MEKELLE UNIVERSITY

ETHIOPIAN INSTITUTE OF TECHNOLOGY – MEKELLE (EiT-M)

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

CRIME PATTERN DETECTION USING DATA MINING TECHNIQUES: CASE OF SHIRE TOWN POLICE OFFICE

TEKLAY LEMA TEKLAHIMANOT

PS/474/11

January, 2025

Mekelle, Ethiopia

MEKELLE UNIVERSITY

ETHIOPIAN INSTITUTE OF TECHNOLOGY – MEKELLE (EiT-M)

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

CRIME PATTERN DETECTION USING DATA MINING TECHNIQUES: CASE OF SHIRE TOWN POLICE OFFICE

A THESIS SUBMITTED TO THE DEPARTMENT OF INFORMTION TECHNOLOGY, SCHOOL OF COMPUTING TO POST GRADUATE STUDIES OF MEKELLE UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

BY:

TEKLAY LEMA

ADVISOR:

GUESH DAGNEW (Phd)

January, 2025

Mekelle, Ethiopia

MEKELLE UNIVERSITY

ETHIOPIAN INSTITUTE OF TECHNOLOGY – MEKELLE (EiT-M)

SCHOOL OF COMPUTING

This is to certify that the thesis prepared by Teklay Lema Teklahimanot, entitled "CRIME PATTERN DETECTION USING DATA MINING TECHNIQUES: CASE OF SHIRE TOWN POLICE OFFICE" and submitted to the school of post graduate studies presented in partial fulfillment of the requirements for the Degree of master of Information Technology complies with the regulations of the university and meets that accepted standards with respect to originality and quality.

Chair Man (Institute Dean).	Signature	Date
1.________________________	________________________	________________________
Adviser	Signature	Date
2.________________________	________________________	________________________
External Examiner	Signature	Date
3.________________________	________________________	________________________
Internal Examiner	Signature	Date
4.________________________	________________________	________________________
Department		
5.________________________	________________________	________________________

January, 2025

Mekelle, Ethiopia

ACKNOWLEDGEMENT

Primary, my deepest gratitude is to the almighty God for giving me the long live and to face challenges and complete this research work.

Next to this, I would like to express my sincerest gratitude to my advisor Dr. Guesh Dagnew (Phd) for his critical guidance and valuable comments.

I would like to thank for Shire Town police office for providing me all the necessary documents and support whenever I need a support from their office.

Finally, I would like to give a special thank for my wife Seble G/medhn for her supporting me on my achievement and being a source of my inspiration and the reason of success in my life.

ACRONYMS AND ABBREVIATIONS

CRISP- DM: Cross Industry Standard Process for Data Mining

DM: Data mining

KDD: Knowledge Discovery in Database

RDBMS: Relational Database Management System

ReCAP: Regional Crime Analysis Program

SEMMA: Sample Explore Modify Model Asses

WEKA: Waikato Environment for Knowledge Analysis

TABLE OF CONTENTS

Contents

page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Shire Town Police Office is not using any technology based system to make analysis of criminals' activities to understand trends of previous years' crimes and to identify the prevalent crime patterns occurred. This problem is not impacted only for the Shire Town Police Office but also for the region and the country. This research investigates the potential of data mining tools and techniques in developing models for crime pattern analysis to support the crime detection activities at the Shire Town Police Office-Tigray-Ethiopia. Out of more than 10,000 offenders' record the researcher used only 9967 offenders' recordS and 11 attributes data for this research. Utilizing clustering and classification algorithms, specifically K-means for clustering, J48 Decision Tree and NaïveBayes for classification, the research analyzes real offenders' data collected from the Shire Town police office. The results demonstrate that the J48 Decision Tree model, achieving an accuracy of 97.48% with 119 Number of Leaves and 157 Size of the tree, outperforms other models in detecting crime patterns based on the criteria that classifiers evaluated. Based on the findings of the J48 Decision Tree one sample is listed like: if the occurred crime type is at 2007 e.c, offenders who are grouped in the age group of Age2 (20 up to 32 years old), their Educational status is illiterate, and the crime occurred time is AM, then 109 offenders (97.32% of them) are classified as Male. There are 112 records. From which 3 records (2.67% of them) are incorrectly classified. This study highlights the significance of data mining in transforming raw crime data into actionable insights, thereby facilitating more effective decision-making in law enforcement. The findings emphasize the necessity of implementing modern data mining techniques to improve crime management strategies, ultimately contributing to enhanced public safety in the region.

CHAPTER ONE

1 INTRODUCTION

1.1 Background of the study

According to [7] data mining is one of the powerful tools that have evolved to play a role as an instrument to discover patterns buried in large databases. Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. As researcher [50] pointed out, Data mining tells important things unknown to the user or what is going to happen in the future. There are a number of modeling techniques and the most commonly used data mining techniques are: Decision tree, naïveBayes, Neural Networks, nearest neighbor method and rule induction, etc.

At this generation it is mandatory that any institution has to use technology based data analysis for the success of the aims and objectives of the institute. Due to that reason utilizing data mining techniques can empowers law enforcement to make informed decisions based on empirical evidence rather than intuition, enhancing the efficiency and effectiveness of crime management efforts.

But as the researcher justified by discussing with the domain experts of the Shire Town police office, the institute is not using any technology based system to make analysis of criminals' activities to understand trends of previous year's crimes and to identify the prevalent crime patterns (based on offender's sex, occupation, age, etc.) occurred. Therefore the researcher is motivated to find solutions for the lack of data mining supported crime analysis activities by developing data mining models for crime pattern analysis to identify prevalent crime patterns and other criminal's information by creating awareness on how the domain experts use data mining techniques and algorithms.

As the researcher reviewed different literatures concerning to data mining based method of detecting crime patterns, few researches were done in some parts of Ethiopia on data mining in crime prevention, however, there are no researches done so far in Shire Town. Therefore, this study is launched to identify appropriate tools as well as to develop models that could extract crime patterns from the criminals' database of Shire Town police office

1.2 Statement of the problem

To control social inequity there is always a need for practical crime detection strategies and policies. accepting and processing of criminal records is one method to learn about both crime and individuals who engage in misdeeds so that police can take crime pattern detection measures accordingly [10].Now a day Crime is a complex social phenomenon and its cost is increasing due to a number of societal changes and the like, and hence, law enforcement organizations like that of police need to learn the factors that constitute higher crime trends [52].

However, in the case of Shire Town Police office there are no modern tools and techniques that can support in managing crime records properly and efficiently. Criminals' information is directly recorded on a paper based way of data documentation. As a result almost all the decision-making processes of the office are not supported by tools and techniques that could help on detecting crime patterns linking to the previous years' crime records. Consequently, training programs, resource operation, crime pattern detection and investigation strategies are being pursued on the basis of crime incidents rather than simply identifying crime patterns and trends. Thus, using data mining techniques it is possible to empower law enforcement to make informed decisions based on empirical evidence rather than intuition, enhancing the efficiency and effectiveness of crime management efforts.

Even though, few researches were done in some parts of Ethiopia on data mining in crime prevention, but there are no research done so far in Shire town. Therefore, this study is launched to identify appropriate tools as well as to develop models that could extract crime patterns from the criminal database which supports the decision making process of crime pattern detection.

In addition to that the research area by itself is different. Which means a crime pattern that found on a particular area may not be the same with a crime pattern that found on other area. Therefore making this research will give a benefit for Shire Town police office to detect the pattern of the crime that occur on the wereda and enables them to prevent crime occurrence by creating awareness for the society concerning to the detected crime patterns.

1.3 Research questions

- What are the limitations of the current data management practices at the Shire Town Police Office in supporting effective crime pattern analysis?
- How can the implementation of data mining techniques enhance decision-making processes in crime prevention strategies for the Shire Town Police Office?
- Which data mining algorithm is better on crime pattern analysis for Shire Town Police Office?
- How can analyze crime type in related to age of the offenders of Shire Town Police Office?
- In which sex is the dominant crime occurrence of the Shire Town Police Office?

1.4 Objective of the Study

1.4.1 General Objective

The general objective of this research is to explore the potential of data mining tools and techniques in developing models for crime pattern analysis to support the crime detection activities at the Shire Town Police Office.

1.4.2 Specific objectives

- To conduct related literature review to fill the gap between this study and the review.
- To identify better data mining algorithm on crime pattern analysis for Shire Town Police Office
- To provide actionable recommendations for the Shire Town Police Office on how to integrate data mining findings into their operational practices for enhanced crime prevention.
- To analyze the age group of the offenders in related to the occurred crime type on Shire Town Police Office.
- To create awareness on how to use data mining techniques for Shire Town police office.
- To suggest areas for further research based on the findings of this study.
- To identify limitations of the current data management practices at the Shire Town Police Office in supporting effective crime pattern analysis.
- To identify dominant crime occurrence based on the offenders' sex.

1.5 Significance of the study

This research provides valuable insights into crime patterns, enabling the Shire Town Police Office to implement more effective crime prevention strategies, ultimately contributing to a safer community. By utilizing data mining techniques, this study empowers law enforcement to make informed decisions based on empirical evidence rather than intuition, enhancing the efficiency and effectiveness of crime management efforts.

Additionally this study adds to the body of knowledge in the fields of data mining and criminal justice, providing a framework for future research and practical applications in crime analysis. The developed models can serve as a foundation for other law enforcement agencies in Tigray and beyond, showcasing the potential of data mining techniques in crime detection.

Results from this research can inform policy-makers about the significance of integrating data analytics into law enforcement practices, advocating for technological advancements in policing. The study highlights the importance of training police personnel in data mining techniques, equipping them with the necessary skills to analyze crime data effectively for ongoing crime pattern assessments.

1.6 Scope and limitation of the study

The study is to explore the potential of data mining tools and techniques in developing models for crime pattern analysis to support the crime detection activities at the Shire Town Police Office. And it concentrate on the offenders' data those are collect from Shire Town police office, which are recorded by police officer that are available in the criminal record database starting from 2007 e.c up to 2017 e.c except 2013 e.c & 2014 e.c Different data mining techniques were applied to the problem domain those are J48 decision tree and NaiveBayes classifier algorithms and K-means clustering algorithm were used to build the model.

The main limitation of this study was almost all of the criminal records are found in hard copy formats. To encode the data in an MS-Excel format it took a lot of time. This research didn't include all of the data available in the police office because of time and financial difficulties. In addition to this victims' demographic data are excluded in this research due to lack of fully recorded victims' data on the criminals' database of the Shire Town police office.

1.7 Thesis Organization

This research is organized into five chapters. The first chapter deals with the basic overview including background, statement of the problem, objective, significance, scope, and this part which is thesis organization. The second and third chapters are devoted to data mining, literature review as well as research methodology respectively. Chapter four contains model building and model evaluation of the research. This includes the data sources, data understanding, training, building and validation of the models. Finally, in the fifth chapter conclusions and recommendations are presented

CHAPTER TWO

2 Data mining Technologies and Related review literatures

This chapter focuses on the potential of data mining to discover knowledge from huge database. This comprises what data mining is, the phases on data mining, variety of models, together with the techniques, tasks and various activities involved in it. In addition, in this chapter the researcher presents the definition of crime and criminals, types of crime and crime pattern detection attempts by the law enforcement organizations are also reviewed.

2.1 Overview of Data Mining

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

As [45] mentioned that, gradually new technologies have begun to play a vital role in storage, facilitating analysis and processing of data. Specially, the advent of computer technology in both hardware and software has revolutionized the way in which data are saved, interpreted and managed. These new methods of looking into data as well as the eagerness to learn from data have brought the chances to evolve disciplines like that of data mining. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% [53]. As the volume of data increases, the proportion of information in which people could understand decreases substantially. This reveals that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which results with increasing information gap. Consequently, people begin to realize this bottleneck and to look into possible remedies.

There are several contributing factors to the proliferation of bulky data. This includes, the use of bar code, the automation of many businesses and other transactions, advances in data collection tools ranging from scanned image documents and image platforms to satellite remote sensing systems [27] As the size of data grows in organizations, there exists a need for new automated methods that can enable them to process and convert the accumulated data into useful information and knowledge. The tremendous amount of data collected and stored in large databases is beyond human capability for comprehension of patterns inside them without using powerful analysis tools. To get benefits from integrated and historical data, there should be a way to identify relevant and useful information [27]

In addition to the above, technological developments that aid to collect and store vast quantities of data have enabled organizations to capture and accumulate huge amount of data in their databases, within which, large amount of valuable information is buried [25] As the volume of data increases, the proportion of information in which people could understand decreases substantially. Consequently, people begin to realize this bottleneck and to look into possible remedies like data mining.

Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data. Machine learning is more properly described as the hybrid of statistics and artificial [25]. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced artificial intelligence heuristics and algorithms to achieve its goals [32] This depicts that the application of machine learning in the study of large volume of data is a radical shift not only from statistics but also from artificial intelligence via merging both fields.

Data mining is the extraction of hidden predictive information from a large database. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Storing the enormous amount of raw data into database will not be able to provide the meaningful information, rather those data should be analyzed and the hidden knowledge must be extracted by the use of data mining which is a main phase of the knowledge discovery process. Data mining is a strong tool because it can provide you with relevant information that you can use to your own advantage. When you have the right data, then you will need to do is apply it in the right manner, and you will be able to get beneficial result. [30]

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

2.2 What is Data Mining?

As [53] mentioned that the convergence of computing and communication has resulted in a society that highly relies on information. Technological developments that help to collect and store vast quantities of data have enabled organizations to capture and cumulate huge amount of

[50] Noted that, data mining is the extraction of hidden patterns and useful trends from large database. It is a robust technology with substantial potential to help organizations concentrate on the most valuable information in their database. Data mining is valuable to discover implicit, potentially useful information from enormous data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns. The opportunity for the application of data mining has increased significantly as databases grew extremely and new machine with searching capabilities evolved [53]

According to [7], Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining usually makes sense when there is large amount of data. For this reason most of the algorithms developed for data mining purpose requires large volume of data so as to build and train models that are responsible for different tasks of data mining such as classification, clustering, prediction, association and the like. The need for bulky data can be explained by a couple of reasons. Primarily, in the case of small databases, it is feasible to capture appealing trends and relationships by introducing traditional tools such as spreadsheets and database query. The second reason is that most data mining tools and algorithms demand large amount of training data in order to generate unbiased models. The rationale is simple and straightforward, small training data results in unreliable generalizations based on chance patterns.

Finding useful patterns in data has been given a variety of names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing [21]. Many People treat data mining as a synonym for another popularly used term Knowledge Discovery and Database. Others view data mining as an essential step in the process of KDD, the term data mining is becoming more popular than the term KDD [28]. According to [21], Knowledge Discovery in Databases is the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." On the other hand, those authors have defined data mining as a single step in the process of Knowledge Discovery that involves the application of appropriate algorithm so as to discover meaningful trends from the database under investigation.

Nevertheless, many authors believe that the term data mining has become more popular in industries, in media and the database researches as a synonym for knowledge discovery and hence the two terms are used interchangeably. This view is also applicable throughout this study. Some of the most frequently used terms in the process of knowledge discovery are instance and attribute. Instance refers to the inputs or example to the data mining process that is going to be investigated based on its features, whereas an attribute refers to the features or characteristics of an input, now to be more specific, an instance [53]. Whenever the data set is presented in a tabular form; usually the instances are the rows of the table, and the attributes are the columns.

2.3 Data mining and Knowledge Discovery

The amount of data stored in databases increases exponentially with time. As a result, the manual analysis of this data is complex and prone to errors. When the amount of data to be analyzed exploded in the mid1990s, knowledge discovery emerged as an important logical tool.

According to [21] mentioned that, the process of extracting suitable knowledge from volumes of data is known as knowledge discovery in databases. KDD is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex datasets. Data mining is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction Knowledge discovery's major objective is to identify effective, novel, potentially valuable, and clear patterns of data.

Knowledge discovery is supported by three tools: massive data collection, powerful multi-processor computers, and data mining. There is some confusion about the terms Data Mining and KDD. Often these two terms are used interchangeably. Many authors agree in that Knowledge Discovery in Data base refers to the overall process of discovering useful knowledge from data and data mining refers to a specific step in this process. KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data whereas Data mining is the application of specific algorithms for extracting patterns from data [21].

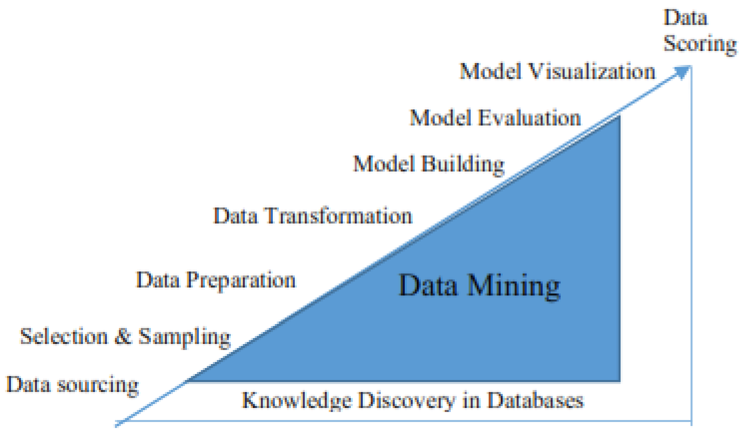


Figure 2.1 The relationship between data mining and knowledge discovery

2.4 Common tasks of data mining

[26] pointed out; Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. The tasks of data mining can be either Predictive or Descriptive in nature. Practically, data mining can accomplish common tasks like; classification, clustering, prediction, association, and description. However, many of the practical problems of scientific, economic and business interests can be mapped into one of these common tasks [7]. All data mining techniques are classified as either of the two models and are not equally applicable to all the above-mentioned tasks. Based on the nature of the problem under consideration and its nearness to the main divisions of data mining tasks, researchers need to choose the suitable techniques among the numerous data mining techniques.

In general, data mining tasks can be classified into two categories: predictive modeling and descriptive modeling.

2.4.1 Predictive modeling

Predictive modeling involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Prediction is arguably the strongest goal of data mining. The aim is to build a model that can permit the value of one variable to be predicted from the known values of other variables. Classification and regression are two good examples of data analysis that can be used to extract models describing important data classes or to predict future data trends [2].

2.4.1.1 Classification

[53] Stated that Classification comprises examining the features of unseen instances and assigning it into one of the predefined classes. In fact, in the case of data mining the objects that are going to be examined comes from a database. The task of classification incorporates updating each record by filling in a field with a class code of some sort. This task starts with instances often called a training

set which consists of predefined classes and are used to train as well as build a data mining model so that the model can be applied to classify unseen objects.

[7] Mentioned that Classification is one of the most common data mining tasks, which is also pervasive in human life. Human beings usually classify or categorize in order to understand and communicate about the world. For any object or instance, classes are predefined according to the value of a specific field. A variety of techniques have been used to obtain good classifiers.

Decision Tree

According [26], a decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions Decision trees can easily be converted to classification rules. Decision trees are a way of representing a series of rules that lead to a class or value. Depending on the algorithm, each node may have two or more branches. For example, CART generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multi way tree. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree can be assigned a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached [26].

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm, the critical distribution of the data is easily understandable [39].

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. It generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy [39].

Trees and Rules

Naive Bayes

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong Naive independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. In simple terms, a Naïve Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature [45].

The Naïve Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event happening given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows. Prob (B/ A) = Prob (A /B) Prob (B)/Prob (A) In possibility theory Bayes theorem shows how one conditional probability such as the Probability of a hypothesis given observed suggestion depends on its inverse in this case, the Probability of that indication given the hypothesis. In more technical terms, the theorem expresses the posterior probability i.e. after evidence B is observed of a hypothesis A in terms of the prior probabilities of A and B, and the probability of B given A. It implies that evidence has a stronger confirming effect if it was more unlikely before being observed.[45]

2.4.2 Descriptive Modeling

2.4.2.1 Clustering

Clustering is a process of grouping a set of physical or abstract objects into classes of similar objects [26]. A cluster is therefore a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis is also used to form descriptive statistics to ascertain whether or not the data consists of set distinct subgroups, each group representing objects with substantially different properties. In this technique the modeling process is unsupervised that is no prior (pre-defined) knowledge is available to exactly guide the clustering process. The clustering algorithm clusters the database autonomously. Clustering plays an outstanding role in data mining applications.

According to [24], a cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups.

Clustering techniques are very useful in data mining because of the speed, reliability, and consistency with which they can organize a huge amount of data into distinct groupings. Despite the availability of a vast collection of clustering algorithms in the literature they are based on two popular approaches: hierarchical clustering and partitioning clustering. The former, which is the most frequently used technique, organizes data in a nested sequence of groups that can be displayed in a tree like structure. Partitioning clustering builds many partitions and then evaluates them by some criteria. K-means clustering and expectancy maximization are the two methods of partitioning clustering. [24]

Depending on the clustering technique, clusters can be expressed in different ways:

- Identified clusters maybe limited any instance belongs to only one cluster.
- They may be overlying an instance may fit to several clusters.
- They may be probabilistic whereby an instance belongs to each cluster with a definite probability.
- Clusters having crude division of instances at highest level of hierarchy which is then refined to sub clusters at lower levels.

2.4.2.2 K-Means clustering algorithm

The K-means algorithm, one of the clustering algorithms proposed for this study, is based on a very simple idea. Given a set of initial clusters K (k-stands for numbers of clusters), assign each point to one of them and then each cluster center is replaced by the mean point on the respective cluster. The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured regarding the mean value of the objects in a cluster, which can be viewed as the cluster's Centroid or center of gravity. The *k*-means algorithm works as follows. First, it randomly selects *k* of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. The *k*-means algorithm is simple, easily understandable and reasonably scalable, and can be easily modified to deal with streaming data. However, one of its drawbacks is the requirement for the number of clusters, k, to be specified before the algorithm is applied. [26].

2.5 Data Mining Models

Data mining modeling cycle involves a number of stages. Initially, it is important to have a clear understanding of the business area in order to understand the operational analytical processes the problems that are to be emphasized, the opportunities that may be realized and to assess the availability of data. Exploring and preparing the data is a crucial stage in the cycle. In preparing data new fields may be derived from one or more existing fields, missing and boundary values identified and processed. in addition, relationships between fields in each column and records in the rows of the entire data is identified to some of the pre-processing tasks that assist in cleaning and made appropriate the data prior to the mining process. Once data has been prepared for mining, the modeling stage can begin. Choosing and developing models involve the results of domain knowledge which are validated against expected results [16].

As [35] mentioned that Model building is a key objective of data mining and data analysis applications. In the past, such applications required only a few models built by a single data analyst. As more data has been collected and real world problems have become more complex, it has become increasingly difficult for that data analyst to build all the required models and manage them manually. Building a system to help data analysts construct and manage large collections of models is an urgent issue.

2.5.1 The Six Step CRISP-DM Model

A. Business Understanding

This initial phase focuses on understanding the objectives and necessities from a business point of view, and then changing this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. At this stage the researcher tried to identify the objective of Shire town police office which prevention or reduction and detection of crimes as well as avoiding opportunistic situations that criminal could get. One such objective could be to decrease crimes over the next year. Assess whether these problems need data mining and are solvable with the help of data mining modeling process.

B. Data Understanding

The data understanding phase beginning with an initial data collection and incomes with activities in order to get aware with the data, to classify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. Prior to any analysis the purpose must be defined and for any such analytical process to take place it is more cost effective to utilize currently stored data. While dealing with the data mining modeling environment like understanding the core business and understanding the data is critically important.

C. Data Preparation

During this stage in the cycle a variety of encoding techniques may be utilized to provide additional fields for analysis and enable fuzzy concepts. An example of missing data could be that some criminal description fields may not contain the criminal's religion. This could either mean that the police officer did not ask the questioner the information was not entered onto the crime report. Experience would propose that the latter option would be correct as it is common practice to take a full written statement at the time of the offence being reported and to transfer brief information onto the crime report. Within policing terms, text-based memo fields are rich in information, however, the text is notably more noisy than other text sources such as news reports etc., containing many spelling errors, typos and grammatical errors [16]. However, in the absence of fully structured data, as in the Shire town Police crime reporting system (hand written memos), the text need to be written and parsed to extract information that can be employed in the analysis of crime.

D. Data Mining

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Here, modeling technique selection, test design generation to validate the model and test its quality, building model and model assessment through interpretation, evaluation, comparison and ranking of models according to the evaluation criteria from a data mining point of view are main tasks of this phase.

E. Evaluation

At this stage in the data mining task the researcher has built a model (or models) that appear to have high quality, from a data analysis point of view. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

A key objective is to determine if there is some important business issue that has not been sufficiently considered. The resultant model may be validated in terms of its clarity, parsimony, generality and testability to assess the degree to which it meets the required objectives.

In case of domain experts evaluation they are crucial to evaluate because they are close to the business problem and they do have expected pattern from which their evaluation is based. This is better to be supplemented with the different evaluation tools. At the end of this phase, a decision on the use of the data mining results should be reached. This is performed based on the domain experts‟ advice and the parameters set and the researchers‟ personal judgment.

In data mining evaluation serves two purposes. First, it helps to imagine how well the final model will work in the future or even whether it should be used at all. Second, as an integral part of many learning methods, it helps to explore the model that best represents the training data.
The model is evaluated together with the domain experts regarding its interestingness and novelty. The different clustering models that are developed in this research are evaluated based on the within cluster sum of squared error values, number of iteration, time taken to build the model that satisfy the threshold, and experts' judgment.
Likewise, classification models that are developed in the research are evaluated using a test dataset based on their classification accuracy.

F. Deployment

Building of the model is generally not the end of the goal. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

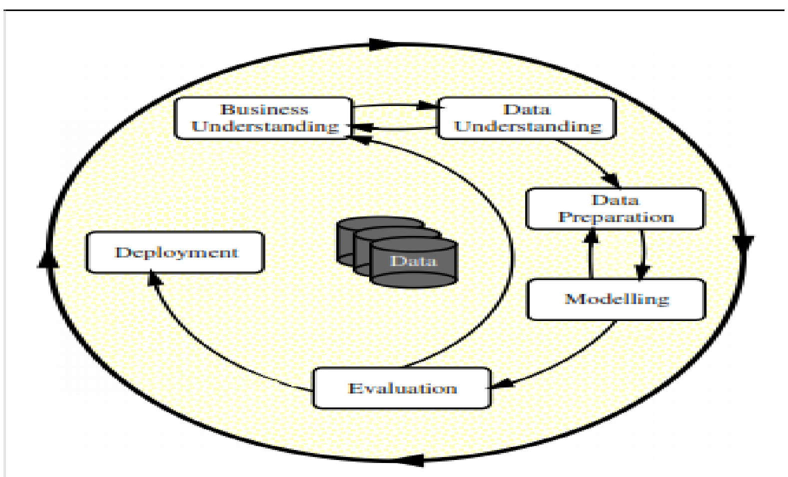


Figure 2.2: The CRISP-DM process model

2.5.2 The Six Step Cios Model

This model was developed by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps [17].

1. Business Understanding

In this step one works closely with domain experts to define the problem and determine the research goals, identifies key people, and learns about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the data mining goals, and include initial selection of data mining tools.

2. Understanding the data

This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, verification of the usefulness of the data in respect to the data mining goals .Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

3. Preparation of the data

This is the key step upon which the success of the entire knowledge discovery process depends it usually consumes about half of the entire research effort. In this step, which data will be used as input for data mining tools of step 4, is decided. It may involve sampling of data, running, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms to reduce dimensionality and by derivation of new attributes say by discretization, and by summarization of data.

The outcome would be new data records, meeting specific input requirements for the planned to be used data mining tools.

4. Data mining

This is another key step in the knowledge discovery process. Although it is the data mining tools that determine new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, clustering, preprocessing techniques, machine learning etc. This step involves the use of several data mining tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen data mining tools; the generated data model is verified by using testing process.

5. Evaluation of the discovered knowledge

This step includes understanding the outcome, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire DM process may be revisited to identify which alternative actions could have been taken to improve the outcome.

6. Using the discovered knowledge

This step is entirely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

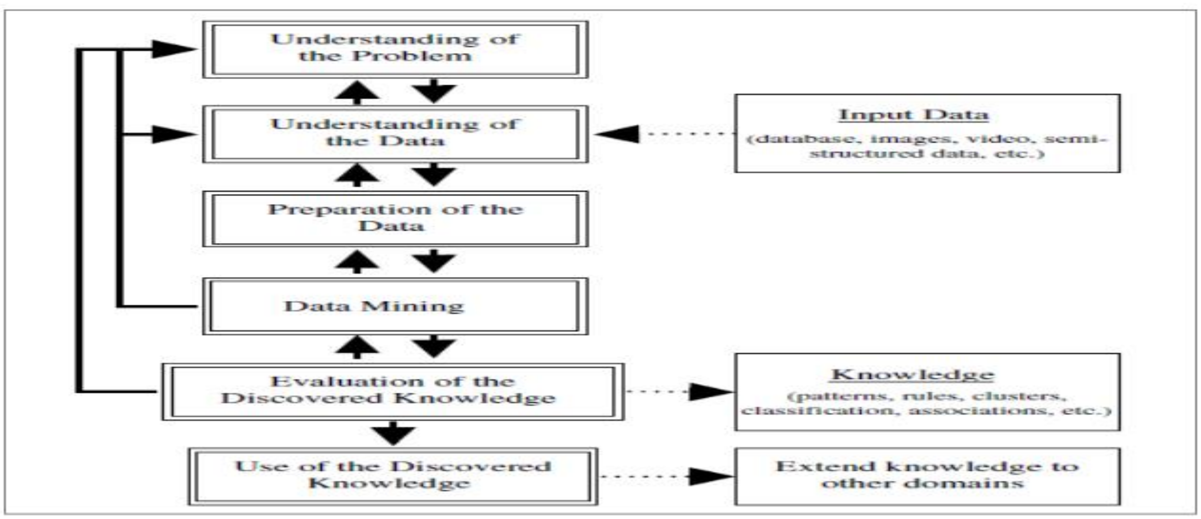


Figure 2.3: The six step Cios et al model

2.5.3 The KDD Process Model

Knowledge Discovery (KDD), process is the process of automatically searching large volumes of data for hidden, interesting, unknown and potentially useful patterns. It is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions [13]. There are five steps in the KDD process.

1. Data Selection: This stage consists on creating a target dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed. The data relevant to the analysis is decided on and retrieved from the data collection.

2. Data Pre-Processing: This step consists on the target data cleaning and pre-processing in order to obtain consistent data.

3. Data Transformation: in this phase the selected data is transformed into forms appropriate for the mining procedure. This stage is consists on the transformation of the data using dimensionality reduction or transformation methods.

4. Data Mining: This is the most important step in which smart techniques are applied to take out potentially useful patterns. It consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective.

5. Interpretation/ Evaluation: This stage focused on the interpretation and evaluation of the data mined patterns.

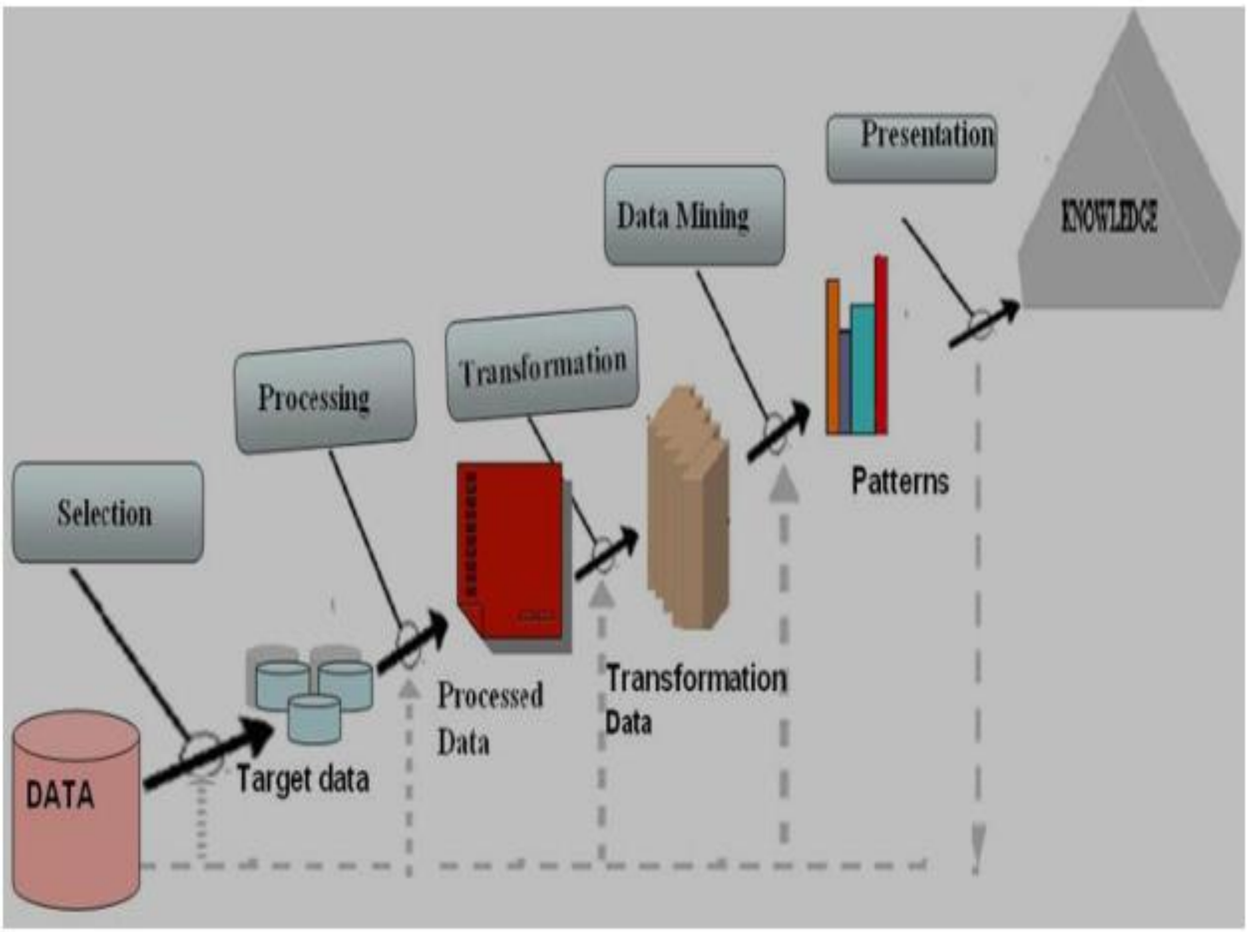


Figure 2.4: The KDD process

2.6 Scope of Data Mining Applications

The applications of data mining can be generic or domain specific. The generic application is required to be an intelligent system that by its own can take certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take on its own these decisions but are guided by users for selection of data, selection of data mining method and for the interpretation of the results.

According to the [6], the multi agent based data mining application has capability of automatic selection of data mining technique to be applied. The Multi Agent System used at different stages: First, at the stage of concept hierarchy definition then at the result level to present the best adapted decision to the user. This decision is stored in knowledge Base to be used in a later decision making. Multi Agent System Tool used for generic data mining system development [9] uses different agents to perform different tasks. Generic systems are required to integrate as many learning algorithms as possible and decide the most appropriate algorithm to use.

2.7 The Application of data mining

Various fields use data mining technologies because of fast access of data and valuable information from vast amount of data. Data mining technologies have been applied successfully in many areas like marketing, telecommunication, fraud detection, and finance, medical and so on [44].

[29] Stated that, Applications of data mining are now increasingly visible in day-to-day life. Wide ranges of companies are employing data mining techniques as an important tool to improve the performance of their everyday activities. According to Gartner Group Advanced Technology Research, cited in [50], data mining and artificial intelligence are at the top of five key technologies and "will clearly have a major impact across a wide range of industries within the next few years."

According to [50], Data mining technology is applicable to any problem if there is massive data and well-defined understanding of the problem at hand. One of the successful data mining applications to practical problems cited frequently is the Soybean Classification, which identifies the type of disease based on some diagnosis [53]. The Labor Negotiation of the Canadian contract negotiations in 1987-88 is another example of the early success of data mining application [53].

Government agencies and private companies are using data mining to enable financial management through analytical fraud detection [23]. There are also several successful systems developed and widely used to support the everyday activities of business organizations). IBM's intelligent Miner is another system which is being applied by retailers to know customers purchasing tendency and product popularity [14].

According to research firm International Data Corporation, cited in [41], worldwide revenue of analytic application solutions, which uses data mining for analysis, was expected to grow at a rate of 28% per annum, from US$2 billion in 1999 to more than US$6 billion in 2004.

Particularly, International Data Corp goes to say that Customer Relations Management applications would be one of the fastest growing data mining application areas.

An upcoming application of data mining is undertaken by higher educational institutions due to the gradual hike in the amount of data along the years. Data mining in this discipline is used to understand student behavior, such as the trends which would indicate student transfer, credit hours trend as well as the skill sets of various clusters of students and their redundant characteristics [28].

2.8 Crime and Criminal

According to [49], Crime is defined as an act or oversight of an act, which is punishable by criminal law. Criminal law, on the other hand, refers to a body of specific rules regarding human conduct, which have been clearly stated by political authority. However, an act that is considered as a crime in one place and time may not hold true in another place or time.

Crime is the act that harms the public, increases the violence, demolishes the assets and denies the respect to people. Distribution of Crime is not even across the globe. The huge available data and hands-on experience helps for investigation .a criminal is an individual person who has violated the legally forbidden act. In fact, there are some factors that have to be taken into account to convict whether a person should be considered as a criminal or not. Among these, an individual should be of competent age in light with the law of the land; and there must be a well-predefined punishment for the act committed [4].

[3] Defined Crime as a comprehensive concept that is defined in both legal and non-legal sense from the legal point of view crime is the breaking or breaching of the criminal law that governs a particular geographical area (jurisdiction) aimed at protecting the lives, property and the rights of citizens of belonging to that jurisdiction. Crime is an offence against a person (for example murder, and sexual assault), or his/her property (for example, theft and property damage) or the State regulation (for example traffic violations) [3], In non-legal terms crime is a set of acts that violate socially accepted rules of human ethical or moral behavior [3], 2004); for example, acting against a custom in some society. There are several causes for the growing rate of crime in a specific place or country. These include unemployment, economic backwardness, over population, illiteracy and inadequate equipment of the police force.

The crime control and criminal suppression are considered as two significant rudiments by the law enforcement in the data analysis. Former uses information to control and later uses history stored in mining to seize a scandalous person. Machine learning methods make the greater impact in the effective and efficient analysis. There are many approaches to crime analysis. The description of the approach followed is the one proposed by G.K. Gupta in the book Introduction to Data Mining with Case Studies. He proposed use of cluster analysis and where the aim is to find groups which are different from each other. He proposed a number of crime data mining techniques which are, cluster analysis and classification prediction and outlier analysis [22].

Crime has adversely affected the societies of both civilized as well as developing countries by declining the quality of life, endangering human right and fundamental freedom and posing a serious challenge to the community. Although the level and intensity of the problem might vary from nation to nation no country has remained unaffected. Crime is an illegal activity which causes physical or moral injuries on others. According to the Ethiopian criminal code of 2006, Article 23, sub article 1, crime is an act which is prohibited and made punishable by law (Ethiopia criminal code). Crime is the omission of a duty that makes the offender liable to punishment by law or a behavior that is prohibited, as well as behavior or an act that is required by law [47].

2.9 Types of crime

Crime occurs in a variety of forms which police informally categorizes as being either major or volume. Major crime consists of the high profile crimes such as murder, armed robbery and rape. These crimes can either be one-offs or serial. In the case of serial crimes it is relatively easy to link crimes together due to clear similarities in terms of descriptions of offenders. This linking is possible due to the comparatively low volume of such crimes. Major crimes usually have a team of detectives allocated to conduct the investigation. In contrast volume crimes such as burglary and shoplifting are far more prevalent. They are usually serial in nature as offenders go on to commit many such crimes. Property theft crimes, such as domestic burglary offences, committed by different individuals are highly similar and it is rare to have a description of the offenders [2].

Generally,murder,aggravated assault ,forcible sex offenses ,nonforcible sex offenses ,kidnapping/abduction ,simple assault ,intimidation ,arson ,bribery ,burglary ,counterfeiting/forgery ,embezzlement , extortion ,fraud ,larceny ,robbery ,stolen property, drugsabuse ,gambling ,child pornography ,prostitution ,weapon law violations ,curfew ,harassment and domestic violations are the main types of crimes.

2.10 Crime Pattern detection

Data mining and machine learning can be a tremendous tool for crime pattern detection, and for predictive policing in general. The police can try to stop the crimes if the crime patterns are identified and used for the planning. Without such tools, it could take weeks or years of sifting through a database to discover a pattern, or it might be missed altogether. Identifying the trends and patterns and making tactical analysis is the duty of criminal investigators across the world [31].

[49], he found that the causes for the growing rate of crime include unemployment, economic backwardness, over population, education illiteracy and inadequate equipment of the police force.

The crime patterns will act as the resource for the detective agencies to take decisions. The profiling data base in interaction makes the matching crime with criminal is achieved. Crime reports comprises of type of crime, date/time, location, suspect, victim and witness etc. were considered along with Clustering. The crime clusters are plotted with legend for significant attributes for that crime pattern detection. The multivariate time series clustering with dynamic time wrapping method is applied on the districts crime dataset. The multivariate time series data and the weight matrix based on domain are taken into consideration [22].

According to [4], understanding the attributes of criminals will be supportive to design and implement prudent crime prevention approaches. The rise in crime both national and transnational is generally supposed as the result of interplay between socioeconomic changes. The situations surrounding the individual criminals such as his character, physical characteristics intelligence, family, environmental surrounding such as peer groups, neighbors etc. have been focus of the study of crime.

The analysis of crime patterns and trends is very significant for police officeers and analysts can learn from historical crime patterns and enhance crime resolution rate. It also helps to prevent future incidents by putting in place preventive mechanisms based on observed patterns. Another possible advantage is, it can reduce the training time for officers assigned to a new location and having no prior knowledge of site-specific patterns to assist them in investigations. In light with the crime patterns extracted from previous records, police can deploy scarce resources to the right place at the right time [36].

Intelligence agencies are actively collecting and analyzing information to investigate criminal's activities. Local law enforcement agencies have also become more alert to criminal activities in their own jurisdictions. When the local criminals are identified properly and restricted from their crimes, then it is possible to considerably reduce the crime rate. Criminals often develop networks in which they form groups or teams to carry out various illegal activities. Data mining task consisted of identifying subgroups and key members in such networks and then studying interaction patterns to develop effective strategies for disrupting the networks. Data is used with a concept to extract criminal relations from the incident summaries and create a likely network of suspects [11].

These patterns will be applied to some data mining algorithms such as classification mining and clustering to classify crime records on the basis of the values of attributes crime. Applying such algorithms will illustrate the overall results of using both algorithms to perform better results rather in association rule mining or in clustering [54].

Without a suspected crime pattern, the detective is less likely to build the complete picture from bits of information from different crime incidents. Automated detection of crime patterns, allows the detectives to focus on crime sprees first and solving one of these crimes results in solving the whole “spree” or in some cases if the groups of incidents are suspected to be one spree, the complete evidence can be built from the different bits of information from each of the crime incidents. For instance, one crime site reveals that suspect has black hair, the next incident/witness reveals that suspect is middle aged and third one reveals there is tattoo on left arm, all together it will give a much more complete picture than any one of those alone [46].

2.11 Data mining and crime detection

According to Brown (2003), Data mining can greatly improve crime analysis and aid in reducing and preventing crime. Most law enforcement agencies today are faced with large volume of data that must be processed and transformed into useful information. “No field is in greater need of data mining technology than law enforcement”. One potential area of application is spatial data mining tools which provides law enforcement agencies with significant capabilities to learn crime trends on where, how and why crimes are committed [51].

[2], in developed countries a majority of crime prevention forces use different types of relational database management systems (RDBMS) for recording and subsequent analysis of crime. Standard or interactive queries are written to produce patterns of crime, offending and various statistics, but it is a common phenomenon in the developing countries to find mainly manual criminal record books used alongside for crime incidence location.

[16], Entity extraction used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports. This is a technique used to extract objects and their attributes with the help of special extraction tools from police reports or narratives. Thus, officers or

analysts can create structured data from free texts. Other data mining techniques such as classification, prediction, cluster analysis, and outlier analysis identify patterns in structured data.

Clustering techniques used to automatically associate different objects such as persons, organizations, vehicles etc. in crime records. Deviation detection and outlier analysis are applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. Classification is used to detect email spamming and find authors who send out unsolicited emails. String comparator is used to detect deceptive information in criminal record. Social network analysis used to analyze criminals‟ roles and associations among entities in a criminal network. [26]

[20] Stated that, Selection of data and methods for data mining is an important task in the process and needs the knowledge of the domain. Different methods of data mining are used to extract patterns and knowledge from variety of databases. However no system is found to be completely generic. For this the domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generated.

2.12 Related Works

The researcher reviewed different literatures related to crime based researches mainly focusing on the four summarized below and others cited on the reference. Therefore the related works listed below are tried to view the data mining techniques they used, the performance they found and the research gap they put as a recommendation as follows.

Leul Woldu (2003) presents application of data mining in crime prevention and he was used Decision Tree and Neural Network Model Building for performance comparison [33].

Letezgi Hagos (2011) was done mining crime data for effective resource allocation and crime prevention and she was used classification, association rule and clustering Model Building for performance comparison [55].

Mugdha Sharma [2014] et al. for detecting suspicious criminal activity e-mails proposed an Enhanced decision Tree Algorithm. For generating better and faster Decision Tree, they applied enhanced feature selection method and attribute importance factor to improved ID3 Algorithm. They are trying to classify emails in various criminal activities [56].

Hailemariam Negussie (2015) was done his research on Crime forecasting by using data mining techniques and he was used hybrid processing model such as classification model (decision tree J48 algorithm and PART rule induction algorithm) for performance comparison [57].

Table 2.1 Summary of reviewed related researches

Performed	Techniques	Tasks	Research Gaps
[33]	Decision Tree and Neural Network Model Building	exploring the applicability of data mining techniques in the process of crime prevention	physical descriptions of offenders, geographic information of crimes were not in incorporated
[55]	classification, association rule and clustering Model Building	identifying which children are exposed to which crimes	Attributes such as sex, habit, education, etc. of the offender were not incorporated.
[56]	improved ID3 Algorithm	Classifying emails with respect to criminal activities	No collections of data, crime prediction were made.
[57]	decision tree J48 algorithm and PART rule induction algorithm	Identifying the most frequent crime types	Still performance improvement is expected.

Based on the above summary of literature reviewed, the researcher motivated to fill the gaps such as, performance improvement of the data mining techniques by applying more training and testing models of the data mining techniques. In addition to that the research area by itself is different. Which means a crime pattern that found on a particular area may not be the same with a crime pattern that found on other area. therefore making this research will give a benefit for Shire town police office to detect the pattern of the crime that occur on the town and enables them to prevent crime occurrence by creating awareness of the criminals concerning to the detected pattern

CHAPTER THREE

3 RESEARCH METHODOLOGY

This study followed different methods in order to develop the good crime pattern detection model.

3.1 Research design

On this research the researcher used the WEKA software due to its ability to provide graphical user interface (GUI) and familiarity of the researcher with the software.

In this study the clustering K-means and classifiers like, J48 Decision Tree and NaïveBayes mining algorithms are used in the experiment.

Reasons that the researcher used the above data mining tools are:

- **K-means algorithm:** It needs no class label, especially crime records lack predefined categories like: discovering unknown crime trends.
- **J48 Decision Tree**: According to [39] decision tree offers many benefits in data mining: for instance it is self-explanatory and easy to follow when compacted, it can be able to handle a variety of input data: nominal, numeric and textual, it also allow to process datasets that may have errors or missing values, it has high predictive performance for a relatively small computational effort,
- **NaïveBayes:** It is fast and simple algorithm for classification of crime datasets.

For this study, Cios et al mode is applied. The reason that the researcher prefers the Cios et al. model is it draws from both academic and industrial models and emphasizes iterative aspects; the Cios et al. model has also an advantage to identify and describe several explicit feedback loops. [17]

Generally, in Cios et al model the life cycle of a data mining process consists of six phases. For the research area a brief outline of the phases is discussed below:

3.1.1 Problem Understanding

At this stage the researcher attempts to understand the objectives of Shire Town police office by discussing with the domain experts like: the police personnel, record officer, and reading the annual plan of the Shire Town police office to understand their needs and challenges in crime detection and prevention. After a detailed discussion between the researcher and the domain experts it is justified that Shire Town police office was not used any technological system to make analysis of criminal activities to understand trends of crime and to identify the prevalent crime type occurred. Therefore this research is aimed to find solutions for the lack of data mining supported crime analysis activities by developing data mining models for crime pattern analysis to identify prevalent crime types and other criminal's information. Finally this research is used the classifiers like: J48 decision tree and NaiveBayes and K-means algorithm for clustering the dataset collected from the Shire Town police office.

3.1.2 Understanding the data

This phase focuses on collecting and exploring the data relevant to the study. The researcher gathers historical crime data from the Shire Town Police Office, covering records from the 5 consecutive years from 2007 e.c. up to 2017 e.c. except 2013 e.c & 2014 e.c. and the data for this study was generated from criminal database of Shire Town Police office contains more than 10,000 records manually.

Finally a data containing 9967 records were taken for this study. The data are classified by the year at which the crime was committed or reported.

Hence, the researcher first encoded all the data changed from paper to an MS-Excel format by changing the language from Tigrigna to English. After the data was encoded, the whole dataset is put in one file having many records and converted the MS-Excel in to .CSV file; lastly the .CSV file converted to compatible file format with the data mining which is the ARFF file format. Each record corresponds to most relevant information of one offender.

Verification of the usefulness of the data in respect to the data mining goals like: checking missing values. Redundancy etc. was applied. The data contains information about offenders (Crime type, Time, Religion, kebele, Occupation, Age, Sex, Educational status, Marital status, Year occurred and Nationality).

Table 3.1 description of Data Sources and Number of records

Source of Data Center	Data Coverage In Year	Total number of crime record	Number of record ,which got final decision from court	Number of Attribute	Data Type
Shire Town police office	From 2007 e.c. Up to 2017 e.c. except 2013 e.c. & 2014 e.c	More than 10,000 records	9967 records	11	Nominal And Numerical

3.1.3 Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. After the data are collected, this includes data selection, handling noisy data, accounting for missing data fields, coding text valued attributes and preparing the processed data in a file format acceptable to the weka software tool.

Data mining tools require suitable data that fulfills the necessary quality to be supported by the selected algorithm. Some algorithms accept only numeric input while others only categorical data or both but not missing attribute values.

This is why the available data by itself does not fulfill everything for developing the model or performing the data mining task rather it needs more processing. These can be summarized as preprocessing, data transformation and data selection.

Data Cleaning:

Almost all of the data preprocessing techniques are done using MS-Excel built-in functions like search and replace, filtering, and auto fill mechanisms.

- ✓ **Outliers:** are misleading data that do not fit to most of the data/facts in the entire data. Since most of the data have nominal value this is not a problem for the data used in this research.

- ✓ **Noisy dat**a: attribute values that might be invalid or incorrect. E.g. typographical errors it means like spelling errors. In this research no noisy data were found.
- ✓ **Missing data:** attribute values might be absent. One can use different strategies such as removing, replacing with estimated values or treating as they are. Removing can be either the entire row that contains missing value or the entire column, when it contains more than 30% missing values. [55] Accordingly, all the 11 attributes of this research satisfy this criterion.

- ✓ **Inconsistent data:** The values of the attribute **Occupation** contains words such as "ስራሕ ዘይብሉ", "መንግስታዊ" and "ናይዉልቀ" similar like that of '**Unemployed**', '**Governmental**' and '**Private**' which talks about the same idea but expressed using two different words even with in the same year of report. Which may contains the problem of lack of standardization or naming conventions.

Data Transformation:

The data needs to transform to improve the effectiveness of pattern discovery. Converting one type of data to another such as numeric ones into categorical or deriving new attributes was reduced the data variables. For instance, age of the offenders was numeric data but for simplicity of grouping it is changed to nominal data (Age=14-19 years old is assigned to Age1, Age =20-32 is assigned to Age2 and Age =more than 33-67 is assigned to Age3)

Attribute selection:

In order to select the best attributes from the initial collected dataset, the researcher evaluates the information content of the attributes using the select attribute techniques of WEKA with Gain Ratio Attribute evaluator and Ranker search method. So, there were variables to be discarded. However, it does not mean that these attributes have no importance rather these variables were believed to provide very little useful information for the problem at hand and since clustering is unsupervised learning they may reduce the accuracy of the clustering algorithm. Therefore, the original table was contained 11 attributes. From this a total of 11 attributes were selected for the research based on their relevance and pre-processing activities. Table 2 shows the final attributes used for model building and their description.

Table 3.2 the final selected attributes and values with their descriptions

NO	ATTRIBUTES	POSSIBLE VALUE	DESCRIPTION
1.	Offenders Crime Type	Theft, Trying to kill/ Murder, Beating/Causing Bodily Harm, Conflict, Destruction of Property, Illegal Driving License, Illegal Migration, illegal weapon, Intimidation/Blasphemy, Killing, preparing illegal document, Suicide, Traffic Accident, Marriage under age, Violation Court Order, Rape	List of crimes
2.	Time	PM AND AM	Time of crime done
3.	Religion	Christian, Muslim and Protestant	Religion of the criminal
4.	Kebele	Adi Ketibay, Dedebit, Hibret, Lekatit and Sihul	Kebele of the offender
5.	Occupation	Governmental, private and unemployed	Occupation of the offender
6.	Age	Age1 = (14 year -19 year) Age2 = (20 year -32 year) Age3= (>=33 year)	Age of the offender

7.	Sex	Male and Female	Sex of the offender
8.	Educational level	College, illiterate, primary, secondary, TTI and University	Education level of offender
9.	Marital status	Married or not married	Marital status of the offender
10.	**Year Occurred**	2007 e.c , 2008 e.c, 2009 e.c, 2010 e.c, 2011 e.c, 2012 e.c, 2015 e.c, 2016 e.c and 2017 e.c	Year of the crime occurred
11.	**Nationality**	Ethiopia	Nation of the offender

```
Attribute selection output
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 Offenders Crime Type):
        Information Gain Ranking Filter

Ranked attributes:
 0.2059    5 Occupation
 0.1882    8 Educational Level
 0.1333   10 Year Occurred
 0.0488    6 Age
 0.0386    4 Kebele
 0.0314    7 Sex
 0.0215    9 Marital Status
 0.0151    2 Time
 0.0132    3 Religion
 0        11 Nationality

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

Figure 3.1 Rank of attribute selection using Gain ratio feature evaluator method

CHAPTER FOUR

4 MODEL BUILDING AND MODEL EVALUATION

4.1 Data mining

At this stage, suitable techniques of data mining were applied to the data set available. Characteristically, there are several data mining techniques for the same problem. Clustering and classification Data mining techniques were applied to the dataset available. After the data was cleaned and prepared, it was analyzed using a data mining tool.

There are varieties of tools available for data mining such as the knowledge Studio, WEKA, SPSS, STATA and others. With those tools, Weka is selected for its familiarity with the researcher and its being open source and availability and also the whole suite of Weka is written in java, so it can be run on any platform. Additionally, the package has three different interfaces: a command line interface, an Explorer GUI interface which allows for 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. Although the choice of data mining techniques for classification tasks seems to be strongly dependent on the application, As it is indicated previously, for the purpose of this research work the researcher experiments the potential applicability of data mining technology in developing a model that detect pattern of the crime in Shire Town in order to work on creating awareness at crime minimization. The Weka software was employed to build the J48 decision tree and NaiveBayes algorithms. This software partitions the dataset prepared for analysis into training and test facts. Where training facts are used to train and build the models and also test facts are used to test the performance of the model. Besides for validation purpose the researcher splits the dataset 80% for training while the remaining 20% for testing. WEKA version 3.9.6 was used in this research and the proposed model was tested using test sets of data. Besides, the validity and performance of the model was tested to check its efficiency and effectiveness. Finally, the confusion matrix was used to evaluate the accuracy and performance of the model built with the J48 decision tree algorithm.

4.1.1 Model building

In this part of research the researcher is interested to explain the models selected to achieve the mining goal. The main objective of this research is, to detect patterns for crime from the criminal dataset. The model-building phase in the data mining process of this investigation is carried out the following two-step process i.e. clustering and then classification data mining approach. As described in the methodology section in chapter three, the model addressed is clustering and classification.

This includes data mining tool selection and the algorithms used for modeling technique. The classification modeling technique has used the clustered dataset as an input and implemented using J48 decision tree and NaiveBayes classifications. The clustering sub phase has been conducted using the K-means algorithm for segmenting the data into the target classes of crimes. Both the clustering and the classification tasks were applied to the training dataset. As described before, these techniques were implemented using Weka 3.9.6 Data Mining tool.

Table 4.1 Abbreviated values

List of abbreviated values	
List of abbreviated values	**Descriptions**
PM	Post meridiem
AM	Ante meridian

4.1.1.1 Cluster modeling

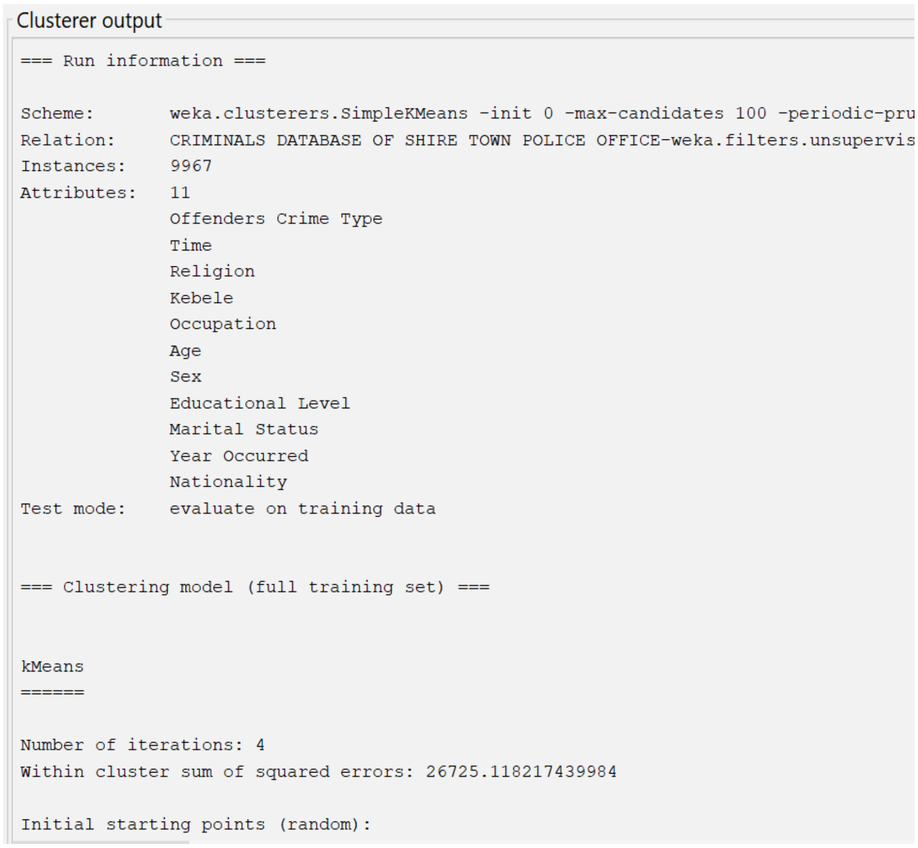
Once the dataset is ready to be used, the next step is building the clustering model using the selected data mining tool. Clustering is a data mining technique for grouping the similar type of crimes. During clustering the whole dataset have been used for the training purpose since clustering is unsupervised learning. In unsupervised learning, algorithms do not need class labels or dependent variable rather learns from the dataset and place objects into their category or clusters based on the underlying data distribution, [55] therefore in this study the researcher cluster the criminal data in to three based on offenders' occupation (Unemployed, Private and Governmental). After clustering the researcher choose the best experiment and select the best cluster for classification model. In this study WEKA K-means algorithm was used to implement the clustering model.

Four experimentations were done for the cluster modeling. These experimentations were analyzed and compared to each other based on different measurements such as Number of iterations, with in cluster sum of squared errors time taken to build the model and validated by the domain expert (police personnel). The models were also compared with regard to the patterns or knowledge discovered and other criteria. This algorithm includes the following configuration parameters as follows:

- **Distance Function:** this option is used to find the similarity and dissimilarity between clusters.
- **The number of clusters:** this option is used to set the K value i.e. the number of clusters that need to be created. To know the distribution percentage of the whole data
- **Seed size:** this option is also used to set the random number of seed to be used.

I. Experimentation-1

This is the first clustering experiment and it was done for K=3, with the default seed value and default distance function. All of the final selected attributes and the instances for the training set is the whole data set that is 9967 records were used as an input for the experiment. In order to cluster the records based on their values the model was trained by using the default values of the algorithm. The algorithm is instructed to segment the dataset into three clusters.



```
Clusterer output
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pru
Relation:     CRIMINALS DATABASE OF SHIRE TOWN POLICE OFFICE-weka.filters.unsupervis
Instances:    9967
Attributes:   11
              Offenders Crime Type
              Time
              Religion
              Kebele
              Occupation
              Age
              Sex
              Educational Level
              Marital Status
              Year Occurred
              Nationality
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 26725.118217439984

Initial starting points (random):
```

Figure 4.1 Run information of the first experiment with value of K=3 seed=10

The table below shows the parameter values and the segmentation of the first experiment and distribution of the dataset for each cluster.

Table 4.2 parameter of experiment –I with K=3 seed=10 and other default values

K	Distance function	Seed value	Cluster distribution	
3	Euclidean Distance	10	C1	2734 (27%)
			C2	5598 (56%)
			C3	1635 (16%)

The above table 4.2 shows the first experiment is conducted with the default values of the K-Means algorithm (K=3, seed=10, and Euclidean distance function). So the 2^{nd} cluster registered high percent 56% with 11 attributes were the first and the 3rd cluster 27% and 16% respectively, in clusters (1 & 3) attributes are similarity in character because they are with approximately the same percent(27% & 16%). The following table 4.3 shows the result of the experiment I.

Table 4.3 Clustering result of experiment I

Cluster no	Distribution no of instances in %	Attribute Names										
		Offenders Crime type	Time	Religion	Kebele	Occupation	Age	Sex	Educational level	Marital	Year Occurred	Nationality
1	2734 (27%)	Beating/Causing Bodily	AM	Orthodox	Dedebit	Unemployed	Age2	M	illiterate	Married	2009.9671	Ethiopian
2	5598 (56%)	Theft	AM	Orthodox	Lekatit	Unemployed	Age2	M	primary	Not Married	22012.6559	Ethiopian
3	1635 (16%)	Theft	AM	Orthodox	Adi Ketibay	Unemployed	Age3	M	illiterate	Married	2010.5284	Ethiopian

As shown in the above table 4.3 the result of the first experiment with the average values of the attributes for each segment. The data had to be segmented using clustering approach to three classes namely: cluster 1, cluster 2 and cluster 3.

Table 4.4: Cluster summary of experiment I

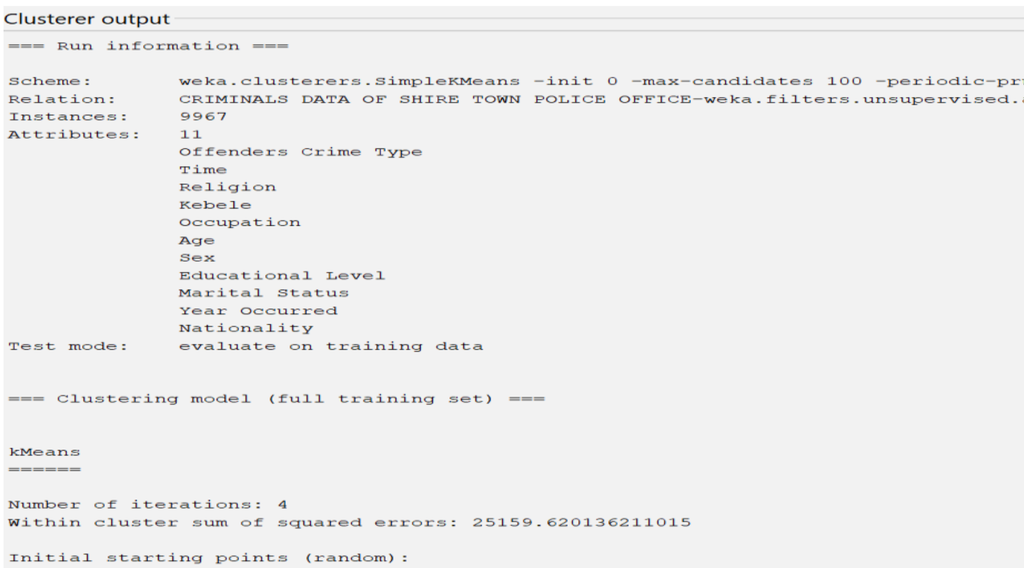
Cluster No	Description
1	Beating/Causing Bodily Harm, AM, Orthodox, Dedebit, Unemployed, Age2, M, illiterate, Married, 2009.9671, Ethiopian
2	Theft, AM, Orthodox, Lekatit, Unemployed, Age2, M, primary, 'Not Married', 2012.6559, Ethiopian
3	Theft, AM, Orthodox, ' Adi Ketibay ', Unemployed, Age3, M,

	illiterate, 'Married', 2010.5284, Ethiopian

the value of cluster 2 shows that type of crime is Theft, the time of crime done AM, the criminal religion Orthodox, Kebele of the offender is in Lekatit, the offender was unemployed, were age of the offender is Age2, sex of the offender is M, educational level of offender is primary, and the offender is No Married, were the year occurred of the crime done is 2012.6559, nationality of the offender is Ethiopian, the crime is assigned under the cluster 2 so in the 2nd cluster those attributes are similar character more than the other two cluster by 56% were the two other clusters (clusters 1 and 3) 27% and 16% respectively.

II. Experimentation-2

The second experiment is conducted with K=3 changed default seed value, (10) to 50 and with default distance function (Euclidean distance).



```
Clusterer output
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pru
Relation:     CRIMINALS DATA OF SHIRE TOWN POLICE OFFICE-weka.filters.unsupervised.a
Instances:    9967
Attributes:   11
              Offenders Crime Type
              Time
              Religion
              Kebele
              Occupation
              Age
              Sex
              Educational Level
              Marital Status
              Year Occurred
              Nationality
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 25159.620136211015

Initial starting points (random):
```

Figure 4.2 Run information of second experiment with value of K=3 seed=50

The table 4.5 below shows that the result of the second experiment with K = 3, seed = 50 and Euclidean distance function is presented.

Table 4.5: Parameters of experiment-II with K=3 seed=50 and other default values

K	Distance function	Seed value	Cluster distribution	
3	Euclidean Distance	50	C1	4215 (42%)
			C2	3501 (35%)
			C3	2251 (23%)

The following table 4.6 shows the clustering result of the model by the second experiment.

Table 4.6: clustering result of experiment II

Cluster no	Distribution no of instances in %	Attribute Names										
		Offenders Crime type	Time	Religion	Kebele	Occupation	Age	Sex	Educational level	Marital Status	Year Occurred	Nationality
1	4215 (42%)	Theft	PM	Orthodox	Adi Ketibay	Unemployed	Age2	M	primary	Not Married	2012.4769	Ethiopian
2	3501 (35%)	Beating/Causing Bodily	AM	Orthodox	Lekatit	Unemployed	Age2	M	illiterate	Married	2010.1617	Ethiopian
3	2251 (23%)	Theft	PM	Orthodox	Adi Ketibay	Unemployed	Age2	M	primary	Not Married	2012.0595	Ethiopian

The above table 4.6 shows the result of the second experiment with the average values of the attributes for each segment. The following table 4.7 exhibits the description for each segment of values.

Table 4.7: Cluster summary of experiment II

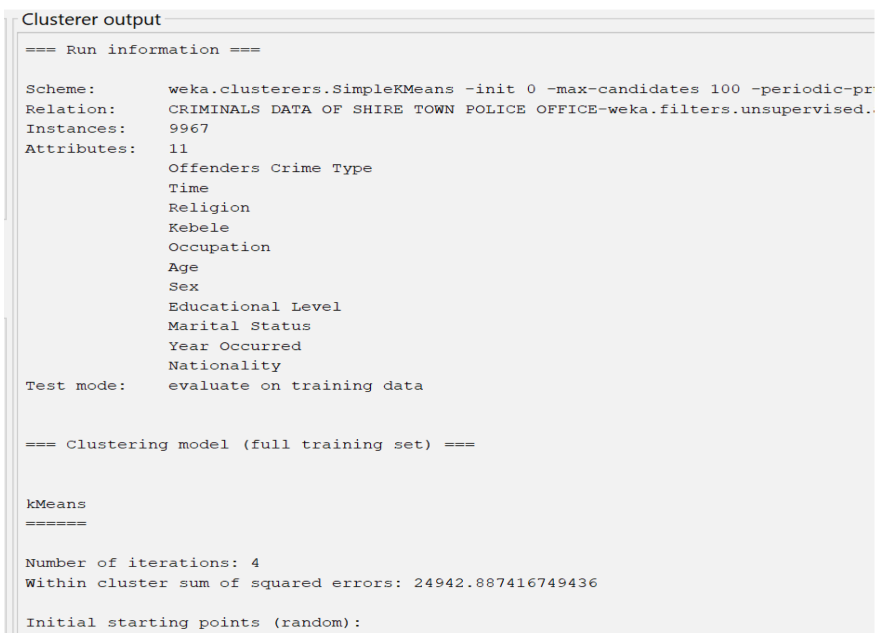
Cluster No	Description
1	Theft, PM, Orthodox, Adi Ketibay, Unemployed, Age2, M, primary, Not Married, 2012.4769, Ethiopian
2	Beating/Causing Bodily Harm, AM, Orthodox, Lekatit, Unemployed, Age2, M, illiterate, Married, 2010.1617, Ethiopian

3	Theft, AM, Orthodox, Adi Ketibay, Unemployed, Age2, M, primary, Not Married, 2012.0595, Ethiopian

From the above table 4.7 the 1st cluster group shows type of crime is Theft, the crime done time is AM, the religion of offender is Orthodox, offender's Kebele is in Adi Ketibay, the offender's work occupation is unemployed, the offender's age is Age2, the offender's gender is Male, educational level of the offender is primary and the offender is Not Married, the Year Occurred of the crime done was 2012.4769, such kind of case is assigned under the cluster 1 (42%) and also the 2nd cluster is registered 35% the attributes have similar characteristics were the left cluster 3 is 23%.

III. Experimentation-3

The third experiment is done for K = 3, changed default seed value to 100 and distance function (Euclidean distance). Similar to the first two runs, all of the final selected attributes and 9967 records are used as an input for conducting the experiment.



```
Clusterer output
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pr
Relation:     CRIMINALS DATA OF SHIRE TOWN POLICE OFFICE-weka.filters.unsupervised.
Instances:    9967
Attributes:   11
              Offenders Crime Type
              Time
              Religion
              Kebele
              Occupation
              Age
              Sex
              Educational Level
              Marital Status
              Year Occurred
              Nationality
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 24942.887416749436

Initial starting points (random):
```

Figure 4.3 Run information of the third experiment with value of K=3 seed=100

Table 4.8 Parameters of experiment III with K=3 & seed=100 and other default values

K	Distance function	Seed value	Cluster distribution	
3	Euclidean Distance	100	C1	4467 (45%)
			C2	3519 (35%)
			C3	1981 (20%)

The table 4.9 below represents the parameters and the cluster distribution of the third experiment. The following table 4.9 shows the result of the model from the third experiment.

Table 4.9: Clustering result of experiment III

Cluster no	Distribution no of instances in %	Attribute Names										
		Offenders Crime type	Time	Religion	Kebele	Occupation	Age	Sex	Educational level	Marital Status	Year Occurred	Nationality
1	4467 (45%)	Beating/Causing Bodily	AM	Orthodox	Sihul	Unemployed	Age2	M	Primary	Not Married	2012.8207	Ethiopian
2	3519 (35%)	Theft	AM	Orthodox	Adi Ketibay	Unemployed	Age2	M	Illiterate	Married	2010.2103	Ethiopian
3	1981 (20%)	Traffic Accident	AM	Orthodox	Adi Ketibay	Private	Age2	M	Secondary	Not Married	2011.162	Ethiopian

The above table 4.9 shows the result of the third experiment with the average values of the attributes for each segment. The following table 4.10 exhibits the description for each segment of values.

Table 4.10 Cluster summary of experiment III

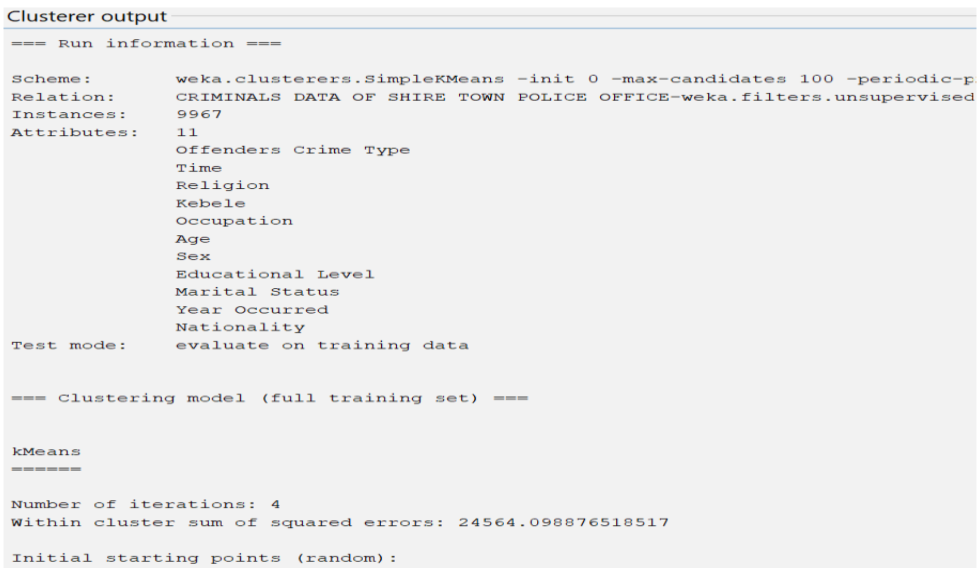
Cluster No	Description
1	Beating/Causing Bodily Harm, AM, Orthodox, Sihul, Unemployed, Age2, M, primary, Not Married, 2012.8207, Ethiopian
2	Theft, AM, Orthodox, Adi Ketibay, Unemployed, Age2, M,

	illiterate, Married, 2010.2103, Ethiopian
3	Traffic Accident, AM, Orthodox, Adi Ketibay, Private, Age2, M, Secondary, Not Married, 2011.162, Ethiopian

From the above table 4.10 the First cluster group shows types of crime is Beating/Causing Bodily Harm, the time of crime done is AM, the offender's religion is Orthodox, offender's Kebele is Sihul y, the offender's work occupation is unemployed, age of the offender is Age2, the offender's gender is Male, education level of the offender is primary and the offender is Not Married, the Year Occurred of the crime done was 2012.8207 the crime is assigned under the cluster 1 and in the First cluster those attributes are similar characteristics more than the other two cluster by 45% were the two other clusters (cluster 2 & 3) 35% and 20% respectively.

IV. Experimentation-4

The fourth experiment was done with K = 3, changed default seed value to 1000 and distance function (Euclidean distance).



```
Clusterer output
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-p
Relation:     CRIMINALS DATA OF SHIRE TOWN POLICE OFFICE-weka.filters.unsupervised
Instances:    9967
Attributes:   11
              Offenders Crime Type
              Time
              Religion
              Kebele
              Occupation
              Age
              Sex
              Educational Level
              Marital Status
              Year Occurred
              Nationality
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 24564.098876518517

Initial starting points (random):
```

Figure 4.4 Run information of experiment IV with value of K=3 seed=1000

Table 4.11 Parameter of experiment-IV with K=3 & seed=1000, and other default values

K	Distance function	Seed value	Cluster distribution	
3	Euclidean Distance	1000	C1	3862 (39%)
			C2	2096 (21%)
			C3	4009 (40%)

Table 4.12 Clustering result of experiment IV

Cluster no	Distribution no of instances in %	Attribute Names										
		Offenders Crime type	Time	Religion	Kebele	Occupation	Age	Sex	Educational level	Marital Status	Year Occurred	Nationality
1	3862 (39%)	Beating/Causing Bodily Harm	AM	Orthodox	Sihul	Unemployed	Age2	M	illiterate	Not Married	2010.4573	Ethiopian
2	2096 (21%)	Traffic Accident	AM	Orthodox	Adi Ketibay	Private	Age2	M	Secondary	Not Married	2010.8454	Ethiopian
3	4009 (40%)	Theft	AM	Orthodox	Adi Ketibay	Unemployed	Age2	M	primary	Not Married	2013.0192	Ethiopian

The above table 4.12 shows the result of the fourth experiment with the average values of the attributes for each segment. The following table 4.13 exhibits the description for each segment of values.

Table 4.13: Cluster summary of experiment IV

Cluster No	Description
1	Beating/Causing Bodily Harm, AM, Orthodox, Sihul, Unemployed, Age2, M, illiterate, Not Married, 2010.4573, Ethiopian
2	Traffic Accident, AM, Orthodox, Adi Ketibay, Private, Age2,

	M, primary, Not Married, 2010.8454, Ethiopian
3	Theft, AM, Orthodox, Adi Ketibay, Unemployed, Age2, M, primary, Not Married, 2013.0192, Ethiopian

From the above table 4.13 the 3rd cluster shows type of crime is Theft, the time of crime done is AM, the offender's religion is Orthodox, offender's Kebele is in Adi Ketibay, the offender's occupation is unemployed, age of the offender is Age2 , the offender's gender is Male, educational level of the offender is primary and the offender is Not Married, the Year Occurred of the crime done was 2013.0192, and his nationality is Ethiopian such kind of case is assigned under the cluster 3 and also the 3rd cluster is registered 40% the attributes have similar characteristics were the left two cluster 1 and cluster 2 are 39% and 21% respectively those attributes have similar characteristics.

Choosing the Best Clustering Model

Four experiments were conducted to come up with the appropriate clustering model. In this study the three criteria's put under consideration when selecting the best cluster model are Within cluster sum of squared error values, Number of iteration and the time which takes to build the model. Within cluster sum of squared error determines the tightness of cluster model, the lower gets its value the better choice it becomes. It's used as a method for assessing the goodness of the clustering model. The number of Iteration of the algorithm tells us how many loops it took to assign the displaced data items in their correct classes. Thus the lower gets the value of the iteration the preferred choice it becomes, since that tells us the convergence of the algorithm was pretty fast.

Table 4.14 Comparison between clustering models

Experimentation	Within cluster sum of squared error values	Number of iteration	Time taken to build the model
I	26725.11	4	0.04
II	25159.62	4	0.03
III	24942.88	4	0.03
IV	24564.09	4	0.03

As we seen from the above table 4.14 the experiment shows the number of iterations in all experiments are the same, additionally the time taken to build the model is approximately the same in all the experiments, but as we compared the value of within cluster sum of squared errors the elbow shows at cluster II, which means when compared the difference value on each cluster the smallest value is between cluster I and cluster II. Therefore, the best cluster model has a value of minimum number of iteration, minimum time to build the model and shows the value elbow at which turns the point. Thus the 2nd experiment shows that it is good in creating clusters.

The 2nd experiment was conducted by setting the value of K=3 with a seed value=50 and the result were shown in table 4.5 and table 4.6. As can be seen from the tables there are three clusters with relatively different behaviors. There exists a relatively better separation between the 3rd cluster and the two clusters (1 & 2) also homogeneity within them is relatively better than the earlier experiments (42%, 35% and 23%) respectively.

The number of iteration is 4 and time took for the model is small (0.03) and with a sum of squared error (25159.62). In addition to the above improvements found by the researcher, the clusters are distinct and meaningful to the domain experts of the Shire Town police office.

Finally the researcher together with the domain experts suggested that the model of the 2nd experiment was selected as good model showing relatively good cluster of the crimes showing a pattern in Shire Town police office in the years 2007-2017 e.c except 2013 e.c. and 2014 e.c. So the output of this cluster model was used as an input for the classification model.

Classification Modeling

The input for this model is output of the clustering model. The resulted clustering model identified on the 2nd experiment was grouped in segments of the criminal data in to cluster 0, cluster 1 and cluster 2. Since the developed clustering model does not classify new instance in to a certain segment, the development of the classification model was necessary. In this case the classifier is used to classify the instances into their already classified cluster index. The cluster index is used as a class label (predictable variable). Unlike clustering classification is supervised learning which divides the whole dataset into training and test sets. Here the classifier used the 10-fold cross-validation. Internal cross-validation is used to determine how well a learning algorithm will fit in independent datasets [61]. The principles of k-fold cross validation are to divided the dataset into k mutually exclusive subsets of approximately equal size, the learning algorithm is then trained on each k-1 subset (the training subset) and its prediction are then verified on the corresponding k subset (the testing subset). The performance measures across all k trials are computed and then averaged to determine the performance of the k-fold cross-validation. The average of the performance measure provides an estimate of the performance of the classifier constructed from the whole dataset [55]. J48 Decision Tree and NaveBayesian Classifier were used to classify the segments using weka tool 3.9.6 machine learning software for classification model. For starting the classification modeling experiments, the J48 decision tree and the NaïveBayes algorithms are selected to classify the records based on their values for the given cluster index.

4.1.1.2 Classification model building using J48 decision tree Algorithm

In this phase of the study, the input for this model is output of the previous clustering model experiment for building the decision tree model. Taken as a whole, a decision tree is a classifier. Any previously unseen record can be fed into the tree. At each model it will be sent either left or right according to some test. Ultimately, it will reach a leaf node and be given the label associated with that leaf.

J48 decision tree algorithm contains some parameters that can be changed to further get better classification accuracy. Initially the classification model is built with the parameter values of the J48 algorithm. The following table summarizes the parameters for J48 decision tree algorithm and their description.

Table 4.15 Some of the J48 algorithms and parameters with their descriptions

Option	Description
Seed	The seed used for randomizing the data when reduced-error pruning is used
unpruned	Whether pruning is performed
Confidence Factor	The confidence factor used for pruning smaller values incur more pruning
NumFolds	Determines the amount of data used for reduced error pruning, one fold is used for pruning and the rest for growing the tree
MinNumObj	The minimum number of instances per leaf
Binary splits	Whether to use binary splits on nominal attributes when building the trees

In this classification experiment, J48 decision tree and Naïve Bayes were selected. There were two experiments to be done for each classification. These experimentations were going to be experimented, analyzed and compared to each other in terms of the performance of matrix value accuracy, number of leaves, size of trees and execution of time. The models also compared with regard to knowledge discovered and the judgment of the domain experts of the Shire Town police office.

A. Experiment 1: J48 decision tree Classification of records using the Class Occupation

This experiment 1 uses the attribute occupation to classify records. The attribute consist of three classes namely Unemployed, private and Governmental. Table 4.16 shows that the experiment has been performed based on two model validation techniques. The first method is 10-fold cross validation techniques. In this case 10 approximately equal proportions, and each in turn was used for testing while the remainder was used for training. This process repeats 10 times and at the end, every instance has been used exactly once for testing. The second is 80-20 percentage split technique to partition the dataset into training and testing data and this parameter was set to 80, which is to mean 80 % for training and 20 for testing. The purpose of using those parameters was to assess the performance of the learning scheme by switching the proportion of testing dataset. The experiment-1 has been tested with 2 scenarios. When the researcher compares the 2 scenarios, scenario-2 has been found to be better model. Out of 1993 records 1869 (93.77%) instances were correctly classified. Besides, in terms of number of leaves and size of tree, both scenarios have the same number of leaves and size of tree (269 and 375 respectively) perform. Time span taken to build the model both scenarios have the same times (Scenario-1=0.04 and Scenario-2=0.03 second), Based on the Average True Positive Rate (AV TPR), Average False Positive Rate (AV FPR), Average Precision Rate (AV.PR), Average Recall Rate (AV.RR), Average Relative Optical Character Curve (ROC) scenario-2 registered better performance. The experiment-1 result of learning schemes are summarized and presented in the following table 4.16.

Table 4.16 Experiment-1 results of J48 using 10-fold cross-validation and 80-20 percentage split

Model Characteristics	**Experiment-1 for Class Occupation**	
	Scenario-1	**Scenario-2**
Model Test option	**10-fold** cross-validation and default parameters	With percentage split of **80-20%** and default parameters
Accuracy (%)	93.74	93.77
Time taken to build model	0.04	0.03
No of leaves	269	269
Size of trees	375	375
AV.TPR (%)	93.7	93.8
AV.FPR (%)	14.1	14.4
AV.PR	0.937	0.937
AV.RR	0.937	0.938
AV.ROC	0.949	0.942
CCI	9344	1869
ICI	623	124

Key: CCI: Correctly Classified Instance, ICI (Incorrectly Classified Instance), Accuracy: registered performance of model, AV: Average, TPR: True Positive Rate FPR: False Positive Rate, ROC: Relative Optical Character Curve, PR: Precision Rate, RR: Recall rate.

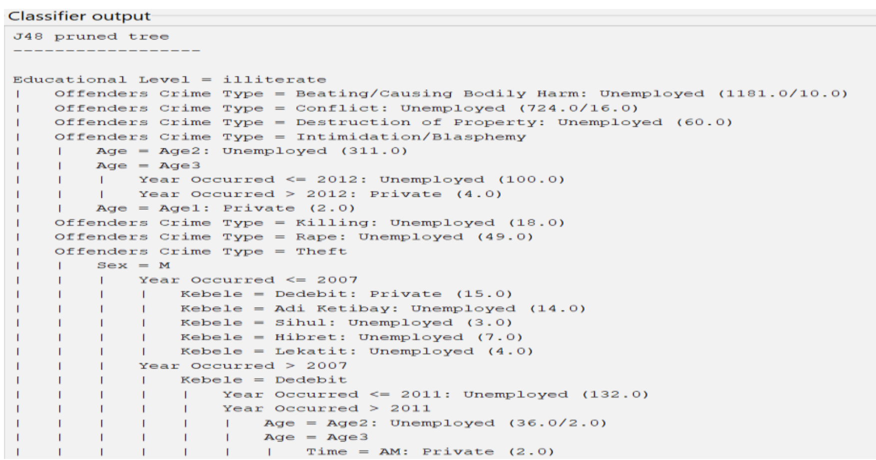
The experiment 1 scenario-2 (J48 with percentage split of 80-20% and default parameters) was performed to evaluate the performance of J48 decision tree classifier in detecting crime patterns. The decision tree algorithm was run on a full training set which contains 1993 instances with 11 attributes. For experiment 1 scenario-2 which is with default value of "Number of objects" (2) the accuracy is found to be 93.77 % which indicates that out of the total 1993 records 1869 (93.77 %) were correctly classified on the other hand, the remaining 124 (6.22 %) records were classified incorrectly. The total number of size of tree is 375 with 269 numbers of leaves and the time taken to build the model is 0.03 seconds.

The following table 4.17 shows the summery result of Confusion matrix for experiment 1 scenario-2 about 97.53**%** of the records in the class of cluster 1 were correctly classified while about 80.57**%** and 84.84**%** of the records in the cluster 2 and cluster 3 classes were classified correctly.

Table 4.17 Confusion matrix of J48 using split 80%/20% and default parameters

Actual	Predicted			Total	Correctly classified
	Unemployed	Private	Governmental		
Unemployed	1505	34	4	1543	97.53%
Private	80	336	1	417	80.57%
Governmental	1	4	28	33	84.84%
Total	1586	374	33	1993	93.77%

On experiment 1 scenario-2 J48 pruned tree model using the percentage split 80%/20% and default parameters is shown below:



```
Classifier output
J48 pruned tree
------------------

Educational Level = illiterate
|   Offenders Crime Type = Beating/Causing Bodily Harm: Unemployed (1181.0/10.0)
|   Offenders Crime Type = Conflict: Unemployed (724.0/16.0)
|   Offenders Crime Type = Destruction of Property: Unemployed (60.0)
|   Offenders Crime Type = Intimidation/Blasphemy
|   |   Age = Age2: Unemployed (311.0)
|   |   Age = Age3
|   |   |   Year Occurred <= 2012: Unemployed (100.0)
|   |   |   Year Occurred > 2012: Private (4.0)
|   |   Age = Age1: Private (2.0)
|   Offenders Crime Type = Killing: Unemployed (18.0)
|   Offenders Crime Type = Rape: Unemployed (49.0)
|   Offenders Crime Type = Theft
|   |   Sex = M
|   |   |   Year Occurred <= 2007
|   |   |   |   Kebele = Dedebit: Private (15.0)
|   |   |   |   Kebele = Adi Ketibay: Unemployed (14.0)
|   |   |   |   Kebele = Sihul: Unemployed (3.0)
|   |   |   |   Kebele = Hibret: Unemployed (7.0)
|   |   |   |   Kebele = Lekatit: Unemployed (4.0)
|   |   |   Year Occurred > 2007
|   |   |   |   Kebele = Dedebit
|   |   |   |   |   Year Occurred <= 2011: Unemployed (132.0)
|   |   |   |   |   Year Occurred > 2011
|   |   |   |   |   |   Age = Age2: Unemployed (36.0/2.0)
|   |   |   |   |   |   Age = Age3
|   |   |   |   |   |   |   Time = AM: Private (2.0)
```

Figure 4.5 Experiment-1 the output of j48 pruned tree rules using split 80%/30% and default parameters

The experiment 1 scenario-1 (10-fold CV and default parameters) was performed to evaluate the performance of J48 decision tree classifier in detecting crime patterns. The decision tree algorithm was run on a full training set which contains 9967 instances with 11 attributes. For experiment 1 scenario-1 which is with default value of "Number of objects" (2) the accuracy is found to be 93.74% which indicates that out of the total 9967 records 9344 (93.74%) were correctly classified on the other hand, the remaining 623 (6.25 %) records were classified incorrectly. The total number of size of tree is 375 with 269 numbers of leaves and the time taken to build the model is 0.04 seconds. The following table 4.18 shows the summery result of experiment 1 scenario-1 about 97.73**%** of the records in the class of cluster 1 were correctly classified while about 80.56**%**and 85.99**%** of the records in the cluster 2 and cluster 3 respectively classes were classified correctly.

Table 4.18 Confusion matrix of J48 using 10-fold CV and default parameters

Actual	Predicted			Total	Correctly classified
	Unemployed	Private	Governmental		
Unemployed	7417	156	16	7589	97.73%
Private	416	1749	6	2171	80.56%
Governmental	10	19	178	207	85.99%
Total	7843	1924	200	9967	93.74%

On experiment 1 scenario-1 J48 pruned tree model using the 10-fold CV and default parameters is shown below:

```
Classifier output
J48 pruned tree
------------------

Educational Level = illiterate
|   Offenders Crime Type = Beating/Causing Bodily Harm: Unemployed (1181.0/10.0)
|   Offenders Crime Type = Conflict: Unemployed (724.0/16.0)
|   Offenders Crime Type = Destruction of Property: Unemployed (60.0)
|   Offenders Crime Type = Intimidation/Blasphemy
|   |   Age = Age2: Unemployed (311.0)
|   |   Age = Age3
|   |   |   Year Occurred <= 2012: Unemployed (100.0)
|   |   |   Year Occurred > 2012: Private (4.0)
|   |   Age = Age1: Private (2.0)
|   Offenders Crime Type = Killing: Unemployed (18.0)
|   Offenders Crime Type = Rape: Unemployed (49.0)
|   Offenders Crime Type = Theft
|   |   Sex = M
|   |   |   Year Occurred <= 2007
|   |   |   |   Kebele = Dedebit: Private (15.0)
|   |   |   |   Kebele = Adi Ketibay: Unemployed (14.0)
|   |   |   |   Kebele = Sihul: Unemployed (3.0)
|   |   |   |   Kebele = Hibret: Unemployed (7.0)
|   |   |   |   Kebele = Lekatit: Unemployed (4.0)
|   |   |   Year Occurred > 2007
|   |   |   |   Kebele = Dedebit
|   |   |   |   |   Year Occurred <= 2011: Unemployed (132.0)
|   |   |   |   |   Year Occurred > 2011
|   |   |   |   |   |   Age = Age2: Unemployed (36.0/2.0)
|   |   |   |   |   |   Age = Age3
|   |   |   |   |   |   |   Time = AM: Private (2.0)
```

Figure 4.6 Experiment-1 the output of j48 pruned tree rules using 10-fold CV and default parameters

B. Experiment 2: J48 decision tree Classification of records using the Class Sex

This experiment 2 uses the attribute Sex to classify records. The attribute consists of two classes namely Male and Female. Table 4.19 shows that the experimentation has been performed based on two model validation techniques. The first method is 10-fold cross validation techniques and the second is 80-20 percentage split technique. The purpose of using those parameters was to assess the performance of the learning scheme by switching the proportion of testing dataset. The experiment 2 has been tested with 2 scenarios. When the researcher compares the 2 scenarios, scenario-1 has been found to be better model. Out of 9967 records 9716 (97.48%) instances was correctly classified. Besides, in terms of number of leaves and size of tree, both scenarios are the same number of leaves and size of tree (119 and 157 respectively) perform. Time span taken to build the model both scenarios have approximately the same times (Scenario-1=0.02 and Scenario-2=0.02 second), Based on the Average True Positive Rate (AV TPR), Average False Positive Rate (AV FPR), Average Precision Rate (AV.PR), Average Recall Rate (AV.RR), Average Relative Optical Character Curve (ROC) scenario-1 registered better performance. The experiment-2 result of learning schemes are summarized and presented in the following table.

Table 4.19 Experiment-2 results of J48 using 10-fold cross-validation and 80-20 percentage split

Model Characteristics	**Experiment-2 for Class Sex**	
	Scenario-1	**Scenario-2**
Model Test option	**10-fold** cross-validation and default parameters	With percentage split of **80-20%** and default parameters
Accuracy (%)	97.48%	97.14%
Time taken to build model	0.02	0.02
No of leaves	119	119
Size of trees	157	157
AV.TPR (%)	97.5	97.1
AV.FPR (%)	47.9	52.2
AV.PR	0.973	0.968
AV.RR	0.975	0.971
AV.ROC	0.921	0.913
CCI	9716	1936
ICI	251	57

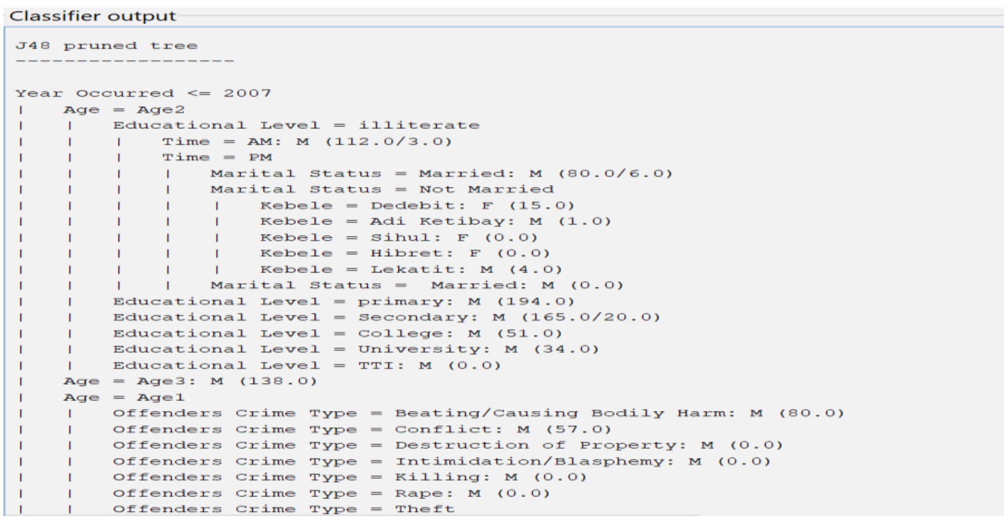
The experiment 2 scenario-1 (10-fold cross-validation and default parameters) was performed to evaluate the performance of J48 decision tree classifier in detecting crime patterns. The decision tree algorithm was run on a full training set which contains 9967 instances with 11 attributes. For experiment 2 scenario-1, which is with default value of "Number of objects" (2) the accuracy is found to be 97.48 % which indicates that out of the total 9967 records 9716 (97.48 %) were correctly classified on the other hand, the remaining 251 (2.51%) records were classified incorrectly. The total number of size of tree is 157 with 119 numbers of leaves and the time taken to build the model is 0.02 seconds.

The following table 4.20 shows the summery result of experiment-2 scenario-1 about 99.69**%** of the records in the class of cluster 1 were correctly classified while about 49.88**%** of the records in the cluster 2 classes were classified correctly.

Table 4.20 Confusion matrix of J48 using 10-fold cross-validation and default parameters

Actual	Predicted		Total	
	Male	Female		
Male	9495	29	9524	99.69%
Female	222	221	443	49.88%
Total	9717	250	9967	97.48%

The experiment 2 scenario-1 J48 pruned tree model using the 10-fold cross-validation and default parameters is shown below:



```
Classifier output
J48 pruned tree
------------------

Year Occurred <= 2007
|   Age = Age2
|   |   Educational Level = illiterate
|   |   |   Time = AM: M (112.0/3.0)
|   |   |   Time = PM
|   |   |   |   Marital Status = Married: M (80.0/6.0)
|   |   |   |   Marital Status = Not Married
|   |   |   |   |   Kebele = Dedebit: F (15.0)
|   |   |   |   |   Kebele = Adi Ketibay: M (1.0)
|   |   |   |   |   Kebele = Sihul: F (0.0)
|   |   |   |   |   Kebele = Hibret: F (0.0)
|   |   |   |   |   Kebele = Lekatit: M (4.0)
|   |   |   |   Marital Status =  Married: M (0.0)
|   |   Educational Level = primary: M (194.0)
|   |   Educational Level = Secondary: M (165.0/20.0)
|   |   Educational Level = College: M (51.0)
|   |   Educational Level = University: M (34.0)
|   |   Educational Level = TTI: M (0.0)
|   Age = Age3: M (138.0)
|   Age = Age1
|   |   Offenders Crime Type = Beating/Causing Bodily Harm: M (80.0)
|   |   Offenders Crime Type = Conflict: M (57.0)
|   |   Offenders Crime Type = Destruction of Property: M (0.0)
|   |   Offenders Crime Type = Intimidation/Blasphemy: M (0.0)
|   |   Offenders Crime Type = Killing: M (0.0)
|   |   Offenders Crime Type = Rape: M (0.0)
|   |   Offenders Crime Type = Theft
```

Figure 4.7 Experiment-2 the output of j48 pruned tree rules using 10-fold cross-validation and default parameters

The experiment 2 scenario-2 of J48 decision tree (percentage split of 80-20% and default parameters) was performed to evaluate the performance of J48 decision tree classifier in detecting crime patterns. The decision tree algorithm was run on a full training set which contains 1993 instances with 11 attributes. For experiment 2 scenario-2, which is with default value of “Number of objects” (2) the accuracy is found to be 97.14% which indicates that out of the total 1993 records 1936 (97.14%) were correctly classified on the other hand the remaining 57 (2.86%) records were classified incorrectly. The total number of size of tree is 157 with 119 numbers of leaves and the time taken to build the model is 0.02 seconds.

The following table 4.21 shows the summery result of experiment 2 scenario-2 about 99.52**%** of the records in the class of cluster 1 and 45.45**%**of the records in the class of cluster 2 were classified correctly.

Table 4.21 Confusion matrix of J48 using percentage split of 80-20% and default parameters

Actual	Predicted		Total	Correctly classified
	Male	Female		
Male	1896	9	1905	99.52%
Female	48	40	88	45.45%
Total	1944	49	1993	97.14%

The experiment 2 scenario-2 J48 pruned tree model using the percentage split 80-20% and default parameters is shown below:

```
Classifier output
J48 pruned tree
------------------

Year Occurred <= 2007
|   Age = Age2
|   |   Educational Level = illiterate
|   |   |   Time = AM: M (112.0/3.0)
|   |   |   Time = PM
|   |   |   |   Marital Status = Married: M (80.0/6.0)
|   |   |   |   Marital Status = Not Married
|   |   |   |   |   Kebele = Dedebit: F (15.0)
|   |   |   |   |   Kebele = Adi Ketibay: M (1.0)
|   |   |   |   |   Kebele = Sihul: F (0.0)
|   |   |   |   |   Kebele = Hibret: F (0.0)
|   |   |   |   |   Kebele = Lekatit: M (4.0)
|   |   |   |   Marital Status =  Married: M (0.0)
|   |   Educational Level = primary: M (194.0)
|   |   Educational Level = Secondary: M (165.0/20.0)
|   |   Educational Level = College: M (51.0)
|   |   Educational Level = University: M (34.0)
|   |   Educational Level = TTI: M (0.0)
|   Age = Age3: M (138.0)
|   Age = Age1
|   |   Offenders Crime Type = Beating/Causing Bodily Harm: M (80.0)
|   |   Offenders Crime Type = Conflict: M (57.0)
|   |   Offenders Crime Type = Destruction of Property: M (0.0)
|   |   Offenders Crime Type = Intimidation/Blasphemy: M (0.0)
|   |   Offenders Crime Type = Killing: M (0.0)
|   |   Offenders Crime Type = Rape: M (0.0)
```

Figure 4.8 Experiment-2 the output of j48 pruned tree rules using percentage split of 80-20% and default parameters

4.1.1.3 Classification model building using NaïveBayes Algorithm

The second data mining technique employed for this study to build the classification model is the naïveBayes. To build the NaïveBayes model, weka software package is used and it employs the NaïveBayes algorithm in developing the model and also to build the model the clustered data set were used as an input to the naïveBayes classifier algorithm. For the purpose of classification model with NaïveBayes algorithm, the following two experimentations were used.

I. Experiment 1: NaiveBayes Classification of records using Occupation class

This experiment 1 uses the attribute Occupation to classify records. The attribute consists of three classes, namely Unemployed, Private and Governmental. Table 4.22 shows that the experiment has been performed based on two model validation techniques. The first method is 10-fold cross validation techniques and the second is 80-20 percentage split technique to partition the dataset into training and testing data and this parameter was set to 80. The purpose of using those parameters was to assess the performance of the learning scheme by switching the proportion of testing dataset. The experiment 1 has been tested with 2 scenarios. When the researcher compares the 2 scenarios, scenario-2 has been found to be better model. Out of 1993 records 1762 (88.40%) instances was correctly classified. Besides, in terms of times span taken, both scenarios have the same time to build the model. (Scenario-1=0.01 second and Scenario-2=0.01 second). Based on the Average True Positive Rate (AV TPR), Average False Positive Rate (AV FPR), Average Precision Rate (AV.PR), Average Recall Rate (AV.RR), Average Relative Optical Character Curve (ROC) scenario-2 registered better performance. The experiment-1 result of learning schemes are summarized and presented in the following table.

Table 4.22 Experiment-1 results of NaiveBayes using 10-fold cross-validation and 80-20 percentage split

Model Characteristics	**Experiment-1 for Class Occupation**	
	Scenario-1	**Scenario-2**
Model Test option	**10-fold** cross-validation and default parameters	With percentage split of **80-20%** and default parameters
Accuracy (%)	87.66	88.40
Time taken to build model	0.01	0.01
AV.TPR (%)	87.7	88.4
AV.FPR (%)	25.2	25.0
AV.PR	0.871	0.879
AV.RR	0.877	0.884
AV.ROC	0.897	0.908
CCI	8738	1762
ICI	1229	231

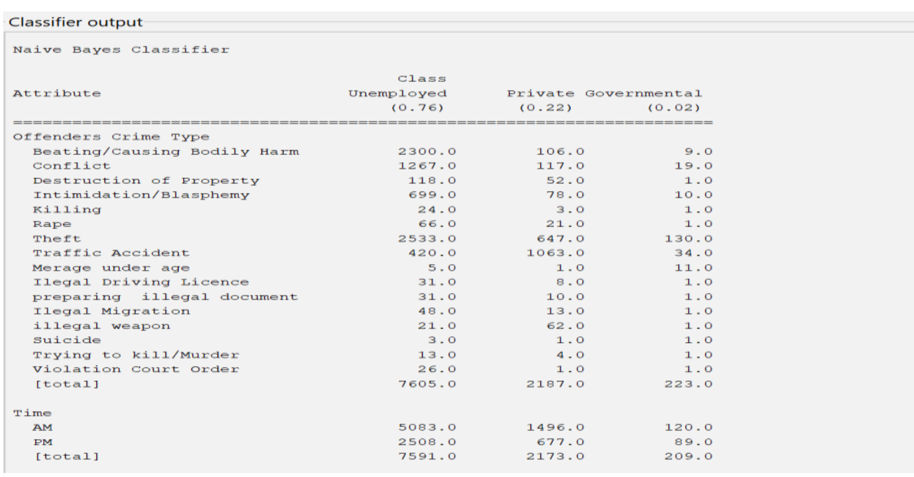
The experiment 1 scenario-1 of the NaiveBayes (10-fold cross-validation and default parameters) was performed to evaluate the performance of NaiveBayes classifier in detecting crime patterns. The NaiveBayes algorithm was run on a full training set which contains 7000 instances with 11 attributes. For experiment 1 scenario-1 which is with default value of “Number of objects” (2) the accuracy is found to be 87.66% which indicates that out of the total number of 9967 records 8738 (87.66%) were correctly classified on the other hand, the remaining 1229 (12.33%) records were classified incorrectly. The time taken to build the model is 0.01 seconds.

The following table 4.23 shows the summary result of experiment 1 scenario-1 about 95.05**%** of the records in the class of cluster 1 were correctly classified while about 65.36**%** and 50.72**%** of the records in the cluster 2 and cluster 3 classes were classified correctly.

Table 4.23 confusion matrix of NaiveBayes using 10-fold cross validation and default parameters

Actual	Predicted			Total	Correctly classified
	Unemployed	Private	Governmental		
Unemployed	7214	358	17	7589	95.05%
Private	717	1419	35	2171	65.36%
Governmental	32	70	105	207	50.72%
Total	7963	1847	157	9967	87.66%

On experiment 1 scenario-1 NaiveBayes pruned model using the 10-fold cross validation and default parameters is shown below



```
Classifier output
Naive Bayes Classifier

                                        Class
Attribute                          Unemployed     Private Governmental
                                       (0.76)      (0.22)       (0.02)
======================================================================
Offenders Crime Type
  Beating/Causing Bodily Harm          2300.0       106.0          9.0
  Conflict                             1267.0       117.0         19.0
  Destruction of Property               118.0        52.0          1.0
  Intimidation/Blasphemy                699.0        78.0         10.0
  Killing                                24.0         3.0          1.0
  Rape                                   66.0        21.0          1.0
  Theft                                2533.0       647.0        130.0
  Traffic Accident                      420.0      1063.0         34.0
  Merage under age                        5.0         1.0         11.0
  Ilegal Driving Licence                 31.0         8.0          1.0
  preparing  illegal document            31.0        10.0          1.0
  Ilegal Migration                       48.0        13.0          1.0
  illegal weapon                         21.0        62.0          1.0
  Suicide                                 3.0         1.0          1.0
  Trying to kill/Murder                  13.0         4.0          1.0
  Violation Court Order                  26.0         1.0          1.0
  [total]                              7605.0      2187.0        223.0

Time
  AM                                   5083.0      1496.0        120.0
  PM                                   2508.0       677.0         89.0
  [total]                              7591.0      2173.0        209.0
```

Figure 4.9 Experiment-1 the output of NaiveBayes pruned tree rules using 10-fold cross-validation and default parameters

The experiment 1 scenario-2 of the NaiveBayes (percentage split of 80-20% and default parameters) was performed to evaluate the performance of NaiveBayes classifier in detecting crime patterns. The

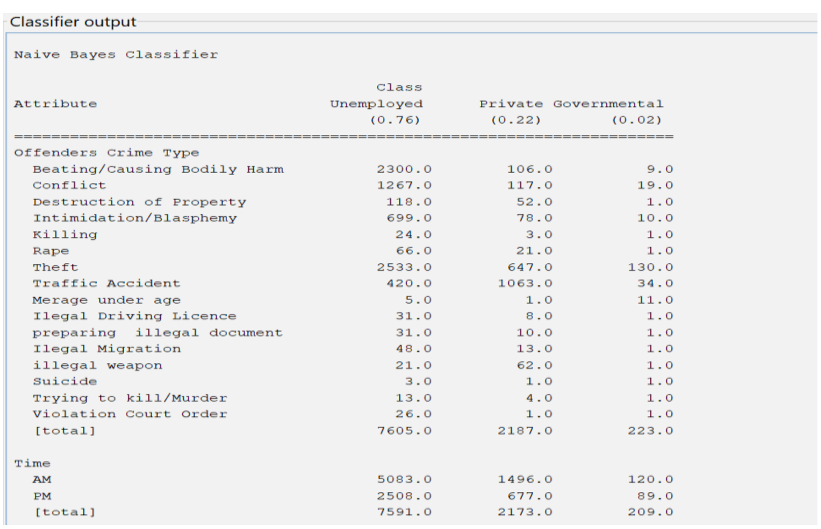
NaiveBayes algorithm was run on a full training set which contains 1993 instances with 11 attributes. For experiment 1 scenario-2 which is with default value of “Number of objects” (2) the accuracy is found to be 88.40% which indicates that out of the 1993 records 1762 (88.40%) were correctly classified on the other hand, the remaining 231 (11.59%) records were classified incorrectly. The time taken to build the model is 0.01 seconds.

The following table 4.24 shows the summery result of experiment 1 scenario-2 about 95.00**%** of the records in the class of cluster 1 were correctly classified while about 65.94**%**and 63.63**%** of the records in the cluster 2 and cluster 3 respectively classes were classified correctly.

Table 4.24 Confusion matrix of NaiveBayes using percentage split of 80-20% and default parameters

Actual	Predicted			Total	Correctly classified
	Unemployed	Private	Governmental		
Unemployed	1466	75	2	1543	95.00%
Private	135	275	7	417	65.94%
Governmental	4	8	21	33	63.63%
Total	1605	358	30	1993	88.40%

On experiment 1 scenario-2 NaiveBayes pruned model using the percentage split of 80-20% and default parameters is shown below:



```
Classifier output

Naive Bayes Classifier

                                        Class
Attribute                          Unemployed      Private Governmental
                                       (0.76)       (0.22)       (0.02)
=======================================================================
Offenders Crime Type
  Beating/Causing Bodily Harm          2300.0        106.0          9.0
  Conflict                             1267.0        117.0         19.0
  Destruction of Property               118.0         52.0          1.0
  Intimidation/Blasphemy                699.0         78.0         10.0
  Killing                                24.0          3.0          1.0
  Rape                                   66.0         21.0          1.0
  Theft                                2533.0        647.0        130.0
  Traffic Accident                      420.0       1063.0         34.0
  Merage under age                        5.0          1.0         11.0
  Ilegal Driving Licence                 31.0          8.0          1.0
  preparing  illegal document            31.0         10.0          1.0
  Ilegal Migration                       48.0         13.0          1.0
  illegal weapon                         21.0         62.0          1.0
  Suicide                                 3.0          1.0          1.0
  Trying to kill/Murder                  13.0          4.0          1.0
  Violation Court Order                  26.0          1.0          1.0
  [total]                              7605.0       2187.0        223.0

Time
  AM                                   5083.0       1496.0        120.0
  PM                                   2508.0        677.0         89.0
  [total]                              7591.0       2173.0        209.0
```

Figure 4.10 Experiment-1 the output of NaiveBayes using percentage split of 80-20% and default parameters

II. Experiment 2: NaiveBayes Classification of records using the Class Sex

This experiment 2 uses the attribute sex to classify records. The attribute consist of 2 classes namely Male and Female. Table 4.25 shows that the experiment has been performed based on two model validation techniques. The first method is 10-fold cross validation techniques and the second is 80-20 percentage split technique. The purpose of using those parameters was to assess the performance of the learning scheme by switching the proportion of testing dataset. The experiment 2 has been tested with 2 scenarios. When the researcher compares the 2 scenarios, the scenario-2 has been found to be better model. Out of 1993 records 1906 (95.63%) instances was correctly classified. Besides, times span taken to build the model, both scenarios have the same time to build the model. (Scenario-1=0.01 second and Scenario-2=0.01 second). Based on the Average True Positive Rate (AV TPR), Average False Positive Rate (AV FPR), Average Precision Rate (AV.PR), Average Recall Rate (AV.RR), Average Relative Optical Character Curve (ROC) scenario-2 registered better performance. The experiment-2 result of learning schemes are summarized and presented in the following table.

Table 4.25 Experiment-2 results of NaiveBayes using 10-fold cross-validation and 80-20 percentage split

Model Characteristics	**Experiment-2 for Class Sex**	
	Scenario-1	**Scenario-2**
Model Test option	**10-fold** cross-validation and default parameters	With percentage split of **80-20%** and default parameters
Accuracy (%)	95.60	95.63
Time taken to build model	0.01	0.01
AV.TPR (%)	95.60	95.60
AV.FPR (%)	91.90	92.3
AV.PR	0.941	0.941
AV.RR	0.956	0.956
AV.ROC	0.819	0.841
CCI	9529	1906
ICI	438	87

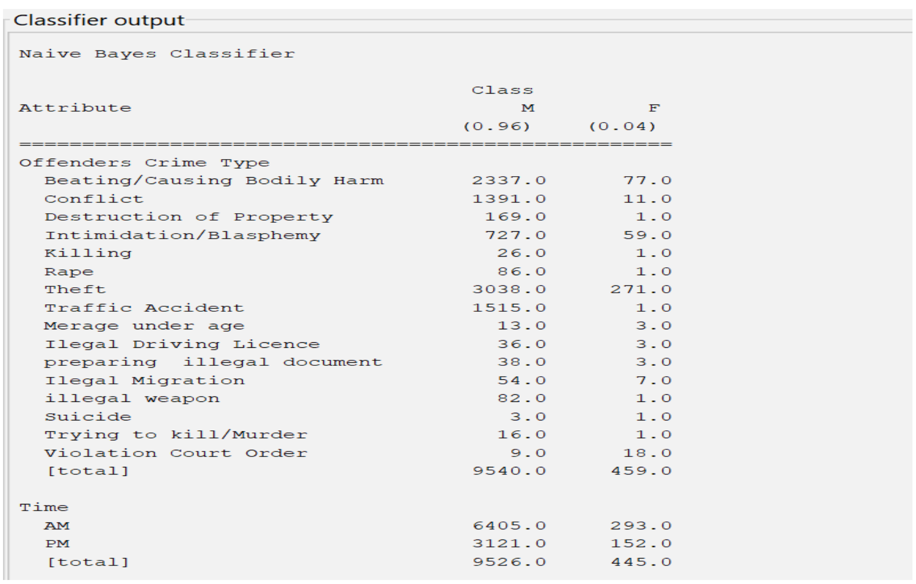
The experiment 2 scenario-1 (10-fold cross-validation and default parameters) was performed to evaluate the performance of NaiveBayes classifier in detecting crime patterns. The NaiveBayes algorithm was run on a full training set which contains 9967 instances with 11 attributes. For experiment 2, scenario-1 which is with default value of "Number of objects" (2) the accuracy is found to be 95.60 % which indicates that out of the total number of 9967 records 9529 (95.60%) were correctly classified on the other hand; the remaining 438 (4.39%) records were classified incorrectly. The time taken to build the model is 0.01 seconds.

The following table 4.26 shows the summery result of experiment 2 scenario-1 about 99.87**%** of the records in the class of cluster 1 were correctly classified while about 3.83**%** of the records in the cluster 2 class were classified correctly.

Table 4.26 confusion matrix of NaiveBayes using 10-fold cross-validation and default parameters

Actual	Predicted		Total	
	Male	Female		
Male	9512	12	9524	99.87%
Female	426	17	443	3.83%
Total	9938	29	9967	92.99%

On experiment 2 scenario-1 NaiveBayes model using the 10-fold cross-validation and default parameters is shown below:



```
Classifier output
Naive Bayes Classifier

                                          Class
Attribute                                     M          F
                                          (0.96)     (0.04)
=====================================================
Offenders Crime Type
  Beating/Causing Bodily Harm             2337.0       77.0
  Conflict                                1391.0       11.0
  Destruction of Property                  169.0        1.0
  Intimidation/Blasphemy                   727.0       59.0
  Killing                                   26.0        1.0
  Rape                                      86.0        1.0
  Theft                                   3038.0      271.0
  Traffic Accident                        1515.0        1.0
  Merage under age                          13.0        3.0
  Ilegal Driving Licence                    36.0        3.0
  preparing  illegal document               38.0        3.0
  Ilegal Migration                          54.0        7.0
  illegal weapon                            82.0        1.0
  Suicide                                    3.0        1.0
  Trying to kill/Murder                     16.0        1.0
  Violation Court Order                      9.0       18.0
  [total]                                 9540.0      459.0

Time
  AM                                      6405.0      293.0
  PM                                      3121.0      152.0
  [total]                                 9526.0      445.0
```

Figure 4.11 Experiment-2 the output of NaiveBayes pruned rules using 10-fold cross-validation and default parameters

The experiment 2 scenario-2 of NaiveBayes (percentage split of 80-20% and default parameters) was performed to evaluate the performance of NaiveBayes classifier in detecting crime patterns. The NaiveBayes algorithm was run on a full training set which contains 1993 instances with 11 attributes. For experiment 2 scenario-2, which is with default value of "Number of objects" (2) the accuracy is found to be 95.63% which indicates that out of the total 1993 records 1906 (95.63%) were correctly classified on the other hand the remaining 87 (4.36%) records were classified incorrectly. The time taken to build the model is 0.01 seconds.

The following table 4.27 shows the summery result of experiment 2 scenario-2 about 98.39**%** of the records in the class of cluster 1 and 16.50**%**of the records in the class of cluster 2 were classified correctly.

Table 4.27 Confusion matrix of NaiveBayes using percentage split of 80-20% and default parameters

Actual	Predicted		Total	Correctly classified
	Male	Female		
Male	1903	2	1905	99.89%
Female	85	3	88	3.40%
Total	1988	5	1993	95.63%

The experiment 2 scenario-2 NaiveBayes pruned rule model using the percentage split 80-20% and default parameters is shown below:

```
Classifier output
Naive Bayes Classifier

                                    Class
Attribute                               M          F
                                   (0.96)     (0.04)
=====================================================
Offenders Crime Type
  Beating/Causing Bodily Harm      2337.0       77.0
  Conflict                         1391.0       11.0
  Destruction of Property           169.0        1.0
  Intimidation/Blasphemy            727.0       59.0
  Killing                            26.0        1.0
  Rape                               86.0        1.0
  Theft                            3038.0      271.0
  Traffic Accident                 1515.0        1.0
  Merage under age                   13.0        3.0
  Ilegal Driving Licence             36.0        3.0
  preparing  illegal document        38.0        3.0
  Ilegal Migration                   54.0        7.0
  illegal weapon                     82.0        1.0
  Suicide                             3.0        1.0
  Trying to kill/Murder              16.0        1.0
  Violation Court Order               9.0       18.0
  [total]                          9540.0      459.0

Time
  AM                               6405.0      293.0
  PM                               3121.0      152.0
  [total]                          9526.0      445.0
```

Figure 4.12 Experiment-2 the output of NaiveBayes using percentage split of 80-20% and default parameters

4.1.1.4 Comparison of J48 decision tree and NaïveBayes models

One of the purposes of this study was come up with a data mining model that is capable to best performance and to select the one, which performs better in handling the crime pattern detection. Accordingly, each experiment carried out in this research has employed j48 decision tree and NaiveBayes. In all experiments the same datasets were used.

Table 4.28 summary of the J48 Decision Tree and NaïveBayes Models

Experiment No	Classification Model	Scenario	No Of leaves	Size of Tree	TPR(%)	FPR(%)	PR	RR	AV. ROC	CCI (%)	ICI (%)	Time
1	J48	1I	269	375	93.8	14.4	0.937	0.938	0.942	93.77	6.23	0.03
2	J48	1	119	157	97.5	47.9	0.973	0.975	0.921	97.48	2.52	0.02
1	NaïveBayes	1I	-	-	88.4	25.0	0.879	0.884	0.908	88.40	11.6	0.01
2	NaiveBayes	1I	-	-	95.60	92.3	0.941	0.956	0.841	95.63	4.37	0.01

The above table 4.28 shows the summary of the result to select the best algorithm, totally two scenarios (scenario-1 and scenario-2) were performed on each the four experiments (two experiments on J48 and two experiments on NaiveBayes). Before the researcher compares the model algorithms above on table 4.28 he compared each experiment for the scenarios they contain. Finally, the researcher had selected the best scenario from each experiments of the model algorithms based on the performance of their Time taken to build the model, No of leaves, Size of trees, AV TPR, AV FPR, AV PR, AV RR, AV ROC, CCI and ICI. Therefore as looks on the above table 4.28 the experiment 2 scenario-1 of the J48 decision tree was selected as the best model algorithm for this research. Because it performs best, based on the criteria listed on the above table 4.28. So, compared all of the experiments with the available criteria the experiment 2 scenario-1 for J48 decision tree with 119 No of leaves, 157 Size of trees, accuracy=97.48%, etc., were chosen. Therefore, the models which were built by these experiments were evaluated by the researcher and discussing with the domain experts of the Shire Town police office i.e. the police personnel and other members of the police institute. Due to that the experiment 2 scenario-1 J48 decision tree was chosen the final working classification model for crime patter detection purpose of Shire Town police office.

4.2 Evaluation of the discovered knowledge

One of the aims of this research is to provide actionable recommendations for the Shire Town Police Office on how to integrate data mining findings into their operational practices for enhanced crime prevention and resource allocation. Due to that a discussion was takes placed among the researcher and the domain experts of the Shire Town police office i.e. with police personnel and other members as this research is help full on making analysis of offenders' information based on the domain experts' need to enhance the institute's crime prevention. The domain experts explain that, there was a challenge to make offenders' analysis from manual based system of crime detection method. But after this research was done the domain experts expressed this research can simplify their practices. Additionally the domain experts explain that, applying data mining based crime analysis enable them to enhance their awareness on how to use data mining techniques for crime detection of Shire Town police office.

This research provides valuable insights into crime patterns, enabling the Shire Town Police Office to implement more effective crime prevention strategies, ultimately contributing to a safer community. By utilizing data mining techniques, this study empowers law enforcement to make informed decisions based on empirical evidence rather than intuition, enhancing the efficiency and effectiveness of crime management efforts.

At the end of this phase, a decision on the use of the data mining results is reached. This is performed based on the domain expert's advice and the parameters set data mining model and the researcher's personal judgment. It is good to see the meaning of the patterns generated by J48 decision tree. As shown in Appendix A, the number of instances for each label is given at the J48 decision tree list as name of the majority class followed by number of instances.

.The following are few of the patterns which were discovered between the attributes. Those patterns have also got an acceptance by the domain experts as consulted informally.

1. Year Occurred <= 2007

| Age = Age2

| | Educational Level = illiterate

| | | Time = AM: M (112.0/3.0)

This rule implies that, if the occurred crime type is at 2007 e.c, offenders who are grouped in the age group of Age2 (20 up to 32 years old), their Educational status is illiterate, and the crime occurred time is AM, then 109 offenders (97.32% of them) are classified as Male. There are 112 records. From which 3 records (2.67% of them) are incorrectly classified.

2. Year Occurred <= 2007

| Age = Age2

| | Educational Level = illiterate

| | | Time = PM

| | | | Marital Status = Married: M (80.0/6.0)

This rule implies that, if the occurred crime type is at 2007 e.c, offenders who are grouped in the age group of Age2 (20 up to 32 years old), their Educational status is illiterate, the crime occurred time is PM, and their marital status is Married then 74 offenders (92.5% of them) are classified as Male. There are 80 records. From which 6 records (7.5% of them) are incorrectly classified.

3. Year Occurred <= 2007

| Age = Age2

| | Educational Level = illiterate

| | | Time = PM

| | | | Marital Status = Not Married

| | | | | Kebele = Dedebit: F (15.0)

This rule implies that, if the occurred crime type is at 2007 e.c, offenders who are grouped in the age group of Age2 (20 up to 32 years old), their Educational status is illiterate, the crime occurred time is PM, their marital status is Not Married, and who are living at Kebele Dedebit then 15 offenders (100% of them) are classified as Female. There are 15 records. From which no records (0% of them) are incorrectly classified.

4. **Year Occurred <= 2007**

   | Age = Age2

   | | Educational Level = primary: M (194.0)

This rule implies that, if the occurred crime type is at 2007 e.c, offenders who are grouped in the age group of Age2 (20 up to 32 years old), and their Educational status is primary, then 194 offenders (100% of them) are classified as Male. There are 194 records. From which no records (0% of them) are incorrectly classified.

5. **Year Occurred <= 2007**

   | Age = Age2

   | | Educational Level = Secondary: M (165.0/20.0)

This rule implies that, if the occurred crime type is at 2007 e.c, offenders who are grouped in the age group of Age2 (20 up to 32 years old), and their Educational status is Secondary, then 145 offenders (87.87% of them) are classified as Male. There are 165 records. From which 20 records (12.12% of them) are incorrectly classified.

6. **Year Occurred > 2007**
   | Offenders Crime Type = Beating/Causing Bodily Harm
   | | Time = AM: M (1244.0/8.0)

This rule implies that, if the occurred crime type is from 2008 e.c, up to 201 e.c, offenders' crime type is Beating/Causing Bodily Harm, and the crime occurred time is AM, and then 1236 offenders (99.36% of them) are classified as Male. There are 1244 records. From which 8 records (0.64% of them) are incorrectly classified.

4.3 Use of the discovered knowledge

On this step of the research the researcher gives awareness and creates consensuses with the owner of the criminals' database i.e. the personnel officer of the police office, on how to use the discovered knowledge using data mining system of crime pattern analysis. It concerns how to link with their crime detection analysis and with their plan, where and how the discovered knowledge will be used. Finally the domain experts promised that the discovered knowledge of this research will be extended to other domains of the Shire Town Police office.

And the researcher discussed with the domain experts of the Shire Town police office about what kind of resources are necessary to apply data mining model which is J48 decision tree algorithm by which hidden knowledge is discovered from their criminals' database based on this research paper.

Therefore simply to apply the data mining based crime pattern detection, materials like: computer, application software (example, WEKA software, etc.), and a person with skill on how to use the data mining based crime pattern analysis are some of the necessary components.

CHAPTER FIVE

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The purpose of this study was to explore the potential of data mining tools and techniques in developing models for crime pattern analysis to support the crime detection activities at the Shire Town Police Office. This research provides valuable insights into crime patterns, enabling the Shire Town Police Office to implement more effective crime prevention strategies, ultimately contributing to a safer community. By utilizing data mining techniques, this study empowers law enforcement to make informed decisions based on empirical evidence rather than intuition, enhancing the efficiency and effectiveness of crime management efforts.

Additionally it is justified that this research is help full on making analysis of offenders' information based on the domain experts' need to enhance the institute's crime prevention plan. The domain experts explain that, there was a challenge to make offenders' analysis from manual based system of crime detection method. But after this research was done the domain experts expressed this research can simplify their practices. Additionally the domain experts explain that, applying data mining based crime analysis enable them to enhance their awareness on how to use data mining techniques for crime detection of Shire Town police office.

In doing so, 9967 sample data were taken randomly from the criminal record database of the Shire Town Police office to build and test the model.

The data was stored in manual unpublished crime record. Data preprocessing and preparing the data for model building was conducted in MS-Excel, CSV, and ARFF filtering and WEKA filters tools. The attribute selection was made with the help of a discussion with domain experts of the police office and WEKA information gain attribute evaluator.

The collection and preparation of data, to make the data suitable for the data mining task, were the major tasks which took considerable time in this research. Then, K-means clustering algorithm was applied to segment the offenders' data into meaningful groups. As a result, four different clustering models were built by using the value of (K =3) and changing the seed size value (seed =10, 50, 100, and 1000). Among the four models, that has best clustering model were selected together with the domain experts. Consequently, from the four clustering model experiments based on their performance ability, the 2nd experiment (K=3) and (seed=50) was selected.

The results of the experiments carried out on this research used classifiers (J48 decision tree and NaiveBayes) have revealed that the technique of data mining is applicable in the process of crime pattern detection. As compared all the experiment of classifiers performed on this research, the J48 decision tree (experiment-2 scenario-1) data mining technique was found to be more appropriate. It is measured based on the criteria those can evaluate performance of classifier algorithms like: Time taken to build the model, No of leaves, Size of trees, AV TPR, AV FPR, AV PR, AV RR and AV ROC, CCI and ICI. The J48 decision tree algorithm scores 97.48% accuracy performance.

5.2 Recommendations

This research was conducted for an academic purpose. However, the results of this study are found to be promising to be applied to address practical problems of crime pattern detection on Shire Town police office. The results of this study have shown that the data mining technology particularly the J48 decision tree algorithm technique is well applicable in the efforts of crime pattern analysis management. Hence, based on the findings of this study, the following recommendations are forwarded.

- ✓ The Shire Town police office should prepare necessary resources to apply data mining model which is J48 decision tree algorithm by which hidden knowledge is discovered from their criminals' database based on this research paper.
- ✓ The researcher recommends that the Shire Town police office should develop a dashboard system to integrating the J48 decision tree algorithm model to aid criminals' real-time access, pattern visualization, and other analyses.
- ✓ The Shire Town police office should give attention to extend the discovered knowledge from this research to other members of the Police office by preparing different data mining skill up training programs.
- ✓ While the researcher used the criminals' data of the Shire Town police office he is confidential on privacy and personal security considerations.
- ✓ The researcher also recommends implementing the discovered classification rules with domain knowledge as knowledge based system that could be helpful for law enforcement agency professionals like police and other researcher who are conducting research on crime prevention and control.
- ✓ Hence, in order to exploit the potential advantage of data mining, police officers should incorporate such additional features of criminals. For instance, if one knows

the specific site where a crime is committed such as nearby school, hotels, market etc., then, a due attention may be given to such sites. Moreover, physical descriptions of an offender could help to search and match suspected person for a specific crime by comparing with the information contained in a database.

- ✓ Although in this study encouraging results were obtained, a sample data was used for training and testing classifiers due to time constraint. Hence, it is appropriate to conduct the experiments with large training and testing datasets as well as making a number of trials to come out with more accurate and better performing classifiers.
- ✓ It would better to use the future researchers on testing continuous data instead of just categorical data.
- ✓ Victims' data is important in the analysis of crime patterns However; due to lack of data availability it is not incorporated on this research. Therefore other researchers should involve such criminals' data on their research.

References

[1]. Abebe, M. (2002). Community policing and ethics, Unpublished BA dissertation, University of Addis Ababa, Ethiopia.

[2]. Adderley, R., William and Musgrove, P. (2001) „Police crime recording and investigation systems, A user's view", International Journal of Police Strategies and Management, Vol. 24, No. 1, pp. 100-114.

[3]. Akpinar E. and Usul N. (2004) „Geographic Information Systems Technologies in Crime Analysis and Crime Mapping".

[4]. Andargachew, T. (1988) „The crime problem and its correction", Vol. I Addis Ababa University, law school training module.

[5]. Angesom T. (2017) An Assessment of Community Policing in Tigray National Regional state.

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

[7]. Berry, M. and Linoff, G. (1997), Data mining techniques and tools.

[8].Berry, M. J. A., Linoff, G.S., (2000) Data Mining Techniques and Algorithms, in Masterin g Data Mining, R.M. Elliott, Editor. John Wiley & Sons, Inc.: USA 2000.

[9]. Botia, J. A., Garijo, M. y Velasco, J. R. and Skarmeta, A. F (1998) A Generic Data Mining System basic design and implementation guidelines", A Technical Project Report of CYCYT project of Spanish Government.

[10]. Brown, D. (1998). The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals.

[11]. Brown, D. (2003). The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals.

[12]. Brown, D.E (2014) the regional crime analysis program: A frame work for mining data to catch criminals".

[13]. Cao. L, 200. Fraud detection using Data mining. Wiley IEEE Press, London.

[14]. Carbone, P. (1997). Data Mining or "Knowledge Discovery in Databases": An Overview.

[15]. Chapman, P Clinton, J Kerber, R Khabaza, T Reinartz, T Shearer, C and Wirth, R. 2000. CRISP- DM 1.0: Step-by-Step data mining guide, SPSS Inc., USA.

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

[17]. Cios. 2005. Data Mining: knowledge Discovery Approach. Springer, New York: USA.

[18]. Cochoran, W.G. (1999). Sampling techniques, 3rd edition, New York: Wiley. Community policing defined [online], 2008 United States of America Department of Justice office for community Oriented Policing.

[19]. D.E. Brown, (1998) "The regional crime analysis program (RECAP): A Frame work for mining data to catch criminals,"

[20]. Deshpande, S. P. and Thakare, V.M., 2010. Data mining system and applications: A review. International Journal of distributed and parallel systems (IJDPS), V.I, pp. 32-44.

[21]. Fayyad, U, Piatesky-Shapiro, G and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework.

[22]. G. Gupta, (2014) Introduction to data mining with case studies: PHI Learning Pvt. Ltd.

[23]. Garvin, M. (1998). Data Mining and the Web: What They Can Do Together.

[24]. Hajizadeh, E. Ardakani, D. and Shahrabi, J. 2010. Application of data mining techniques in stock markets: A survey, Journal of Economics and International Finance, V. 2(7), pp. 109-118.

[25]. Han, J. and Kamber, M (2001). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann publishers.

[26]. Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques, Second Edition.

[27]. Hand, D, Manila, H. and Smyth, P. (2001) Principles of Data Mining", Cambridge, MA, the MIT Press.

[28]. J. Luan, "Data Mining and Its Applications in Higher Education," Wiley Periodicals, pp. 17-36, 3 June 2002.

[29]. Javid, S. (1999). Data Mining in the Next Millennium.

[30]. Kavitha g and Elango n.m (2017) an overview of data mining techniques and its applications school of information technology and engineering, it, Vellore, tamilnadu, India.

[31]. Lalitha Thota and Suresh Babu, (April 2017) Cluster based Zoning of Crime Info, IEEE,

[32]. Larose, D. T. (2005) "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc.

[33]. Leul Woldu (July 2003) the application of data mining in crime prevention: the case of oromia police office

[34]. Seema, M. Rathi and Mamta, "Decision Tree: Data Mining Techniques," International Journal of Latest Trends in Engineering and Technology (IJLTET), vol. 1, no. 3, pp. 150-155, 2012.

[35]. Liu, B and Tuzhilin, A. 2008. Communications of the ACM, February/2008. Journal of Mechanical Engineering Science. Vol. 51, No. 2. London.

[36]. Megaputer Intelligence, Inc. (2002). Crime Pattern Analysis: Megaputer Case Study.

[37]. Mena, J. (2003), Investigative Data Mining for Security and Criminal Detection", Butterworth Heinemann, ISBN 0-7506-7613.

[38]. Mesfin, G/Michael (1999). A computer based criminal records management system case study of the Addis Ababa police. Master's Thesis Addis Ababa University (SISA), Addis Ababa.

[39]. Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., (April 2011)"Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol.6, no. pp.161, 165, 810.

[40]. O`Block, R.L. Donnermeyer, J.F. & Doeren, E. 1991. Security and crime prevention. 2nd editions. US: Butterworth-Heinemann.

[41]. Perez, B. (2001). Data mining technology use grows.

[42]. Qunney (1970) the social reality of crime. Boston: Little, Brown.

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

[44]. S.D.Gheware, A.S.Kejkar, S.M.Tondare (2015) Data Mining: Task, Tools, Techniques and Applications.

[45]. Saygin, Y. and Ulusoy, O. (2002) „Exploiting data mining techniques for broadcasting data in mobile computing environments", IEEE Trans, Knowl, Data Eng. pp1387–1399.

[46]. Shyam Varan Nath (2004) Crime Pattern Detection Using Data Mining.

[47]. Stevens, (2003). Applied community policing in the 21st century. Boston: Pearson.

[48]. Taye Kejia, (2011). "Assessing Factors that affect the implementation of community policing in Awassa, FDRE". University of South Africa.

[49]. Thakur, C. (2003). Crime Control. http://ncthakur.itgo.com/chand3c.htm

[50]. Thearling, K. (2003) „An introduction to data mining", McGraw- Hill, New York.

[51]. Veenendaal and Houweling (2003).Gut feelings, crime data and GIS.

[52]. Wilson, O.W. (1963). Police Administrator. USA, McGraw Hill Company. With special regard to economic factors, International Journal of Data Mining.

[53]. Witten, I. H. and Frank, E. (2000) Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations.

[54]. Zakaria suliman zubi and ayman altaher mahmmud (2014) Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention.

[55]. Letezgi, (2011). mining crime data for effective resource allocation and crime prevention. The case of addis ababa police office

[56]. Mugdha Sharma —Z - CRIME: A Data Mining Tool for the Detection of Suspicious Criminal Activities Based on Decision Tree 2014 IEEE.

[57]. Hailemariam Negussie (2015), Crime forecasting by using data mining techniques, the case of addis ababa police office

[58].K. sharma, S. Vashisht and R. Dhiman, "A hybrid Approach Using Rule Induction And Clustering Techniques In Terms Of Accuracy And Processing Time In Data Mining," International Journal of Soft Computing and Engineering (IJSCE), vol. 3, no. 1, pp. 146-148, 2013a.

[59].M. Anunachalam and S. Baboo, "Enhanced Algorithms to Identify Change in Crime Patterns.," International Journal of Combinatorial Optimization Problems and Informatics, vol. 2, pp. 32-38., 2011.

[60]. Y. Fu, Data Mining: Tasks, Techniques and Applications, Rolla : University of Missouri-Rolla ., 1997.

[61]. Kohavi, R. (1995) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo, CA, Morgan Kaufmann.

Appendix A: View of the J48 Experiment-1 Scenario-1 Pruned Tree

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: CRIMINALS DATA OF SHIRE TOWN POLICE OFFICE-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.RemoveUseless-M99.0

Instances: 9967

Attributes: 11

Offenders Crime Type

Time

Religion

Kebele

Occupation

Age

Sex

Educational Level

Marital Status

Year Occurred

Nationality

Test mode: 10-fold cross-validation

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

J48 pruned tree

Year Occurred <= 2007

| Age = Age2

| | Educational Level = illiterate

| | | Time = AM: M (112.0/3.0)

| | | Time = PM

| | | | Marital Status = Married: M (80.0/6.0)

| | | | Marital Status = Not Married

| | | | | Kebele = Dedebit: F (15.0)

| | | | | Kebele = Adi Ketibay: M (1.0)

| | | | | Kebele = Sihul: F (0.0)

| | | | | Kebele = Hibret: F (0.0)

| | | | | Kebele = Lekatit: M (4.0)

| | | | Marital Status = Married: M (0.0)

| | Educational Level = primary: M (194.0)

| | Educational Level = Secondary: M (165.0/20.0)

| | Educational Level = College: M (51.0)

| | Educational Level = University: M (34.0)

| | Educational Level = TTI: M (0.0)

| Age = Age3: M (138.0)

| Age = Age1

| | Offenders Crime Type = Beating/Causing Bodily Harm: M (80.0)

| | Offenders Crime Type = Conflict: M (57.0)

| | Offenders Crime Type = Destruction of Property: M (0.0)

| | Offenders Crime Type = Intimidation/Blasphemy: M (0.0)

| | Offenders Crime Type = Killing: M (0.0)

| | Offenders Crime Type = Rape: M (0.0)

| | Offenders Crime Type = Theft

| | | Educational Level = illiterate: M (11.0)

| | | Educational Level = primary

| | | | Kebele = Dedebit: F (2.0)

| | | | Kebele = Adi Ketibay

| | | | | Marital Status = Married: M (12.0)

| | | | | Marital Status = Not Married: F (25.0)

| | | | | Marital Status = Married: F (0.0)

| | | | Kebele = Sihul: F (10.0)

| | | | Kebele = Hibret: F (48.0/1.0)

| | | | Kebele = Lekatit: M (17.0/5.0)

| | | Educational Level = Secondary: M (1.0)

| | | Educational Level = College: F (0.0)

| | | Educational Level = University: F (0.0)

| | | Educational Level = TTI: F (0.0)

| | Offenders Crime Type = Traffic Accident: M (0.0)

| | Offenders Crime Type = Merage under age: M (0.0)

| | Offenders Crime Type = Ilegal Driving Licence: M (0.0)

| | Offenders Crime Type = preparing illegal document: M (0.0)

| | Offenders Crime Type = Ilegal Migration: M (0.0)

| | Offenders Crime Type = illegal weapon: M (0.0)

| | Offenders Crime Type = Suicide: M (0.0)

| | Offenders Crime Type = Trying to kill/Murder: M (0.0)

| | Offenders Crime Type = Violation Court Order: M (0.0)

Year Occurred > 2007

| Offenders Crime Type = Beating/Causing Bodily Harm

| | Time = AM: M (1244.0/8.0)

| | Time = PM

| | | Kebele = Dedebit: M (38.0)

| | | Kebele = Adi Ketibay: M (199.0/8.0)

| | | Kebele = Sihul

| | | | Age = Age2: M (134.0)

| | | | Age = Age3: F (13.0)

| | | | Age = Age1: M (5.0)

| | | Kebele = Hibret: M (163.0/1.0)

| | | Kebele = Lekatit: M (297.0/31.0)

| Offenders Crime Type = Conflict: M (1164.0/2.0)

| Offenders Crime Type = Destruction of Property: M (167.0)

| Offenders Crime Type = Intimidation/Blasphemy

| | Marital Status = Married

| | | Year Occurred <= 2012: M (237.0/21.0)

| | | Year Occurred > 2012

| | | | Age = Age2: F (37.0/9.0)

| | | | Age = Age3: M (7.0)

| | | | Age = Age1: F (0.0)

| | Marital Status = Not Married: M (421.0)

| | Marital Status = Married: M (12.0)

| Offenders Crime Type = Killing: M (24.0)

| Offenders Crime Type = Rape: M (74.0)

| Offenders Crime Type = Theft

| | Age = Age2

| | | Kebele = Dedebit

| | | | Time = AM

| | | | | Educational Level = illiterate

| | | | | | Marital Status = Married

| | | | | | | Year Occurred <= 2009: M (2.0)

| | | | | | | Year Occurred > 2009: F (22.0/5.0)

| | | | | | Marital Status = Not Married: M (34.0)

| | | | | | Marital Status = Married: M (1.0)

| | | | | Educational Level = primary: M (115.0)

| | | | | Educational Level = Secondary: M (35.0/8.0)

| | | | | Educational Level = College: M (0.0)

| | | | | Educational Level = University: M (0.0)

```
| | | | | Educational Level = TTI: M (0.0)
| | | | Time = PM: M (186.0/6.0)
| | | Kebele = Adi Ketibay
| | | | Educational Level = illiterate
| | | | | Occupation = Unemployed: M (170.0/8.0)
| | | | | Occupation = Private
| | | | | | Year Occurred <= 2010: F (23.0/7.0)
| | | | | | Year Occurred > 2010: M (12.0)
| | | | | Occupation = Governmental: M (0.0)
| | | | Educational Level = primary: M (149.0)
| | | | Educational Level = Secondary
| | | | | Time = AM: M (129.0)
| | | | | Time = PM
| | | | | | Occupation = Unemployed: M (54.0/9.0)
| | | | | | Occupation = Private: M (3.0)
| | | | | | Occupation = Governmental: F (9.0)
| | | | Educational Level = College: M (32.0)
| | | | Educational Level = University: M (16.0)
| | | | Educational Level = TTI: M (0.0)
| | | Kebele = Sihul: M (468.0/31.0)
| | | Kebele = Hibret: M (411.0/2.0)
| | | Kebele = Lekatit: M (376.0/8.0)
| | Age = Age3: M (307.0)
| | Age = Age1
| | | Year Occurred <= 2010: M (264.0)
| | | Year Occurred > 2010
| | | | Year Occurred <= 2011
| | | | | Educational Level = illiterate: M (5.0)
| | | | | Educational Level = primary: F (44.0/2.0)
```

| | | | | Educational Level = Secondary: F (0.0)

| | | | | Educational Level = College: F (0.0)

| | | | | Educational Level = University: F (0.0)

| | | | | Educational Level = TTI: F (0.0)

| | | | Year Occurred > 2011

| | | | | Kebele = Dedebit: M (28.0)

| | | | | Kebele = Adi Ketibay: M (43.0)

| | | | | Kebele = Sihul: M (7.0)

| | | | | Kebele = Hibret: M (4.0)

| | | | | Kebele = Lekatit

| | | | | | Educational Level = illiterate: M (2.0)

| | | | | | Educational Level = primary: F (14.0/1.0)

| | | | | | Educational Level = Secondary: F (0.0)

| | | | | | Educational Level = College: F (0.0)

| | | | | | Educational Level = University: F (0.0)

| | | | | | Educational Level = TTI: F (0.0)

| Offenders Crime Type = Traffic Accident: M (1437.0)

| Offenders Crime Type = Merage under age

| | Year Occurred <= 2016: M (12.0)

| | Year Occurred > 2016: F (2.0)

| Offenders Crime Type = Ilegal Driving Licence

| | Marital Status = Married: F (3.0/1.0)

| | Marital Status = Not Married: M (34.0)

| | Marital Status = Married: M (0.0)

| Offenders Crime Type = preparing illegal document: M (39.0/2.0)

| Offenders Crime Type = Ilegal Migration

| | Time = AM

| | | Marital Status = Married: F (7.0/2.0)

| | | Marital Status = Not Married: M (27.0/1.0)

| | | Marital Status = Married: M (0.0)

| | Time = PM: M (25.0)

| Offenders Crime Type = illegal weapon: M (81.0)

| Offenders Crime Type = Suicide: M (2.0)

| Offenders Crime Type = Trying to kill/Murder: M (15.0)

| Offenders Crime Type = Violation Court Order

| | Age = Age2: F (17.0)

| | Age = Age3: M (8.0)

| | Age = Age1: F (0.0)

Number of Leaves : 119

Size of the tree : 157

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	9716	97.4817 %
Incorrectly Classified Instances	251	2.5183 %
Kappa statistic	0.6258	
Mean absolute error	0.0429	
Root mean squared error	0.1509	
Relative absolute error	50.4066 %	
Root relative squared error	73.2354 %	
Total Number of Instances	9967	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.997	0.501	0.977	0.997	0.987	0.653	0.921	0.995	M
	0.499	0.003	0.884	0.499	0.638	0.653	0.921	0.621	F
Weighted Avg.	0.975	0.479	0.973	0.975	0.971	0.653	0.921	0.979	

=== Confusion Matrix ===

a	b	<-- classified as
9495	29	| a = M
222	221	| b = F