low

	Lafto and Gulele sub-cities.	
4	This cluster has similar number of objects with cluster #3. 77% of the victims are exposed to offence1, 73% of the are at the age of 9-18 years, almost all of the victims (98%) are male and 94% are at grades 1-6, 90% of the whole victims in this cluster are orthodox, within the six recorded years the highest percentage of crime (30%) was committed in the year 1999 and within the ten sub-cities in the city highest percentage (28%) was committed in Arada sub-city.	Good

Table 5. 6 ***Description of the four clusters***

Accordingly, when the researcher with domain experts evaluate the four clusters, they have tried to show patterns that can help police offers, planners and awareness creators deal with crimes proactively. The clusters try to show which age group of children are highly exposed to offences against children, which sub-cities are places comfortable to commit crimes, which sex category is highly affected by offences against children and so on. But there are clusters with relatively similar pattern. Good clustering algorithms should have to maximize within cluster similarity and minimize between cluster similarities. Based on the domain experts expectation the clusters are ranked as “Good”, “Moderate” and “Low” to mean matches, averagely matches and do not match respectively. In the case of these two, cluster #2 and cluster #4 do have similar patterns which rather maximize between cluster similarities.

Only differences between these two clusters lie on the attributes education and offender age. Having similar patterns for the rest six attributes shows about 75% similarity. Due to this, these two clusters are not good enough to describe two different categories. That’s why the researcher believes that it is better to merge these two clusters and reduce the value of K to 3 rather than 4.

Experiment #3

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 20
Relation:     CRIME-weka.filters.unsupervised.attribute.Remove-R6,9-weka.filters.unsupervised.attribute.RemoveUseless-M99.0
Instances:    5355
Attributes:   8
              typeofcrime
              ageofvictim
              Sex
              Education
              Year
              Religion
              Offenderage
              subcity
Test mode:    evaluate on training data

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


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 15438.0
Missing values globally replaced with mean/mode
```

Figure 5. 8 ***Run information from the third experiment with value of K set to 3***

Cluster index	Frequ. of Records	Type of crime	Age of victim	Sex	Education	Year	Religion	Offender age	Sub city
1	2753	O2	A2	F	E4	2000	R1	G2	5
2	1600	O1	A2	M	E5	1999	R1	G1	3
3	1002	O1	A2	F	E4	1998	R1	G2	9

Table 5. 7 ***Output of the third experiment without displaying standard deviation***

The first and the third cluster have similar patterns with about 62.5% relatively better dissimilarity than the above two experiments' clusters. Let me describe each of the three clusters in detail below.

Cluster index	Frequ. of Records	Type of crime	Age of victim	Sex	Education	Year	Religion	Offender age	Sub city
1	1936 (36%)	67%O1 8%O2 13%O3 3%O4 4%O5 1%O6	9%A1 90%A2	71%M 28%F	2%E1 0%E2 0%E3 41%E4 54%E5 0%E6	6%1997 23%1998 35%1999 13%2000 8%2001 12%2002	87%R1 11%R2 0%R3	58%G1 36%G2 4%G3	6% 1 6% 2 32% 3 7% 4 16% 5 4% 6 6% 7 3% 8 11% 9 5% 10
2	1523 (28%)	71%O1 14%O2 4%O3 4%O4 1%O5 3%O6	57%A1 42%A2	60%M 39%F	6%E1 4%E2 0%E3 84%E4 4%E5 0%E6	6%1997 11%1998 11%1999 41%2000 16%2001 12%2002	84%R1 13%R2 1%R3	14%G1 80%G2 5%G3	7% 1 9% 2 5% 3 8% 4 8% 5 25% 6 6% 7 5% 8 11% 9 10% 10
3	1896 (35%)	13%O1 71%O2 6%O3 1%O4 2%O5 3%O6	16%A1 83%A2	3%M 96%F	10%E1 1%E2 0%E3 67%E4 19%E5 0%E6	8%1997 14%1998 12%1999 19%2000 31%2001 14%2002	87%R1 10%R2 1%R3	22%G1 70%G2 7%G3	10% 1 8% 2 3% 3 10% 4 11% 5 8% 6 12% 7 6% 8 6% 9 21% 10

Table 5. 8 ***Output of the third experiment with displaying standard deviation***

In this experiment it is relatively possible to distinguish between the clusters and also to classify each of the victims to a different level according to their expected exposure to the different offences against children. Based on the above experiment the following behaviors are detected from each cluster.

Cluster index	Cluster description	Remark
1	This is the highest cluster that constitutes 36% of the total victims. Most of them (90%) are at the age of 9-18 years. The age group that dominates the dataset (74%) is 9-18 years. To see the ratio of the two age groups in the dataset, it is 74% by 25% where as in cluster #1 the ratio is 90% by 9%.	Moderate

	The sex proportion in the dataset is 44% male and 55% female. In this cluster (cluster #1) the proportion of male and female is 24% male and 71% female. A great number of victims 67% are affected by offence1 while the remaining 33% are affected by the other four crime categories. In the whole dataset offence1 constitutes 49%. Almost all of the victims are at the level of education ED4 and ED5 41% and 54% respectively. The proportion of the six education level categories in the dataset is 6% ED1, 2% ED2, (2 in number of records) 0% ED3, 63%ED4, 27% ED5 and (3 in number of records) 0% ED6. The proportion of both ED4 and ED5 that constitute 90% of the dataset 63% ED3 and 27% ED5 in the dataset is 41% ED4 and 54% ED5 in cluster #1. In this cluster the highest percentage of crimes (35%) are committed in the year 1999 while in the dataset the highest percentage (23%) was committed in the year 2000. more than half 58% was committed in two years 1998 and 1999. The rest 39% was committed in the rest four years. The percentage of crimes committed in the six years starting from 1997-2002 are 7%, 17%, 20%, 23%, 19%, 13% respectively. The percentages in cluster #1 are 6%, 23%, 35%, 13%, 8% and 12% respectively. More than half 58% of the offenders are at the age 9-18 years. The percentage of the three age groups (9-18, 18-45 and > 45) in the dataset is 33%, 60% and 4% respectively. In cluster #1 their percentage is 58% 9-18years, 36% 18-45 years and 4% above 45 years. To see the percentage of crimes committed in the ten sub-cities in the	

	dataset and this cluster, 8%, 8%, 14%, 8%, 17%, 6%, 9%, 5%, 9% and 11% crimes are committed in the dataset and 6%, 6%, 32%, 7%, 16%, 4%, 6%, 3%, 11%, and 5% crimes records are classified in this cluster. In this cluster, highest percentage of crimes (32) is committed in Arada sub-city. The next highest percentage (16) of crimes is committed in Gulele sub-city followed by Nifa Silk Lafto sub-city at the percentage of 11% while the lowest percentage was committed in Lideta sub-city.	
2	This cluster consists of 28% of the dataset. More than half (60%) of them are male. The percentages of male and female in the dataset are 44% male and 55% female. In contrast to the dataset percentage of male and female, the percentage of male and females of this cluster percentage of males is greater than female. In this most of the victims 57% are the age of 0-9 and the rest 42% are at the age of 9-18. Almost all 84% of the victims in this cluster are at education level of 1-6 followed by illiterate which is 6%. None of the victims in this cluster are at KG level and above 12. Almost (2/3) 71% of the records in this cluster are exposed to offence1 which is a crime category that includes all types of physical chastisements. Boys are more exposed to such crimes than girls. All age categories are relatively have similar exposure to such crimes. Except the two years with highest percentage 41% and lowest percentage 6% 2000 and 1997 respectively, all the four years are with similar percentage of crimes committed.	Moderate
3	This cluster constitutes 35% of the dataset. Out which 71% of them are exposed to sexual assault which is encoded as offe2. The next offence that is risk to this croup is offence1 which constitutes 13% of the records in this dataset. Offence1 is a category that contains	Good

	offences like corporal punishment, bodily injury with all types of physical chastisements. The least one in this cluster is offence4 which constitutes only 1%. The age of victim that dominates this cluster is the age between 9 and 18 years old which accounts for 83%. The rest 13% are at the age of 0-9 years old. Most 86% of the victims in this cluster are female this is why they are the victims of offence2 which is a category that includes rape, attempt for rape, sexual harassment, buggery, Abduction, etc. Only 3% of them are male. More than half 67% are at the level of education 1-6 and followed by the level 7-12 with 19% and 10% illiterate. Highest percentage of crime was committed in the year 2001. Similar percentage of crimes 14%, 12%, 19% and 14% are committed in the years 1998, 1999, 2000 and 2002 respectively. The lowest percentage of crime was committed in the year 1997. From the ten sub-cities the highest percentage of crimes was committed at Yeka sub-city. The next percentages of crimes are 12%, 11%, 10% and 10% committed in Kolfe Keranew, Gulele, Bole and Addis Ketems sub-cities respectively. The lowest percentage of crimes was committed at Arada sub-city which constitutes 3%.	

Table 5. 9 ***Description of the three clusters***

5.3. Choosing the Best Clustering Model

Three experiments were conducted to come up with the appropriate clustering model. Finally the clustering model that satisfies the criteria of good clustering model more than any other clusters was selected. The best set of clusters is those that have high intra class similarity and low inter class similarity and those that show expected patterns by domain experts. Some of the criteria of good clustering model one that could be easily understood

and interpreted by domain experts. Other criteria are objectively from the application tool like small number of iterations and minimum sum of square errors,

The first experiment which has five clusters indicated that two of the clusters have similar patterns and a very few records are allocated is one of the clusters. Due to this the researcher together with the domain experts tried to modify and reduce the value K to four (4) to merge the two clusters and increase the number of records in a given cluster as well as to have clusters with relatively different patterns. A cluster should contain enough numbers of records in order to describe the pattern with the victims profile, the crime trends through time and the places where crimes are committed. There should be other cluster options to compare with.

Thus, the second experiment with the value of K reduced to four was conducted. As the result summarized in tables 5.4 and table 5.5 some of the problem still exists. Two clusters do have similar patterns. From previous discussion with the domain experts the researcher was forced to conduct another experiment.

The third experiment was conducted by setting the value of K to 3 and the result was shown in table 5.7 and table 5.8. As can be seen from the tables there are three clusters behaving relatively different. There exists a relatively better separation among the three clusters and also homogeneity within them is relatively better than the previous experiments. Even the proportion of records in the three clusters is relatively better (36%, 28% and 37%) respectively.

The number of iteration is three and sum of squared error is smaller for the third experiment. In addition to the above improvements found by the researcher, the clusters are distinct and meaningful to the domain experts. Crime events are sometimes categorized as serious, medium and low. While segmenting crime events in to these three classes to handle them in different department for investigation purpose, there are some demarcation problems. Even the proportion is different in different situations. Thus three cluster segments could be used to represent them based on the underlying distributions in

the crime dataset. Hence it is possible to develop crime protection strategies for each of the three crime cluster.

Finally the researcher together with the domain experts decided that the appropriate number of clusters is three and the third experiment was selected as good model showing relatively good cluster of the crimes (offences committed against children) in Addis Ababa in the years 1997-2002 EC. The output of this cluster model was used as an input for the classification model.

5.4. Modeling Building and Analysis of Classification

The input for this model is output of the clustering model in section 5.3. The algorithm used is a decision tree called J48. In this case the classifier is used to classify the instances into their already classified cluster index. The cluster index is used as a class label (predictable variable). Unlike clustering classification is supervised learning which divides the whole dataset into training and test sets. Here the classifier used the 10-fold cross-validation.

Internal cross-validation is used to determine how well a learning algorithm will fit in independent datasets (Kohavi, 1995). The principles of k-fold cross validation are to divided the dataset into k mutually exclusive subsets of approximately equal size, the learning algorithm is then trained on each k-1 subset (the training subset) and its prediction are then verified on the corresponding k subset (the testing subset). The performance measures across all k trials are computed and then averaged to determine the performance of the k-fold cross-validation. The average of the performance measure provides an estimate of the performance of the classifier constructed from the whole dataset.

Decision tree performs best when all of the attribute contain non-continuous values. This is why the age attribute and education attribute are converted into nominal. In this classification sub task two experiments were done. The first one is using the default value for the parameter “Number of Objects” that the leaf node should contain. The second experiment is conducted by changing the value the parameter “Number of Objects” at

each leaf node. The researcher found that when the value increases the accuracy decreases. As we know the size and number of leaves decreases when the value for the parameter of "Number of Objects" increases.

Analysis of these decision tree models are made in terms of detailed accuracy of the classifier on the dataset based on the confusion matrix of the two models' results. Confusion matrix is useful tool for analyzing how well the classifier can recognize records of different cluster index (class).

The two experiments are analyzed and compared to each other in terms of performance of matrices value accuracy, number of leaves, size of the tree and execution time. The models also compared with regard to knowledge (rules) discovered.

For experiment #1 which is with default value of "Number of objects" (2) the accuracy is found to be 97.89%. The Number of Leaves of the tree is 248, the Size of the tree is 307 and the Time taken to build model is 0.05 seconds.

Actual	Predicted			Total	Score (Actual Rate)
	Cluster #1	Cluster #2	Cluster #3		
Cluster #1	1893	36	20	1949	97.13%
Cluster #2	11	1452	3	1466	99.05%
Cluster #3	26	17	1897	1940	97.78%
Total	1930	1505	1920	5355	97.89%

Table 5. 10 ***Output from the J48 decision tree learner by using the default value of the parameter "Number of Objects"***

A portion of the J48 pruned tree model using the default value of the parameter of the "Number of objects" is shown below:

```
Sex = F
|   ageofvictim = ag1
|   |   typeofcrime = offe1
|   |   |   Education = ed1
|   |   |   |   Offenderage = oag1: cluster1 (8.07/1.91)
```

```
|   |   |   |   Offenderage = oag2: cluster2 (32.49)
|   |   |   |   Offenderage = oag3: cluster2 (3.72)
|   |   |   Education = ed2
|   |   |   |   Offenderage = oag1: cluster1 (4.15/1.07)
|   |   |   |   Offenderage = oag2: cluster2 (16.15)
|   |   |   |   Offenderage = oag3: cluster2 (2.43)
|   |   |   Education = ed3: cluster2 (0.0)
|   |   |   Education = ed4
|   |   |   |   Year = 1997: cluster2 (12.8)
|   |   |   |   Year = 1998: cluster2 (27.87/0.27)
|   |   |   |   Year = 1999
|   |   |   |   |   Offenderage = oag1: cluster1 (7.87)
|   |   |   |   |   Offenderage = oag2: cluster2 (15.27/0.27)
|   |   |   |   |   Offenderage = oag3: cluster2 (1.0)
|   |   |   |   Year = 2000: cluster2 (47.77)
|   |   |   |   Year = 2001: cluster2 (38.17)
|   |   |   |   Year = 2002: cluster2 (20.47/0.27)
```

Figure 5. 9 ***A portion of the rules generated using default parameter for the value of "Number of objects".***

The second experiment is conducted by increasing the value for the parameter "Number of Objects" to 20. In this experiment the accuracy was lowered to 93.63%. The Number of Leaves of the tree is 109, the Size of the tree is 136 and the Time taken to build model is 0.05 seconds.

Actual	Predicted			Total	Score (Actual Rate)
	Cluster #1	Cluster #2	Cluster #3		
Cluster #1	1797	96	56	1949	92.2%
Cluster #2	34	1413	19	1466	96.38%
Cluster #3	70	68	1804	1940	92.99%
Total	1901	1577	1877	5355	93.63%

Table 5. 11 ***Output from the J48 decision tree learner by setting the value of the parameter "Number of Objects" to 20***

Portion of the rules generated in this experiment are presented below(see for detail in appendix C):

```
Sex = M
|   ageofvictim = ag1
|   |   Education = ed1
|   |   |   Offenderage = oag1: cluster1 (26.46/8.97)
|   |   |   Offenderage = oag2: cluster2 (31.23/0.04)
|   |   |   Offenderage = oag3: cluster2 (7.25/0.0)
|   |   Education = ed2: cluster2 (47.55/7.39)
|   |   Education = ed3: cluster2 (0.0)
```

```
|   |   Education = ed4
|   |   |   Offenderage = oag1
|   |   |   |   Year = 1997: cluster2 (12.71)
|   |   |   |   Year = 1998: cluster2 (11.44/0.02)
|   |   |   |   Year = 1999: cluster1 (21.89)
|   |   |   |   Year = 2000: cluster2 (38.19)
|   |   |   |   Year = 2001: cluster2 (10.95/0.14)
|   |   |   |   Year = 2002: cluster2 (16.39/0.26)
|   |   |   Offenderage = oag2: cluster2 (217.23/0.7)
|   |   |   Offenderage = oag3: cluster2 (12.63/0.0)
|   |   Education = ed5
|   |   |   Year = 1997: cluster2 (0.79)
|   |   |   Year = 1998: cluster2 (33.49/14.68)
|   |   |   Year = 1999: cluster1 (54.07/0.4)
|   |   |   Year = 2000: cluster2 (9.15/2.0)
|   |   |   Year = 2001: cluster2 (9.79/1.05)
|   |   |   Year = 2002: cluster2 (4.87/2.26)
|   |   Education = ed6: cluster2 (0.0)
```

Figure 5. 10 ***Rules generated by setting the value of the parameter Number of objects to 20***

Some of the rules extracted from J48 pruned tree model are presented below.

- `If (Sex = M and ageofvictim = ag1 and Education = ed1 and Offenderage = oag1 then cluster1 (26.46/8.97))`

This rule implies that male, at the age of 0-9 years, who are illiterates and offended by offenders of age 9-18 years are victims classified as cluster1. There are around 27 records having this property. From which only 18 records are correctly classified and the rest 8 records are miss classified.

- `If (Sex = M and ageofvictim = ag1 and Education = ed1 and Offenderage = oag2 then cluster2 (31.23/0.04))`

  As can be from the above rule if a victim in the record is male and is at the age of 0-9 years and is illiterate and is offended by offenders of age 18-45 years he/she is classifies as cluster2.

- `If (Sex = M and ageofvictim = ag1 and Education = ed1 and Offenderage = oag3 then cluster2 (7.25/0.0))`

  Some records in cluster2 are with the rule, if the victim is male and the victim is at the age of 0-9 years and the victim is illiterate and offended by offenders of age above 45 then they are classified as cluster2.

- `If (Sex = F and ageofvictim = ag1 and typeofcrime = offe1 and Education = ed1 then cluster2 (44.27/6.15))`

  The above rule implies that some records in cluster2 are with characteristics like female in gender, 0-9 years in age of victim and are exposed to physical injuries.

- `If (Sex = F and ageofvictim = ag1 and typeofcrime = offe2 and Year = 1997 then cluster3 (42.0))`

  The above rule implies that 42 victims in cluster3 have the following characteristics:

  All of them are female, age of victim is from 0-9 years and the crime type to which they are exposed is sexual offences.

Finally from the two experiments we have preferred to see the models' results and analysis of each result and compare their results to find the most out performing model based on the criteria of evaluation. The criteria of evaluations are accuracy from confusion matrix, size of the tree, number of leaves of the tree, and time taken to build the model. Based on these criteria the one with the value of parameter "number of Objects" set to twenty is better than the default value.

5.5. Evaluation

During evaluation the degree to which the model meets the business objective was assessed. As indicated in different parts of this research, the business goals are to come up with models that could extract the profile of victims, could differentiate and predict the number of clusters of victim children according to their likelihood of being vulnerable to a particular crime category and also assign a new victim to which it is exposed and their appropriate cluster index. This works better in predicting the victims' vulnerability and showing what does profile of victims look like.

The basic criteria to evaluate segmentation of victims' records are the behavior of victims, the behavior of offenders and environmental setting. These three things make them highly affected by particular crime.

The final analysis which was undertaken by the domain experts reveals that the final segmentation experiment indeed discover patterns that are interesting for them. In addition

to domain experts comment and the above measures there are other measures of good cluster. These include low sum of square errors, low number of iterations.

5.6. Interpretation and discussion (Findings)

Association rule

From the different list of rules generated over various experiments in association rule using different set of attributes, a number of rules with satisfactory objective measure (high support and confidence) and most importantly meeting the subjective judgment of domain experts on their interestingness and applicability were evaluated.

Summary of the discovery task done on the 5355 records using association rule, and the subsequent interpretation and discussion of the discovered interesting rules is presented below.

Experiment 1

In this experiment all the 13 attributes are used.

Attributes: “Type of offence”, victim’s level of education”, “Age of victim”, “sex”, “year”, “religion”, “occupation”, “living”, “marital status”, “special habit the victim has”, “offender age” “sub-city” and “decision”.

Rule

If (Habit=no and Religion=rel1 and Living=famyes then MaritalStatus=single)

This is to mean that the victims in the database have the following characteristics. Whenever the victim has no special habit and is orthodox and living with his/her family then he/she is single in marital status.

According to domain experts, this rule is a generalization of the fact that instances in the database are characterized by the occurrence of very few instances having SPECIAL HABIT such as smoking cigarettes, drinking alcohol, drug addiction, or chewing Chat.

This finding is in line with the popular conception that such special bad habits are more often characteristics of child offenders than child victims.

Similarly for the attribute LIVING the number of instances in the database having values different from living with family like living “with relative”, “child care institution”, “street” or “with some who is not relative”. These living conditions may expose children n to offences like sexual assaults and willful injury but they are few in number to generate pattern. Such an output of the discovery task is an indication that apart from discovering surprising or hidden rules, the learning scheme also results in rules that confirm facts existing in the real world

Experiment 2

This experiment is conducted using 10 attributes (Type of offence”, victim’s level of education”, “Age of victim”, “sex”, “year”, “religion”, “occupation”, “offender age” “sub-city” and “decision”).

Rule

If (Age_Of_victim=ag2 and Sex=F and Job=student and Decision=not then Religion=rel1)

This is to mean whenever there is a victim at the age of 9-18 and female, student in occupation, decision is on progress then she is orthodox.

If (Sex = F and Education = Ed4 then Job = Student).

Whenever the victim is female, at education level 1-6 she is a student.

According to domain experts, the rules stated above represent interesting regularity within the crime database.

Experiment 3

This model is build using 9 attributes.

Rules

If (typeofcrime=sexual assault and ageofvictim=9-18 then sex=F)

This is to mean whenever a victim is exposed to offence of sexual assault and are at the age of 9-18 they are female.

If (type of offence=sexual assault and sex=F and job=student then decision = not (on progress)).

The decision given to cases related with sexual assault which victimize female students is always delayed.

According to domain experts, the rules stated above represent interesting regularity within the crime database. For example female children at the age of 9-18 years are exposed to sexual assaults. Female children who are students and exposed sexual assault do not get quick decision since it is time taking to collect evidence.

Experiment 4

In this experiment only 7 attributes are used.

If (typeoscrime=sexual assault and ageofvictim=9-18 and offenderage=18-45 then sex=F) which is to mean:

A victim at the age of 9-18 years, exposed to sexual assault and offended by age group 18-45 years are female.

If (typeofcrime=sexual assault and ageofvictim=9-18 and Education=1-6 then Sex=F) which is to mean: Victims at the age of 9-18 years, at education level of 1-6 and exposed to sexual assault are female.

Like the rules in experiment three the rules experiment 4 are also interesting rules. Because three of the rules sexual assaults are potential offences for victims at the age of 9-

18 years and are offended by offenders at the age of 18-45 and are female in gender not male.

Comparison of the rules generated from both association rule and classification.

This discussion is to compare and present similar rules generated from association rule and classification model. Most of the rules generated in association rule do appear in the list of rules generated by the classification model. For example the rule “If (sex=F and ageofvictim=9-18 and offenderage=18-45 and typeoffence= sexual assault then cluster1)” similar with the rule generated in “experiment 4” of the assossiation model which is “If typeoscrime=sexual assault and ageofvictim=9-18 and offenderage=18-45 then sex=F)”. The rule generated by the classification tree has additional dependent attribute which is the cluster as a class label. The classification model predicts better when its rule overlaps with the association rule model.

Those that do not appear in the classification model are those that contain attributes that are not selected as bets attributes in the clustering and classification models. Fore example the rule like, If (Age_Of_victim=ag2 and Sex=F and Job=student and Decision=not then Religion=rel1) and If (Habit=no and Religion=rel1 and Living=famyes then MaritalStatus=single) some of the examples that do not appear in the classification model due to their inclusion of the attributes “Living”, “MaritalStatus” and “Habit”.

The other difference in the two models is that there are rules that are not generated by the association rule model but appear in the classification model. This difference comes because of the values of the attributes which are relatively many valued. These include year and sub-city. Many valued attributes fail to generate pattern in association rule model. But in the classification model, all the testing records are classified and placed into one of the predefined classes based on the training set if they are not pruned.

CHAPTER SIX

Conclusion and Recommendations

6.1. Conclusion

Law enforcement agencies like police in general are data rich but knowledge poor. This situation resulted from the behavior of law enforcement. Whenever crime event is committed, police officer visits the place and records detail evidences that can help for the purpose of investigation. They may record it in structured or unstructured format. They don't worry about the structure and format rather about their detail information that gives those clues that serve as evidence. Due to this, the data grows exponentially from time to time requiring analysis tool to extract knowledge from it.

The most compelling and promising tools for knowledge extraction are data-mining tools. However, the use of data mining is scarce in the domain of law enforcement. Data mining techniques are solutions in discovering non- trivial, hidden and potentially useful patterns out of large volume of data collected overtime from many sources. Since it is impossible that trained researchers examine all possible interesting patterns in such huge amounts of data, one requires an intelligent assistant to process the data and to autonomously (or at least with very little guidance) analyze it.

The role of law enforcement agencies is to secure legal order. Securing legal order implies two tasks. First, public safety should be established and maintained to reduce the growing unsafe feeling of citizens. Second crime should be tracked down and the offenders should be prosecuted subsequently. This research is basically conducted to achieve these two roles with respect to offences against children.

For this research, the data was collected from an NGO FSCE and APCO. The data was stored in MS-Excel database categorized based on year of report. Data preprocessing preparing the data for model building was conducted in both MS-Excel filtering and WEKA filters tools. The attribute selection was made with the help of WEKA information gain attribute evaluation and the domain experts as well as preprocessing assumptions.

The modeling methodology applied was CRISP-DM model. Different literatures were reviewed with the aim to bring the problem into a data mining problem.

The selected data mining techniques to be implemented for this research are classification, association rule and clustering. For classification model J48 decision tree is used, for clustering K-means algorithm is used and for association rule the appriori algorithm was used. Eight experiments were undergone for the three techniques. These experiments were performed as three with apriori, three with K-means and two with J48 decision tree.

From the two decision tree experiments the one with default value of the parameter "Number of Objects" has high accuracy. But due to the other criteria the number of leaves of the tree, the size of the tree, and the time taken to build the model, the second one with number of objects set twenty (20) outperforms better. The size of the tree with "Number of Objects" increased to 20 in reduced and has fewer number of leaves than the first one even though still there exists a problem to visualize its tree structure by WEKA in a readable form. Generally, the classification over clustering helps us to convert clustering description model in to prediction one through model development.

The clustering task is performed to help the classifier to predict the class of a particular record. The dependent variable for the classifier is the cluster index which serves as the class values. The prediction is done to show the potential risks with children knowing some attributes in the crime records or to generate rules.

Similarly the association rule mining helps us to identify the profile of the victim children in relation to the offender age and year. It shows inter attribute relationship. Whenever there is a victim with the profile, having any special habit, orthodox and living with families then he/she is single.

In general, as commented by domain experts the results from this research are encouraging. It is important to determine attributes and their values to understand the profile of victim children and identify which children are exposed to which crimes. The associations rule mining shows what profiles are common with the victim children. The clustering technique shows which attributes are common in a given cluster since it is

finally cluster index serves as class label. This helps to generate rules to identify the potential victim children for a specified crime category. Having this knowledge FSCE in collaboration with APCO can educate and counsel victims as well as work on awareness creation for children.

6.3. Recommendations

Even though this research is done for academic purpose; its output would help law enforcements to identify potential victim children, help them up on call for help and protect children from offences against them proactively. This research has identified the profile of victim children by generating rules like If (typeofcrime=sexual assault and ageofvictim=9-18 and offenderage=18-45 then sex=F) which is to mean: A victim at the age of 9-18 years, exposed to sexual assault and offended by age group 18-45 years are female. Additionally it predicts which children are exposed to which crime category using classification over clustering models. It is helpful for planners to use such rules to design programs to protect children (e.g. selecting target areas and allocating resources).

To perform these tasks more efficiently and effectively, more emphasis should be given to service like reporting crime events, modifying the database records, accessing these recoded data without violating privacy issues of the victims and offenders. A thorough discussion among technical experts and domain experts is done for the purpose of database design, on the patterns discovered to identify meaningful ones, to deploy the model to the benefit of the society is important. In the existing database the available recoded attributes are not enough to generate different analysis results. For example the attributes special habit, education, sex and religion are equally important to describe the profile of offenders to that of victims.

Results can be further enhanced and expanded by improving the detailed content of the database. It can incorporate other attributes such as sex, habit, education, etc. of the offender. Future researches can test other mining techniques such as artificial neural network, time series, summarization, etc. to improve the performance of the model. Continuous data can also be used instead of just categorical data. Another area is to use text mining on the vast amount of unstructured data available in crime records.

Reference:

1. Adderley, R. and Musgrove, P.B. (2001) *'Data mining: case study modeling the behavior of offenders who commit serious sexual assaults'*, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA, ACM , pp. 215-220.
2. Adderley, R., and Musgrove, P.B., (2003) *'Modus operandi modelling of group offending, a data mining case study'*, International Journal of Police Science and Management, Vol. 5 no. 4 pp 265-276.
3. Adderley, R., William and Musgrove, P. (2001) *'Police crime recording and investigation systems, A user's view'*, International Journal of Police Strategies and Management, Vol. 24, No. 1, pp. 100-114.
4. Agrawal, R. and Srikant, R. (1994) 'Fast *algorithms for mining association rule in large databases'*, IBM Almaden Research Center, San Jose, California, Research report.
5. Agrawal, R., Imielinski, T. and Swami, A. (1993) 'Mining association rules between sets of items in large databases', ACM-SIGMOD, International Conference Management of Data, Washington, D.C. pp 207-216
6. Akpinar E. and Usul N. (2004) '*Geographic Information Systems Technologies in Crime Analysis and Crime Mapping*'
7. Andargachew, T. (1988) '*The crime problem and its correction'*, Vol. I Addis Ababa University, Law School training module (unpublished).
8. Anjewierden, A., Koll¨Offel, B. and Hulshof, C. (2007) '*Towards educational data mining, Using data mining methods for automated chat analysis to understand and support inquiry learning processes*', International Workshop on Applying Data Mining in e-Learning, ADML'07.
9. *Antonie, M. L., Zaiane, O. R.,Coman, A.(2001), 'Application of Data Mining Techniques for Medical Image Classification'*, Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco.

10. Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H. (2005) *'A Framework for Data Mining Based Multi-Agent', An Application to Spatial Data*, Vol. 5, ISSN 1307-6884, Proceedings of World Academy of Science, Engineering and Technology.

11. Baeza-Yates and Ribeiro-Neto, B. (1999) '*Modern Information Retrieval'*, ACM Press, New York.

12. Bishop, C. M. (1995) '*Neural Networks for Pattern Recognition'*, Oxford University Press, Oxford.

13. Botia, J. A., Garijo, M. y Velasco, J. R. and Skarmeta, A. F.,(1998) 'A Generic Data mining System basic design and implementation guidelines*', A Technical Project Report of CYCYT project of Spanish Government*

14. Brachman, R., J., Anand, T., (1996) '*The Process of Knowledge Discovery in Databases', In, 'Advances in Knowledge Discovery and Data Mining*', Usama

15. Brantingham P.L. and Brantingham P.J. (1981) '*Notes on the geometry of crime*'. In, P.J. Brantingham & P.L. Brantingham (eds), Environmental Criminology, Waveland Press, Inc., Prospect Heights, IL, pp. 27-54

16. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C. J. (1984) *'Classification and regression trees'*, Monterey, CA, Wadsworth.

17. *Cai, W. and Li, L. (2004), 'Anomaly Detection using TCP Header Information'*, STAT753 Class Project, (unpublished).

18. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinhertz, T., Shearer, C., Wirth, R., (2000) '*CRISP-DM 1.0 Step-by-step data mining guide'*, USA, SPSS Inc. CRISPWP-0800 2000.

19. Chau, M., Xu, J., Chen, H. (2002) '*Extracting meaningful entities from police narrative reports*', *Proceedings of the National Conference for Digital Government Research.* Los Angeles, California, USA pp 271-275.

20. Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J. J., Wang, G., Zheng, R., Atabakhsh, H. *(2003) 'Crime Data Mining, An Overview and Case Studies', A project under NSF Digital Government Programme, USA, 'COPLINK Center, Information and Knowledge Management for Law Enforcement*, Decision Support Systems (DSS), Special Issue "Digital Government: technologies and practices", Vol. 34, No. 3, pp 271-285

21. Chen, H., Chung, W., Xu Jennifer, J., Wang, G., Qin, Y., Chau, M. (2004) *'Crime Data Mining, A General Framework and Some Examples'*, IEEE Computer Society, Technical Report.
22. Clarke, R.V., Felson M. (1993) *'Introduction, Criminology, Routine activity, and rational choice in Routine activity and rational choice, Advances in criminological theory', volume 5*, Clarke, R.V., Felson, M. (eds.) New Jersey, USA, Transaction Publishers.
23. Cohen, L.E. And Felson, M. (1979) *'Social Change and Crime Rate Trends, A Routine Activity Approach', American Sociological Review*.
24. Cope, N. (2004), *'Intelligence Led Policing or Policing Led Intelligence? Integrating Volume Crime Analysis into Policing'*, British Journal of Criminology Vol. 44, pp 188-203.
25. Dinan K. Owed justice (2000) 'Thai women trafficked into debt bondage in Japan', New York, NY, Human Rights Watch.
26. Dragoon, A. (2003*) 'Business intelligence gets smart(er)'*, CIO, 15th September 2003.
27. Elizabeth R. Groff and Nancy G. La Vigne (nd), *'Forecasting the future of Predictive crime mapping'*, Crime Prevention Studies, volume 13, pp.29-57.
28. Ellsberg MC. Candies in hell (1997) 'domestic violence against women in Nicaragua', Umea˚, Umea˚ University.
29. Foster, D. P. and Stine, R. A. (2004) *'Variable Selection in Data Mining, Building a Predictive Model for Bankruptcy'*, Journal of the American Statistical Association, Alexandria, VA, ETATS-UNIS, vol. *99, **ISSN** 0162-1459, pp. 303-313.*
30. Garner, S., R. (nd) 'WEKA, *The Waikato Environment for Knowledge Analysis'*, Department of Computer Science, University of Waikato, Hamilton.
31. Gordon, A.D. (1981) 'Classification', published by Chapman and Hall HALKIDI.
32. Gupta M., B. Chandra and M. P. Gupta (2008) 'Crime Data Mining for Indian Police Information System' Indian Institute of Technology Delhi, Hauz Khas, New Delhi - 110 016, India

33. Hand, D., Mannila, H. and Smyth, P. (2001) '*Principles of Data Mining'*, Cambridge, MA, The MIT Press.
34. Hipp, J. Ulrich, H. (2000) '*Algorithms for Association Rule Mining- A General Survey and Comparison'*, Tubingen, University of Tubingen.
35. Jonathan C. and Andrew W. (nd) '*Forecasting Crime'*, An Ethical Conundrum, Available from, http,//www.aic.gov.au/conferences/mapping/muscat.pdf (accessed on April 20, 2011)
36. Kilonzo, N., Ndung'u, N., Nthamburi, N., Ajema, C., Taegtmeyer, M. and Theobald, S. (2009) '*Sexual violence legislation in sub-Saharan Africa, the need for strengthened medico-legal linkages***'** *Reproductive Health Matters 0968-8080*
37. Kohavi, R. (1995) *'A study of cross-validation and bootstrap for accuracy estimation and model selection'*, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo, CA, Morgan Kaufmann.
38. Krause, P and Clarke, D. (1993) *'Uncertain reasoning, an artificial intelligence approach'*, Intellect Books.
39. Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., and Tseng, T.L. (2000) *'Data Mining, Medical and Engineering Case Studies'*, Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, pp. 1-7.
40. Larose, D. T. (2005) 'Discovering Knowledge in Data: An Introduction to Data Mining', *ISBN 0-471-66657-2, John Wiley & Sons, Inc.*
41. Leul, W. (2003) 'T*he application of data mining in crime prevention: the case of Oromia police commission'*, Addis Ababa University, M.Sc. thesis.
42. Mena, J. (2003) '*Investigative Data Mining for Security and Criminal Detection*', Butterworth Heinemann, ISBN 0-7506-7613.
43. Mitchell, K. J., Finkelhor, D. and Wolak, J. (2001) *'Risk factors for and impact of online sexual solicitation of youth'*, JAMA, Vol.285, No.23, pp.3011-14.
44. Deshpande, S. P., Thakare, V. M. (2010) *'Data mining system and applications'*, a review Thakare2 International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1. pp 173-179.
45. Rogers, J. (2001) 'Data Mining Using the EM Clustering Algorithm on Places Rated Almanac Data', INFT.

46. Saygin, Y. and Ulusoy, O. (2002) 'Exploiting data mining techniques for broadcasting data in mobile computing environments', IEEE Trans, Knowl, Data Eng. pp1387–1399.
47. Schultz, M. G., Eskin, Eleazar, Zadok, Erez, and Stolfo, Salvatore, J.(2001) *'Data Mining Methods for Detection of New Malicious Executables'*, Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society Washington, DC, USA.
48. Sherman, R., (2005) Data integration advisor, 'set the stage with data preparation, *DM Review'*.
49. Thearling, K. (2003*) 'An introduction to data mining',* McGraw- Hill, New York.
50. Thibault, E., Lynch, L., and McBride, R. (2006) Proactive Police Management, Seventh Edition. London, Prentice Hall.
51. Thomsen, E., (1998) Presentation, '*Very Large Data Bases / Data Mining' Summit,* Beverly Hills, California.
52. Tilley, N. (2005), '*Community Policing, Problem-Oriented Policing and Intelligence-Led Policing'* In T. Newburn (Ed.), Pleymouth Devon, Willian Publishing, Handbook of Policing pp. 311-339.
53. Wang, G., Chen, H. and Atabakhsh, H. (2004) '*Automatically detecting deceptive criminal identities*, Comm ACM.
54. Weiss, S., M. and Indurkhaya, N. (1998) '*Predictive Data Mining - a Practical Guide'*, Morgan Kaufman Publishers.
55. Witten, I. H. and Frank, E. (2000) Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations.
56. Woldekidan, K. (2003) '*Application of KDD on crime data to support the advocacy and awareness raising program of forum on street children Ethiopia', Addis Ababa University*, M.Sc. thesis.
57. World Health Organization (1996) *'Violence against women',* WHO consultation. FRH/WHD/96.27, Geneva, WHO.

Appendix A: Description of some of the attributes in the FSCE database

S.No	Attribute	Data type	Description
1	Registration number	Date	Date in which the crime was reported
2	Crime type	Text-nominal	Crime type which is considered as crime and reported to the police.
3	Code of crime	Text-nominal	Code used instead of the crime at the time of investigation.
3	Sex	Text-nominal	Sex of the victim
4	Victim age	Number	Age of the victim
5	Education	Text-nominal	Level of education of the victim
6	Habit	Text-nominal	Special habit that the victim has
7	Job	Text-nominal	Job of the victim
8	Marital status	Text-nominal	Marital status of the victim
9	Religion	Text-nominal	The religion of the victim
10	Lining	Text-nominal	Living situation of the victim
11	Offender age	Text-nominal	Age of offenders
13	Decision	Text-nominal	Decision given by court

14	Special name of places	Text-nominal	Special places where the crime is committed
15	Sub-city	Text-nominal	The sub-city where the crime was committed
16	Report date		
17	Victim name	Text-nominal	Name of the victim
19	Offender name	Text-nominal	Name of the offender
20	Victim kebele	Text-nominal	Address of the victim’s kebele
21	Offender kebele	Text-nominal	Address of the offender’s kebele

Appendix B: attributes and their values in the dataset

Sex
- a. Male
- b. Female

Age
- a. 0-9 (ag1)
- b. 9-18 (ag2)

Education
- a. Illiterate (Ed1)
- b. Read and write only (Ed2)
- c. Kg (Ed3)
- d. 1-6 (Ed4)
- e. 7-12 (Ed5)
- f. above 12(Ed6)

Marital status of the victim
- a. single
- b. not

Religion of the victim)
- a. Orthodox (Rel1)
- b. Muslim (Rel2)
- c. Protestant (Rel3)

Job of the victim
- a. Student
- b. Not

Habit (special habit of the victim)
- a. No bad habit
- b. Yes

LIVING (Living status of the child before the offence)
- a. Family yes
- b. with relative (family not)
- c. with someone not relative as a servant (family not)
- d. on the street (family not)
- e. child care institution (family not)

Offender age
- a. 9-18 (oag1)
- b. 18-45 (oag2)
- c. Above 45 (oag3)

Sub-city
- a. Addis ketema (1)
- b. Akaki (2)
- c. Arada (3)
- d. Bole (4)
- e. Gulele (5)
- f. Kirkos (6)
- g. Kolfe (7)

h. Lideta (8)
i. Nifas silk lafto (9)
j. Yeka (10)

Decision

a. Closed
b. Not(on progress)

Appendix C: Partial view of the pruned decision tree

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

Sex = M
|   ageofvictim = ag1
|   |   Education = ed1
|   |   |   Offenderage = oag1: cluster1 (26.46/8.97)
|   |   |   Offenderage = oag2: cluster2 (31.23/0.04)
|   |   |   Offenderage = oag3: cluster2 (7.25/0.0)
|   |   Education = ed2: cluster2 (47.55/7.39)
|   |   Education = ed3: cluster2 (0.0)
|   |   Education = ed4
|   |   |   Offenderage = oag1
|   |   |   |   Year = 1997: cluster2 (12.71)
|   |   |   |   Year = 1998: cluster2 (11.44/0.02)
|   |   |   |   Year = 1999: cluster1 (21.89)
|   |   |   |   Year = 2000: cluster2 (38.19)
|   |   |   |   Year = 2001: cluster2 (10.95/0.14)
|   |   |   |   Year = 2002: cluster2 (16.39/0.26)
|   |   |   Offenderage = oag2: cluster2 (217.23/0.7)
|   |   |   Offenderage = oag3: cluster2 (12.63/0.0)
|   |   Education = ed5
|   |   |   Year = 1997: cluster2 (0.79)
|   |   |   Year = 1998: cluster2 (33.49/14.68)
|   |   |   Year = 1999: cluster1 (54.07/0.4)
|   |   |   Year = 2000: cluster2 (9.15/2.0)
|   |   |   Year = 2001: cluster2 (9.79/1.05)
|   |   |   Year = 2002: cluster2 (4.87/2.26)
|   |   Education = ed6: cluster2 (0.0)
|   ageofvictim = ag2
|   |   Offenderage = oag1: cluster1 (689.09/18.43)
|   |   Offenderage = oag2
|   |   |   Education = ed1: cluster1 (20.25/8.81)
|   |   |   Education = ed2: cluster1 (4.88/2.78)
|   |   |   Education = ed3: cluster1 (0.0)
|   |   |   Education = ed4
|   |   |   |   Year = 1997: cluster2 (40.64/1.21)
|   |   |   |   Year = 1998
|   |   |   |   |   typeofcrime = offe1: cluster2 (70.45)
|   |   |   |   |   typeofcrime = offe2: cluster3 (6.0)
|   |   |   |   |   typeofcrime = offe3: cluster2 (22.16)
|   |   |   |   |   typeofcrime = offe4: cluster2 (4.4/0.13)
|   |   |   |   |   typeofcrime = offe5: cluster2 (2.63)
```

```
|   |   |   |   |   typeofcrime = offe6: cluster2 (5.0)
|   |   |   |   Year = 1999: cluster1 (159.9)
|   |   |   |   Year = 2000: cluster2 (143.6/0.16)
|   |   |   |   Year = 2001: cluster2 (102.74/23.84)
|   |   |   |   Year = 2002: cluster2 (112.57/3.71)
|   |   |   Education = ed5: cluster1 (365.93/19.67)
|   |   |   Education = ed6: cluster1 (0.0)
|   |   Offenderage = oag3: cluster1 (78.79/21.54)
Sex = F
|   ageofvictim = ag1
|   |   typeofcrime = offe1
|   |   |   Education = ed1: cluster2 (44.27/6.15)
|   |   |   Education = ed2: cluster2 (22.73/3.08)
|   |   |   Education = ed3: cluster2 (0.0)
|   |   |   Education = ed4: cluster2 (171.23/8.7)
|   |   |   Education = ed5
|   |   |   |   Offenderage = oag1: cluster1 (21.08/1.91)
|   |   |   |   Offenderage = oag2: cluster2 (25.11/10.0)
|   |   |   |   Offenderage = oag3: cluster1 (1.68/0.68)
|   |   |   Education = ed6: cluster2 (0.0)
|   |   typeofcrime = offe2
|   |   |   Year = 1997: cluster3 (42.0)
|   |   |   Year = 1998: cluster3 (66.0/7.0)
|   |   |   Year = 1999
|   |   |   |   Education = ed1: cluster3 (0.0)
|   |   |   |   Education = ed2: cluster3 (0.0)
|   |   |   |   Education = ed3: cluster3 (0.0)
|   |   |   |   Education = ed4: cluster3 (32.44)
|   |   |   |   Education = ed5: cluster1 (30.65/13.65)
|   |   |   |   Education = ed6: cluster3 (0.0)
|   |   |   Year = 2000: cluster2 (125.0/3.0)
|   |   |   Year = 2001: cluster3 (84.0)
|   |   |   Year = 2002: cluster3 (60.27/2.0)
|   |   typeofcrime = offe3: cluster2 (33.0/15.0)
|   |   typeofcrime = offe4: cluster2 (31.06/14.85)
|   |   typeofcrime = offe5: cluster2 (3.0)
|   |   typeofcrime = offe6: cluster2 (19.0/9.0)
|   ageofvictim = ag2
|   |   Offenderage = oag1
|   |   |   Education = ed1: cluster3 (24.43/8.23)
|   |   |   Education = ed2: cluster3 (3.16/1.03)
|   |   |   Education = ed3: cluster1 (0.0)
|   |   |   Education = ed4
|   |   |   |   typeofcrime = offe1
|   |   |   |   |   Year = 1997: cluster1 (14.56)
|   |   |   |   |   Year = 1998: cluster1 (25.73)
|   |   |   |   |   Year = 1999: cluster1 (26.39)
|   |   |   |   |   Year = 2000: cluster1 (21.76/1.2)
|   |   |   |   |   Year = 2001: cluster3 (24.01)
|   |   |   |   |   Year = 2002: cluster1 (20.56)
|   |   |   |   typeofcrime = offe2: cluster3 (193.83)
|   |   |   |   typeofcrime = offe3: cluster3 (18.43/5.0)
|   |   |   |   typeofcrime = offe4: cluster1 (3.86/1.13)
|   |   |   |   typeofcrime = offe5: cluster3 (1.0)
|   |   |   |   typeofcrime = offe6: cluster3 (9.0)
|   |   |   Education = ed5: cluster1 (259.08/27.37)
|   |   |   Education = ed6: cluster1 (0.0)
```

```
|   |   Offenderage = oag2
|   |   |   typeofcrime = offe1
|   |   |   |   Education = ed1: cluster3 (25.45/6.36)
|   |   |   |   Education = ed2: cluster2 (3.18/1.14)
|   |   |   |   Education = ed3: cluster3 (0.0)
|   |   |   |   Education = ed4
|   |   |   |   |   Year = 1997: cluster3 (18.63)
|   |   |   |   |   Year = 1998: cluster3 (35.52)
|   |   |   |   |   Year = 1999: cluster3 (41.61)
|   |   |   |   |   Year = 2000: cluster2 (76.28)
|   |   |   |   |   Year = 2001: cluster3 (61.54)
|   |   |   |   |   Year = 2002: cluster3 (31.35)
|   |   |   |   Education = ed5
|   |   |   |   |   Year = 1997: cluster1 (4.3/0.3)
|   |   |   |   |   Year = 1998: cluster1 (17.21/1.21)
|   |   |   |   |   Year = 1999: cluster1 (29.91/0.91)
|   |   |   |   |   Year = 2000: cluster1 (26.45/1.82)
|   |   |   |   |   Year = 2001: cluster3 (29.61)
|   |   |   |   |   Year = 2002: cluster1 (20.3/0.3)
|   |   |   |   Education = ed6: cluster3 (0.0)
|   |   |   typeofcrime = offe2: cluster3 (737.86)
|   |   |   typeofcrime = offe3: cluster3 (73.6/3.0)
|   |   |   typeofcrime = offe4: cluster3 (23.23/6.0)
|   |   |   typeofcrime = offe5: cluster3 (33.63)
|   |   |   typeofcrime = offe6: cluster3 (57.63)
|   |   Offenderage = oag3
|   |   |   typeofcrime = offe1
|   |   |   |   Education = ed1: cluster3 (6.25/3.0)
|   |   |   |   Education = ed2: cluster1 (0.0)
|   |   |   |   Education = ed3: cluster1 (0.0)
|   |   |   |   Education = ed4: cluster3 (44.25/14.29)
|   |   |   |   Education = ed5: cluster1 (25.29/1.22)
|   |   |   |   Education = ed6: cluster1 (0.0)
|   |   |   typeofcrime = offe2: cluster3 (63.08)
|   |   |   typeofcrime = offe3: cluster3 (7.18)
|   |   |   typeofcrime = offe4: cluster3 (0.06)
|   |   |   typeofcrime = offe5: cluster3 (2.07)
|   |   |   typeofcrime = offe6: cluster3 (6.07)

Number of Leaves  :     109

Size of the tree :      136


Time taken to build model: 0.05 seconds
```