ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

APPLICATION OF KDD ON CRIME DATA TO SUPPORT THE ADVOCACY AND AWARENESS RAISING PROGRAM OF FORUM ON STREET CHILDREN ETHIOPIA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFROMATION SCIENCE

By

WOLDEKIDAN KIFLE

JULY 2003

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

APPLICATION OF KDD ON CRIME DATA TO SUPPORT THE ADVOCACY AND AWARENESS RAISING PROGRAM OF FORUM ON STREET CHILDREN ETHIOPIA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFROMATION SCIENCE

By

WOLDEKIDAN KIFLE

JULY 2003

Acknowledgment

First and foremost I owe it to my parents for having come this far. Their unfailing support and belief in me have been a constant source of strength and enthusiasm.

I would also like to extend my gratitude to my advisors, Dr. Gashaw Kebede and Ato Shegaw Anagaw, for their constructive comments on the research document.

I would like to express my appreciation to the wonderful friends I have had at the Department of Information Science most of whose company I can count on 24/7. I will particularly miss the afternoon and night sessions we have had.

Last but not least, I would like to thank the FSCE staff, Ato Tedla G/Mariam in particular, without whose good will to supply me with the data, I would not have accomplished this research.

Table of Content

List of abbreviations

AI	Artificial Intelligence
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

List of Tables

List of Appendices

List of Figures

Abstract

This thesis work gives an account of the process followed to determine the application of KDD to support the advocacy and awareness raising program of FSCE and Addis Ababa Police Commission, and the potential of a data mining learning scheme to discover regularities that underlie the crime dataset.

The KDD process as described by Fayyad, Piatetsky-Shapiro and Gregory (1996) that consists of five major phases, namely understanding of the problem domain, data selection, data preprocessing, data mining, and discussion and interpretation was adopted.

The discovery task was run on the crime database that consists of 10,878 records/tuples in 17 tables describing a total of 25 attributes. Association rule mining, an exploratory data mining technique was applied to accomplish the goal of the research. To this effect, the Apriori algorithm, which is an implementation of the Association rule in the Weka software, was used.

The KDD process can be applied on the crime database to good effect since it can result in rules that can serve as input for the advocacy and awareness raising program. On the basis of subjective (opinions of domain experts) and objective (support and confidence) measures of interestingness, a number of rules having practical relevance or that can add to the current knowledge in the problem domain were identified.

This thesis describes the above work in detail.

Chapter One

Introduction

This thesis focuses on investigating the application of KDD to explore and discover regularities within the crime dataset. An overview of the research field and the specifics of this particular research are presented below.

1.1 Background

The amount of information in the world is estimated to double every 20 month. As a result, large number of scientific, government and corporate information systems are being overwhelmed by a flood of data that are produced and stored routinely, developing into large databases amounting to giga (and even tera) bytes of data. These databases comprise potentially highly valuable information, but it is not within the capacity of human beings to analyze massive amounts of data and draw meaningful patterns (Deogun, 1997).

This explosive growth in data and databases has resulted in an acute need for novel techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Consequently, data mining has become a research area with increasing importance (Chen et al, 1996).

Data mining has emerged as visible research and development area in the 1990s (Mannila et al, 1997). It is a promising and an interdisciplinary research area spanning several disciplines such as database systems, machine learning, statistics, and expert systems (Deogun et al, 1997).

Data mining, which is also referred to as knowledge discovery in databases (KDD), means a process of non-trivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases. There are also many other terms, appearing in some articles and documents, carrying a similar or slightly different meaning, such as knowledge extraction from databases, knowledge extraction, data archeology, data dredging, data analysis, and so on (Chen et al, 1996).

However, according to prominent researchers in the field (Fayyd et al. b, 1996), despite their popular interchangeable usage, there is significant difference between data mining and KDD. While KDD refers to the whole process of changing low level data into high level knowledge, data mining constitutes one of the phases in this process, namely "the application of specific algorithms for extracting patterns from data" (Fayyd et al., 1996).

Several typical kinds of knowledge can be discovered through KDD, including association rules, characteristic rules, classification rules, clustering, evolution, and deviation analysis (Chen et al, 1996).

Different organizations have been adapting KDD to their own environment, in their own way. As it is the case in other sectors of the society, the NGO community also depends on the availability of

information for decision-making purpose. They also play a role in the increase in the amount of data generated. There is also recognition among non-governmental and international organizations to make use of improved data management and data analytical tools, to rip the benefits that can be accrued from the enormous data/database that is at their disposal. As a case in point, UNESCO develops, maintains and disseminates two interrelated software packages for database management (CDS/ISIS) and data mining/statistical analysis (IDAMS). These software systems are available on the *"UNESCO Information Processing Tools"* CD-Rom, distributed on a non-commercial basis, to agreed distributors and Institutions upon request (UNITeS, nd).

NGO is a broad term encompassing a wide array of diverse organizations. According to the World Bank, the term NGO refers to *"private organizations that pursue activities to relieve suffering, promote the interests of the poor, protect the environment, provide basic social services or undertake community development"* (World Bank, 2001).

It is almost 30 years since many of the NGOs first began working in Ethiopia. The leading ones (both national and international) originally became involved in order to mitigate the effects of the droughts of 1973-1974 and 1984-1985 (The Code of Conduct for NGOs, 1998).

Among the wide variety of roles that NGOs play, Advocacy for and with the Poor is the major one. This refers to the role of NGOs where they become spokespersons or ombudsmen for the poor and attempt to influence government policies and programs on their behalf. NGOs can also assume the

role of providing technical assistance and training whereby they try to develop the capacity of community-based organizations and government institutions.

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.

In line with this, as a component of its awareness raising and child protection program, FSCE in collaboration with the Addis Ababa Police Commission collects stores and maintains data about offences committed against children in the city. The crime data has been gathered for the past four years from the 28 Woreda police stations in the city. The database was initiated with the purpose to introduce a systematic way of keeping data, and to be able to carry out data analysis for preparing periodic (quarterly and annual) reports to donors about the number of victim children reported to the police and who have received certain services such as counseling or medication by FSCE *(see appendix 1 to see the format for the periodic reports)*. The data regarding victim children is collected in relation to 25 attributes *(See appendix 2 to see the metadata for the attributes contained in the database).*

1.2 Statement of the Problem

This experimental research was undertaken in order to perform exploratory data analysis on the data, and discover regularities, which are pertinent in the prevention and/or control of offences committed against children in Addis Ababa. The research also attempts to determine if association rule mining software, which is the principal algorithm chosen for the purpose, can be effectively utilized on the crime data.

To the best of the researcher's knowledge, there has not been any data mining study applied on crime data in Ethiopia. However, there are experiences of applying data mining technology on crime database in developed countries in order to get a clearer picture of how, when and where crimes were being committed in order to best deploy their limited resources and to see if predictable patterns in crime could be identified (Veenendaal, nd) *(see Chapter three: Application of KDD in crime prevention and control for further detail).*

What is more, in the case of the crime database, which is the focus of this study, except for occasional quarterly and annual statistical reports that are prepared for donors to communicate the number of children that received certain services such as counseling and medication, no significant application of the Crime Database at the Addis Ababa Police Commission has been made to date. Despite its current underutilization, the database has the potential to offer valuable knowledge that can particularly be used as a source of input in the awareness raising activities of FSCE. It is, therefore, with this understanding that this experimental research was conducted.

1.3 Significance of the research

During the past four years (10,878) records have been gathered and stored in the database. However, the use of the database has been confined to date to the preparation of quarterly and annual statistical reports that depict the distribution of the records over a combination of related fields *(see appendix 1 to see the list of combinations of attributes used to prepare periodic reports).*

The database has the potential to result in valuable information provided that good data analysis techniques such as KDD are applied. The records in the database can serve as an input for the development of full-fledged data mining application using association rule algorithm, which can be employed for the purpose of supporting the activities of the police and organizations like FSCE in the area of preventing and controlling crime committed against children. This research aims to accomplish this goal by contributing to the existing knowledge by carrying out exploratory data mining on the crime data.

The results of the KDD process could give FSCE and the Addis Ababa Police Commission insight in making decisions, for instance, in relation to which group of society to focus their awareness raising activity on, and what programs/services to design to best meet the needs of the victim children. The results could also give insight on issues to emphasize during awareness raising and training programs, and allocation of their limited human resource in areas where crime is fraught.

1.4 Objectives of the Study

1.4.1 General Objective

The general objective of the study is to examine the application of the KDD process in discovering the different regularities that underlie the dataset in the crime database maintained by FSCE and Addis Ababa Police Commission.

1.4.2 Specific Objectives

To accomplish the above stated general objective, the following specific objectives were formulated.

- To develop an understanding of the application domain through review of relevant documents and interview with domain experts from FSCE and Addis Ababa Police Commission
- To identify the problem for the KDD process in consultation with domain experts
- To conduct preprocessing of data including tasks like selecting attributes relevant to the problem domain/goal of the project, removing noise, and handling missing values,
- To apply the data mining tool on the selected data,
- To look for regularities in the data in terms of segmenting the data and preparing the profile of different groups of victim children
- To evaluate generated rules by employing objective (confidence and support) and subjective measures of interestingness (domain experts' opinion on relevance of discovered regularities)
- To deploy the discovered regularities by documenting and reporting the pertinent discovered regularities

1.5 Research Methodology

Undertaking a KDD project using a step-by-step methodology has gained wide acceptance, as it ensures repeatability of the project. There is a wide acceptance of the steps that constitute the KDD process (Hipp & Nakhaeizadeh, nd), and they have been adopted in this project. The various steps followed and tasks accomplished are discussed below.

1.5.1 Understanding the Problem Domain

A close look of the problem environment was the first step taken in this KDD project. On the basis of the insights gained from this phase, the data-mining problem was defined. Information was secured by looking at periodic reports of the organization and discussion with relevant staff of the organization. *(See "Chapter Two: Understanding the Problem Domain" to refer to the findings of the analysis).*

1.5.2 Understanding the Data

After getting familiar with the problem area, I went on to accomplish the second step of the KDD process, namely understanding of the data. At this phase of the process, attempt was made to understand the attributes and their corresponding values.

Generally speaking, the Crime Database, which is the focus of this study, comprises three different and unrelated files where records about victimized children, persons committing offenses against children, and juvenile delinquents are stored. Although two of the files, namely the file on victimized

children and the one on persons committing offences against children, could have been related, in the absence of a unique common matching field, this has not been possible. Hence, the three files are considered to be independent and are used for preparing different reports to donors.

Of the three files contained in the database, the file on victimized children which contains 10,878 records and having 25 fields is the focus of this research. The data is stored in 17 different tables. Except for being stored separately and referring to data collected at different times, the 17 tables hold information about the 25 attributes. The file is in a relational database format *(See the Meta data for the attributes at Appendix 2).*

In this phase it was learned that there were more attributes in the crime dataset than actually required for this analysis. The attributes can be characterized as redundant or non-variant.

1.5.3 Tool Selection

Two data mining tools were chosen namely, Weka and Knowledge STUDIO to implement the discovery task

1.5.3.1 Weka

Weka, a machine-learning algorithm in Java, was adopted for undertaking the experiment. Weka constitutes several machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. Weka is also well suited for developing new machine

learning schemes. Weka is open source software issued under the GNU General Public License. It incorporates an association rule learner. Also included there are about ten different methods for classification, three for clustering, and six for numeric prediction and several so called "meta-schemes" (bagging, stacking, boosting . . .). In addition to the learning schemes, Weka also comprises several tools that can be used for datasets pre-processing (Palous, nd).

1.5.3.2 Knowledge Studio

Knowledge Studio was designed for both the business user and the professional analyst. With a user interface based on industry standard design, knowledgeSTUDIO has the look and feel of other business software applications commonly used in enterprise environments (ANGOSS, 2002).

KnowledgeSTUDIO contains a comprehensive and growing portfolio of advanced data mining algorithms. It supports eight decision-tree, three neural-network, two time-series, two clustering, logistic and linear regression, time series and covariance analysis (ANGOSS, 2002).

KnowledgeSTUDIO has extensive profiling and visual exploration techniques that accelerate the time required to quickly understand relationships in the data. Colorful and decisive reports can be easily created through tight integration with standard business productivity applications such as Microsoft Office (ANGOSS, 2002).

KnowledgeSTUDIO interoperates seamlessly with enterprise applications, data warehouse and data mart environments and other statistical tools by easily reading data from all major databases and statistical tools such as SAS (ANGOSS, 2002).

The association rule mining algorithm from the Weka software was used to carry out the main discovery task in the study- discovering regularities that underlie the crime dataset. On the other hand, the K-meansalgorithm fro the Knowledge studio software is used to divide the dataset into homogenous clusters that facilitate the learning process by minimizing the complexity of the data by grouping them into classes of similar instances.

1.5.6 Data Preparation

At this step, the dataset on which the algorithm is to be run on is constructed. For one thing, the data, which was stored in 17 different tables, which are distinct from one another on the basis of the quarter they were gathered, were merged into one large table.

For another, since Association rule mining, which is the selected data mining algorithm, is often used in situations where attributes are nominal, the data was transformed to fit this algorithm. Other tasks addressed include mapping data to a single naming convention, uniformly representing and handling missing data, and handling noise and errors.

1.5.7 Data Mining

The actual data mining took place at this stage. Based on the identified goals and the assessment of the available data, appropriate mining algorithm was chosen and run on the prepared data.

In many applications of machine learning to data mining, the explicit knowledge structures that are acquired, namely the structural descriptions, are at least as important, and often very much more important, than the ability to perform well on new examples. The use of data mining to gain knowledge- discover regularities in the data- and not just predictions is common (Witten & Frank, 2000).

As one can tell from the objectives of the research, gaining knowledge, discovering regularities with in the crime dataset in particular, is certainly the purpose of this study. Having this purpose in mind, unsupervised learning technique, namely association rule and clustering technique was adopted. The association rule algorithm was selected, for instance over classification rules, for it allows the prediction of any attribute, not just the class. This also provides with the freedom to predict combinations of attributes. Besides, association rules are not intended to be used together as a set, as classification rules are. Different association rules express different regularities that underlie the dataset, and they generally predict different things (Witten & Frank, 2000).

Since its introduction in 1993 the task of association rule mining has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern discovery methods in KDD (Hipp et al, 2000).

The association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems (Hipp et al, 2000).

Thus, in this study the association rule technique was applied on the cleaned and transformed data to extract valuable regularities within the crime dataset. This was made iteratively until acceptable results were obtained.

1.5.8 Evaluation and Interpretation

As it is often the case, with the application of the learning algorithm, several association rules were discovered from the dataset. Considering the large number of discovered rules, it was imperative to select only those rules that are interesting in relation to the purpose of the research- discovering regularities that are relevant to the prevention and/or control of crimes committed against children.

To accomplish this task, two different types of measure of interestingness were adopted, namely objective and subjective.

The objective measures of interestingness of a rule used in the project include confidence and support measures. While Support of an association rule refers to the number of instances the rule predicts correctly, Confidence of an association rule refers to the same number expressed as a proportion of the number of instances that the rule applies to (Witten & Frank, 2000).

According to Liu et al. (1997), however, although the objective measures of interestingness are useful in many respects, they often fail to capture all the complexities of the pattern discovery process. Accordingly, subjective measures of interestingness were also employed. These measures depend mainly on the user who examines the pattern. The use of subjective measures is considered even more important in many data mining application due to the fact that one can discover a large number of rules that are interesting "objectively" but of little interest to the user (Silberschhatz & Tuzhilns, 1996).

Subjective measures of interestingness are classified into actionable and unexpected. According to the unexpectedness measure, a pattern is interesting if it is "surprising" to the user. On the other hand, the actionability measure, which is deemed as the essential subjective measure of interestingness, considers a pattern as interesting if the users can act on it to their advantage. Actionability is an important subjective measure of interestingness because users are mostly interested in the knowledge that permits them to do their jobs better by taking some actions in response to the newly discovered knowledge (Silberschhatz & Tuzhilns, 1996). Hence, the researcher together with domain experts attempted to determine the subjective interestingness of the discovered regularities based on knowledge about the problem domain.

1.5.9 Deployment

Following the mining of data and assessing and interpreting the output, the results were documented in the form of report that can be referred by the people in the concerned organizations, FSCE and the

Addis Ababa Police Commission. What is more, the whole KDD process followed in the research was incorporated in the documentation to allow repeatability of the KDD process by the organizations.

Furthermore, the documentation of the KDD process followed in this research can be employed to develop a working KDD process that can be used to support decision making by serving as a source of input particularly for the awareness raising program of FSCE.

1.6 Scope of the Study

Among the various purposes of a KDD process, this research focuses on data exploration. Hence, supervised learning schemes such as classification schemes were not used. That is, discovering the underlying regularities within the crime dataset was chosen over prediction/classification of new examples. Among the different exploratory or unsupervised data mining techniques, Association rule mining and clustering were the chosen techniques.

The Crime Database comprises three different and unrelated files where records about victimized children, persons committing offenses against children, and juvenile delinquents are stored. Of the three files contained in the database, the file on victimized children which contains 10,878 records is the focus of this study.

1.7 Organization of the Study

The thesis is organized into five chapters. The first chapter introduces the research project by defining the problem that constitutes the basis of the paper, significance and scope of the research. It is also in

this chapter that the general and specific objectives of the study, and the methodology employed to accomplish them are discussed.

The second chapter of the thesis focuses on reviewing the problem domain by describing the mission and activities of the organization responsible for initiating and maintaining the crime database. A brief description of the crime database was also provided in this section.

The third chapter of the thesis focuses on reviewing relevant literature on data mining and/or knowledge discovery in general and association and clustering rule algorithms in particular. The application of data mining to address problems similar to the one focused by this study- discovery of patterns to prevent and control crimes- are also reviewed.

The fourth chapter of the thesis dwells on the experimentation aspect of the project, where emphasis was put on illustrating the process followed to discover the regularities using association and clustering rule algorithms. The various parameters employed/adopted in the experiment are also discussed in this section of the thesis. What is more, this chapter includes a presentation and interpretation of the rules/regularities discovered after a series of experiments.

The fifth chapter of the thesis presents the concluding remarks and recommendations that are forwarded on the basis of the outcomes of the experiment.

Chapter Two

Understanding the Problem domain

2.1 FSCE and its Advocacy and Awareness Raising Program

In this part of the thesis the researcher attempts to describe the problem domain, namely the organization that is responsible for initiating and maintaining the database, and whose activities are directly related to this database. Attempts are made to point out the role the database could play in the effectiveness of the activities of the organization if a KDD process is to be applied on it.

Children are the future of any country. What is being done to them now determines, to a large extent, what the future of the society will be. If children grow being victims of offenses/ abuse, they are, most likely, to end up as offenders and abusers. With recognition of this fact, creating a child friendly environment everywhere becomes imperative.

Ethiopia has signed and ratified the UN Convention on the Rights of the Child. The Constitution of the Federal Democratic Republic of Ethiopia has also enumerated the fundamental rights of children. In addition, the provisions prescribed in the ordinary laws of the country pertaining to children are also considered to be sufficient to protect children's right.

Despite the above mentioned legal instruments which came into force to protect the rights children in Ethiopia, the attitude prevalent among the general public and law enforcement bodies is not conducive for the realization of these rights. The situation is further worsened by poverty, rapid urbanization,

drought and famine, armed conflict, and destabilization of families that have left millions of children in Ethiopia without proper care and protection.

In addition, the country constitutes a place where traditional values have existed for centuries and are deeply rooted in the day-to-day lives of the people. Many of the cultural values related to children consider them as properties of their parents, having no rights of their own. These and other related socio-economic and cultural factors have led to the abuse and neglect of children in almost all settings, including homes, schools and neighborhoods. As a case in point, cultural practices encourage physical punishment of children, particularly by parents, teachers and other adults, who are close to the child.

In response to this, one of the preoccupations of FSCE during the past twelve years has been advocacy and awareness raising program, which was aimed at creating public awareness on the rights of children in general and urban disadvantaged children in particular. The program targeted among others government policy makers and planners, officials and professionals of pertinent government and non-government organizations, members of the police force, the school community, religious and community leaders and grassroots government administrative institutions.

FSCE has been implementing the program in collaboration with pertinent government and non-government organizations, namely, Ministry of Labor and Social Affairs, federal and regional police commissions, national television and radio enterprises, schools, religious and community organizations and grass root administrative organs. Major strategies used for conducting the program include:

- Organizing workshops, seminars, meetings for the target groups
- Transmitting sensitization program through TV, Radio, and print media,
- Publishing Information, Education and Communication materials
- Organizing training programs on various aspects of child right issues for police recruits.
- Establishing networking with organizations working on child service locally and abroad for sharing knowledge and experience on the field.

FSCE's advocacy and awareness raising program focused particularly on the police force since it is one of the law enforcement institutions that have a professional duty to protect children's right. There prevails lack of awareness among members of the police force on the rights of children and the provision prescribed in the working laws of the country pertaining to children. Having realized this, FSCE has been conducting awareness-raising and training programs for police officers, cadets and recruits on child right issues.

The awareness-raising program conducted has brought significant changes among the police, in terms of raising their awareness on child right issues and increasing their involvement in protecting and caring for children. The positive change achieved through the awareness-raising program has motivated FSCE to initiate a practical program with the police.

Towards this end, FSCE has launched what is called the Child Protection Program in Addis Ababa and nine other cities/towns in collaboration with the respective regions and zone police commission and departments to handle cases of children reported to the police stations either committing an offence or being abused. Two police staff and one community worker were assigned to work in the

child protection units by the police and FSCE respectively. Besides, a coordinating office was established in the premises of the Addis Ababa City Administration Police Commission to supervise the activities of the Child Protection Units and maintaining the interests of the police in the program.

2.2 The Crime Database

As part of its Advocacy and Child Protection Program, the organization set up a Database Center at the Addis Ababa Police Commission. The center documents crime data in relation to the cases of juvenile delinquents, child victims and victimizers thereof. The data is reported from all the 28 Woreda Police Stations found in Addis Ababa. Every time an offence against children is reported to the Woreda police stations, officers in the respective Woredas capture the data on a form that is prepared for this purpose.

This crime database has been maintained for four years (since 1987 E.C.) by FSCE in collaboration with the Addis Ababa Police Commission. The database consists of 10,847 records/tuples in 17 tables describing a total of 25 attributes *(see Appendix 2 to view the 25 attributes, their data type, description, and possible value they can take, and appendix 3 to see categories of attributes selected for the discovery process).* The database was initiated to introduce a systematic way of record keeping and ease the generation of simple statistical analysis while preparing reports to donors regarding the profile of victim children who received the services of the organization such as counseling and medication *(see appendix 1 to view format used for preparing periodic reports).*

Chapter Three

Review of Literature

In this section the researcher attempts to highlight the major issues and concepts in the field of KDD. Hence, brief descriptions of the difference between KDD and data mining, the development of the field, the KDD process, the interdisciplinary nature of the field and the core fields that contributed to its growth, the different approaches of learning algorithms, and the association rule and clustering algorithms.

3.1 Overview of KDD

Conventionally the idea of searching pertinent patterns in data has been referred using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely (Fayyd et al, 1996).

Statisticians, data analysts, and management information system and database communities have commonly used the term data mining. On another front, the phrase knowledge discovery in databases was introduced during the first KDD workshop in 1989 to stress that knowledge is the ultimate output of a data-driven discovery. KDD was then widely accepted in AI and machine-learning fields (Fayyd et al, 1996).

There still prevails misunderstanding about the terms KDD and data mining. More often than not, these terms have been used interchangeably. However, the two terms have distinct meanings. The

term KDD refers to the whole process of changing low level data into high level knowledge. KDD can simply be defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Worth noting that data in the definition refers to a set of facts (for instance, cases in a database), and pattern refers to an expression in certain language describing a subset of the data or a model applicable to the subset (Fayyd et al. b, 1996).

Data mining is also defined as "the application of specific algorithms for extracting patterns from data." Despite the fact that data mining constitutes the center of the KDD process, it only accounts for the small proportion of the total endeavor (estimated at 15% to 25 %) (Fayyd et al. b, 1996).

The KDD process is interactive and iterative, comprising a number of phases requiring the user to make several decisions. A general outline of the central phases is presented below (Fayyd et al. b, 1996).

- Building an understanding of the application domain and the pertinent prior knowledge and identifying the objective of the KDD process from the user's point of view.
- Creating a target dataset by selecting a dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- Data cleaning and preprocessing that involves certain core tasks like removing noise if appropriate, collecting the necessary information to model or account for noise, choosing strategies for handling missing data fields, and accounting for time-sequence information and known changes

- Data reduction and projection, which refers to finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.
- Matching the objective of the KDD process (step 1) to a specific data-mining method. For instance, summarization, classification, regression, and clustering.
- Exploratory analysis and model and hypothesis selection including choosing the data mining algorithm(s) and selecting method(s) to be employed to find data patterns. This phase involves choosing which models and parameters might be appropriate and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).
- Data mining includes looking for patterns of interest in a specific representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.
- Interpreting mined patterns, possibly backtracking to any of the preceding phases for further iteration. This phase can also include visualization of the extracted patterns and models or visualization of the data given the extracted models.
- Acting on the discovered knowledge refers to deploying the knowledge directly, integrating the knowledge into another system for further use, or simply documenting and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

The KDD process can involve significant iteration and can contain loops between any two steps. Despite the stress by previous works on the data mining phase of the process, all the other phases are indispensable for the successful application of KDD in the real world.

Carbone (nd) attributed the growing attention that data mining continued to get from commercial and scientific communities to three reasons.

1. The number and size of databases in several organizations have increased at an incredible rate. Terabyte databases, once difficult to imagine, have presently become a reality in a variety of domains, including marketing, sales, finance, healthcare, earth science, molecular biology (e.g., the human genome project), and various government applications.
2. Organizations have acknowledged the prevalence of valuable knowledge which is hidden in the data which, if uncovered, could provide those organizations with competitive advantage.

The large proportion of data collected is placed on archive files with perhaps little chance of being analyzed ever. With the growing recognition that there is a valuable data veiled within these masses of data accelerated the allocation of a great deal of effort and resources into the development of Knowledge Discovery in Databases or Data Mining (Anand b, 1995).

3. Some of the enabling technologies have only recently become mature enough to make data mining possible on large datasets.

Despite the rapid increase in the amount of digital information stored in scientific and business domains, the techniques for using the data in advantageous ways are slow to develop (Fayyad et al., 1996).

Data mining has developed, and continues to develop, from the intersection of research in fields like databases, machine learning, pattern recognition, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization (Carbone, nd), information retrieval, and high-performance computing. Data mining software systems incorporate theories, algorithms, and methods from all of these fields (Fayyd, 1996).

The principal components of data mining technology have been under development for over a decade, in research areas such as statistics and machine learning. Presently, the growth of these techniques, combined with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments (An Introduction to Data Mining, nd). The relation KDD has with two of its core disciplines, namely statistics and machine learning, is briefly described below.

Statistics is considered as the foundation of most technologies on which data mining is built. Classical statistics include concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These constitute the building blocks with which more advanced statistical analyses depend. (Data Mining Software and Solutions, nd).

Looking into their similarity and difference, both statistics and data mining pursue the same goal of building compact and understandable models incorporating the relationships ("dependencies") between the description of a situation and a result (or a judgment) concerning this description (PMsi, 2001).

The core difference is the fact that while data mining techniques construct the models automatically, classical statistics tools need to be guided by a trained statistician with a good - or possibly preconceived - idea of what to look for (a "dependency hypothesis"). To sum up, data mining techniques result in huge gains in terms of performance, user-friendliness and time expenditure (Science Tribune (1997) in PMsi). Summarizing the difference between the two fields, an article entitled ”an Introduction to data mining” described statistics relative to data mining as being “Ill-suited for nominal and structured data types, completely data driven - incorporation of domain knowledge not possible, interpretation of results is difficult and daunting, and requires expert user guidance.”

On the other hand, machine learning, the other ancestor of data mining, is described as the union of statistics and AI. Machine learning could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. 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 AI heuristics and algorithms to achieve its goals (Data Mining Software and Solutions, nd)**.**

The goal of both data mining and machine learning technologies is learning from data. Although the frameworks for data mining and machine learning in general may appear very similar, there are important distinctions. Primarily, in the case of data mining, the database is often designed for purposes different from data mining. That is, the representation of the real world objects in the database has been chosen to meet the needs of applications rather than the needs of data mining. Hence, properties or attributes that would simplify the learning task are not necessarily present. Moreover, these properties can not be requested from the real world (Holsheimer, 1991).

The second important distinction is that databases are usually contaminated by errors. Whereas in machine learning the algorithm is often supplied with judiciary selected laboratory examples, in data mining the algorithm has to deal with noisy and sometimes contradictory data (Holsheimer, 1991).

3.2 Applications of KDD

The vast majority of organizations can presently be considered as being ‘data rich’, as they engage in gathering of an ever increasing data in terms of facts and figures about their business process and resource. However, facts and figures by themselves (as they are) do not represent knowledge. What is worse, they can result in what is called information overload. Considering knowledge as being represented in the form of patterns/regularities in the data, a large proportion of such organizations can safely be considered as ‘knowledge poor’ (Al-Attar, nd).

KDD is a recent technology with great potential to enable organizations direct their attention on the most relevant information in their data warehouses (An Introduction to Data Mining, nd).

The significance of KDD has gained recognition by information intensive industries that maintain large databases of customer transactions, namely banking, health care, insurance, marketing, retail and telecommunications.

According to Goebel & Gruenwald (1999), the popular applications of KDD include:

Prediction: Given a data item and a predictive model, predict the value for a specific attribute of the data item. For example, given a predictive model of credit card transactions, predict the likelihood that a specific transaction is fraudulent. Prediction may also be used to validate a discovered hypothesis.

Regression: Given a set of data items, regression is the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic production of a model that can predict these attribute values for new records. For example, given a dataset of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions.

Classification: Given a set of predefined categorical classes, determine to which of these classes a specific data item belongs. For example, given classes of patients that correspond to medical treatment responses, identify the form of treatment to which a new patient is most likely to respond.

Clustering: Given a set of data items, it partitions this set into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar. For example, given a dataset of customers, identify subgroups of customers that have a similar buying behavior.

Link Analysis (Associations): Given a set of data items, identify relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. These relations may be associations between attributes within the same data item ('Out of the shoppers who bought milk, 64% also purchased bread') or associations between different data items ('Every time a certain stock drops 5%, a certain other stock raises 13% between 2 and 6 weeks later'). The investigation of relationships between items over a period of time is also often referred to as 'sequential pattern analysis'.

Market/shopping basket analysis: Such an application analyses the combinations of products purchased by individual buyers to uncover dependencies. Supermarkets employ KDD to learn about the profile of their customer and their respective behavior. Other common applications are for promotion effectiveness, customer vulnerability analysis, cross-selling, portfolio creation and fraud detection. It is also employed in healthcare, where it looks for relationships between patient histories, illnesses and surgical operations. It is also applied in manufacturing processes to monitor quality and spot machine wear. In marketing, if an organization wants to cross-sell one product to another, it cannot target all customers, because the volume may be too large. Therefore it is necessary to mine the

database of existing customers to identify patterns which describe the characteristics of purchasers of the product. These patterns can then be applied to the database of customers who have not purchased the product to segment and predict those who are more likely to purchase the product. These are then targeted in a very specific marketing campaign (Newing, nd).

Model Visualization: Visualization plays an important role in making the discovered knowledge understandable and interpretable by humans. Besides, the human eye-brain system itself still remains the best pattern-recognition device known.

Visualization techniques may range from simple scatter plots and histogram plots over parallel coordinates to 3D movies.

Exploratory Data Analysis (EDA): Exploratory data analysis (EDA) is the interactive exploration of a dataset without heavy dependence on preconceived assumptions and models, thus attempting to identify interesting patterns. Graphic representations of the data are used very often to exploit the power of the eye and human intuition.

3.2.1 Application of KDD in crime prevention and control

The application of data mining is widely popular among the business community. Lately, however, data mining is being effectively employed in the public sector as well. According to Brown (nd) and SPSS INC (2003), the vast proportion of law enforcement bodies are facing an overwhelming amount

of data that need to be processed and changed into useful information. Data mining has been recognized as one of the technologies that provide the means to turn data into information.

There is a growing report in the media lately about the active role data mining can play in fighting crime, fraud, and terrorism. The applications are rising and an increasing number of government and law enforcement experts are implementing programs and adopting new technologies in the area of collecting information and evidence, and conducting investigations (SPSS INC, 2003).

Invariably someone gets hurt when crimes are committed and society has the moral duty to safeguard the most exposed group. Predictive techniques constitute one way of accomplishing this objective since they facilitate the projection of the temporal-geography of crime, thus allowing the deployment of preventative measures (Corcoran & Ware, nd).

To supplement their crime analytic ability, law enforcement agencies are incorporating database management systems (DMBS) and geographic information systems (GIS). However, the application of these technologies to discover spatial relationships and associations between crimes has been inadequate. Data mining has been acknowledged as providing a way to leverage GIS and DBMS technology to support a broader range of crime analytic functions. Data mining has to do with the automatic discovery of patterns and relationships in large databases. Brown reiterated the issue claiming that "No field is in greater need of data mining technology than law enforcement."

Successful analysis of crime requires looking into the combinations of spatial, demographic, victim and other data (Brown, nd). Spatial analysis in particular has been considered very essential in making

decisions about law enforcement resource allocation. Effective planning requires an understanding of where and when resources are needed. In a nutshell, the spatial data mining tools will endow law enforcement agencies with considerable capabilities that they currently miss. In the absence of these tools, they will be forced to deal with the enormous available data through manual analyses only, even the most significant crimes (Brown, nd)

As a case in point the researcher likes to briefly describe a computer application known as the Regional Crime Analysis Program (ReCAP) system that is designed to assist local police forces (e.g. University of Virginia (UVA), City of Charlottesville, and Albemarle County) in the analysis and prevention of crime (Brown, nd).

The individual components of the system include a database, geographic information system (GIS), and data mining tools. The database component of the system stores all necessary information in relation to police related incidents. The database query tools allow the extraction of required information. The GIS component of the system through its spatial tools maps out the crime incident data and provides spatial data mining and analysis. The data mining and temporal tools provide insight into the nature of crime incident data through control charts, clustering analysis and forecasting (Brown, nd)

Two of the capabilities and functions of this system are: (Brown, nd)

- Hot Spot analysis using Kernel Density Estimation (KDE): Kernel density projection is applied to determine spatial patterns of crime within an area. Contour and color plots show the

high density areas. A threshold density level can also be chosen, where higher densities are labeled at “Hot-Spots” and law enforcement can take action to reduce crime in that area. This analysis can be performed for any subset of crimes produced by a user’s query.

- Cluster Analysis: Clustering algorithms search for statistically significant groupings of crimes in an area and alert the crime analyst user to potential problem areas. Currently partitional methods, K-meansand nearest neighbor, are used to perform the clustering. ReCAP allows users to vary the resolution of the clusters found. As a user zooms into an area, the system will divide existing clusters into smaller clusters. In this way, the user can establish statistical significance at whatever level of detail is desired.

The accurate forecasting of the temporal-geography of crime (predicting where and when crime is likely to take place) can have immense benefits, for accurate prediction if acting upon should lead to effective prevention (Corcoran & Ware, nd).

In relation to learning algorithms, association rule mining algorithm has also a proven application in the area of law enforcement. According to Chen (nd), visual crime mining techniques such as association rule mining could become the catalyst for assisting intelligence and law enforcement agencies in capturing knowledge and creating transformation

3.3 Learning Algorithms

From a logical perspective, two inference techniques can be distinguished- deductive and inductive. Deduction is a technique to infer information that is a logical consequence of the information in the

database. Most database management systems (DBMSs), such as relation DBMSs, offer simple operators for the deduction of information. For example, the join operator applied to two relational tables where the first administrates the relation between employees and departments and the second the relation between departments and managers, infers a relation between employees and managers (Holsheimer, 1991).

Induction is a technique to infer information that is generalized from the information in the database. For example, from the employee-department and the department-manager tables from the example above, it might be inferred that each employee has a manager (Holsheimer, 1991).

This is higher level information, or knowledge, meaning general statements about properties of objects. We search the database for regularities- combinations of values for certain attributes, shared by facts in the database. We can also formulate such regularity as a rule, predicting the value of an attribute in terms of other attributes (Holsheimer, 1991).

The most important difference between deduction and induction is that the former results in provably correct statements about the real world provided that the database is correct, while the latter only results in statements that are supported by the database, but not necessarily true in the real world. One of the most important aspects of the induction process is therefore the selection of the most plausible rules and regularities, supported by the database (Holsheimer, 1991).

Inference of information from a database is beyond human capabilities, if only because of the ever growing size of databases. Hence, the inference-process should be supported by the DBMS. However, although all DBMSs support deduction of information, none supports induction (Holsheimer, 1991).

Humans and other intelligent creatures attempt to understand their environment by using a simplification of this environment- called a model. The creation of such a model is called inductive learning. During the learning phase the cognitive system observes its environment and recognizes similarities among objects and events in this environment. It groups similar objects in classes and constructs rules that predict the behavior of the inhabitants of such a class (Holsheimer, 1991).

Two learning techniques are of special interest in inductive learning: supervised and unsupervised learning. In supervised learning, an external teacher defines classes and provides the cognitive system with examples of each class. The system has to discover common properties in the examples for each class- the class description. This technique is also known as learning from examples (Holsheimer, 1991). A supervised algorithm is first trained on a set of labeled data. A set of data whose classes are already known is used to build up profiles of the classes, and this information can then be used to predict the class of new data (Clare, 2003). A class, together with its description forms a classification rule “if (description) then (class)” that can be used to predict the class of previously unseen objects (Holsheimer, 1991).

In unsupervised learning, which is also known as learning from observation and discovery, the system has to find its own classes in a set of states, without any help of a teacher. Practically, the system has to find some clustering of the set of states S. The data mine system is supplied with objects, as in

supervised learning, but now, no classes are defined. Te system has to observe the examples, and recognize patterns (i.e. class description) by itself. Hence, this learning form is also called learning by observation and discovery. The result of an unsupervised learning process is a set of class descriptions, one for each discovered class, that together cover all objects in the environment. These descriptions form a high-level summary of the objects in the environment (Holsheimer, 1991).

The following section focuses on two of the unsupervised learning algorithms, namely association rule and clustering.

3.3.1 Association rule mining

These days there are several efficient algorithms that cope with the popular and computationally expensive task of association rule mining (Hipp & Ulrich, nd).

Association rule mining refers to finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. Frequent pattern here refers to pattern (set of items, sequence, etc.) that occurs frequently in a database (Stiles, nd).

Since its introduction by Agrawal, Imielinski, and Swami (Toivonen et al., 1995), the task of association rule mining has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern discovery methods in KDD (Hipp & Ulrich, 2000). It has attracted tremendous interest among data mining researchers and practitioners. It has an elegantly simple

problem statement, that is, to find the set of all subsets of items (called itemsets) that frequently occur in many database records or transactions, and to extract the rules telling us how a subset of items influences the presence of another subset (Zaki & Hsiao, 2002).

The idea of mining association rule originates from the analysis of market-basket data where rules like "a customer who buys products x_1 and x_2 will also buy product y with probability c%." are found. Their direct applicability to business problems together with their inherent understandability- even for non data mining experts – made association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems (Hipp & Ulrich, 2000).

One of the important problems in data mining is discovering association rules from databases of transactions, where each transaction contains a set of items. The most time consuming operation in association rules discovery process is the computation of the frequencies of the occurrence of subsets of items, also called candidates, in the database of transactions. Since usually such transaction-based databases contain extremely large amounts of data and large number of distinct items, the total number of candidates is prohibitively large. Hence current association rule discovery techniques try to prune the search space by requiring a minimum level of support for candidates under consideration. Support is a measure based on the number of occurrences of the candidates in database transactions (Han, nd).

3.3.3.1 Apriori Algorithm

Apriori is perhaps the best known association rule mining algorithm is (Clare, 2003). The algorithm is used for mining association rules. 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 (Agawal, 1994).

Thus, two notions characterize the association rule, namely support and confidence. While Support of an association rule refers to the number of instances they predict correctly, Confidence of an association rule refers to the same number expressed as a proportion of the number of instances that the rule applies to (Witten & Frank, 2000).

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 related (Witten & Frank, 2000).

The Apriori algorithm works in the following manner (Witten & Frank, 2000; Agawal, 1994).

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 itemsets will be at least equal to a pre-determined minimum support.

Apriori iteratively searches frequent itemsets: at each iteration k, Fk, it identifies the set of all the itemsets of k items (k-itemsets) that are frequent. In order to generate Fk, a candidate set Ck of potentially frequent itemsets is first built. By construction, Ck is a superset of Fk. 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 Fk, and the next iteration is started. The algorithm terminates when Fk 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 Ck 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.

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

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 (Witten & Frank, 2000).

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 (Witten & Frank, 2000).

The brute-force method will be excessively computation-intensive unless item sets are small, because the number of possible subsets grows exponentially with the size of the item set. Apriori algorithm can only handle nominal attributes (Witten & Frank, 2000).

Association rules are really different from classification rules except that they can predict any attribute, not just the class, and this gives them the freedom to predict combinations of attributes too. Also association rules are not intended to be used together as a set, as classification rules are. Different association rules express different regularities that underlie the dataset, and they generally predict different things (Witten & Frank, 2000).

3.3.2 Clustering

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. From a machine learning perspective, clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, Customer Relation Management, marketing, medical diagnostics, computational biology, and many others (Berkhin, 2002).

Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. These clusters presumably reflect some mechanism at work in the

domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods (Witten & Frank, 2000).

There are different ways in which the result of clustering can be expressed. The groups that are identified may be exclusive, so that any instance belongs to only one group. Or they may be overlapping, so that an instance may fall in several groups. Or they may be probabilistic, whereby an instance belongs to each group with a certain probability. Or they may be hierarchical, such that there is a crude division of instances into groups at the top level, and each of these groups is refined further-perhaps all the way down to the individual instances. The choice between these possibilities should be dictated by the nature of the mechanisms that are thought to underlie the practical clustering phenomenon. The choice is usually dictated by the clustering tools that are available (Witten & Frank, 2000).

There are different clustering methods. The classic is the K-meansalgorithm, which forms clusters in numeric domains, portioning instances into disjoint clusters (Witten & Frank, 2000).

3.3.2.1 K-meansalgorithm

This algorithm has as an input a predefined number of clusters, which is the k from its name. Means stands for an average, an average location of all the members of a particular cluster. When dealing with clustering techniques, one has to adopt a notion of a high dimensional space, or space in which orthogonal dimensions are all attributes from the table of data we are analyzing. The value of each

attribute of an example represents a distance of the example from the origin along the attribute axes. Of course, in order to use this geometry efficiently, the values in the dataset must all be numeric (categorical data must be transformed into numeric ones!) and should be normalized in order to allow fair computation of the overall distances in a multi-attribute space (Rudjer Boskovic Institute, 2001).

The classic K-meansalgorithm forms clusters in numeric domains, partitioning instances into disjoint clusters. It is a simple and straightforward technique that has been used for several decades (Clare, 2003).

K-meansalgorithm is a simple, iterative procedure, in which a crucial concept is the one of centroid. Centroid is an artificial point in the space of records, which represents an average location of the particular cluster. The coordinates of this point are averages of attribute values of all examples that belong to the cluster. The steps of the K-meansalgorithm are: (Ruder Boskovic Institute, 2001).

- Select randomly k points (it can be also examples) to be the seeds for the centroids of k clusters.
- Assign each example to the centroid closest to the example, forming in this way k exclusive clusters of examples.
- Calculate new centroids of the clusters. For that purpose average all attribute values of the examples belonging to the same cluster (centroid).
- Check if the cluster centroids have changed their "coordinates". If yes, start again form the step 2). If not, cluster detection is finished and all examples have their cluster memberships defined.

- Usually this iterative procedure of redefining centroids and reassigning the examples to clusters needs only a few iterations to converge. (Ruder Boskovic Institute, 2001).

The process of assigning points to cluster and then re-calculating centroids continues until the cluster boundaries stop changing. The cluster boundaries are set after a handful of iterations for most dataset (Berry & Linoff, 1997).

If the number of clusters k in the K-means method is not chosen so to match the natural structure of the data, the results will not be good. The proper way to alleviate this is to experiment with different values for k. In principle, the best k value will exhibit the smallest intra-cluster distances and largest inter-cluster distances. More sophisticated techniques measure these qualities automatically, and optimize the number of clusters in a separate loop (AutoClass) (Ruder Boskovic Institute (2001).

The original choice of the value for k determines the number of clusters that will be found. Furthermore, if this number does not match the natural structure of the data, the technique will not obtain good results. Unless the data miner suspects the existence of a certain number of clusters, one has to experiment with different values for k.

A large number of variants of the basic K-means procedure have been developed. Some produce a hierarchical clustering by applying the algorithm with k=2 to the overall dataset and then repeating, recursively, within each cluster. Others concentrate instead on speeding up clustering. The basic algorithm can be rather time-consuming because a substantial number of iterations may be necessary each involving finding the distance of the k-cluster centers from every instance to determine the

closest. There are simple approximations that will speed it up considerably, for example by dealing with projections of the dataset and making cuts along selected axes instead of the arbitrary hyperplane divisions implied by choosing the nearest cluster center, but they inevitably compromise the quality of the resulting clusters.

The K-means algorithm iterates over the whole dataset until convergence is reached.

Every set of clusters will then have to be evaluated. Berry & Linoff (1997) believe that, in general the best set of clusters is the one that does the best job of keeping the distance between members of the same cluster small and the distance between members of adjacent clusters large. They further state that, the best set of clusters in descriptive data mining may be the one showing some unexpected pattern in the data.

Automatic cluster detection using the K-means algorithm is an undirected knowledge discovery process.

According to Bounsyathip (2001), K-means is based on a concept of distance, which requires a metric to determine distances. Euclidean distance can be used for continuous attributes. Since choosing a suitable metric is a very delicate task, a business expert is needed to help determine a good metric.

Chapter Four

Experimentation

In this section of the thesis, the researcher discusses the experimentation process by recounting the steps followed, the choices made, the tasks accomplished, the results obtained, and the feedbacks on evaluation of results.

4.1 Data Understanding

Domain experts were consulted to have insight into the problem domain. The domain experts constitute two individuals from FSCE and Addis Ababa Police Commission that are in charge of running the Advocacy and Awareness raising program, and responsible for the establishment of the database (see Chapter Two on FSCE and its Advocacy and Awareness Raising Program). On the basis of the insight gained from discussion with domain experts and review of relevant documents, the goal/purpose of the KDD process was defined, which is the discovery of regularities within the crime dataset.

4.2 Tool selection

Following the definition of the KDD goal, data mining tools, Weka and the Knowledge Studio Software, were selected with the assumption that they can adequately serve the purpose- discovering regularities within the crime data, and illustrating the application of KDD to good effect on the data. In addition to serving the purpose of the KDD process, the two tools were easily accessible. Weka is available freely while Knowledge Studio was obtained from individuals. On the other hand, obtaining

other data mining software that are commercially available that are equally competitive was beyond the cost of the project.

The Waikato Environment for Knowledge Analysis (Weka) was utilized to perform the association rule mining 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).

Weka consists 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).

Typically the modules in the Weka system fall into three categories: dataset processing, machine learning schemes, and output processing. The processing of datasets involves extracting information about a dataset for the user, splitting datasets into test and training sets, filtering out features in the data not required by the user, and translating the dataset into a form suitable for a machine learning scheme to work with (Garner, nd).

Interaction by the user with Weka will result in the modules being combined in such a way as to produce the desired output. For example, a typical task might involve selecting a dataset to train on, selecting a dataset to test with, excluding features not required from the datasets, choosing a machine learning scheme, running the scheme on the training data and then looking at the rules produced and how well they did on the test data (Garner, nd).

As one of the functionalities of the Weka software, Association rules and the Apriori algorithm in particular is supported (Witten & Frank, 2000).

Weka has proved itself to be a useful and even essential tool in the analysis of real world datasets. It reduces the level of complexity involved in getting real world data into a variety of machine learning schemes and evaluating the output of those schemes. It has also provided a flexible aid for machine learning research and a tool for introducing people to machine learning in an educational environment (Garner, nd). The obvious advantage of a package like Weka is that a whole range of data preparation, feature selection and data mining algorithms are integrated. This means that only one data format is needed, and trying out and comparing different approaches becomes really easy. The package also comes with a GUI, which should make it easier to use (Witten & Frank, 2000). The following figure depicts the graphical user interface chooser.

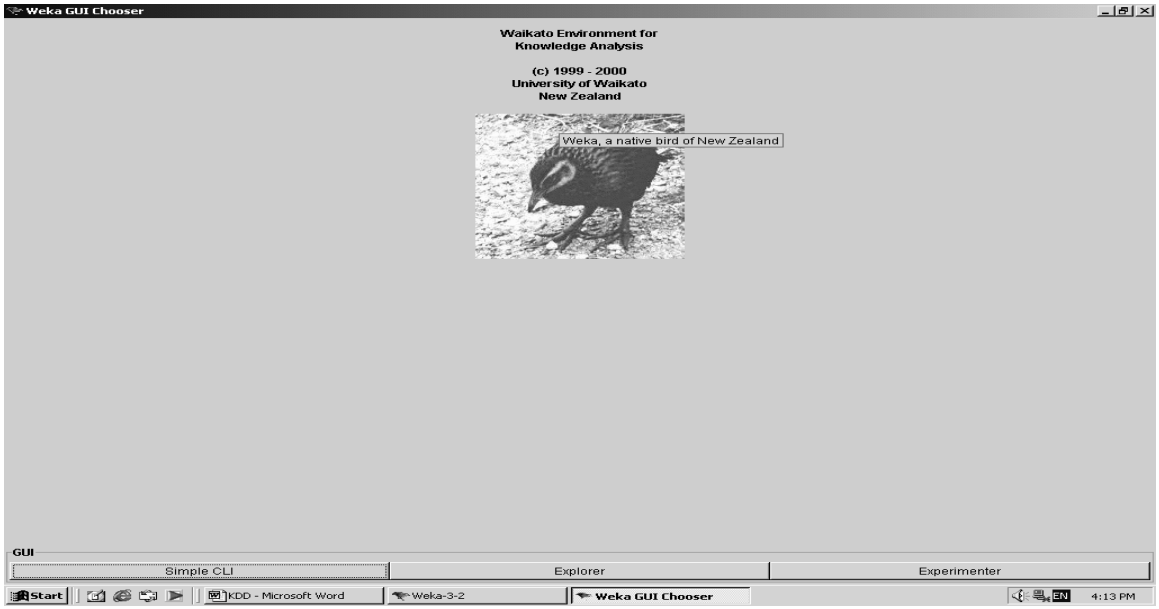

Figure 1: Weka GUI Chooser

4.2.1 Dataset Format

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. 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).

4.3 The Knowledge discovery process

To accomplish the objectives of the research, I have adopted the KDD process as described by Fayyad, Piatetsky-Shapiro and Gregory (1996). According to Fayyd et al, the KDD can be viewed as constituting four major phases/activities. These are data selection, data preprocessing, data mining, and post-processing/interpretation.

4.3.1 Data Selection

Data selection is important as the data consists of features that are not related to the problem at hand such as registration number since they are redundant, irrelevant, or invariable throughout the dataset. Taking into account such features in the automated analysis might not result in meaningful patterns. Hence, 15 attributes out of the original 25, that best suit the objectives of the research were selected *(see Appendix 2 to view the metadata for the 25 attributes, and see appendix 3 to view the selected 15 attributes and their respective categories.)*

To effect scalability of the learning algorithm, data reduction tasks were carried out. Relevant features for the research project were selected following consultation with domain experts, two individuals from Addis Ababa Police Commission and FSCE, who are responsible for the planning and supervising the implementation of the advocacy and awareness raising and child protection program, which is the program that keeps the crime database.

Association rule mining is computationally expensive and thus sampling of the dataset is often recommended to ensure scalability of the learning algorithm. However, in this research the whole

dataset in the database was used since the dataset size did 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.

4.3.2 Data Preprocessing

At this phase the dataset on which the association rule-mining algorithm is to be used is cleaned and prepared. The preparation task includes handling missing values, redundant or irrelevant features, and non-variant fields.

The crime data, which was stored in different tables, which are distinct from one another on the basis of the quarter they were gathered, were merged into one large table.

Converting into .ARFF format

The fields in the database were tab separated. The database was opened in Excel. Since Association rule mining, which is the selected data mining algorithm, is often used in situations where attributes are nominal, the numeric values (age and time the offence was committed) were converted to nominal, and certain similar categories were merged. There were missing data, that is, fields that were left unfilled. Missing values were replaced with the most frequent value (mode), as it is one of the proper ways of representing such values in Weka. The values “others”, 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. The database was then opened in Word, header information added. That is, the @ symbol was placed in front of the relation name and each attribute. The @DATA symbol was placed before the data. Finally, prior to saving the file extension was changed to .arff. The following figure indicates the representation of the data in ARFF format.

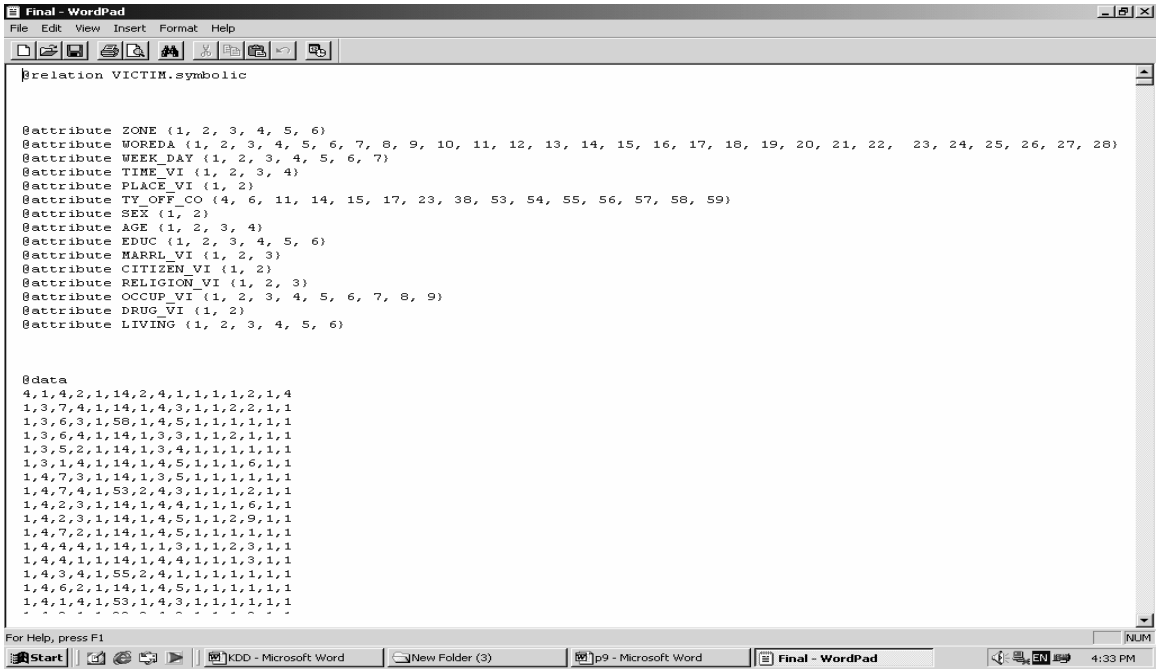

Figure 2: Representation of the data in ARFF format

Data reduction

During consultation with the domain experts, persons directly responsible for designing and implementing the awareness raising program at the Addis Ababa Police commission and FSCE, it was found out that the database has more attributes than actually necessary for the problem at hand. Hence, certain attributes were excluded from being considered in the data mining. The attributes that were removed can be viewed as belonging to one of the following three groups.

a. Fields with same information content

When fields with the same information content were encountered, only one of the fields was considered. For instance, Code of crime type was chosen over Name of crime type, which holds the same information.

b. Non-variant fields

Attributes that take a value that holds true for all the records in the database were also dropped, for instance, the attribute Region. The crime data is collected from all the 28 Woredas in Region 14. The database does not include data for other regions. Since Region 14 applies to all the cases, the attribute was considered as being irrelevant.

c. Fields taking many different values

The following text fields were also disregarded since they take many different values. This is done with the intention to improve the speed, accuracy of analysis and training. Examples include:

- Registration number
- Registration book number
- 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. 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.

- Name of the victim- in order to keep confidentiality and privacy of the data the names of the victim children were excluded from the data exploration process.
- Kebele- there are over 300 Kebeles in the Region 14 administration. This is too detail a data for association rule mining, which is computationally expensive. Having more attributes will make the exploration task difficult, if possible. Besides, the purpose of this research is to discover general rules that depict regularities or summarize the records in the database, and thus other address related attributes such as Woreda and Zone could serve the purpose.
- Application date- although difference is bound to prevail, the value of this attribute is, more often than not, similar to the value for the Time offence committed attributed for individual records. The time offence committed attribute was chosen, and it was mapped into an attribute with only 4 nominal categories.

In this research the standard of considering fields as being complete as long as at least 70% of the records comprise values is adopted (SPSS INC., 1995). Accordingly, all the 15 attributes satisfy this criterion. Although the attribute WEEK DAY appears to be relevant, the attribute constitutes 28% of missing values. What is more, the proportions of the values are almost close to each other and assigning the mode value to the missing value was thought to result in distorted data. The missing value amounts to twice any of the attribute values. Hence, the attribute WEEK DAY was left out.

This reduces the total number of remaining attributes to 14. These 14 attributes were taken up as they were deemed to be relevant for the exploration purpose. The selected 14 attributes include: ZONE,

WOREDA, TIME OFFENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, OCCUPATION, LIVING ARRANGEMENT, PLACE THE OFENCE WAS COMMITTED, MARITAL STATUS, SPECIAL HABIT, and CITIZENSHIP.

Merging categories

Merging was essential because there were a few attributes that have categories similar enough to be merged under a broad category and considered as one. The other reason for merging categories was the scanty distribution of the categories over the dataset. This measure was taken considering the computational advantage this will bring about. Accordingly, 14 categories of the attribute TYPE OF OFFENCE were merged and reduced into five categories *(see appendix 3 for the reduced number of categories).*

Twelve categories of the Educational status feature of children who had been to formal school were merged into 3 categories, namely 1 to 6 (primary), 7 to 8 (junior secondary), and 9 to 12 (secondary school).

Although the attribute SPECIAL HABIT has seven attributes, the NO BAD HABIT category accounts for the vast proportion of the values in the database. The rest of the attributes were scantly distributed, hence, they were merged into a category that generally depicts the presence of a special habit.

Two categories of the OCCUPATION attribute, namely government and NGO employee, were also merged into employee of a formal organization.

The two attributes with real/numerical data type, namely AGE and TIME THE OFFENCE WAS COMMITTED, were converted into a nominal/binary data type, since the chosen algorithm dictates the use of attributes with such data type. The *age* attribute was converted into four nominal categories each representing the age interval 0 to 4, 5 to 9, 10 to 14, and 15 to 18. The TIME OFFENCE WAS COMMITTED was converted into four categories representing late night, early day, late day, and early night.

4.3.3 Data mining

With the aim to discover regularities in the data, the data that had undergone preprocessing was subject to the Weka data mining software.

The dataset include a total of 10878 records having 25 attributes each. Early in the preprocessing stage 10 of the attributes were dropped since they were deemed irrelevant, and/or redundant.

Experiment 1

The association rule learning algorithm was applied on the whole dataset that had been cleaned and transformed into the .arff format. The information on running the algorithm on the database is presented in the table below.

Table 1: Run information of Apriori (10847 instances and 14 attributes)

Scheme	weka.associations.Apriori
-N(required number of rules output)	20
-T(metric type by which to rank rules)	0 (confidence)
-C (the minimum confidence of a rule)	0.9
-D (delta at which the minimum support is decreased at each iteration)	0.05
-U (upper bound for minimum support)	1.0
-M (the lower bound for the minimum support)	0.1
-S (significance of a rule at a given level)	-1.0
Relation	VICTIM.symbolic
Instances	10847
Attributes (see appendix 3 to view the list of attributes)	15

According to Agrawal in Stiles (nd), the support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all others are referred as small itemsets. Large itemsets are used to generate the desired rules.

At the end of this experiment, several rules that satisfy the above metrics were generated. The following table depicts the size and frequency of the large item sets generated.

Table 2: Size and frequency of generated rules (10847 instances and 14 attributes)

Size of Generated large item sets	Frequency of large item sets
One	4
Two	6
Three	4
Four	1

As shown in the above table, there were four one-item frequent itemset, six two-item frequent itemsets, four three-item frequent itemsets and one four-item frequent itemsets generated.

Best rules found

Apriori generated a number of rules that satisfy the above set minimum metrics of support and confidence. If a rule has a confidence above the minimum set confidence, then the rule holds.

The major characteristics of large Itemset include the fact that any subset of a frequent itemset must be frequent (Stiles, nd). Hence, there are relationships between particular association rules, i.e. some rules imply others, as in the above case where supersets implying subsets that constitute them. In addition, the rule of contrapositive also works here, i.e. If an itemset is not large, none of its supersets are large.

For the purpose of illustration five of the 20 set of rules produced were presented below *(See appendix4 to view the complete list of best 20 set of rules generated by Apriori):*

- PLACE_VI=1 MARRL_VI=1 HABIT_VI=1 9455 ==> CITIZEN_VI=1 9408 conf:(1)

(If Place = Urban and Marital Status = Single and Special habit = No bad habit then Citizenship = Ethiopian) This is the rule with the largest itemset size, i.e. 4. Being a superset, the subsets of this rule also hold true. The number preceding the '==>' symbol indicates the rule's support, that is the number of items covered by its antecedent. Following the rule is the number of those items for which the rule's consequent holds as well. In parenthesis is the confidence of the rule, i.e. the second number divided by the first.

Apriori orders rules according to their confidence and uses support as a tiebreaker. Although Apriori tries to generate ten rules, in this research the number of rules generated by the algorithm was specified to be double this default size (i.e. 20).

- MARRL_VI=1 CITIZEN_VI=1 HABIT_VI=1 9617 ==> PLACE_VI=1 9408 conf:(0.98)

(If Marital status = Single, Citizenship = Ethiopian, Special habit = No bad habit then Dwelling place = Urban)

This rule has an 86% support (i.e. 9617/10847) and a 98% accuracy/confidence.

- PLACE_VI=1 CITIZEN_VI=1 HABIT_VI=1 9713 ==> MARRL_VI=1 9408 conf:(0.97)

(If Place = Urban and citizenship = Ethiopian, Special habit = No special habit then Martial status = Single)

- PLACE_VI=1 MARRL_VI=1 CITIZEN_VI=1 10229 ==> HABIT=1 10024 conf:(0.98)

(If Place = Urban and Marital status = Single and Citizenship = Ethiopian then Special habit = No bad habit)

The above generated rules constitute those attributes (itemsets) with large frequency throughout the dataset. Despite the fact that the rules scored high in terms of the objective measures of interestingness (high support and confidence), they were found to be less interesting in the eyes of users/domain experts and the purpose of the research, which is discovering interesting rules/regularities. The rules can be considered as trivial since anyone with the knowledge of the distribution of the respective values of these attributes can tell the possible relationship. That is in the above attributes, particular values highly dominate the value for that specific attribute. In the case of the attribute PLACE the value URBAN account for the vast majority of the instances in the database (i.e. 10623/10847). Similarly for the attribute CITIZENSHIP, the value ETHIOPIAN account for the lions share of the instances in the database (i.e. 10739/10847). The situation is the same for MARITAL STATUS where SINGLE account for 10427 of the instances, and SPECIAL HABIT where there are 9993 value of NO BAD HABIT.

Among the generated 20 rules identified by the Apriori as best (on the basis of confidence and support), one of them seems interesting.

- PLACE_VI=1 MARRL_VI=1 CITIZEN_VI=1 10229 ==> HABIT=1 10024 conf:(0.98)

  (If Place = Urban and marital Status = Single and Citizenship = Ethiopian then Special habit = No bad habit)

This rule indicates that the instances in the database are characterized by the occurrence of very few instances with SPECIAL HABIT such as smoking cigarettes, drinking alcohol, 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. Apart from discovering surprising or hidden rules, the learning scheme also results in rules that confirm facts existing in the real world.

In the effort to discover relatively more interesting rules that underlie the crime dataset, continuing with the experimentation was imperative. During subsequent experimentation, leaving out the most frequent or invariable attributes that dominate the rules in the first experiment was considered to be a proper measure, since the use of these attributes resulted in more or less uninteresting or trivial rules.

Experiment 2

A few of the attributes, namely PLACE THE OFFENCE WAS COMMITTED, MARITAL STATUS, SPECIAL HABIT, and CITIZENSHIP, were progressively removed to give chance for other attributes to be considered in the construction of the rules. The above attributes can be considered as invariant attributes since their values are highly dominated by a single value. Since invariant attributes apply to almost all instances, the use of such attributes results in easily predictable or trivial rules such as the ones indicated previously. Accordingly, an experiment was run over the following 9 attributes: ZONE, WOREDA, TIME OFFENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, OCCUPATION, and LIVING ARRANGEMENT.

The information from running the algorithm is presented below.

Table 3: Run information of Apriori (10878 instances and 10 attributes)

Scheme	Weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation	VICTIM.symbolic
Instances	10878
Attributes	10

The following table depicts the size and frequency of the large item sets generated.

Table 4: Size and frequency of generated itemsets

Size of Generated large item sets	Frequency of large item sets
One	9
Two	22
Three	17
Four	6

Best rules found

Apriori was made to generate 20 rules that best satisfy the set confidence and support metrics *(see appendix 5 to view the complete list of the best 20 rules generated).* The researcher together with domain experts selected those rules that were considered to be interesting. These include:

- TY_OFF_CO=14 RELIGION_VI=1 OCCUP_VI=1 3440 ==> LIVING=1 3293 conf:(0.96)

  (If Type of crime = willful injury and Religion = Christian and Occupation = Student then Living arrangement = Lives with both parents)

- TY_OFF_CO=14 SEX=1 RELIGION_VI=1 3718 ==> LIVING=1 3448 conf:(0.93)

(Type of offence = willful injury and Sex = Male and Religion = Christian then Living Arrangement = lives with both parents)

- TY_OFF_CO=14 AGE=4 OCCUP_VI=1 2874 ==> LIVING=1 2731 conf:(0.95)

(If type of crime = willful injury and Age = 15 to 18 years and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 SEX=1 OCCUP_VI=1 2392 ==> LIVING=1 2266 conf:(0.95)

(If Type of crime = Willful Injury and Sex = Male and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 RELIGION_VI=1 OCCUP_VI=1 3215 ==> LIVING=1 3042 conf:(0.95)

(If Type of crime = Willful injury and Religion = Christian and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 AGE=4 RELIGION_VI=1 4767 ==> LIVING=1 4321 conf:(0.91)

(If Type of crime = willful injury and Age = 15 to 18 years and Religion = Christian then Living Arrangement = with both parents)

The above rules can depict interesting regularity within the crime database. As cases in point, the rule "Type of offence = willful injury and Sex = Male and Religion = Christian then Living Arrangement = Lives with both parents" state that children living with both parents are the children who are exposed to willful injury. According to these rules, the other features that make up the children's profile

include being Christian, male, student, and age between 15 to 18 years. According to domain experts, these rules are interesting because they are contrary to popular conception that it is children with both parents and not orphan children, males and not females, older and not younger children that are the subject of the offence willful injury. That is, children presumed to be relatively more protected or more able to take care of themselves are the ones who make up the profile of the victim.

Despite the discovery of these interesting rules, the experimentation process was sustained in search of more and better rules. This time another attribute, namely LIVING ARRANGEMENT, which is characterized by a single highly frequent value, namely LIVE WITH BOTH PARENTS (9165 out of 10878 instances), was left out.

Experiment 3

The third experiment was conducted over 8 attributes leaving out 1 attribute from the previous 9 attributes, which were used in the second experiment. The attributes include: ZONE, WOREDA, TIME OFFENCE COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, and OCCUPATION.

The information from running the algorithm is presented below.

Table 5: Run information on Apriori (10878 instances and 9 attributes)

Scheme	weka.associations.Apriori -N 20 -T 0 –C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation	VICTIM.symbolic
Instances	10878
Attributes	9

The following table depicts the size and frequency of the large item sets generated.

Table 6: Size and frequency of generated itemsets (10847 instances and 9 attributes)

Size of Generated large item sets	Frequency of large item sets
One	23
Two	78
Three	90
Four	40
Five	5

Best rules found

The Apriori algorithm generated 20 best rules on the basis of preset confidence and support *(See appendix 6 to see the complete list of the 20 best generated rules).* Among the generated rules those rules that were thought to be interesting were selected together with domain experts.

- TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1140 ==> AGE=4 1094 conf:(0.96)

(If Type of offence = willful injury and education = 9 to 12 grade and Religion = Christian and Occupation = Student then Age = 15 to 18 years)

- TY_OFF_CO=14 AGE=4 EDUC=5 OCCUP_VI=1 1192 ==> RELIGION_VI=1 1094 conf:(0.92)

(If Type of offence = willful injury and Age = 15 to 18 years and Education = 9 to 12 grade and Occupation = Student then Religion = Christian)

The above sets of rules supplement the previously discovered regularities regarding those children who experienced willful injury. According to the rules generated in this experiment, children who were exposed to Willful injury include students who are between 9 to 12 grades.

With the hope to discover still more and better rules, the researcher experimented by leaving out one of the attributes, namely ZONE from the previous experiment. The reason for leaving out the ZONE attribute is because of the fact that the address information (including ZONE) is contained in the attribute WOREDA. Certain WOREDAS belong to a specific ZONE, and if need be the ZONE attribute can be derived.

Experiment 4

The fourth experiment was conducted over 8 attributes leaving out 1 attribute from the previous 9 attributes, which were used in the second experiment. The attributes include: WOREDA, TIME OFFENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, and OCCUPATION.

The information from running the algorithm is presented below.

Table 7: Run information of Apriori (10878 instances and 8 attributes)

Scheme	weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.05 -S -1.0
Relation	VICTIM.symbolic
Instances	10878
Attributes	8

The following table depicts the size and frequency of the large item sets generated.

Table 8: Size and frequency of generated itemsets (10878 instances and 8 attributes)

Size of Generated large item sets	Frequency of large item sets
One	18
Two	57
Three	73
Four	37
Five	5

Best rules found

Apriori generated 20 best rules on the basis of confidence and support criteria *(see appendix 7 to view the complete list of the 20 best rules found).* Those rules that were considered interesting were selected:

- TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1143 ==> AGE=4 1096 conf:(0.96)

(If type of Crime = willful injury and Education = 9 to 12 grades and Religion = Christian and occupation then age between 15 and 18.)

- TY_OFF_CO=14 AGE=4 EDUC=5 OCCUP_VI=1 1194 ==> RELIGION_VI=1 1096 conf:(0.92)

(If Type of offence = willful injury and Age = 1 to 18 and Education = 9 to 12 grade then Religion = Christian)

The interesting rules from this experiment are not different from the previous ones perhaps because the attribute ZONE that is left out from the previous experiment did not play any significant role in the construction of rules previously. That is why the absence of the attribute in this experiment did not result in different rules.

As a measure to see more and better rules, the discovery task was confined to five of the attributes which were deemed to be most relevant and interesting.

Experiment 5

The fifth experiment was run on the following attributes: WOREDA, TIME VIOLENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, and AGE.

Setting the support and confidence values at 0.1 and 0.9 would not result in any rule. Hence, the support and confidence thresholds were reduced to allow the generation of rules.

The information from running the algorithm is presented below.

Table 9: Run information of Apriori (10878 instances and 5 attributes)

Scheme	weka.associations.Apriori -N 20 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.05 -S -1.0
Relation	VICTIM.symbolic
Instances	10878
Attributes	5

The following table depicts the size and frequency of the large item sets generated.

Table 10: Size and frequency of generated itemsets (10878 instances and 5 attributes)

Size of Generated large item sets	Frequency of large item sets
One	17
Two	27
Three	18
Four	5

Best rules found

Best rules were generated by Apriori *(see appendix 8 to view the complete list of all the generated rules).* The rules selected from the generated rules on the basis of interestingness include:

- TY_OFF_CO=53 1097 ==> AGE=4 993 conf:(0.91)

(If Type of offence = Violation of municipal regulations then Age = 15 to 18 years)

This is one of the interesting regularities in the dataset, children between the ages of 15 and 18 are the ones who often experienced violation of municipal regulations.

- TY_OFF_CO=58 755 ==> AGE=4 683 conf:(0.9)

(If Type of offence = violation of municipal regulations then Age = 15 to 18 years)

Those children who were reported as victims of theft also belong to the age group 15 to 18. This is perhaps because children within this age category have better access or are assigned with the responsibility to spend money or look after gadgets/material than younger children.

- TY_OFF_CO=55 616 ==> SEX=2 549 conf:(0.89)

(If Type of crime = sexual abuse then Sex = Female)

In line with the popular conception, it is female children who were the ones exposed to sexual abuse.

- TIME_VI=4 TY_OFF_CO=14 SEX=1 1229 ==> AGE=4 1084 conf:(0.88)

(If Time offence was committed = Early night and Type of offence = Willful injury and Sex = male then Age = 15 to 18)

According to this rule, male children, who were identified in the previous experiments as being exposed to the offence Willful injury, were exposed to the specific offence Early in the night (from 7 to 12 in the evening).

- TIME_VI=2 TY_OFF_CO=14 SEX=2 671 ==> AGE=4 554 conf:(0.83)

(If Time offence was committed = Early Day and Type of Offence = Willful Injury and Sex = Female Then Age = 15 to 18 years)

- TIME_VI=3 TY_OFF_CO=14 SEX=2 877 ==> AGE=4 722 conf:(0.82)

(If Time offence was committed = Late day and Type of offence = Willful Injury and Sex = Female Then Age = 15 to 18 years)

According to these rules, unlike male children who were identified in previous experiments, as being exposed to Willful Injury early in the night, female children are exposed to the offence early in the day. Like male children, however, female children that were victims of his offence were between the ages of 15 and 18.

Keeping in mind the goal of producing an improvement in the support and confidence level of generated rules, and discovering subjectively interesting rules within the problem domain, the nominal

attributes (data) with multiple categories were converted to binary format so that they can be processed by the data mining software that handls numeric attributes only, namely the K-means algorithm.

Experiment 6

The 14 nominal attributes were converted into 99 binary attributes. The Weka association rule miner was not able to run on this converted data, as it might have found it computationally infeasible/expensive to run on such large number of attributes and records.

Attempt was made to leave out less relevant attributes, and the 99 attributes were progressively reduced and the learning algorithm was made to run on the data with 76 attributes.

The following figure depicts the binary representation of the 99 attributes using the Weka knowledge explorer.

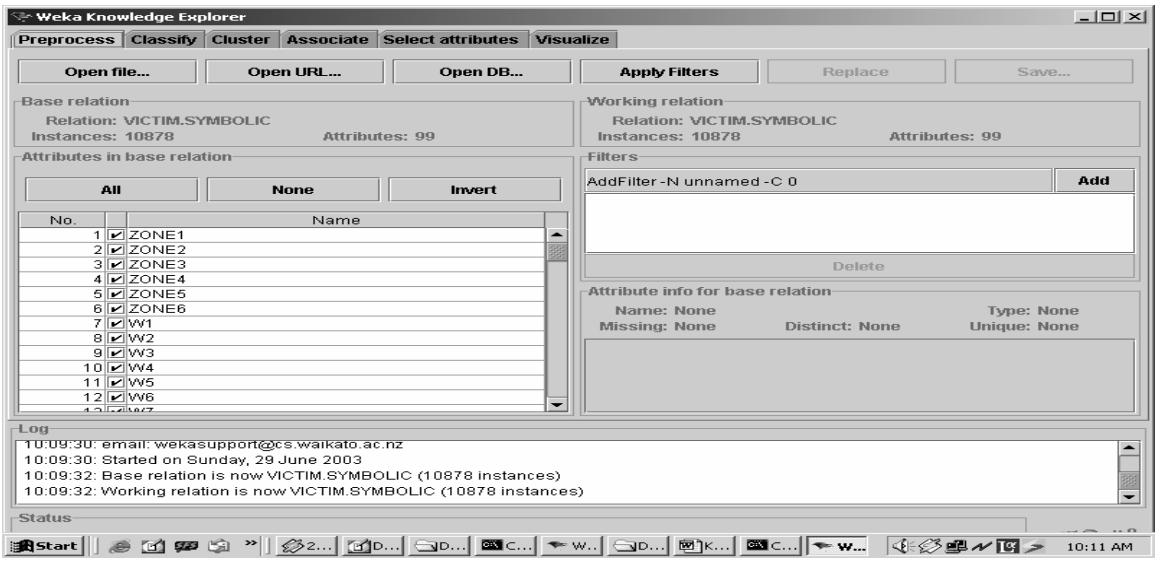

Figure 3: The Weka knowledge explorer window displaying the binary data

The information from running the algorithm is presented below.

Table 11: Run information of Apriori (10878 instances and 76 attributes)

Scheme	weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation	VICTIM.SYMBOLIC
Instances	10878
Attributes (see appendix 7)	76

The following table depicts the size and frequency of the large item sets generated.

Table 12: Size and frequency of generated itemsets (10878 instances and 76 attributes):

Size of Generated large item sets	Frequency of large item sets
One	38
Two	299
Three	793
Four	733
Five	223
Six	14

Best rules found

The Apriori algorithm generated 20 best rules on the basis of confidence and support metrics *(See Appendix 9 to view the compete list of 20 best rules generated)*. Interesting rules were selected in consultation with domain experts.

Except a few, almost all of the generated rules were found to be uninteresting and trivial. To illustrate a few rules are presented below:

- W18=0 YHABIT=0 10420 ==> LIVING3=0 10399 conf:(1)

(If Woreda 18= No and Special Habit = No then Living with father = No)

Although not conclusive, this rule implies that children from Woreda 18 and having a special bad habit of one form or another are those that live with their father only.

Experiment 7

To further improve the quality of the rules generated, the use of clustering algorithm prior to applying association rules were thought to be an appropriate measure. Thus, clustering of the crime dataset into disjoint groups was done using the K-means algorithm from the Knowledge Studio software produced by ANGOSS. Hence, the methodology involved here can be considered as a combination of clustering and association rule mining, which were used sequentially.

Clustering is often used in the early iterations of the data exploration phases of the mining process and is intended to uncover previously unknown categories or types of observations. Creating clusters prior to application of some other data mining technique (decision trees, neural networks) might reduce the complexity of the problem by dividing the space of examples. This space partitions can be mined separately and such two step procedure might exhibit improved results (descriptive or predictive) as compared to data mining without using clustering" (Rudjer Boskovic Institute, 2001).

Accordingly, in this research employing clustering algorithm was believed to reduce the heterogeneity of the records by grouping the records into groups of similar instances. This way the records were

segmented into similar groups, and the association rule miner was applied on the segmented data to build the profile of each group, on its own and in relation to other clusters.

The binary representation of the dataset allows the data to be manipulated by both algorithms that operate on numeric or nominal data. Accordingly, the association rule mining algorithm that works only on nominal data was able to work on it. Similarly the clustering algorithm that operates on numeric data also was able to run on it.

Although the Weka workbench has both association rule and clustering algorithms, only the association rule algorithm was used. This is because one of the clustering implementations of the algorithm in Weka, K-means, is not very adept for descriptive purposes. The user does not have access to look into the records that constitute the clusters. For this reason, the K-means algorithms form the KnowledgeSTUDIO software was used, as it provides a complete list of the records that belong to any of the clusters.

Although the K-means clustering algorithm does not have any problem dividing the dataset into groups, the size and the relevance of the attributes considered for the data mining purpose greatly affects the quality of the division.

Accordingly, although clusters were constructed using all the 99 binary attributes, it was too much for the association rule algorithm to work on. Hence, the attributes which were deemed to be less relevant, non-variant, and with little information content were removed. Consequently, the 99 attributes were reduced to 47 attributes and the learning algorithms were applied on the whole dataset.

The data was partitioned into 2, 3, 4, and 5 groups using the K-means algorithm. The best results were achieved by clustering into two groups on the basis of 47 attributes. Each of the two clusters has a vast majority of children from one sex than the other. What is more, except in the case when the cluster number is two, the association rule mining failed to generate rules for clusters, as it finds a cluster/two too large for computation.

In the first cluster the offence WILLFUL INJURY was reported by 4485 and 6626 of the 6628 of the children in the group were males. What is more, 5550 of the children are between 15 and 18 years of age. On the other hand, in cluster 2, the offence SEXUAL ABUSE was reported by 549 of the children. None of the children in the group were males. WILLFUL INJURY was reported by 2110 of the children. The following figure indicates the instances and attributes of the first cluster.

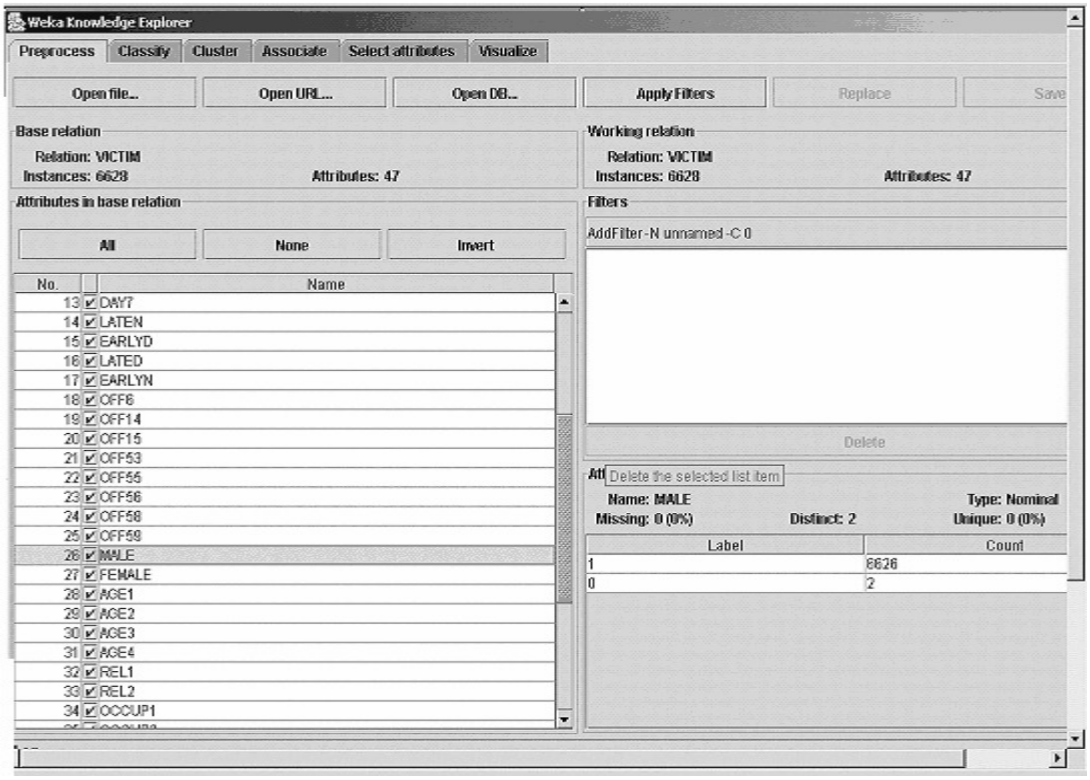

Figure 4: The WEKA explorer window depicting the instances and attributes of the first cluster

The information from running the algorithm is presented below.

Table 13: Run information of Apriori (6628 instances and 47 attributes)

Scheme	weka.associations.Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation	VICTIM.SYMBOLIC
Instances	6628
Attributes (see appendix 7)	47

The following table depicts the size and frequency of the large item sets generated.

Table 14: Size and frequency of generated itemsets:

Size of Generated large item sets	Frequency of large item sets
One	15
Two	96
Three	342
Four	683
Five	761
Six	460
Seven	134
Eight	12

Best rules found

Apriori generated 20 best rules on the criteria of confidence and support *(See appendix 10 to view the complete list of the generated rules).* The interesting rules generated after the association rule was applied on one of the two clusters are presented below.

- OFF55=0 LIVING3=0 6549 ==> FEMALE=0 6549 conf:(1)

(If Sexual abuse = No and Living with father only = No then Female = No)

Although not conclusive, the above rule implies that sexual abuse happens to those female children living with their father only.

- OFF55=0 LIVING3=0 LIVING6=0 6514 ==> FEMALE=0 6514 conf:(1)

(If sexual abuse = No and Living with father only = No and Living on the street = No

then Female = No)

Similarly, among other things, this rule also implies that sexual abuse occurs on those female children living on the street or living with their father only.

Experiment 7

Apriori was also applied on the second cluster.

The information from running the algorithm is presented below.

Table 15: Run information of Apriori (4249 instances and 47 attributes)

Scheme	weka.associations.Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation	VICTIM.SYMBOLIC
Instances	4249
Attributes *(see appendix 7)*	47

The following table depicts the size and frequency of the large item sets generated.

Table 16: Size and frequency of generated itemsets (4249 instances and 47 attributes)

Size of Generated large item sets	Frequency of large item sets
One	15
Two	92
Three	310
Four	617
Five	711
Six	451
Seven	157
Eight	34
Nine	2

Best rules found

Following the application of Apriori on the second cluster, 20 best rules were generated on the basis of confidence and support metric *(See appendix 11 to view the complete list of the best 50 rules generated).*

- LIVING3=0 LIVING6=0 4215 ==> MALE=0 4215 conf:(1)

(If living with father only = No and Living on the street = No then male = No)

- OCCUP7=0 LIVING3=0 LIVING6=0 4202 ==> MALE=0 4202 conf:(1)

(If Farmer = No and Living with father only = No and Living on the street = No then male = No)

- OFF6=0 OCCUP5=0 4195 ==> MALE=0 4195 conf:(1)

(If Breach of trust = No and Street vendor = No then male = No)

This rule could also imply that it is female children who work as street vendors that are exposed to the offence BREACH OF TRUST.

- OFF6=0 LIVING6=0 4187 ==> MALE=0 4187 conf:(1)

(If Breach of trust = No and Living on the street = No then Male = No)

This rule implies, among other things, the fact that BREACH OF TRUST happens to male children living on the street.

While using the data that has been converted to the binary format for the generation of association rule, rarely was found a rule that is conclusive, since the non-existent state of the attributes constitute the rules. Hence, the value of the rules can only be appreciated for what they imply rather than what they state/conclude.

4.3.4 Interpretation and discussion

From the different list of rules generated over various experiments using different set of data and attributes, a number 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 selected.

Summary of the input and output of the discovery task, and the subsequent interpretation and discussion of the discovered interesting rules is presented below.

Experiment 1

Input

The input and output for the first experiment are

Instances: 10878

Attributes: ZONE, WOREDA, TIME OFFENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, OCCUPATION, LIVING ARRANGEMENT, PLACE THE OFENCE WAS COMMITTED, MARITAL STATUS, SPECIAL HABIT, and CITIZENSHIP.

Output

PLACE_VI=1 MARRL_VI=1 CITIZEN_VI=1 10229 ==> HABIT=1 10024 conf:(0.98)

(If Place = Urban and Marital Status = Single and Citizenship = Ethiopian then Special habit = No bad habit)

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, 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. 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

Input

Instances: 10878

Attributes: ZONE, WOREDA, TIME OFFENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, OCCUPATION, and LIVING ARRANGEMENT.

Output

- TY_OFF_CO=14 RELIGION_VI=1 OCCUP_VI=1 3440 ==> LIVING=1 3293 conf:(0.96)

(If Type of crime = willful injury and Religion = Christian and Occupation = Student then Living arrangement = Lives with both parents)

- TY_OFF_CO=14 SEX=1 RELIGION_VI=1 3718 ==> LIVING=1 3448 conf:(0.93)

(Type of offence = willful injury and Sex = Male and Religion = Christian then Living Arrangement = lives with both parents)

- TY_OFF_CO=14 AGE=4 OCCUP_VI=1 2874 ==> LIVING=1 2731 conf:(0.95)

(If type of crime = willful injury and Age = 15 to 18 years and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 SEX=1 OCCUP_VI=1 2392 ==> LIVING=1 2266 conf:(0.95)

(If Type of crime = Willful Injury and Sex = Male and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 RELIGION_VI=1 OCCUP_VI=1 3215 ==> LIVING=1 3042 conf:(0.95)

(If Type of crime = Willful injury and Religion = Christian and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 AGE=4 RELIGION_VI=1 4767 ==> LIVING=1 4321 conf:(0.91)

(If Type of crime = willful injury and Age = 15 to 18 years and Religion = Christian then Living Arrangement = with both parents)

The rules stated above represent interesting regularity within the crime database. For example, the rule "Type of offence = Willful injury and Sex = Male and Religion = Christian then Living Arrangement = Lives with both parents" holds that children living with both parents are the ones who are exposed to Willful injury. According to these rules, the other features that make up the children's profile include being Christian, male, student, and age between 15 to 18 years. According to domain experts, these rules are interesting because they are contrary to popular conception that it is children with parents and not orphan children, males and not females, older and not younger children that are the subject of the Willful injury. That is, children presumed to be relatively more protected or more able to take care of themselves are the ones who make up the profile of the victim.

Experiment 3

Input

Instances: 10878

Attributes: ZONE, WOREDA, TIME OFFENCE COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, and OCCUPATION.

Output

- TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1140 ==> AGE=4 1094 conf:(0.96)

(If Type of offence = willful injury and education = 9 to 12 grade and Religion = Christian and Occupation = Student then Age = 15 to 18 years)

- TY_OFF_CO=14 AGE=4 EDUC=5 OCCUP_VI=1 1192 ==> RELIGION_VI=1 1094 conf:(0.92)

(If Type of offence = willful injury and Age = 15 to 18 years and Education = 9 to 12 grade and Occupation = Student then Religion = Christian)

The above sets of rules supplement the discovered regularities in experiment 2. Here also the rules generated, children who were exposed to Willful injury include grade 9 to 12 students.

Experiment 4

Input

Instances: 10847

Attributes: WOREDA, TIME OFFENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, AGE, EDUCATION, RELIGION, and OCCUPATION.

Output

- TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1143 ==> AGE=4 1096 conf:(0.96)

(If type of Crime = willful injury and Education = 9 to 12 grades and Religion = Christian and occupation then age between 15 and 18.)

- TY_OFF_CO=14 AGE=4 EDUC=5 OCCUP_VI=1 1194 ==> RELIGION_VI=1 1096 conf:(0.92)

(If Type of offence = willful injury and Age = 1 to 18 and Education = 9 to 12 grade then Religion = Christian)

The output of this experiment is, more or less, the same as that of experiment two and three.

Experiment 5

Input

Instances: 10878

Attributes: WOREDA, TIME VIOLENCE WAS COMMITTED, TYPE OF OFFENCE, SEX, and AGE.

Output

- TY_OFF_CO=53 1097 ==> AGE=4 993 conf:(0.91)

(If Type of offence = Violation of municipal regulations then Age = 15 to 18 years)

This is one of the interesting regularities in the dataset, children between the ages of 15 and 18 are the ones who often experienced violation of municipal regulations.

- TY_OFF_CO=58 755 ==> AGE=4 683 conf:(0.9)

  (If Type of offence = theft then Age = 15 to 18 years)

Those children who were reported as victims of theft also belong to the age group 15 to 18. This is perhaps because children within this age category have better access or are assigned with the responsibility to spend money or look after gadgets/material than younger children.

- TY_OFF_CO=55 616 ==> SEX=2 549 conf:(0.89)

  (If Type of crime = sexual abuse then Sex = Female)

In line with the popular conception, it is female children who were the ones exposed to sexual abuse.

- TIME_VI=4 TY_OFF_CO=14 SEX=1 1229 ==> AGE=4 1084 conf:(0.88)

  (If Time offence was committed = Early night and Type of offence = Willful injury and Sex = male then Age = 15 to 18)

According to this rule, male children, who were identified in the previous experiments as being exposed to the offence Willful injury, were exposed to the specific offence Early in the night (from 7 to 12 in the evening).

- TIME_VI=2 TY_OFF_CO=14 SEX=2 671 ==> AGE=4 554 conf:(0.83)

(If Time offence was committed = Early Day and Type of Offence = Willful Injury and Sex = Female Then Age = 15 to 18 years)

- TIME_VI=3 TY_OFF_CO=14 SEX=2 877 ==> AGE=4 722 conf:(0.82)

(If Time offence was committed = Late day and Type of offence = Willful Injury and Sex = Female Then Age = 15 to 18 years)

According to the above two rules, unlike male children who were identified in previous experiments, as being exposed to Willful Injury early in the night, female children are exposed to the offence early in the day. Like male children, however, female children that were victims of this offence were between the ages of 15 and 18. As pointed out by domain experts, there are a number of male children that come to the police station after having suffered willful injury as a result of an individual or group fight. In relation to female children, several domestic servants report Willful injury. This perhaps sheds some light on the difference in the timing male and female children are exposed to offence. The difference in time can also be explained in terms of the fact that unlike male children, female children often do not leave home during the night.

Experiment 6

Input

Instances: 10878

Attributes: 76

Output

- W18=0 YHABIT=0 10420 ==> LIVING3=0 10399 conf:(1)

(If Woreda 18= No and Special Habit = No then Living with father = No)

Although not conclusive, this rule implies that children from Woreda 18 and having a special bad habit of one form or another are those that live with their father only.

Experiment 7

Input for cluster 1

Instances: 6628

Attributes: 47

Output

- OFF55=0 LIVING3=0 6549 ==> FEMALE=0 6549 conf:(1)

(If Sexual abuse = No and Living with father only = No then Female = No)

Although not conclusive, the above rule implies, among other things, that sexual abuse happens to those female children living with their father only.

- OFF55=0 LIVING3=0 LIVING6=0 6514 ==> FEMALE=0 6514 conf:(1)

(If sexual abuse = No and Living with father only = No and Living on the street = No

then Female = No)

Similarly, among other things, this rule also implies that sexual abuse occurs on those female children living on the street or living with their father only.

Input for cluster 2

Instances: 4249

Attributes: 47

Output

- OCCUP7=0 LIVING3=0 LIVING6=0 4202 ==> MALE=0 4202 conf:(1)

(If Farmer = No and Living with father only = No and Living on the street = No then male = No)

This rule implies, among other things, that it is male children who either live with their father only or live on the street, or are farmers. If the rule holds, it represents a high level summary of the instances in the database.

- OFF6=0 OCCUP5=0 4195 ==> MALE=0 4195 conf:(1)

(If Breach of trust = No and Street vendor = No then male = No)

This rule could also imply that it is female children who work as street vendors that are exposed to the offence BREACH OF TRUST.

- OFF6=0 LIVING6=0 4187 ==> MALE=0 4187 conf:(1)

(If Breach of trust = No and Living on the street = No then Male = No)

This rule implies, among other things, the fact that BREACH OF TRUST happens to male children living on the street.

The rules generated over a number of experiments constitute those attributes (itemsets) occurring in large frequency in the dataset. Despite the fact that most of the generated rules scored high in terms of the objective measures of interestingness (high support and confidence), most were found to be less

interesting in the eyes of users/domain experts and the purpose of the research, which is discovering interesting rules/regularities.

In the effort to discover relatively more interesting rules that underlie the crime dataset, a series of experiments were conducted. During subsequent experimentation, leaving out the most frequent or invariable attributes that dominate the rules in the first experiments was considered to be a proper measure, since the use of these attributes resulted in more or less uninteresting or trivial rules.

While using the data that has been converted to the binary format, rarely was found a rule that is conclusive, since the non-existent state of the attributes constitute the rules. Hence, the value of the rules can only be appreciated for what they imply rather than what they state/conclude.

From the different list of rules generated over various experiments using different set of data and attributes, those rules with satisfactory objective measure (high support and confidence) and most importantly the subjective judgment of domain experts on their interestingness and applicability were selected.

The rules that were found to be interesting by the domain experts include a rule where the instances in the database are characterized by the occurrence of very few instances having special bad habit such as smoking cigarettes, drinking alcohol, or chewing Chat. According to them, this finding is in line with the popular conception in the domain area that such special bad habits are more often characteristics of child offenders than child victims.

The other interesting rule drawn is the rule that indicates children living with both of their parents constitute the children who are exposed to willful injury. This rule particularly works for Christian males between the ages of 15 to 18 years. In another rule male children who were identified in the above rule as being exposed to the offence of willful injury were exposed to the particular offence early in the night (from 7 to 12 in the evening).

Although not conclusive, one of the rules generated using the binary data stated that if female children are living with their fathers only then they are exposed to sexual abuse. Similarly, another rule implies that if female children are living on the street then they are exposed to sexual abuse.

The above stated rules have practical relevance as they contribute to the increase in knowledge about the profile of children exposed to certain types of offences and the details on the time and place the offences are committed. Hence, the result from the KDD process can serve as a source of input for the awareness raising program of FSCE and the Addis Ababa Police Commission for the purpose of prevention and control of crimes committed against children.

Chapter Five

Conclusion and Recommendation

In this section the conclusions drawn from the findings of the research and the recommendations forwarded in light of the findings and conclusions are presented.

5.1 Conclusion

In this thesis an effort is made to examine the application of the KDD process to support the advocacy and awareness raising program of FSCE and the Addis Ababa Police Commission, and to discover regularities that underlie the crime database.

As it is often the case, the KDD process was undertaken in phases. The process adopted in this research can be described as constituting five phases: Understanding of the problem domain, understanding of the data, data preprocessing, data mining, and evaluation and interpretation of data mining results.

Understanding of the problem domain: The problem domain was explored to have insight into the area and to be able to define the problem to focus on. During this phase close interaction with the domain experts and review of documents was made to good effect.

Understanding of the data: The discovery task was run on the crime database that consists of 10,878 records/tuples in 17 tables describing a total of 25 attributes. Two of the attributes were numeric and the rest were text or nominal. Following consultation with domain experts, 10 of the attributes were

excluded from the discovery task since they either carry little information, are redundant or invariant over instances in the database.

Data mining: Association rule mining, an exploratory data mining technique was applied to accomplish the goal of the research. To this effect, the Apriori algorithm, which is an implementation of the Association rule in the Weka software, was used. With the aim to improve quality of discovered rules, the nominal data with multiple category was transformed into a binary form and the K-means clustering algorithm from the KnowledgeStudio software was applied on it late in the experimentation stage. The clustering scheme was used prior to applying the association rule mining on the binary data to reduce the complexity of the mining task by segmenting the dataset into homogeneous groups, which were then given to the association rule algorithm.

Data preprocessing: Data cleaning and preparation tasks were carried out to handle missing value and noise. Redundant, irrelevant, and invariant attributes were also excluded in this stage of the research. What is more, values of an attribute that communicate more or less similar information were merged and considered as one. The K-means algorithm, which is one of the implementations of the clustering algorithm, was employed late in the experimentation stage as a data preprocessing endeavor to enable the generation of good rules by the association rule algorithm.

Data evaluation and interpretation: 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.

Implications of the results: Although not conclusive, the rules appear to have practical relevance, as they can contribute to the existing knowledge about the profile of victim children. The results of the research support the fact that the KDD process can be applied to the crime dataset and with modest success.

In addition to using the nominal data with multiple values, rules with seemingly interesting implications were uncovered using sequentially a combination of clustering and association rule mining. The generated rules were found to be satisfactory from the point of view of the objective and subjective measures of interestingness. List of the discovered interesting rules is presented below.

- PLACE_VI=1 MARRL_VI=1 CITIZEN_VI=1 10229 ==> HABIT=1 10024 conf:(0.98)

(If Place = Urban and Marital Status = Single and Citizenship = Ethiopian then Special habit = No bad habit)

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, 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.

- TY_OFF_CO=14 SEX=1 OCCUP_VI=1 2392 ==> LIVING=1 2266 conf:(0.95)

  (If Type of crime = Willful Injury and Sex = Male and Occupation = Student then Living Arrangement = with both parents)

- TY_OFF_CO=14 AGE=4 RELIGION_VI=1 4767 ==> LIVING=1 4321 conf:(0.91)

  (If Type of crime = willful injury and Age = 15 to 18 years and Religion = Christian then Living Arrangement = with both parents)

These rules are interesting because they are contrary to popular conception that it is children with parents and not orphan children; males and not females; older and not younger children that are the subject of the Willful injury. That is, children presumed to be relatively more protected or more able to take care of themselves are the ones who make up the profile of the victim.

- TY_OFF_CO=53 1097 ==> AGE=4 993 conf:(0.91)

  (If Type of offence = Violation of municipal regulations then Age = 15 to 18 years)

This is one of the interesting regularities in the dataset, children between the ages of 15 and 18 are the ones who often experienced violation of municipal regulations.

- TY_OFF_CO=58 755 ==> AGE=4 683 conf:(0.9)

  (If Type of offence = theft then Age = 15 to 18 years)

Those children who were reported as victims of theft belong to the age group 15 to 18. This is perhaps because children within this age category have better access or are assigned with the responsibility to spend money or look after gadgets/material than younger children.

- TIME_VI=4 TY_OFF_CO=14 SEX=1 1229 ==> AGE=4 1084 conf:(0.88)

  (If Time offence was committed = Early night and Type of offence = Willful injury and Sex = male then Age = 15 to 18)

- TIME_VI=2 TY_OFF_CO=14 SEX=2 671 ==> AGE=4 554 conf:(0.83)

(If Time offence was committed = Early Day and Type of Offence = Willful Injury and Sex = Female Then Age = 15 to 18 years)

- TIME_VI=3 TY_OFF_CO=14 SEX=2 877 ==> AGE=4 722 conf:(0.82)

(If Time offence was committed = Late day and Type of offence = Willful Injury and Sex = Female Then Age = 15 to 18 years)

According to the above three rules, unlike male children who were identified in previous experiments, as being exposed to Willful Injury early in the night, female children are exposed to the offence early or late in the day. The difference in time can be explained in terms of the fact that unlike male children, female children often do not leave home during the night. Furthermore, according to domain experts, domestic servants comprise a significant proportion of the victims, and children victimized by individual or group fight also constitute a significant proportion of the male children.

- W18=0 YHABIT=0 10420 ==> LIVING3=0 10399 conf:(1)

(If Woreda 18= No and Special Habit = No then Living with father = No)

Although not conclusive, this rule implies that children from Woreda 18 and having a special bad habit of one form or another are those that live with their father only.

- OFF55=0 LIVING3=0 6549 ==> FEMALE=0 6549 conf:(1)

(If Sexual abuse = No and Living with father only = No then Female = No)

Although not conclusive, the above rule implies, among other things, that sexual abuse happens to those female children living with their father only.

- OFF55=0 LIVING3=0 LIVING6=0 6514 ==> FEMALE=0 6514 conf:(1)

(If sexual abuse = No and Living with father only = No and Living on the street = No

then Female = No)

Similarly, among other things, this rule also implies that sexual abuse occurs on those female children living on the street or living with their father only.

- OFF6=0 OCCUP5=0 4195 ==> MALE=0 4195 conf:(1)

(If Breach of trust = No and Street vendor = No then male = No)

This rule could also imply that it is female children who work as street vendors that are exposed to the offence BREACH OF TRUST.

- OFF6=0 LIVING6=0 4187 ==> MALE=0 4187 conf:(1)

(If Breach of trust = No and Living on the street = No then Male = No)

This rule implies, among other things, the fact that BREACH OF TRUST happens to male children living on the street.

5.2 Recommendation

On the basis of the findings of the study and the experience gained from the research, the following recommendations were suggested.

Incorporating the KDD process

The crime database has the potential to grow immensely with increase in the awareness of the society, and techniques of this sort are very important. The KDD process can play a crucial role in making available important information that can serve as input for planning and implementation of an effective advocacy and awareness raising program directed to prevent and control crimes committed against children. In this regard FSCE should appreciate the benefits that can be accrued from adopting the KDD process and should create access for relevant staff to have knowledge about the process.

Data integration

Although the crime database consists of three files, no attempt was made to relate these files. The file on adults that commit offences against children and the data for victim children could have been related using a common unique field. This will provide the opportunity to explore the relationship between victim children and adult offenders.

Improving data quality

More attributes should be added to allow complete analysis of the profile of the children. Possible values of an attribute should be less ambiguous. Hence such categories need to be redesigned.

Further research

The research focused on applying exploratory/descriptive data mining techniques. Having learned about the regularities that underlie the crime data, attention should now focus on experimenting on the application of classification/predictive data mining on the data, perhaps, having type of offence as the class label .

Deployment/Practical Implications of the results

Although the findings of the research are not conclusive, they, however, can be considered as giving insight to the bigger picture of the phenomena of crimes committed against children. Accordingly, the findings can be incorporated for the advocacy and awareness raising efforts of FSCE and the Addis Ababa Police Commission, preferably after being substantiated and supplemented by qualitative research.

The following recommendations are forwarded on the basis of the findings of the study, and the conclusion drawn from them.

Willful Injury

With regard to willful injury older children (15 to 18 years) living with both their parents constitutes the majority of the victims. Hence, this can serve as an input in the effort to prevent this particular type offence. For one thing, the awareness raising program can target children within the given age

category, or can even incorporate those children between the ages of 10 to 14, as an early intervention. Since the children are living with their parents, the program can also target parents. With supplementary information that could be gained from qualitative research, insight can be gained about the contributory factors for the offence to occur and the profile of persons inflicting the offence, which enrich the awareness creation program.

The findings of the research also disclosed that male children are exposed to willful injury early in the night (7 to 12 pm) while female children are exposed to the offence early (7 to 12 am) and late (1 to 6 pm) in the day. This can also prove to be an important finding to be used in the awareness raising program since it gives insight on the frequent timing of the offence for children of both sexes.

Furthermore, according to domain experts, while domestic servants comprise a significant proportion of the female victims, male children constitute a significant proportion of children victimized by individual or group fight. Hence, the awareness raising effort can focus in lobbying for the respect of the right of female domestic servants. Male children can also be informed to keep away from disputes that could lead to physical confrontation.

The interesting rules discovered in this research can also serve as good/interesting research problems or hypotheses. For instance, a research can look into the reason why “children with parents and not orphan children; males and not females; older and not younger children experienced Willful injury more?”

Sexual Abuse

Despite lacking conclusiveness, there are findings that imply, among other things, that sexual abuse is committed against female children living with their father only. Similarly, another rule implies, among other things, that sexual abuse is committed against female children living on the street. Hence, the awareness raising program could focus on forwarding and promoting ways to promoting female children living on the street or with their fathers only from sexual offences.

Breach of Trust

Some of the rules imply, among other things, that it is female children who work as street vendors, and male children living on the street that are exposed to the offence BREACH OF TRUST. Hence better protection for these groups of children should be advocated. This is particularly important in the case of male children living on the street since they are considered, more often than not, as perpetrators/cheaters than victims by the general public.

Theft

Those children who were reported as victims of theft belong to the age group 15 to 18. This is perhaps because children within this age category have better access or are assigned with the responsibility to spend money or look after gadgets/material than younger children. In this regard children that meet the above profile can be informed on how to protect themselves from being exposed to such problem.

References

1. Data Mining Software and Solutions (nd). A Brief History of Data Mining. Available From: Http://Www.Data-Mining-Software.Com
2. Al-Attar, Akeel (Nd). White Paper: Data Mining - Beyond Algorithms, Attar Software. Available from: http://www.attar.com/tutor/mining.htm
3. An Introduction To Data Mining: Discovering Hidden Value In Your Data Warehouse. Available from: http://www.thearling.com/text/dmwhite/dmwhite.htm
4. Anand. Sarabjot S. (B), Bell, David A. & Hughes, John G. (1995). The Role of Domain Knowledge in Data Mining. Northern Ireland: School Of Information and Software Engineering. Faculty of Informatics. University of Ulster (Ordanstown). Available from: http://citeseer.nj.nec.com/anand95role.html
5. Berkhin, Pavel (2002). Survey of clustering data mining techniques. Accrue Software, Inc.
6. Berry, Michael and Linoff, Gordon (1997). Data mining techniques: for marketing, sales, and customer support. New York: Wiley computer publishing
7. Brown, Donald E. (nd). The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals. Virginia: Department of Systems Engineering, University of Virginia. Available from:
8. Carbone, Patricia L. (Nd) Data Mining or "Knowledge Discovery in Databases": An Overview. Available from: http://people.cs.uct.ac.za/~agelderb/MastersPapers/Research Proposal Draft Two.doc

9. Chen, Hsinchun (nd). From Digital Library to Digital Government: A Case Study in Crime Data Mapping and Mining. Artificial Intelligence Lab and Hoffman E-Commerce Lab. Arizona: University of Arizona. Available from: http://vijis.sys.virginia.edu/publication/RECAP.pdf

10. Chen, Ming-Syan et al. (1996). Data Mining: An Overview from Database Perspective. Taipei: National Taiwan University. Available from http://www.citeseer.nj.nec.com/chen97data.html

11. Clare, Amanda (2003). Machine learning and data mining for yeast functional genomics. Aberystwyth: The University of Wales. (Doctor of Philosophy)

12. Corcoran, Jonathan and Ware, Andrew (nd). Forecasting Crime: An Ethical Conundrum. Available from: http://www.aic.gov.au/conferences/mapping/muscat.pdf

13. Deogun, Jitender S. et al (1997). Data Mining: Research Trends, Challenges, and Applications. Lincoln: University of Nebraska. Available from: http://citeseer.nj.nec.com/deogun97data.html

14. Fayyad, Usama, Piatetsky-Shapiro, Gregory, & Smyth, Padhraic (1996). From Data Mining To Knowledge Discovery In Databases. Available from: http://citeseer.nj.nec.com/fayyad96from.html

15. Fayyd, Usama Et Al. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Available from: http://citeseer.nj.nec.com/fayyad96from.html

16. Garner, Stephen R. (nd). WEKA: The Waikato Environment for Knowledge Analysis. Department of Computer Science, University of Waikato, Hamilton. Available from

17. Goebel, Michael & Gruenwald, Le (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. Oklahoma/New Zealand. Available from: http://citeseer.nj.nec.com/goebel99survey.html

18. Han, Eui-Hong; Karypis, George; Kumar, Vipin (1997). Scalable Parallel Data Mining for Association Rules. Minneapolis: University Of Minnesota. Available from: http://citeseer.nj.nec.com/199683.html

19. Hipp, Jochen and Nakhaeizadeh, Gholamreza (Nd). Data Mining Of Association Rules and the Process of Knowledge Discovery In Databases. Ulm: Daimlerchrysler Ag, Research & Technology. Available from: http://citeseer.nj.nec.com/534971.html

20. Hipp, Jochen et al (2000) Algorithms for Association Rule Mining- A General Survey and Comparison. Tubingen: University of Tubingen. Available from: http://citeseer.nj.nec.com/hipp00algorithms.html

21. Hipp, Jochen, G¨Untzer, Ulrich, & Grimmer, Udo (Nd). Integrating Association Rule Mining Algorithms with Relational Database Systems. Ulm/Germany: Daimlerchrysler Ag, Research & Technology. Available from: http://www-db.informatik.uni-tuebingen.de/forschung/papers/iceis01.pdf

22. Holsheimer, Marcel & Siebes, Amo (1991). Data Mining: The Search For Knowledge In Databases. Amsterdam. Available from: http://citeseer.nj.nec.com/holsheimer91data.html

23. Joshi, Karuna Pande (1997). Analysis Of Data Mining Algorithms. Available from: http://gl.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm

24. KNOWLEDGE Studio/Knowledge Server. White Paper, Version 4. KnowledgeSTUDIO Workstation| KnowledgeSERVERData Mining Engine. A new generation of data mining technologies from ANGOSS Software Corporation (2002). Available from:

25. Kok, J. N. & Kosters, W.A. (1991) Natural Data Mining Techniques. Netherlands: Leiden Institute of Advanced Computer Science Universiteit Leiden. Available from: http://citeseer.nj.nec.com/kok91natural.html

26. Liu, Bing et al (1997). Using General Impressions To Analyze Discovered Classification Rules. Department Of Information Systems And Computer Science. Available from: http://citeseer.nj.nec.com/liu97using.html

27. Mannila, Heikki (1997). Methods and Problems in Data Mining. Helsinki: University of Helsinki. Available from: http://citeseer.nj.nec.com/mannila97methods.html

28. Mannila, Heikki (nd). Theoretical Frameworks for Data Mining. Finland: Nokia Research Center. Available from: http://www.acm.org/sigkdd/explorations/issue1-2/mannila.pdf

29. Newing, Rod (Nd). Understanding Data Mining, In Management: Overview Pc Network Advisor (Pcna). Issue 74 Page 16. Available from: http://www.pcsupportadvisor.com/nasample/M0481.pdf

30. Palous, Jiri (Nd). Machine Learning and Data Mining. Prague: Gerstner Laboratory for Intelligent Decision Making And Control Czech Technical University. Available from: http://citeseer.nj.nec.com/506615.html

31. Pmsi. Data Mining: An Attempt to Clear Up the Confusion. Last Update: August 20, 2001. Available from: http://www.pmsi.fr/dminit2a.htm

32. Rogers, Jim (2001). INFTData Mining Using the EM Clustering Algorithm on Places Rated Almanac Data. Available from:

33. Rudjer Boskovic Institute (2001). About DM Tutorial.

34. Science Tribune (1997). Data Mining: Data Mining Vs Statistics. Pmsi. Available from: http://www.pmsi.fr/dminita.htm

35. Silberschhatz, Avi & Tuzhilns, Alexander (1996). What Makes Patterns Interesting In Knowledge Discovery System. New York University. Available from: http://citeseer.nj.nec.com/silberschatz96what.html

36. SPSS INC. Rapid Response in Government: Fight Crime and Improve Security with Data Mining. Tuesday, February 25, 2003 Seminar] Availabelf from: http://www.spss.com/RR-InvestigativeDM

37. Stiles, Eric (nd). DATA MINING: Introductory and Advanced Topics: Association Rules. **Available from:**

38. The Code of Conduct for NGOs in Ethiopia (1998)

39. Toivonen Et Al. (1995). Pruning And Grouping Discovered Association Rules. Finland: University Of Helsinki. Available from: http://citeseer.nj.nec.com/toivonen95pruning.html

40. UNITeS. Research Center (nd) Database Support available at http://WWW.unites.org

41. Veenendaal, Bert et al (nd). Gut Feelings, Crime Data and GIS. Western Australia: Curtin University of Technology. Available from: http://www.aic.gov.au/conferences/mapping/houweling.pdf

42. Webb, Geoffery I. (1996). A heuristic covering algorithm has higher predictive accuracy than learning all rules. Australia: Deakin University. Available from: http://www.cm.deakin.edu.au/webb/Papers/lar.pdf.

43. Witten, Ian H. and Frank, Eibe (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations

44. World Bank (2001). The World Bank and Civil Society: Key Documents/Working with NGOs. Available from: http://wbln0018.worldbank.org/esd.nsf/0/169bd0c0c618852567ed004c5114?OpenDocument

45. Zaki, Mohammed J. & Hsiao, Ching-Jui (2002). Charm: An Efficient Algorithm for Closed Association Rule Mining. Troy Ny: Computer Science Department Rensselaer Polytechnic Institute. Available from: http://citeseer.nj.nec.com/zaki02charm.html

46. Zaki, Mohammed J. (1999). Parallel and Distributed Association Mining: A Survey. Rensselaer Polytechnic Institute, Ieee. Available from: http://www.cs.rpi.edu/~zaki/PS/concurrency.pdf

Appendices

Appendix 1: List of combinations of attributes for preparing reports

- Type of offense by place of residence (Woreda) of child victim
- Type of offense by place of child victim
- Type of offense by sex of child victim
- Type of offense by age of child victim
- Type of offense by educational level child victim
- Type of offense by marital status child victim
- Type of offense by citizenship of child victim
- Type of offense by religion of child victim
- Type of offense by occupational status child victim
- Type of offense by special habit of child victim
- Type of offense by living condition of child victim
- Type of offense by current status of child victim

Appendix 2: Metadata of the attributes in the database

Attribute	Description	Format	Domain (possible values)
Registration No		Text- nominal	-
Registration book NO		Text- nominal	-
Application date	The date in which the case was reported to the police	Date	-
Region	Refers to the region where the victim child dwells	Text- nominal	14
Zone	Refers to the zone where the victim child dwells	Text- nominal	Zone 1, 2, 3, 4, 5, and 6
Woreda	Refers to the Woreda where the victim child dwells	Text- nominal	From 1 to 28
Kebele	Refers to the Kebele where the victim child dwells	Text- nominal	-
Special names for the areas	Refers to the special name of the place where the victim child dwells	Text- nominal	-
Name of the victim	The child victim's name	Text-nominal	-
Week_day	Refers to the name of the date of the week the offence was committed	Text-nominal	days of the week Monday through Sunday
TIME_VI	Refers to the time the offence was committed	Numeric-integer	GMT
Special names of the areas	Refers to the special name of the site where the offence was committed	Text- nominal	-
Place_VI	Refers to the general description of the site the offence was committed	Text-nominal	Urban, rural
TY_OFF_CO	Refers to the code representation of the	Text-nominal	A list of codes from 1 to 54

	crime committed against the child		
Type of offence (name)	Refers to the name of the crime committed against the child	Text-nominal	Name of the offence
Sex	The victim child's sex	Text-nominal	Male, female
Age	The victim child's age	Numeric—integer	-
Education	The level of education of the child	Text-nominal	Illiterate, able to read and write. Grades from 1 to 12, above grade 12
MARRL_VI	The marital status of the child	Text-nominal	Single, married, divorced/separated, widowed
CITIZEN_VI	The child's citizenship	Text-nominal	Ethiopian, another
RELIGION_VI	The child's religion	Text-nominal	Christian, Muslim, pagan
OCCUP_VI	The major daily engagement of the child	Text-nominal	Student, unemployed, Private work, employee of individual, Daily laborer, Farmer, Street vendor, Government employee, Owner of private enterprise, Employee of NGO, others
Habit_VI	Refers to any harmful drug using habit of the child	Text-nominal	No special (bad) habit, Drug and substance abuse, Alcohol drinking, Smoking cigarettes, Sniffing benzene, Gambling, "Chat" chewing, others
LIVING	Refers to the persons/situations the child used to live with/in during the offence	Text-nominal	Living with both parents, Living with mother alone, Living with father alone, Living with relatives, Living with non-relatives, Living on the street, Others
Living arrangement (status) after offence	Refers to the persons/situation the child used to live with/in after the offence	Text-nominal	Living with parents, Living on street, Sent to orphanages, Locally adopted, Internationally adopted, Others

Appendix 3: The 15 attributes and their respective categories

1. Zone: 1 to 6
2. Woreda: 1 to 28
3. Week_day

    1. Monday
    2. Tuesday
    3. Wednesday
    4. Thursday
    5. Friday
    6. Saturday
    7. Sunday

4. TIME_VI (Time violence committed)

    1. Late night (7-12)

    2. Early day (1-6)

    3. Late day (7-12)

    4. Early night (1-6)

5. Place_VI
    1. Urban
    2. Rural

6. TY_OFF_CO (crime type)

Code	Description
4	Counterfeit currency
6	Breach of trusty
11	Negligent homicide
14	Willful injury
15	Injuries caused by negligence
17	Exposure or abandonment of an infant
23	Burglary
53	Violation of municipal regulations
55	(merged 19, 20, and 22) sexual offences
56	(merged 10 and 13) homicide
57	(merged 16 and 18) abortion
58	(merged 24, 25, and 31) theft
59	(merged 32, 33, 34, and 35) robbery

7. Sex
    1. Male
    2. Female

8. Age
    1. 0-4
    2. 5-9
    3. 10-14
    4. 15-18

9. Education
    1. Illiterate
    2. Read and write only
    3. 1-6
    4. 7-8
    5. 9-12
    6. above 12

10. MARRL_VI (marital status of the victim)
    1. single
    2. married
    3. widowed/separated

11. CITIZEN_VI (citizenship of the victim)
    1. Ethiopian
    2. another/non-Ethiopian

12. RELIGION_VI (religion of the victim)
    1. Christian
    2. Muslim
    3. Pagan

13. OCCUP_VI (occupation of the victim)
    1. Student
    2. Unemployed
    3. Employee of individual
    4. Daily laborer
    5. Street vendor
    6. Private work
    7. Farmer
    8. Owner of private enterprise
    9. Employed in organizations (Government employee, Employee of NGO)

14. Habit_VI (special habit of the victim)
    1. No bad habit
    2. Bad habit

15. LIVING (Living status of the child before the offence)
    1. with parents
    2. only with mother
    3. only with father
    4. with relative
    5. with someone not relative
    6. on the street

Appendix 4: Best 20 rules generated using 14 attributes

1. MARRL_VI=1 HABIT=1 10273 ==> CITIZEN_VI=1 10234 conf:(1)
2. MARRL_VI=1 10486 ==> CITIZEN_VI=1 10446 conf:(1)
3. PLACE_VI=1 MARRL_VI=1 HABIT=1 10063 ==> CITIZEN_VI=1 10024 conf:(1)
4. PLACE_VI=1 MARRL_VI=1 10269 ==> CITIZEN_VI=1 10229 conf:(1)
5. HABIT=1 10648 ==> CITIZEN_VI=1 10592 conf:(0.99)
6. PLACE_VI=1 HABIT=1 10432 ==> CITIZEN_VI=1 10376 conf:(0.99)
7. PLACE_VI=1 10654 ==> CITIZEN_VI=1 10595 conf:(0.99)
8. PLACE_VI=1 MARRL_VI=1 CITIZEN_VI=1 10229 ==> HABIT=1 10024 conf:(0.98)
9. PLACE_VI=1 MARRL_VI=1 10269 ==> HABIT=1 10063 conf:(0.98)
10. HABIT=1 10648 ==> PLACE_VI=1 10432 conf:(0.98)
11. MARRL_VI=1 CITIZEN_VI=1 10446 ==> HABIT=1 10234 conf:(0.98)
12. MARRL_VI=1 10486 ==> HABIT=1 10273 conf:(0.98)
13. CITIZEN_VI=1 HABIT=1 10592 ==> PLACE_VI=1 10376 conf:(0.98)
14. MARRL_VI=1 HABIT=1 10273 ==> PLACE_VI=1 10063 conf:(0.98)
15. MARRL_VI=1 CITIZEN_VI=1 HABIT=1 10234 ==> PLACE_VI=1 10024 conf:(0.98)
16. CITIZEN_VI=1 10818 ==> PLACE_VI=1 10595 conf:(0.98)
17. PLACE_VI=1 CITIZEN_VI=1 10595 ==> HABIT=1 10376 conf:(0.98)
18. MARRL_VI=1 10486 ==> PLACE_VI=1 10269 conf:(0.98)
19. MARRL_VI=1 CITIZEN_VI=1 10446 ==> PLACE_VI=1 10229 conf:(0.98)
20. PLACE_VI=1 10654 ==> HABIT=1 10432 conf:(0.98)

Appendix 5: Best 20 rules generated using 10 attributes

1. SEX=1 OCCUP_VI=1 3615 ==> LIVING=1 3490 conf:(0.97)
2. TY_OFF_CO=14 RELIGION_VI=1 OCCUP_VI=1 3440 ==> LIVING=1 3293 conf:(0.96)
3. AGE=4 RELIGION_VI=1 OCCUP_VI=1 4363 ==> LIVING=1 4176 conf:(0.96)
4. TY_OFF_CO=14 OCCUP_VI=1 3879 ==> LIVING=1 3712 conf:(0.96)
5. AGE=4 OCCUP_VI=1 4894 ==> LIVING=1 4682 conf:(0.96)
6. RELIGION_VI=1 OCCUP_VI=1 5544 ==> LIVING=1 5280 conf:(0.95)
7. OCCUP_VI=1 6252 ==> LIVING=1 5954 conf:(0.95)
8. SEX=1 AGE=4 RELIGION_VI=1 4599 ==> LIVING=1 4273 conf:(0.93)
9. SEX=1 RELIGION_VI=1 5469 ==> LIVING=1 5074 conf:(0.93)
10. TY_OFF_CO=14 SEX=1 RELIGION_VI=1 3718 ==> LIVING=1 3448 conf:(0.93)
11. TIME_VI=3 RELIGION_VI=1 3539 ==> LIVING=1 3266 conf:(0.92)
12. SEX=1 AGE=4 5581 ==> LIVING=1 5112 conf:(0.92)
13. SEX=1 6659 ==> LIVING=1 6095 conf:(0.92)
14. TY_OFF_CO=14 SEX=1 AGE=4 3778 ==> LIVING=1 3453 conf:(0.91)
15. TY_OFF_CO=14 SEX=1 4556 ==> LIVING=1 4163 conf:(0.91)
16. TIME_VI=3 4167 ==> LIVING=1 3802 conf:(0.91)
17. TY_OFF_CO=14 RELIGION_VI=1 5696 ==> LIVING=1 5169 conf:(0.91)
18. TY_OFF_CO=14 AGE=4 RELIGION_VI=1 4767 ==> LIVING=1 4321 conf:(0.91)
19. AGE=4 RELIGION_VI=1 7605 ==> LIVING=1 6881 conf:(0.9)
20. RELIGION_VI=1 9135 ==> LIVING=1 8263 conf:(0.9)

Appendix 6: Best rules generated using 9 attributes

1. EDUC=5 2700 ==> AGE=4 2600 conf:(0.96)
2. EDUC=5 RELIGION_VI=1 2478 ==> AGE=4 2384 conf:(0.96)
3. EDUC=5 OCCUP_VI=1 2058 ==> AGE=4 1979 conf:(0.96)
4. TY_OFF_CO=14 EDUC=5 1635 ==> AGE=4 1572 conf:(0.96)
5. TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 1504 ==> AGE=4 1446 conf:(0.96)
6. EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1890 ==> AGE=4 1816 conf:(0.96)
7. TY_OFF_CO=14 EDUC=5 OCCUP_VI=1 1242 ==> AGE=4 1192 conf:(0.96)
8. TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1140 ==> AGE=4 1094 conf:(0.96)
9. SEX=1 EDUC=5 1686 ==> AGE=4 1614 conf:(0.96)
10. SEX=1 EDUC=5 RELIGION_VI=1 1544 ==> AGE=4 1476 conf:(0.96)
11. SEX=1 EDUC=5 OCCUP_VI=1 1260 ==> AGE=4 1203 conf:(0.95)
12. SEX=1 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1155 ==> AGE=4 1102 conf:(0.95)
13. ZONE=4 AGE=4 OCCUP_VI=1 1415 ==> RELIGION_VI=1 1324 conf:(0.94)
14. ZONE=4 OCCUP_VI=1 1784 ==> RELIGION_VI=1 1667 conf:(0.93)
15. OCCUP_VI=6 1399 ==> AGE=4 1294 conf:(0.92)
16. TY_OFF_CO=14 EDUC=5 1635 ==> RELIGION_VI=1 1504 conf:(0.92)
17. TY_OFF_CO=14 AGE=4 EDUC=5 1572 ==> RELIGION_VI=1 1446 conf:(0.92)
18. EDUC=5 OCCUP_VI=1 2058 ==> RELIGION_VI=1 1890 conf:(0.92)
19. TY_OFF_CO=14 EDUC=5 OCCUP_VI=1 1242 ==> RELIGION_VI=1 1140 conf:(0.92)
20. TY_OFF_CO=14 AGE=4 EDUC=5 OCCUP_VI=1 1192 ==> RELIGION_VI=1 1094 conf:(0.92)

Appendix 7: Best rules found using 8 attributes

1. EDUC=5 2700 ==> AGE=4 2600 conf:(0.96)
2. EDUC=5 RELIGION_VI=1 2479 ==> AGE=4 2385 conf:(0.96)
3. EDUC=5 OCCUP_VI=1 2060 ==> AGE=4 1981 conf:(0.96)
4. EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1892 ==> AGE=4 1818 conf:(0.96)
5. TY_OFF_CO=14 EDUC=5 1636 ==> AGE=4 1572 conf:(0.96)
6. TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 1506 ==> AGE=4 1447 conf:(0.96)
7. TY_OFF_CO=14 EDUC=5 OCCUP_VI=1 1245 ==> AGE=4 1194 conf:(0.96)
8. TY_OFF_CO=14 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1143 ==> AGE=4 1096 conf:(0.96)
9. SEX=1 EDUC=5 1686 ==> AGE=4 1614 conf:(0.96)
10. SEX=1 EDUC=5 RELIGION_VI=1 1544 ==> AGE=4 1476 conf:(0.96)
11. SEX=1 EDUC=5 OCCUP_VI=1 1262 ==> AGE=4 1205 conf:(0.95)
12. SEX=1 EDUC=5 RELIGION_VI=1 OCCUP_VI=1 1157 ==> AGE=4 1104 conf:(0.95)
13. OCCUP_VI=6 1399 ==> AGE=4 1294 conf:(0.92)
14. TY_OFF_CO=14 EDUC=5 1636 ==> RELIGION_VI=1 1506 conf:(0.92)
15. TY_OFF_CO=14 AGE=4 EDUC=5 1572 ==> RELIGION_VI=1 1447 conf:(0.92)
16. EDUC=5 OCCUP_VI=1 2060 ==> RELIGION_VI=1 1892 conf:(0.92)
17. EDUC=5 2700 ==> RELIGION_VI=1 2479 conf:(0.92)
18. TY_OFF_CO=14 EDUC=5 OCCUP_VI=1 1245 ==> RELIGION_VI=1 1143 conf:(0.92)
19. TY_OFF_CO=14 AGE=4 EDUC=5 OCCUP_VI=1 1194 ==> RELIGION_VI=1 1096 conf:(0.92)
20. AGE=4 EDUC=5 OCCUP_VI=1 1981 ==> RELIGION_VI=1 1818 conf:(0.92)

Appendix 8: Best rules found using 5 attributes

1. TY_OFF_CO=53 1097 ==> AGE=4 993 conf:(0.91)
2. TY_OFF_CO=58 755 ==> AGE=4 683 conf:(0.9)
3. TY_OFF_CO=55 616 ==> SEX=2 549 conf:(0.89)
4. TIME_VI=4 TY_OFF_CO=14 1806 ==> AGE=4 1605 conf:(0.89)
5. TIME_VI=4 SEX=1 1808 ==> AGE=4 1599 conf:(0.88)
6. TIME_VI=4 TY_OFF_CO=14 SEX=1 1229 ==> AGE=4 1084 conf:(0.88)
7. TIME_VI=4 2896 ==> AGE=4 2530 conf:(0.87)
8. WOREDA=7 767 ==> AGE=4 668 conf:(0.87)
9. TIME_VI=4 SEX=2 1088 ==> AGE=4 931 conf:(0.86)
10. TIME_VI=1 673 ==> AGE=4 574 conf:(0.85)
11. TY_OFF_CO=14 SEX=2 2263 ==> AGE=4 1921 conf:(0.85)
12. SEX=1 6659 ==> AGE=4 5581 conf:(0.84)
13. TY_OFF_CO=14 6821 ==> AGE=4 5701 conf:(0.84)
14. TY_OFF_CO=14 SEX=1 4558 ==> AGE=4 3780 conf:(0.83)
15. TIME_VI=2 TY_OFF_CO=14 SEX=2 671 ==> AGE=4 554 conf:(0.83)
16. TIME_VI=2 SEX=1 1948 ==> AGE=4 1607 conf:(0.82)
17. TIME_VI=2 3141 ==> AGE=4 2586 conf:(0.82)
18. TIME_VI=3 TY_OFF_CO=14 SEX=2 877 ==> AGE=4 722 conf:(0.82)
19. SEX=2 4218 ==> AGE=4 3467 conf:(0.82)
20. TIME_VI=2 SEX=2 1193 ==> AGE=4 979 conf:(0.82)

Appendix 9: Best 20 rules found using 78 binary attributes

1. MARRL1=1 10427 ==> MARRL2=0 10427 conf:(1)
2. MARRL1=1 LIVING3=0 10400 ==> MARRL2=0 10400 conf:(1)
3. MARRL1=1 LIVING6=0 10373 ==> MARRL2=0 10373 conf:(1)
4. AGE1=0 MARRL1=1 10359 ==> MARRL2=0 10359 conf:(1)
5. MARRL1=1 LIVING3=0 LIVING6=0 10346 ==> MARRL2=0 10346 conf:(1)
6. MARRL1=1 OCCUP9=0 10334 ==> MARRL2=0 10334 conf:(1)
7. W18=0 YHABIT=0 10420 ==> LIVING3=0 10399 conf:(1)
8. W18=0 YHABIT=0 LIVING6=0 10374 ==> LIVING3=0 10353 conf:(1)
9. AGE1=0 AGE2=0 YHABIT=0 10369 ==> LIVING3=0 10348 conf:(1)
10. W18=0 AGE1=0 AGE2=0 10357 ==> LIVING3=0 10336 conf:(1)
11. AGE2=0 YHABIT=0 10452 ==> LIVING3=0 10430 conf:(1)
12. W18=0 AGE2=0 10442 ==> LIVING3=0 10420 conf:(1)
13. W11=0 W18=0 10436 ==> LIVING3=0 10414 conf:(1)
14. W10=0 YHABIT=0 10427 ==> LIVING3=0 10405 conf:(1)
15. W10=0 W18=0 10414 ==> LIVING3=0 10392 conf:(1)
16. AGE2=0 YHABIT=0 LIVING6=0 10405 ==> LIVING3=0 10383 conf:(1)
17. W11=0 AGE1=0 AGE2=0 10399 ==> LIVING3=0 10377 conf:(1)
18. W22=0 AGE1=0 YHABIT=0 10394 ==> LIVING3=0 10372 conf:(1)
19. W18=0 AGE2=0 LIVING6=0 10390 ==> LIVING3=0 10368 conf:(1)
20. W11=0 W18=0 LIVING6=0 10383 ==> LIVING3=0 10361 conf:(1)

Appendix 10: Best 20 rules found using 76 binary attributes

1. W18=0 YHABIT=0 10420 ==> LIVING3=0 10399 conf:(1)
2. W18=0 YHABIT=0 LIVING6=0 10374 ==> LIVING3=0 10353 conf:(1)
3. AGE1=0 AGE2=0 YHABIT=0 10369 ==> LIVING3=0 10348 conf:(1)
4. W18=0 AGE1=0 AGE2=0 10357 ==> LIVING3=0 10336 conf:(1)
5. AGE2=0 YHABIT=0 10452 ==> LIVING3=0 10430 conf:(1)
6. W18=0 AGE2=0 10442 ==> LIVING3=0 10420 conf:(1)
7. W11=0 W18=0 10436 ==> LIVING3=0 10414 conf:(1)
8. W10=0 YHABIT=0 10427 ==> LIVING3=0 10405 conf:(1)
9. W10=0 W18=0 10414 ==> LIVING3=0 10392 conf:(1)
10. AGE2=0 YHABIT=0 LIVING6=0 10405 ==> LIVING3=0 10383 conf:(1)
11. W11=0 AGE1=0 AGE2=0 10399 ==> LIVING3=0 10377 conf:(1)
12. W22=0 AGE1=0 YHABIT=0 10394 ==> LIVING3=0 10372 conf:(1)
13. W18=0 AGE2=0 LIVING6=0 10390 ==> LIVING3=0 10368 conf:(1)
14. W11=0 W18=0 LIVING6=0 10383 ==> LIVING3=0 10361 conf:(1)
15. W10=0 YHABIT=0 LIVING6=0 10383 ==> LIVING3=0 10361 conf:(1)
16. W11=0 AGE1=0 YHABIT=0 10379 ==> LIVING3=0 10357 conf:(1)
17. W18=0 W22=0 AGE1=0 10377 ==> LIVING3=0 10355 conf:(1)
18. W10=0 W11=0 AGE1=0 10374 ==> LIVING3=0 10352 conf:(1)
19. W10=0 AGE1=0 AGE2=0 10370 ==> LIVING3=0 10348 conf:(1)
20. W10=0 W18=0 LIVING6=0 10364 ==> LIVING3=0 10342 conf:(1)

Appendix 11: Best 50 rules generated from using 47 binary attributes on the first cluster

1. MALE=1 6626 ==> FEMALE=0 6626 conf:(1)
2. LIVING3=0 6616 ==> FEMALE=0 6616 conf:(1)
3. MALE=1 LIVING3=0 6614 ==> FEMALE=0 6614 conf:(1)
4. LIVING6=0 6590 ==> FEMALE=0 6590 conf:(1)
5. MALE=1 LIVING6=0 6588 ==> FEMALE=0 6588 conf:(1)
6. AGE1=0 6579 ==> FEMALE=0 6579 conf:(1)
7. LIVING3=0 LIVING6=0 6578 ==> FEMALE=0 6578 conf:(1)
8. MALE=1 AGE1=0 6577 ==> FEMALE=0 6577 conf:(1)
9. MALE=1 LIVING3=0 LIVING6=0 6576 ==> FEMALE=0 6576 conf:(1)
10. AGE1=0 LIVING3=0 6567 ==> FEMALE=0 6567 conf:(1)
11. MALE=1 AGE1=0 LIVING3=0 6565 ==> FEMALE=0 6565 conf:(1)
12. OFF55=0 6561 ==> FEMALE=0 6561 conf:(1)
13. OFF55=0 MALE=1 6559 ==> FEMALE=0 6559 conf:(1)
14. AGE2=0 6555 ==> FEMALE=0 6555 conf:(1)
15. MALE=1 AGE2=0 6553 ==> FEMALE=0 6553 conf:(1)
16. OCCUP9=0 6552 ==> FEMALE=0 6552 conf:(1)
17. MALE=1 OCCUP9=0 6550 ==> FEMALE=0 6550 conf:(1)
18. OFF55=0 LIVING3=0 6549 ==> FEMALE=0 6549 conf:(1)
19. LIVING2=0 6549 ==> FEMALE=0 6549 conf:(1)
20. OFF55=0 MALE=1 LIVING3=0 6547 ==> FEMALE=0 6547 conf:(1)
21. MALE=1 LIVING2=0 6547 ==> FEMALE=0 6547 conf:(1)
22. AGE1=0 LIVING6=0 6544 ==> FEMALE=0 6544 conf:(1)
23. AGE2=0 LIVING3=0 6543 ==> FEMALE=0 6543 conf:(1)
24. MALE=1 AGE1=0 LIVING6=0 6542 ==> FEMALE=0 6542 conf:(1)
25. MALE=1 AGE2=0 LIVING3=0 6541 ==> FEMALE=0 6541 conf:(1)
26. OCCUP9=0 LIVING3=0 6540 ==> FEMALE=0 6540 conf:(1)
27. OCCUP7=0 6539 ==> FEMALE=0 6539 conf:(1)
28. MALE=1 OCCUP9=0 LIVING3=0 6538 ==> FEMALE=0 6538 conf:(1)
29. LIVING2=0 LIVING3=0 6537 ==> FEMALE=0 6537 conf:(1)
30. MALE=1 OCCUP7=0 6537 ==> FEMALE=0 6537 conf:(1)

Appendix 12: Best rules found from using 47 binary attributes on the second cluster

1. OCCUP5=0 4238 ==> MALE=0 4238 conf:(1)
2. OCCUP7=0 4236 ==> MALE=0 4236 conf:(1)
3. LIVING3=0 4234 ==> MALE=0 4234 conf:(1)
4. LIVING6=0 4230 ==> MALE=0 4230 conf:(1)
5. OCCUP5=0 OCCUP7=0 4225 ==> MALE=0 4225 conf:(1)
6. OCCUP5=0 LIVING3=0 4223 ==> MALE=0 4223 conf:(1)
7. OCCUP7=0 LIVING3=0 4221 ==> MALE=0 4221 conf:(1)
8. OCCUP5=0 LIVING6=0 4220 ==> MALE=0 4220 conf:(1)
9. FEMALE=1 4218 ==> MALE=0 4218 conf:(1)
10. OCCUP7=0 LIVING6=0 4217 ==> MALE=0 4217 conf:(1)
11. LIVING3=0 LIVING6=0 4215 ==> MALE=0 4215 conf:(1)
12. AGE1=0 4212 ==> MALE=0 4212 conf:(1)
13. OCCUP5=0 OCCUP7=0 LIVING3=0 4210 ==> MALE=0 4210 conf:(1)
14. OCCUP9=0 4209 ==> MALE=0 4209 conf:(1)
15. OCCUP5=0 OCCUP7=0 LIVING6=0 4207 ==> MALE=0 4207 conf:(1)
16. FEMALE=1 OCCUP5=0 4207 ==> MALE=0 4207 conf:(1)
17. OFF6=0 4206 ==> MALE=0 4206 conf:(1)
18. OCCUP5=0 LIVING3=0 LIVING6=0 4205 ==> MALE=0 4205 conf:(1)
19. FEMALE=1 OCCUP7=0 4205 ==> MALE=0 4205 conf:(1)
20. FEMALE=1 LIVING3=0 4203 ==> MALE=0 4203 conf:(1)
21. OCCUP7=0 LIVING3=0 LIVING6=0 4202 ==> MALE=0 4202 conf:(1)
22. AGE1=0 OCCUP5=0 4201 ==> MALE=0 4201 conf:(1)
23. AGE1=0 OCCUP7=0 4200 ==> MALE=0 4200 conf:(1)
24. FEMALE=1 LIVING6=0 4199 ==> MALE=0 4199 conf:(1)
25. OCCUP5=0 OCCUP9=0 4198 ==> MALE=0 4198 conf:(1)
26. AGE1=0 LIVING3=0 4198 ==> MALE=0 4198 conf:(1)
27. OCCUP7=0 OCCUP9=0 4196 ==> MALE=0 4196 conf:(1)
28. OFF6=0 OCCUP5=0 4195 ==> MALE=0 4195 conf:(1)
29. OFF56=0 4195 ==> MALE=0 4195 conf:(1)
30. FEMALE=1 OCCUP5=0 OCCUP7=0 4194 ==> MALE=0 4194 conf:(1)

Appendix 13: Code and description of crime type

1	Outrages against the constitution or the constitutional authorities
2	Offenses against organizations
3	Bribery corruption
4	Counterfeit currency
5	Offenses against official stamps or instruments
6	Breach of trusty
7	Offenses against public office
8	Grave endangering or sabotage of communications or transport
9	Attempt or damage to services and installations
10	Intentional homicide
11	Negligent homicide
12	Self - defense
13	Attempted homicide
14	Willful injury
15	Injuries caused by negligence
16	Abortion or attempt to produce abortion
17	Exposure or abandonment of an infant
18	Abortion
19	Rape
20	Unnatural carnal offenses
21	Sexual offenses on adult women
22	Sexual outrages on children
23	Burglary
24	Pick pocketing
25	Fraudulent misrepresentation
26	Theft of motor car
27	Theft of parts of a motor car
28	Theft of goods from motor car

29 Theft of livestock
30 Receiving of stolen goods and misappropriation
31 Other kinds of theft
32 Organized robbery
33 Armed and organized robbery
34 Snatching/Extortion/
35 Attempted robbery or attempt to snatch
36 Attempt to or setting fire on crops
37 Arson or attempt thereto
38 Attempt to or damage to property
39 Attempt to or setting fire on forestry
40 Illegal hunting
41 Falsification of industrial products
42 Falsification of agricultural products
43 Falsification of trademarks
44 Importation, acquisition and selling of illicit property
45 Contraband
46 Illegal handling and trafficking of weapons
47 Production of narcotic substances
48 Distribution and hiding of narcotic substances
49 Use/consumption of narcotic substances
50 Offenses against public health
51 Terrorists
52 Press offenses
53 Violation of municipal regulations
54 Matrimonial offenses