ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

CRIME FORECASTING BY USING DATA MINING TECHNIQUES: THE CASE OF ADDIS ABABA POLICE COMMISSION

BY

HAILEMARIAM NEGUSSIE GIFFAW

October 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

CRIME FORECASTING BY USING DATA MINING TECHNIQUES: THE CASE OF ADDIS ABABA POLICE COMMISSION

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

BY

HAILEMARIAM NEGUSSIE GIFFAW

October 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

CRIME FORECASTING BY USING DATA MINING TECHNIQUES: THE CASE OF ADDIS ABABA POLICE COMMISSION

BY

HAILEMARIAM NEGUSSIE GIFFAW

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
Ato Kidus Menfes	Chairman	_____	_____
Dr.Gashaw Kebede	Advisor	_____	_____
Dr.Tibebe Beshah	Examiner	_____	_____
Ato Getachew Jemaneh	Examiner	_____	_____

DEDICTION

I would like to dedicate this thesis work to Ethiopian victims who lost their life at Libyan Desert without justice.

ACKNOWLEDGEMENT

First, and foremost, I would like to thank the almighty God for giving me the opportunity to pursue my graduate study. I owe the deepest gratitude to Dr. Gashaw Kebede, my thesis advisor for his valuable and constructive comments and encouragements throughout my study. I would also thank all the staffs of Addis Ababa Police commission for their good will for letting me use the data for this study and provide me with information about crime and related concept.

My deepest thanks go to my new baby Birikti Hailemariam and my wife Yemata Kefelegne, who have supported me all the way to fulfill my dream, and I am highly indebted to my friends Daniel Mamo, Eyob Alemayehu, Almaz Abuhaye, Genet Getanehe, Mahelete Teferawork, Messay Yohannis, Messeret Assefa, Engidaw Serawitu, Aytenew Sewunet, Yisehak Adenew, Aklilu Yilma, Dessalegn Amsalu and all my precious instructors for their unreserved support and encouragement throughout my study.

Finally I am also grateful to Addis Ababa University especially Dr. Jelu Umer, Dr. Ahemed Hassen and all I.E.S staffs for all encouragement and support.

LIST OF ABBREVIATIONS

AAPHQ	Addis Ababa police head quarter
AGPS	'Addisu Gebeya' Police Station
CART	Classification and regression Trees
CIA	Central intelligent agency
CRISP -DM	CRoss-Industry Standard Process-Data Mining
CRM	Customer Relationship Management
CSV	Comma separated values
FBI	Federal bureau of investigation
GCPS	'Guelele' Command police Station
KDD	Knowledge discovery in database
KDP	Knowledge discovery process
KDPM	Knowledge discovery process models
KPS	'Kechene' Police station
MPS	**'**Menene' police station
NN	Neural network
OSAC	Overseas Security Advisory council
PART	Projective Adaptive Resonance Theory

PPS	‘Paster’ police station
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
SAS	Statistical analysis system
SEMMA	Sample, Explore, Modify, Model and Assess.
SPS	‘Sheromeda’ police station
SPSS	Statistical Package for the Social Sciences
SVM	Support vector machine
WEKA	Waikato Environment for knowledge Analysis

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

APPENDICES

ABSTRACT

Law enforcement agencies like that of police today are faced with large volume of data that must be processed and transformed into useful information and hence data mining can greatly improve crime analysis and aid in reducing and preventing crime. Knowledge discovery process has come across a variety of approaches. Hybrid methodology of Knowledge discovery process is one of the popular approaches used in Knowledge discovery process models. Hybrid knowledge discovery process model is a problem solving strategy for this study.

The purpose of this study is to identify the relation between offenders, victims and offences and to develop crime prediction model from the available data of Addis Ababa police commission. With this objective, J48 decision trees and PART rule induction algorithms were employed to generate patterns and classify crime records on the basis of the values of attributes CrimeLevel, Time, OffenderJob, Victimjob, OffenderMartialStatus and VictimMaritalStatus. Results of the experiments have shown that PART rule induction algorithm has classified crime records at an average accuracy rate of 88.6% when the above stated attributes were used as a basis for classification. In the experiments, the output indicated that PART rule induction algorithm performed better. The model has been evaluated on the testing dataset for crime level target class and scores a prediction accuracy of 97.2%. Besides, PART generated understandable rules that could be easily presented in human language. The research also demonstrates that crime type, age, educational level, job, marital status, sex, and particular place were the common demography factors of victims and offenders who were exposed to crime. The challenge in this study was, since the demographic similarity of both victims and offenders. So, further study has to be done for each alone.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Crime has been a part of society ever since laws were first introduced. Various scholars define crime in different ways. Crime is a behavior disorder that is an integrated result of social, economic and environmental factors [2].It is also defined by other scholars as an act committed or omitted in violation of a law forbidding or commanding it and for which punishment is imposed upon conviction [1].The later definition has been supported by the Webster's Dictionary [3], which define it as an act or the committing of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law.

Anunachalam and Baboo [4] stated that “An act of crime encompasses a wide range of activities, ranging from simple violation of civic duties (e.g., illegal parking) to internationally organized crimes (e.g., the 9/11 attacks)”.Different scholars categorized crime differently. For instance, Anunachalam and Baboo[4] categorized crime as: property (murder for gain, dacoity, robbery, burglary and theft), violent (murder, attempt to commit murder, hurt and riots), crime against women and child (rape, dowry death, molestation, sexual harassment, and cruelty by husband and her relatives).Other kinds of crime include kidnapping and abduction of others, criminal breach of trust, arson, cheating etc. Concern about national security has increased significantly since the terrorist attacks on 11 September 2001[5].Chen et al [5] stated that following the attack the Central Intelligence Agency (CIA), Federal Bureau of Investigation (FBI), and other federal agencies are

actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated local authorities to more closely monitor criminal activities in their own jurisdictions.

Regarding the necessity of crime analysis Zubi and Mahmmud [6] mentioned some important reasons as follows: The first reason is to analyze crime to inform law enforcers about general and specific crime trends, patterns, and series in an ongoing, timely manner. Second, to analyze crime to take advantage of the abundance of information existing in law enforcement agencies, the criminal justice system, and public domain. Third, to analyze crime to maximize the use of limited law enforcement resources. Fourth, to analyze crime to have an objective means to access crime problems locally, regionally, nationally within and between law enforcement agencies. Fifth, to analyze crime to be proactive in detecting and preventing crime. Six, to analyze crime to meet the law enforcement needs of a changing society. Seventh and the last reason is, to analyze crime to understand the criminal behaviors. There may be more other reasons depending on the community culture, but the most valuable reasons are those listed.

Historically, solving crimes has been the responsibility and duty of the criminal justice and law enforcement specialists. With the increasing use of the computerized systems to track crimes, computer data analysts have started helping the law enforcement officers and detectives to speed up the process of solving crimes [7].Criminology is an area that focuses the scientific study of crime and criminal behavior and law enforcement and is a process that aims to identify crime characteristic [8]. Criminology is one of the most important fields where the application of data

mining techniques can produce important results. Crime analysis, a part of criminology, is a task that includes exploring and detecting crimes and their relationships with criminals [9].

1.1.1 Data Mining and Crime Forecasting

Crime, in all its facets, is a well-known social problem affecting the quality of life and the economic development of a society [10]. Sharma and Kumar [2] stated that it is the social problem that costs our society greatly in several ways. For instance, victim costs (direct economic losses suffered by crime victims, including medical care costs, loosing earnings, and property loss or damage), criminal justice system costs (local, state, and federal government funds spent on police protection, legal and adjudication services, and corrections programs, including imprisonment), crime career costs (opportunity costs associated with the criminal's choice to engage in illegal rather than legal and productive activities), intangible costs (indirect losses suffered by crime victims, including pain and suffering, decreased quality of life, and psychological effects) etc.

Crime analysis involves exploiting data about crimes to enable law enforcement to better arrest criminals and prevent crimes [11]. Data used by crime analysts include data related with crime (date, type, location, time etc.), victims (name, job, gender, age etc.) and offenders (age, gender, name, job, etc.) [2]. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques [8]. Since solving crimes is a complex task that requires human intelligence and experience and data mining is a technique that can assist police force with crime detection problems [4].The Criminals have become technologically sophisticated in committing crimes and that is why crime is among the major social problems threatens the well-being of the society. It is the police forces and law enforcement bodies that are expected to fight against it. The knowledge that is

gained from data mining approaches is a very useful tool which can help and support police forces [10]. Therefore, law enforcement bodies (the police) supposed to use the current technologies (crime analysis tools) to fight crimes and to remain ahead of the everlasting race between criminals and the law enforcement bodies. In addition, to provide adequate service and protect the society from criminals in advance it is advised that the concerned body better use and apply technology that enable them to identify and detect crime patterns efficiently and take action in future before a crime is committed. Locating these patterns automatically is a challenge that machine learning tools and data mining analysis may be able to handle in a way that directly complements the work of human crime analysts [12].

1.2 Statement of the Problem

Criminals are active throughout Ethiopia. For example, homicide, pick pocketing, “snatch and run” thefts and other petty crimes are common in Addis Ababa. These are generally crimes of opportunity rather than planned attacks. These incidents have occurred in both the day and night times. Most of the time especially, Addis Ababa police do not give attention to the relation between profile of victims and offenders who were mostly affected by crime [56]. In addition, most of the new staffs of the police commission have no adequate idea about the place where and the times when most likely crime concentrated and committed. Often, they implement traditional responses such as increased police hiring and presence to provide rapid 991 (Free telephone number of Addis Ababa Police) response, random patrol, reactive arrests and analyze one crime or incident at a time. Some staffs of the commission do not have sufficient knowledge or their knowledge is not systematically organized for ease of use. As a result, they are not able to identify areas within a city

that have disproportionate amounts of crime and employ responses in those specific areas. They use and allocate resources without adequate analysis of crime distribution in the city.

On the other hand, police organizations in Addis Ababa have been handling a large amount of crime information and huge volume of records. Those records are extremely useful for improving their understanding of a range of crime and policing related issue. Therefore, police needs data mining techniques and tools which can help in analyzing criminal data and to control criminals as well as to remain ahead in the everlasting race between criminals and the law enforcement.

The researcher follows the most common quote to forecast crime that is simply to assume that the hot spots of yesterday are the hot spots of tomorrow [8].Therefore, the purpose of this study is to identify and investigate the demography factors of victims and offenders who were exposed to crime and to develop crime prediction model from the available data of Addis Ababa police commission. To this end, the study attempts to obtain answer for the following research questions:

- What type of relation can exist between offense and offenders?
- What type of demographic relation can exist between victims and the offenders?
- Which time and place where most of the offenders prefer to commit crimes?
- What are the most interesting patterns or rules generated using the determinant factor (attribute) of crime, victim and offenders?

1.3 Objective of the Study

The general and specific objectives of the research are described below.

1.3.1 General Objective

The general objective of the study is to identify the relation between offenders, victims and offences and to develop crime prediction model that could help the police commission of Addis Ababa in decision making for crime detection and prevention.

1.3.2 Specific Objectives

To accomplish the above stated objective the following specific objectives were accomplished:

- To conduct a thorough review of literature on the existing data mining techniques and methods in general, and their application in crime prevention in particular.
- To identify appropriate data mining algorithms and select the best algorithms.
- To select and extract the dataset required for analysis from the commission manual and computerized crime records.
- To Prepare the data for analysis which includes adjusting inconsistent data encoding, accounting for missing values, and deriving other fields from existing ones;
- To compare and suggest the best model for crime prediction for the study area.
- To report the result and forward recommendation.

1.4 Methodology of the study

Research methodology explains how the data is collected and analyzed so as to answer the stated research questions. This sub-section study contains what research method was used and the

choice of the study design and a strategy of data collection, management and analysis. The method used to attain the desired goal on a particular topic is one of the important tools in an academic research.

1.4.1 Research design

Research design is the strategic plan for a research project or research program, setting out the broad outline and key features of the work to be undertaken, including the methods of data collection and analysis to be employed, and showing how the research strategy addresses the specific aims and objectives of the study. This study follows Hybrid methodology of KDP (Knowledge discovery process) to achieve the goal of building predictive model using data mining techniques. It is the hybrid of the development of academic and industrial models. This is a model that combines aspects of both. It was developed based on the CRISP DM model by adopting it to academic research. The main differences and extensions include providing more general, research-oriented description of the steps, introducing a data mining step instead of the modeling step, introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains [13]. Hybrid process model is selected since it combines best features of CRISP-DM (Cross-Industry Standard Process for Data mining) and KDD (Knowledge Discovery in Database) methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives.

1.4.1.1 Tasks and methods used at each phase of the hybrid methodology

1.4.1.1.1. Domain understanding

In order to define the research problem properly, primary data was collected by interviewing concerned officers who are working in the police commission as well as through observation. Then based on the information obtained from these attempts, the overall crime prevention processes of the Addis Ababa police Commission was described. Relevant literatures on data mining techniques and crime were reviewed from books, journals, proceedings and the internet. The potential of data mining in general and particularly successful data mining application in crime prevention was assessed and described.

1.4.1.1.2. Data Understanding

The criminal records datasets and unpublished criminal records of the Addis Ababa Police Commission, particularly focusing on 'Guelele sub city', was the principal target data set to this study. From the Commission's dataset and unpublished crime records, 5106 records were used for the study. It only includes crime record from the year 2003 up to 2006 E.C. The data contains information about offenders (name, nationality, gender, age, educational status religion, and marital status), victims (name, nationality, gender, age, job, religion, marital status) and crime (date, time, name of unique location, 'kebele').

1.4.1.1.3. Data preparation

After the data was collected, orders such as processing and cleansing were performed in order to make the data more suitable for the particular data mining software used in the study. These comprised attributes selection, defining target classes, handling noisy data, accounting for missing data fields, coding text valued attributes and preparing the data to be processed in a file

format acceptable to the Waikato Environment for Knowledge Analysis (WEKA) software. Then, to partition the data to training and test dataset, cross-validation and percentage split methods was applied.

1.4.1.1.4. Model building

One of the critical tasks that were performed at model building stages is selection of software that supports the data mining techniques that were employed in the study. In conducting this research WEKA software was employed for reasons of accessibility and familiarity. WEKA software package has different programs for different techniques and algorithms. However, only two programs (sub packages) were employed in this study. The first one,J48 decision tree classifier and used for decision tree construction. The second, PART rule induction algorithm and produced a set of rules called decision lists which are ordered set of rules. Feature selection and model building was made iteratively by modifying the values of the parameters of decision tree and PART rule induction algorithm in order to improve the performance of the model.

1.4.1.1.5. Analysis and evaluation of the discovered Knowledge

In data mining evaluation has two primary functions. Primarily, it helps to predict how well the final model will work in the future. Secondly, evaluation is an integral part of many learning methods and helps to explore the model that best represents the training data [14].The different classification model developed in this research were evaluated using a test dataset based on their classification accuracy. In addition, evaluation of the performance of the classifier can also made in terms of different confusion matrices (True positive Rate (TPR), False Positive Rate (PRT), True Negative Rate (TNR), False Negative Rate (FNR), Relative Operating Characteristic (ROC), the number of correctly classified instances, number of leaves and the size

of the tree, execution time. Han et al [14] suggested the performance of the experiments as depicted below in Table 1.1.

Table 1.1 Performance measurement matrixes

Actual class		Predicted class		Total
		Negative	Positive	
	Negative	TN	FP	TN + FP
	Positive	FN	TP	TP + FN
Total		TN+FN	FP+TP	TN+FP+FN+TP

Sensitivity (TPR): referred to as the true positive (recognition) rate (i.e., the proportion of positive tuples that are correctly identified) i.e. TPR=TP/TP+FN. Specificity (TNR): is the true negative rate (i.e., the proportion of negative tuples that are correctly identified) i.e. TNR=TN/TN=FN.

False negative rate (FNR): the proportion of positive tuples that are erroneously classified as negative i.e. FNR=1-TPR or FNR=FN/TP+FN hence, the sum of TPR plus FPR equals 1.True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. True negatives (TN) are the negative tuples that were correctly labeled by the classifier. False positives (FP) are the negative tuples that were incorrectly labeled as positive. False negatives (FN) are the positive tuples that were mislabeled as negative.

Correctly classified instances (accuracy) compute the proportion of tuples that are correctly classified using this formula i.e. Accuracy= (TP+TN)/ (TN+FP+FN+TP).Incorrectly Classified Instance (Error Rate) are the proportion of tuples that are incorrectly classified. Error Rate= (FP+FN)/ (TN=FP=FN=TP).

Furthermore, we can also compute the effectiveness and efficiency of the model in terms of recall and precision. As a result, recall can be computed for positive and negative class. Thus, the formula of the recall for negative class will be TN/TN + FP and for positive class is TP/TP+FN. Similarly to compute the perception of the model, the following formula can be used for negative and positive class respectively, TN/TN+FN and TP/FP=TP.

But for this particular study the researcher focused on some parameters to evaluate the models. Some of them includes accuracy, execution time, number of leaves, size of the tree, average true and false positive rate, average precision and recall, average ROC,CCI, ICI and number of rules generated.

1.5 Significance of the Study

This study attempts to identify and investigate the demography factors of victims and offenders who were exposed to crime and to develop crime prediction model and hence law enforcement body can make use of these patterns in their day-to-day battle against crime. Police and other law enforcement officers in the crime prevention and investigation authority of the respective area can make use of the results of this study in order to make optimal use of resources in crime prevention and to protect the society in advance before accidents committed by the offender. Moreover, the output of the study can be used for designing appropriate training program and crime prevention and investigation strategies. It can also be used as a benchmark for police officials as well as a source of methodological approach for studies dealing on the application of data mining on crime management, forecasting or pattern detection in other similar areas. In general a crime prediction technique helps to visualize criminal network in advance, to reduce risk and increase crime analysts work productively.

1.6 Scope and limitation of the study

The scope of this study is delimited to Addis Ababa Police Commission 'Guelele Kefle Ketema' Command Post. It only includes the classification and analysis of crime data from the year 2003 up to 2006 E.C using offenders and victim's related information and deployment of appropriate data mining techniques for predictive data mining. Other data mining techniques such as association are not accomplished due to the time and budget allotted for the research. Similarly the study also delimited to those crimes committed by men whose minimum age was18. In addition, the study does not include crime records which had not got final judgment by court. Since the majority of crime records were found in hard copy formats, the study tooks more time from expected.

1.7 Thesis Organization

This thesis is organized into five chapters. The first chapter deals with the basic overview including background, statement of the problem, objective, significance, scope and limitation of the study, methodology and thesis organization. In the second chapter review of data mining literature and concept related with data mining and crime were presented. Overview of data mining, knowledge discovery process models, data mining functions, application areas of data mining technology, review of related works and type of crime was discussed here.

The third chapter deals with data preprocessing tasks. In this chapter how the major data preprocessing tasks were applied to the current data were shows. Data preprocessing data description, statistical description of attributes data cleaning and transformation. The fourth chapter deals with experimentation and result interpretation. In this chapter building of model with training dataset and validating the result with testing data set, and interpretation of the result

of the experimentation were the major concern. Finally, a comparisons of the algorithms used for reasonable accuracy was made. The last chapter is devoted for the final conclusion and recommendation based on the research findings.

CHAPTER TWO

LITERATURE REVIEW

2.1 Definition and overview of Data Mining

The fast developing computer science and engineering techniques have made information easy to capture process and store in databases [15]. Databases today can range in size into more than terabytes- 1,000,000,000,000 bytes of data [16]. Computerization of many business, scientific and governmental transactions, advances in data collection tools ranging from scanned text and image platform to satellite remote sensing systems popular use of the web as a global information system, are some of the contributing factors for the availability of large collection of data in different institutions.

Hanet et al [14] and Zaïane [17] stated that the fast growing tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human capability for comprehension that requires a powerful tool in order to process the data and gain the advantage. To get benefit from the collected data, there should be a way to identify relevant and useful information. As a result, data mining has become a research area with increasing importance. There are several definitions for data mining, but the following are the most used ones by the scientific community:

- Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data [18].
- Data mining refers to extracting or mining" knowledge from large amounts of data [14].

- Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data [19].
- Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data [19].

To sum up, data mining is a technology that is used to explore the hidden knowledge or pattern from huge data set and many other terms carry a similar meaning to data mining such as knowledge mining from data, knowledge extraction, data or pattern analysis, data archaeology, and data dredging.

2.2 Knowledge discovery process models (KDPMs)

In view of the fact that knowledge discovery is explained as a process, a thorough understanding of those sequences of steps that should be followed to discover knowledge in data is necessary before one goes into extracting knowledge from data. There are many process models designed to provide a roadmap to follow while planning and executing knowledge discovery projects. Effective use of these models helps to save cost and time, enables to better understand the project and leads to acceptable results [20]. In general, these process models are systematic approaches essential for successful data mining [21].Since then, several different KDPM models have been developed in both academia and industry.

2.2.1 Knowledge Discovery in Databases (KDD) Process Model

Fayyad et al [22] and Kahlon & Kaur [23] stated that KDD process is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It is iterative and interactive, consisting of several steps. Note that the process is iterative at each step, meaning that moving back to previous steps may be required. The process has many artistic aspects in the sense that one cannot present one formula or make a complete taxonomy for the right choices for each step and application type [24]. Thus it is required to understand the process and the different needs and possibilities in each step. As described by Fayyad et al [22] the KDD process consists of five steps.

The first step is Selection: it involves collecting data from possible sources to get the target data. It includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed. The second is preprocessing: it concerns with cleaning the target data so as to get preprocessed data. It includes data cleaning (such as dealing with missing data or error) and deciding on methods for modeling information, accounting for noise, or dealing with change overtime. The third is transformation: it is all about any sort of rearrangement of the preprocessed data so that it would be easy to apply the data mining tool. The fourth is Data mining**:** it involves applying appropriate data mining techniques and algorithms onto the transformed data to get interesting patterns of the data. The fifth and the last step is interpretation/evaluation: It includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by users. Finally, the discovered pattern is evaluated to give recommendations.

2.2.2 CRISP-DM

The CRISP-DM process was developed by the means of the effort of a consortium initially composed with Daimler Chryrler, SPSS and NCR. CRISP-DM stands for CRoss-Industry Standard Process for Data Mining [26].CRISP-DM is most popular process model that has proven application is Cross Industry Standard Process Data Mining (CRISP-DM).It is a comprehensive data mining methodology and process model that provides any one from beginners to data mining experts with a complete blueprint for conducting a data mining project[25]. CRISP-DM as a process model is the de facto standard for developing data mining and knowledge discovery projects [21]. It is a standard process model in industries which consisting of a sequence of steps that are usually involved in a data mining study. In CRISP-DM, the sequences of the phases are adaptive i.e. the next phase in the sequence often depends on the outcomes associated with the preceding phase.

Business understanding is the initial phase in the CRISP–DM standard process which focuses on understanding business area in which DM objectives and project requirements are assessed as a whole from business perspective points of view. Once the business is well defined and understood, data understanding phase begins with collecting the initial data and continues with several activities like collecting initial data for modeling; data description to get insights through descriptive statistics available in the statistical tools, exploration of data to capture an overall sense of the dataset through computing summary and lastly, verification, and visualization of data quality is checked if any unnecessary data fields with incomplete, inconsistent, noisy, and redundant values existed. The data preparation step contains all activities needed to construct the final dataset. It includes data selection, data cleaning, data construction which attempts to produce derived attribute, new records and transformed values for existing attributes, data

integration (combine data from multiple sources) , and, finally data formatting for reconstructing data values without changing its meaning. After data being ready to apply data mining tools in proceeding step, various modeling techniques are selected and applied at this phase. Since some data mining tools may require specific formatting for input, it may needs reiteration into the previous phases for improvement.

The modeling step selects first modeling techniques based on data mining objectives and also generates test set to evaluate and validate model performance. This is followed by model building using the modeling tool. Finally, assess and interpret the pattern according to the domain knowledge. After models have been built, they need to be evaluated to check whether they fulfill the requirements and objectives set at the beginning of the project. The model is also evaluated from the point of business objectives. Reviewing of steps executed to build the model and evaluating the models for quality and effectiveness before generating them for end users in the field are also performed. At the end, decision regarding the deployment and use of the data mining results is reached.

2.2.3 SEMMA

The SEMMA process was developed by the SAS (Statistical Analysis System) Institute [26]. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. Statistical Analysis System Institute [27] defines data mining as the process used to reveal valuable information and complex relationships that exist in large amounts of data. According to SAS institute to provide a methodology in which the data mining process can operate, it divides data mining phases into five stages that are represented by the acronym SEMMA.

The first phase in SEMMA process model is sampling. In this phase, a portion of a large dataset is extracted in order to take reliable and statistically representative sample from the huge data for optimal cost and computational performance. Sample data selection is followed by explore. This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas. This help to refine the data set and redirect the discovery process. The third phase in SEMMA is Modify. This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process. It also manipulates data to include information and handle outlier to increase significance of variables and focus on model selection process. The fourth phase is Model. This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

The last phase of SEMMA process model is Assess. This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

2.2.4 Hybrid Model

The development of academic particularly the KDD and industrial oriented (CRISP-DM) models has led to the development of hybrid models, that is, models that combine aspects of both [28]. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include providing more general, research-oriented description of the steps, introducing a data mining step instead of the modeling step and has several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and modification of the last step, since in

the hybrid model, the knowledge discovered for a particular domain may be applied in other domains [13]. Cios et al [28] stated the six steps of Hybrid model as follows:

Phase 1: Understanding of the problem domain. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

Phase 2: Understanding of the data. This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

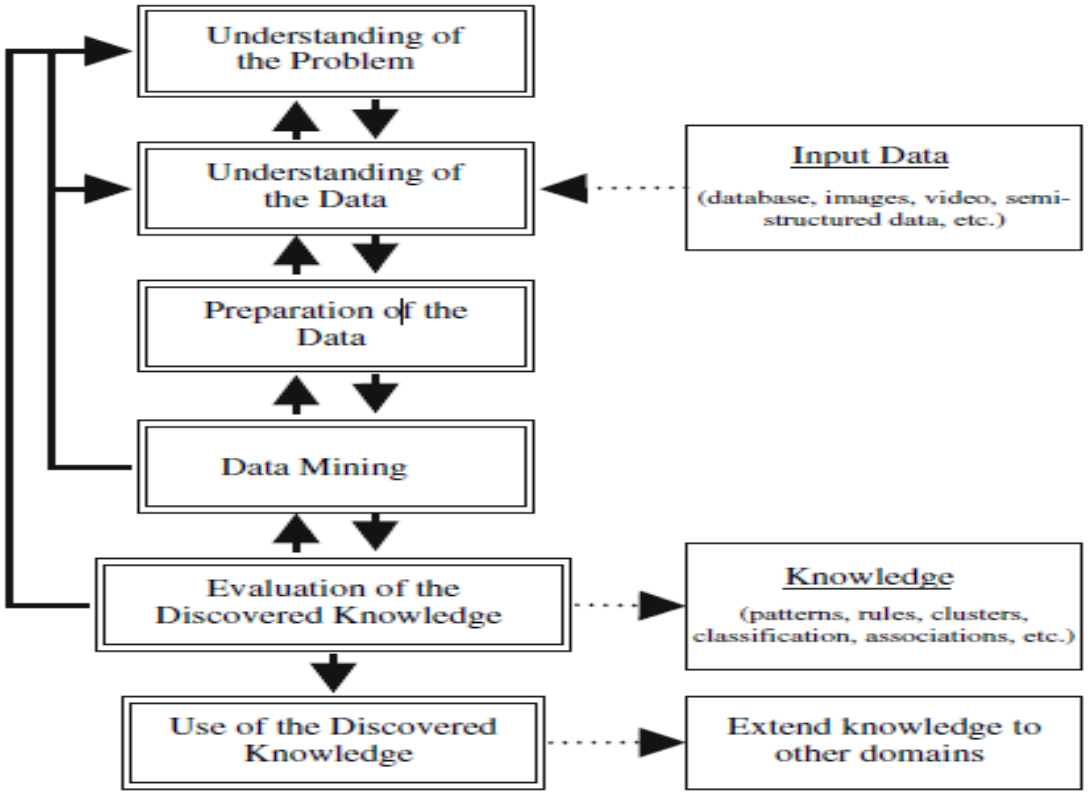
Phase 3: Preparation of the data. This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in phase 1.

Phase 4: Data mining. Here the data miner uses various DM methods to derive knowledge from preprocessed data.

Phase 5: Evaluation of the discovered knowledge. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

Phase 6: Use of the discovered knowledge. This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed. The following figure shows the six-step KDP model.

Figure 2.1 the six-step KDP model adopted from: Cios et al [28]



2.3 Data Mining Functions

The tasks of data mining are very diverse and distinct because there are many patterns in a large database. Different kinds of methods and technique are needed to find different kinds of patterns. Those data mining functionalities also used to specify the kind of patterns to be found in data mining [14].Based the above concept the goal of any data mining effort can be divided in one of the following two types. The first is using data mining to generate descriptive models to solve problems. The second is using data mining to generate predictive models to solve problems [29]. Thus, data mining tasks can be classified as descriptive data mining and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea [30].The descriptive data mining describe the data set in a concise

and summary manner and presents interesting general properties of the data. The descriptive data mining tasks characterize the general properties of the data in the database. It also focus on finding patterns describing the data that can be interpreted by humans, and produces new, nontrivial information based on the available data set. The goal of a descriptive data mining model is therefore to discover patterns in the data and to understand the relationships between attributes represented by the data.

The second category of data mining function is predictive data mining tasks. This data mining task constructs one or a set of models performs inference on the available set of data, and attempts to predict the behavior of new data sets. Predictive data mining tasks perform inference on the current data in order to make prediction. It involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of a predictive data mining model is to predict the future outcomes based on passed records with known answers. For this purpose it produces a model that can be used to perform tasks such as classification, prediction or estimation, regression and time series analysis.

2.3.1 Predictive Models

Predictive modeling is the event in which a model is made or chosen to accurately predict or forecast an outcome. It is therefore the process of analyzing the current and past states of the attribute and forecasting of its state.

2.3.1.1 Classification and Prediction

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models

describing important data classes or to predict future data trends [31]. Classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions.

Classification maps data into predefined groups or classes [32]. It is often referred to as supervised learning because the classes are determined before examining the data. They often describe these classes by looking at the characteristics of data already known to belong to the classes. Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes [33].

Prediction is also another mining activity, the goal is to predict, for a new record, the value of one of the attributes (i.e. target attribute) based on the values of the other attributes. The relationship between the target attribute and the other attributes is learned from a set of data in which the target attribute is already known (i.e. training data) [34]. The training data captures an experimental dependency between the normal attributes and the target attribute; so that the data-mining technique builds an explicit model of the observed dependency. This model can then be used to generate a prediction of the target attribute from the values of the other attributes for new records. There are many different methods, which may be used to predict the appropriate class for the objects or stations. Among the most popular are: logistic regression, decision tree, rule induction, case-based reasoning, and neural network [35].But the most widely used techniques for classification are decision trees, neural network, Bayesian Classification, and Support Vector Machines [36]. Even if, there are many data mining tasks available and commonly used in various applications, the researcher used decision tree and PART techniques in this research.

2.3.1.2 Decision Trees

Decision tree is a tree-like graph which is used for classification of data sets and for taking decisions in decision making system [37]. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label [38].A decision tree is a predictive modeling technique that used in classification, clustering and predictive task [39].The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called root that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes) [40].Decision tree uses a divide-conquer technique to split the problem search space into subsets. The most important feature of decision tree classifier is their ability to break down a complex decision making process into collection of simpler decision, thus providing solution which is easier to interpret [39].

According to Seema et al [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, it also available in many data mining packages over a variety of platforms and it is very useful for various tasks, such as classification, regression, clustering and feature selection.

I. Decision tree construction techniques

All decision tree construction methods are based on the principles of recursively portioning the data set until homogeneity is achieved [20].The construction of the decision tree involves the following three main phases [37].

Construction Phase: The initial decision tree is constructed in this phased, based on the entire training data set. It requires recursively partitioning the training set into two or more, sub partition using a splitting criterion, until a stopping criteria is met. The basic strategy of construction phase is described as follows. The tree starts as a single node representing the training sample. If the samples are of the same class, then the node become a leaf and is labeled with that class. Otherwise, the algorithm uses entropy-based measure known as information gains as a heuristic for selecting the attribute that best separate the samples into individual's classes. The information gain measure is used to select the test attribute at each node in the tree. The attribute with the highest information gain is chosen as the test attribute for the current node.

Pruning Phase: the tree constructed in the previous phase may not result in the best possible set of rules due to over-fitting. The pruning phase removes some of the lower branches and nodes to improve its performance.

Processing the Pruned tree: in this stage decision tree is proposed to improve understandability. Various listed decision tree algorithm such as CHAID, C4.5/C5.0,CART, ID3, J48 and many other with less familiar algorithms produce tree that differ from one another in the number of split allowed at each level of tree, how those splits are chosen when the tree is built.

II. Decision Tree Algorithms

Perhaps the most important thing to remember is that no one model or algorithm can or should be used exclusively. For any given problem, the nature of the data itself will affect the choice of models and algorithms we choose. There is no “best” model or algorithm. Consequently, you will need a variety of tools and technologies in order to find the best possible model [16].For creation of decision tree, many algorithms are used which gives good results. Data mining techniques are used for systematic analysis of large data sets. Decision tree is one of the most popular and efficient technique in data mining [16]. There are many algorithms for generation of classification trees [24, 38]. Some of them are: C4.5generates classifiers as decision trees in addition to that it also create classifiers in more logical rule set form. CART algorithm is Classification and Regression Trees which is used in the field of Artificial Intelligence, data mining and Machine Learning. J48 is an implementation of C4.5 algorithm. C4.5 was a version ofJ48. J48 uses two pruning techniques with bottom up strategy where nodes are replaced by leaf i.e. start from leaves and move towards root node. M5P algorithm is commonly used to develop regression trees whose leaves are combination of multivariate linear models. The nodes of the tree are chosen over the attribute by which error reduction can be done as a function of the standard deviation of output parameter. For this research J48 decision tree algorithm is applied.

To sum up, the decision tree algorithm was selected due to: its simplicity to understand, decision tree are easy to convert to a set of production rules, it can classify both categorical and numerical data, even if the output attribute must be categorical and there is no prior assumption about the nature of the data.

III. J48 Decision tree Algorithm

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 [41]. 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 [42].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. This algorithm 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. Figure 2.2shows basic pseudo code for J48 decision tree algorithm.

Figure 2.2 Basic pseudo codes for J48 decision tree algorithms [14]

Algorithm: Generate decision tree.

Input: Sets of training dataset (D), Attribute list, Attribute method;

Output: A decision tree.

Procedures:

(1) Create a Node N

(2) If tuples in D are of the same class, C then

(3) Return N as a leaf node labeled with the class C;

(4) If attribute list is empty then

(5) Return N as a leaf node labeled with the majority class in D; //majority voting

(6) Apply attribute selection method (D, attribute list) to find the "best" splitting criterion;

(7) Label node N with Splitting criterion;

(8) If splitting attribute is discrete-valued and multi way split allowed then//not restricted to binary trees

(9) Attribute list attribute list-splitting _ attribute;//remove splitting attribute

(10) For each outcome j of splitting criterion //partition the tuples and grow sub tree for each partition

(11) Let Dj be the set of data tuples in D satisfying outcome j;//a partition

(12) If Dj is empty then:

(13) attach a leaf labeled with the majority class in D to node N;

(14) Else attach the node returned by generate decision tree (Dj, attribute list) to node N; End for

(15) Return N;

Basic Steps in the Algorithm [42]:

- In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.
- The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

- Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

2.3.1.3 Naïve Bayes Classifier

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem [14].

Naïve Bayes classifier method is based on probabilistic knowledge. This method goes by the name Naïve Bayes, because it is based on Bayes's rule and naively assumes independence-it is only valid to multiply probabilities when the events are independent [43]. Thus, the Naïve Bayes rule output probability for the predicted class of each member of the set of test instance. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data and to perform statistical classification. In Naïve Bayes classifier, the probability of the attribute are calculated based on normal distribution's mean, standard deviation, weighted sum, and precision. So, the researcher tried to show the experiments on Naïve Bayes algorithm in order to get the best fitted model for the classification, prediction and statistical information of the crime dataset.

Moreover, Bayesian classifiers are statistical classifier. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Studies comparing classification algorithms have found a simple Bayesian classifier known as the Naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers [14]. Han et al [14] found that Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes.

This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve”.

2.3.1.4 Neural Network

Neural networks, also called artificial neural network (ANNs), are one of the most famous predictive models used for classification [44]. Neelamegam and Ramaraj [45]defined artificial neural network as a system that works based on the operation of biological neural networks, in other words, is an emulation of biological neural system. Artificial neural networks born after McCulloc and Pitts introduced a set of simplified neurons in 1943. These neurons were represented as models of biological networks into conceptual components for circuits that could perform computational tasks. Voznika and Viana, [46] stated that artificial neural network is developed with a systematic step-by-step procedure which optimizes a criterion commonly known as the learning rule. The input/output training data is fundamental for these networks as it carries the information which is necessary to discover the optimal operating point. According to Singh and Chauhan [30] there are various neural networks architectures, the most successful applications in classification and prediction have been multilayer feed forward networks. The layer where input patterns are applied is the input layer; the layer from which an output response is desired is the output layer. Layers between the input and output layers are known as hidden or transfer layers, because their outputs are not readily observable.

2.3.1.5 Rule induction

Another important machine learning techniques is the induction of rule sets. According to Sharma et al [47] rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of

the data, or merely represent local patterns in the data. Some major rule induction paradigms are: Association rule algorithms: In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Decision rule algorithms: Decision rules play an important role in the theory of statics and economics. In order to evaluate the usefulness of a decision rule, it is necessary to have a loss function detailing the outcome of each action under different states.

The learning of rule based models has been a main research goal in the field of machine learning since its beginning in the early 1960's [48]. Rule induction is an area of machine learning in which formal rules are extracted from a set of observations [47].A rule based classification models consists of a set of if-then rules. Each rule has a conjunction of attribute values in the conditional part of the rule, and a class label in the consequent.

Girma [20] mentioned that usually rules are expressions of the form if (attribute -1; value-1) and (attribute-2; value-2) and ….and (attribute-n; value-n) then (decision; Value). As Furnkranz et al [48] stated that the main difference between the rules generated by a decision tree and the rules generated by a rule learning algorithm is that the former rule set consists of non-overlapping rules that span the entire instance space (i.e. each possible combination of attribute values will be covered by exactly one rule). Relaxing this constraint by allowing for potentially overlapping rules that need not span the entire instance space that may often result in smaller rule sets, however, in this case we need a mechanisms for tie breaking (which rule to choose when more than one covers the example to be classified) and default classification (what classification to choose when no rule covers the given example).Typically, one prefers rules with a higher ratio of correctly classified examples from the training set.

To make the rule useful, we require two important information including accuracy and coverage. This is because the pattern in the database is expressed as rule does not mean that it is true all the time. Similarly to other data mining algorithms it is important to recognize and make explicit the uncertainty in the rule, i.e. how often is the rule correct. This is the "accuracy" of the rule means. The coverage of the rule has to do with how much of the dataset the rule "cover" or how often the rule applies [4]. Table 2.1 shows rule coverage versus accuracy.

Table 2.1 Rule coverage versus accuracy adapted from [49].

	Accuracy Low	Accuracy High
Coverage High	Rule is rarely correct but can be used often.	Rule is often correct and can be used often.
Coverage Low	Rule is rarely correct and can Only be used rarely.	Rule is often correct but can only be used rarely.

2.3.1.5.1 Rule induction for prediction

The goal of data mining is to extract valuable information from one's data, to discover the 'hidden gold'. In decision support management terminology, data mining can be defined as 'a decision support process in which one search for patterns of information in data' [4]. Rule induction is a data mining techniques used to extract classification rule of the form IF(conditions) THEN (predict class) from data.

Data mining techniques are based on data retention and data distillation. Rule induction models belong to the logical, pattern distillation based approaches of data mining. These technologies extract patterns from data set and use them for various purposes, such as prediction of the value of a dependent field.

After the rules are created and their interestingness is measured there is also a call for performing prediction with the rules. Each rule by itself can perform prediction; the consequent is the target and the accuracy of the rule is the accuracy of the prediction. But because rule induction systems produce many rules for a given antecedent or consequent there can be conflicting prediction with different accuracy [4].This can be done in a variety of ways by summing the accuracies as if they were weight or just by taking the prediction of the rule with the maximum accuracy [49].

2.3.1.5.2 Rule Induction Algorithm

The term rule induction algorithm is often used to refer to an algorithm which discovers a rule set somewhat more “flexible” than a decision tree, in a sense that the discovered (induced) rules cover data space regions that can have some overlapping [50]. Hence a data instance can be covered by more than one rule. In contrast; a decision tree produces rules representing data partitions that are mutually exclusive and collectively exhaustive. Hence each data instance is covered by exactly one rule.

In other words, there is a certain difference in the kind of model induced by the two kind of algorithm. Although in both cases the model is essentially a rule set, the rule set induced by rule induction algorithms can be considered more flexible than the rule set induced by decision tree building algorithm. Thearling [49] stated that in particular, the latter can be considered a special case of the former where there is no overlapping between rules.

An induction algorithm takes as input specific instance and produces a model that generalizes beyond these instances. There are two major approaches by induction algorithm: supervised and unsupervised. In the supervised approach, specific examples of a target concept are given, and the goal is to learn how to recognize members of the class using the description attribute. In the

contrary, unsupervised approach is provided with a set of example without any prior classification, and the goal is to discover underlying regularities or patterns by identifying clusters or subsets of similar examples [50].

This research is deal with supervised induction algorithm method because of simplicity and the goal of the research is prediction about the labeled and known class.

Rule induction seeks to go from the bottom up and collect all possible patterns that are interesting and then later use those patterns for some prediction target. It also retains all possible patterns even if they are redundant or do not aid in predictive accuracy. Hence, in a rule induction system if there were two columns of data that were highly correlated (or in fact just simple transformations of each other) they would result in two rules [49].

Rule induction is also known as **Separate-And-Conquer method**. The term separate-and-conquer has been coined by Pagallo and Haussler in 1990 because of the way of developing a theory that characterizes this learning strategy: learn a rule that covers a part of a given training examples, remove the covered example from the training set (the separate part) and recursively learn another rule that cover some of the remaining examples (the conquer part) until no example remain [47]. That is to say this method apply an iterative process consisting of first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process [47]. Some examples of these kinds of systems which are supported by WEKA software are discussed below:

OneR: OneR or “One Rule” is a simple algorithm proposed by Holt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its ‘one rule’. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random [52].

PART: PART is a separate-and-conquer rule learner proposed by Witten and Frank [53].The algorithm producing sets of rules called ‘decision lists’ which are ordered set of rules. Anew data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds partial C4.5 decision tree in each iteration and makes the “best” leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning [53].

Decision Table: It is the method used to build a complete set of test cases without using the internal structure of the program in question. In order to create test cases we use a table to contain the input and output values of a program. It summarizes the dataset with a ‘decision table’ which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates as maller and condensed decision table.

2.3.1.6 Support vector machine (SVM)

Support Vector Machines (SVM) is one of the most recent Data mining techniques used for classification, developed by Cortes and Vapnik in 1995 for binary classification [54]. Support vector machine have been developed in the framework of statistical learning theory, and have been successfully applied to a number of applications, ranging from time series prediction, to face recognition, to biological data processing for medical diagnosis [55]. SVM classification finds the hyper -plane where the margin between the support vectors is maximized. If all classifications contain two-class dependent variables with two predictors, then the points of each class could be easily divided by a straight line. Support vector machine is an algorithm for the classifications of both linear and nonlinear data [54].

2.3.1.7 Regression

The regression involves the learning of functions that maps data item to real valued prediction variables. Given a set of data items, regression is the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic production of a model that can predict these attribute values for new records [14].

2.3.1.8 Time series analysis

In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measure are used to determine the similarity between different time series, the structure of the line is examine its behavior and the historical and time series plot is used to predict future values of the variables [56].Time series forecasting predicts unknown future values based on a time-varying series of predictors [16]. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time,

especially the hierarchy of periods (including such varied definitions as the five or seven day work week, the thirteen month year, etc.), seasonality, calendar effects such as holidays, date arithmetic, and special considerations such as how much of the past is relevant.

2.3.2. Descriptive Modeling

Descriptive data mining model is the unsupervised learning functions. These functions do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness etc., of the data. It presents the main feature of the data or a summary of the data. A descriptive modeling techniques, such as summarization, Association rule, sequence analysis and clustering which produces classes (or categories) which are not known in advance.

2.3.2.1 Clustering

Clustering is the identification of classes, also called clusters or groups, for a set of objects whose classes are unknown [29]. Given a set of data items, partition this set into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar. For example, given a data set of customers, identify subgroups of customers that have a similar buying behavior.

2.3.2.2 Summarization

It maps data into subsets with associated simple descriptions. Basic static such as mean, standard deviation, variance, mode and median can be used as summarization approach [57]. Before building good predictive model one must understand the data. Summarization involves methods for finding a compact description for a subset of data by gathering a variety of numerical summaries (including descriptive statistic such as averages, standard deviation and graph) and looking at the distribution of the data [29].

2.3.2.3 Association rules

Association is the discovery of togetherness or connection of objects. Such kind of togetherness or connection is termed as association rule [29]. An association rules is a popular technique for market basket analysis because all possible combination of potentially interesting product groupings can be explored [56].The investigation of relationships between items over a period of time is also often referred to as sequence analysis.

2.3.2.4 Sequence analysis

It is used to determine sequential pattern in data. The pattern in the dataset is based on time sequence of actions, and they are similar to association data, however the relationship is based on time. In market basket analysis, the item are to be purchased at the same time, on the other hand, for sequence analysis the item are purchased over time in some order [14].

2.4 Application Areas of Data Mining Technology

Many businesses and scientific communities are currently employing data mining technology. Their number continues to grow, as more and more data mining success stories become known[54].Many business companies have been exploiting the advantage of data mining for many years. As a result it is becoming increasingly popular. The possible benefit of data mining can be control costs as well as contribute to income increases [16].

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. By determining characteristics of customers by their profile, an organization can

focus prospects with similar characteristics. After profiling the characteristics of customer, the company can identify those who have bought a particular product.

Data mining offers value across a broad spectrum of industries. In the banking industry, data mining is used heavily in the areas of modeling and predicting credit fraud, in evaluating risk, in performing trend analyses, in analyzing profitability, as well as in helping with direct marketing campaigns [54].

Two Crows Corporation [16] stated that telecommunications and credit card companies are two of the leaders in applying data mining technologies. The telecommunication industry has quickly evolved from offering local and long distance telephone services to providing many other comprehensive communication services including voice, fax, pager, cellular phone, images, e - mail, computer, and Web - data transmission, and other data traffic. The integration of telecommunications, computer networks, Internet, and numerous others means of communication and computing is under way [54].

Therefore, development of telecommunication services creates a great demand for data mining to help understand the new business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of services. In general, the telecommunications industry is interested in answering some strategic questions through data - mining applications such as: How does one retain customers and keep them loyal as competitors offer special offers and reduced rates? What characteristics indicate high - risk investments, such as investing in new fiber optic lines? How does one predict whether customers will buy additional products like cellular services, call waiting, or basic services? What characteristics differentiate our products from those of our competitors?

Moreover, the potential applications of data mining technology can be applied in other sectors such as airlines, retail industry, medical area etc.

2.5 Review of related works

Different scholars conduct research on crime forecasting and data mining techniques. Yu et al [58] on their research article entitled “crime forecasting using data mining techniques “tried to compare some techniques that enable them to forecast crime adequately. The main objective of their research was to develop a crime forecasting model in collaboration with the police department of United State city in the North-east. On their studies they used classification approach and they were tried to test and compare five algorithms namely SVM, J48, neural, naïve approach and One Nearest Neighbor (lNN).At last they were proposed the best forecasting approach to achieve the most stable outcome. Regarding the overall classifier performance is concerned they noted that J48, Neural, and SVM classifiers consistently outperform the naive and 1NN approach. Moreover, Neural often performs slightly better than J48 and SVM. They were suggesting that neural networks are better when modeling complex systems.

The other study was conducted by Nath [59].His article was entitled “crime pattern detection using data mining”. The main objective of his article was to use clustering algorithm for a data mining approach to detect the crime patterns and speed up the process of solving crime. Here, he was applied k-means clustering technique; it is one of the most widely used data mining clustering techniques. This algorithm groups hundreds of crimes into some small groups or related crimes; it makes the job of the detective much easier to locate the crime pattern. The scholar use real crime data from sheriff’s office (Law enforcement agencies of California).As a future work the researcher stated that he will create models for predicting the crime hot-spots

that will help in the deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources.

Researchers Zubi et al [60] on their article tried to study how to analyze crime pattern by using data mining. Their article entitled as “using data mining techniques to analyze crime patterns in the Libyan national crime data.” the main objective of their study was to help the Libyan government to make a strategically decision regarding prevention the increasing of the high crime rate during the period. Their paper presents a proposed model for crime and criminal data analyzes using simple k-means algorithm for data clustering and Aprior algorithm for data association rules. According to those scholars data for both crimes and criminals were collected from police departments’ dataset to create and test the proposed mode (350 records with 7 attributes), and then these data were preprocessed to get clean and accurate data using different preprocessing techniques (cleaning, missing values and removing inconsistency).The preprocessed data were used to find out different crime and criminal trends and behaviors, and crimes and criminals were grouped into clusters according to their important attributes. WEKA mining software and Microsoft excel were used to analyze the given data. Finally they conclude that the attributes for crime and criminal and the results of k-means algorithm shows a promising result of their proposed model. But they did not state any confusion matrix result that means accuracy, precision, recall and f- measure etc.

The other local researcher who conducts research on application of data mining in crime prevention was Leul in 2003[61].The purpose of his study was to explore the applicability of data mining technique in the efforts of crime prevention with particular emphasis to the Oromia Police Commission and to build a model that could help to extract crime patterns. With this objective he used decision trees and neural network data mining techniques to classify crime

records on the basis of the values of attributes crime label and crime scene (SceneLabel).His experiments result show that decision tree has classified crime records at an accuracy rate of 94 percent when the attribute CrimeLabel is used as a basis for classification. Whereas, in the same experiment, the accuracy rate of neural networks is 92.5 percent. On the other hand, in the case of classification of records on the values of the attribute SceneLabel decision tree has shown an accuracy rate of 85 percent while neural network revealed 80 percent. In both experiments his work indicated that decision tree performed better. Besides, decision tree seem more appropriate for the domain problem. Finally the researcher reviewed other local researcher Letezgi Hagos (2011), who conducts research on mining crime data for effective resource allocation and crime prevention in the case of Addis Ababa Police commission The purpose of the study was to extract meaningful crime trends regarding offences against children from the data in existing police records with the help of data mining techniques. With this objective she used J48 decision trees for classification, K-means algorithm (clustering) and appriori algorithm for association rule). Her experiments result show that from all the crime categories in the crime records sexual assault has the highest number and best rules are generated related with sexual assault. The researcher used six-phase CRISP-DM for data mining process.

2.6. Types of Crime

Various scholars tried to categorize crimes in different way. The common and major types of crimes are categorized as follows:

2.6.1 Crimes against Persons

Anunachalam and Baboo [4] and Douglas et al [62] mentioned seven crimes under the category of crimes against persons. The first is Murder; it is the willful killing of one human being by

another. Second is aggravated assault**;** it is the unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury. Third is forcible sex offenses**;** it is any sexual act directed against another person, forcibly and/or against that person's will. Fourth is non-forcible sex offenses**;** it is the unlawful, non-forcible sexual intercourse. This includes incest; where persons are related to each other and statutory rape where the victim is under the statutory age of consent. Fifth is kidnapping or abduction; it is the unlawful attack, transportation and/or imprisonment of a person against his or her will or a minor without the consent of a legal guardian or parent. Six is simple assault**:** it is the unlawful physical attack by one person upon another where no weapons are involved and the victim does not have severe bodily injury. The last is intimidation or threats; it is unlawfully place another person in reasonable fear of bodily harm through words or conduct but without displaying a weapon or attacking the victim.

2.6.2 Crimes against Property

Under the category of crimes against property Anunachalam and Baboo [4] and Douglas et al [62] describe thirteen crime types. First is arson/inflammable: it is the willful or malicious burning or attempting to burn, with or without intent to defraud, a house, public building, motor vehicle or aircraft, personal property of another, etc. Second is bribery, it is the offering, giving, receiving or soliciting of anything of value to influence the judgment or action of a person in a position of trust or influence. Third is burglary; it is the unlawful or forcible entry or attempted entry of organization with the intent to commit an offense therein. Fourth is counterfeiting or forgery; it involve the altering, copying, or imitating of something, without authority or right, with the intent to mislead or defraud by passing the copy or thing altered or imitated as if it were original or genuine; or the selling, buying, or possession of an altered, copied, or imitated thing with the intent to deceive or defraud. Fifth is criminal mischief or damaged property**;** involves

acts that willfully or maliciously destroy, injure, disfigure, or deface any public or private property, real or personal, without the consent of the owner or person having custody or control by cutting, tearing, breaking, marking, painting, drawing, covering with filth, or any other such means as may be specified by local law. Six is embezzlement; it is the unlawful misappropriation or misapplication by an offender to his or her own use or purpose of money, property, or some other thing of value entrusted to his or her care, custody, or control.

The seven is extortion; it is to unlawfully obtain money, property or any other thing of value either tangible or intangible through the use or threat of force, misuse of authority, threat of criminal prosecution, and threat of destruction of reputation or social standing or through other coercive means. Eighth is fraud; it is the intentional perversion of the truth for the purpose of inducing another person or other entity in reliance upon it to part with something of value or to surrender a legal right. Ninth is larceny; it is the unlawful taking, carrying, leading, or riding away of property from the possession or constructive possession of another. This crime category includes shoplifting, pocket-picking, purse-snatching, bicycle thefts, and so forth, in which no use of force, violence, or fraud occurs. Ten is theft from motor vehicle; it refers to the theft of parts from a motor vehicle, whether locked or unlocked or the theft of any part or accessory affixed to the interior or exterior of a motor vehicle in a manner that would make the item an attachment of the vehicle or necessary for its operation. Eleventh is motor vehicle theft; it is the theft or attempted theft of a motor vehicle. The twelve is robbery; it is the taking or attempting to take anything of value from the care, custody, or control of a person or persons by force or threat of force or violence and/or by putting the victim in fear. The last is stolen property**;** it include the buying, receiving, possessing, selling, concealing, or transporting of any property with the

knowledge that it has been unlawfully taken, as by burglary, embezzlement, fraud, larceny, robbery, etc.

2.6.3 Crimes against Society

Douglas et al [62] mentioned five crime types under this category. The first crime type is drugs or narcotics violations; this kind of crime includes the production (cultivation and/or manufacture), transportation or importation, distribution or sale, purchase, possession, or use of any controlled drug or narcotic substance. The second is gambling; it is illegally bet or wager money or something else of value, assists, promote or operate a game of chance for money or some other stake. The third is child pornography; it is the violation of law prohibiting the manufacture, publishing, sale, purchase or possession of sexually explicit material of children. The fourth is prostitution; it is the unlawful promotion of or participation in sexual activities for profit, including attempts. The last is weapon law violations; it include the violation of laws or ordinances prohibiting the manufacture, sale, purchase, transportation, possession, concealment, or use of firearms, cutting instruments, explosives, incendiary devices, or other deadly weapons.

2.6.4. Internet-Related Crime/Cyber Crime

Internet-related crime is a term used to describe a range of different crime types that are committed or facilitated online, including: paedophilia, internet fraud, junk email or spam, viruses, and Hacking. This sort of crime is also referred to as cybercrime, e-crime and hi-tech crime. The cost of Internet crime in human and economic terms is high, and it‘s still growing. \

2.6.5 All Other Offenses

Anunachalam and Baboo [4] stated six crime types under this category. The first is fraud -non sufficient fund-closed account; it is to report a check written or other payments made on a closed or non-sufficient funds account. The second is curfew or embargo; it involves violations by juveniles of local curfew ordinances. The third is disturbing the peace; it is any behavior that tends to disturb the public peace or shock the public sense of morality. The fourth is family Offenses: it is unlawful, nonviolent act by a family member (or legal guardian) that threaten the physical, mental, or economic well-being or morals of another family member and that are not classifiable as other offenses, such as sex assault or sex offenses. The fifth is drunkenness; it is the violation of laws prohibiting the manufacture, sale, purchase, transportation, possession or use of alcohol beverages. The last is other sex offenses; which includes fondling, offensive exposure, window peeping, failing to register as a sex offender and child enticement. There are also others crimes like violation of court order, harassment, and any violations of state or local laws.

CHAPTER THREE

DATA UNDERSTANDING AND PREPROCESSING

3.1 Overview

This chapter discusses about the data understanding and preprocessing tasks of the data mining. In addition, to solve the problems mentioned in chapter one, the following approaches have been implemented. Those are embraces detail description of the dataset: such as the data cleaning, transformation and integration activities is explained in this chapter.

3.2. Understanding of the problem

Addis Ababa police commission has primary and secondary goals and objectives. On the aspect of primary goals and objectives, maintaining order and protecting life and property are the most basic functions. The secondary goals and objectives are intended to meet the primary goals. It include:-preventing crime, arresting and prosecuting offenders, recovering stolen and missing property, assisting the sick and injured, enforcing non-criminal regulations and delivering services not available elsewhere in the community.

In one direction or another, the police have vast jurisdictions and it seems no clear demarcation of the police functions. However, the argument is that the police should attempt to prevent crime through routine patrol, responding to calls and establishing partnership with the community to prevent crime. On the other hand, arresting offenders and assisting prosecutors in bringing charges against defendant is also the primary functions. But, to accomplish its duty Addis Ababa Guelele police command post has a number of problems. Some of which were delay in crime

investigation process, work load and inadequate knowledge and skill lack of sufficient budget, lack of trained detectives, misconducts and unfair treatment by the Police and lack of infrastructure capacity.

In a normal circumstance the police provide crime prevention and control services to community in police stations and off police stations. They were always tried to analyze and respond for a crime or incident at a time. Beyond that there is no way of identifying the demographical factors of offenders and victims rather keep their profile in manual record and computer. Basically Guelele police analyze reports by using traditional method of data analysis such as MS Excel. However, such method of data analysis has limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional database [16]. Besides, identifying those patterns of crime attribute is difficult because the dataset contains or involves too many attribute or parameter. So studying the demographical behavior and pattern of offender and victims attribute that have relationship among each other rather than listing are more crucial and support decision making.

Consequently, applying DM technology can handle and generate such pattern of variables that offenders and victims and used the output by concerned body to facilitate decision making process. Hence, is main factor that the researcher initiated to conduct this research.

3.3. Understanding of the Data

According to Cios et al [28] model, the next phase after understanding the business and problem domain has been data understanding. Hence, prerequisite for understanding Data Mining (DM) research was dealing with data itself. For this reason, Berry and Linoff [64] described that the identified good source of data to attain the DM goals has been the corporate data warehouse.

3.3.1 Data Source and Collection

There are two main tasks in data mining. The first task is coming up with precise formulation of the problem we are trying to solve. The second task is using the right data. From the above main tasks analyzing and understanding the content and structure of the collected data is one of the most important tasks that need attention in hybrid Data Mining process model [29].

The data for this research has been collected from Addis Ababa police Commission unpublished crime records as well as from their cyber software which was installed by Ministry of justice. Their cyber software system is a local developed system, which is suited for handling crime related data. It has user interface that allows the data encoder to enter crime's information. But while this research was conducted the cyber systems only works and uses at 'Menene' and 'Sheromeda' police stations the rest was under maintenance. The data includes both the victims as well as the criminal's profile. It also contains the time and the place where the crime occurred. The data includes crimes which have got the final judgments or decisions from the year 2003 - 2006 E.C. This data does not include crimes which have not got final decision by court or prosecution and the crime committed by the criminals whose age is less than 18.

3.3.2 Description and Quality of Data

The number of crime records available in Guelele sub city from 2003-2006 E.C was 10,483. From those records around 5,106(48.7%) records already got their finale judgments by court and prosecution. This research was conducted based on the crime data which got the final decision by the intended court and prosecution. The data has 42 attributes and the researcher added one additional attribute based on the literature of crime category. However, with the discussion of domain experts, only relevant attributes were selected to perform data mining tasks.

As Han et al [14] stated, data quality can be measured in terms of accuracy, completeness consistency, timelines, believability and interpretability. Although, Addis Ababa police Commission relatively uses the same crime record form and have the same information requirement; the commission's data suffer from incompleteness and inconsistency problem. Since data quality can also be affected by the structure of the data being analyzed, the researcher gave due emphasis to this aspect of the data too. The lack of data standard in using abbreviation and human error are significantly available and need to be cleaned and standardized before applying data mining techniques. Table 3.1 shows the description of police station and number of crime record it holds.

Table 3.1 Description of Data Sources and Number of records

Source of Data Center		Data Coverage E.C	Total number of crime record 2003-2006	Number of record ,which got final decision from court	Number of Attribute	Size of the data	Data Type
AddisAbaba police Commission (Head Quarter)		2003-2006	227	158	42	1.26 MB (1,327,497 bytes	Nominal and Numerical
Guele Command Post		2003-2006	1,023	620	42		Nominal and Numerical
Police Stations	Paster	2003-2006	3,204	1,650	42		Nominal and Numerical
	Menene	2003-2006	1,908	808	42		Nominal and Numerical
	Sheromeda	2003-2006	1,492	631	42		Nominal and Numerical
	AddisuGebeya	2003-2006	1,530	858	42		Nominal and Numerical
	Kechene	2003-2006	899	381	42		Nominal and Numerical
Total			10,483	5106			

3.3.3. Attribute Selection and Descriptive Statistical summary

Even though the majority of the data was collected manually from unpublished record, the initial dataset has been described and visualized using Microsoft Excel to examine the properties of the dataset. Simple statistical analysis has been performed to verify the quality of the data set such as missing values, error values and to obtain high level information regarding the data mining questions. Hence the selected attributes used for model building are statistically described below.

3.3.3.1 Attribute Selection

Attribute can be ranked either manually with the help of domain experts or automatically by using the application tools like WEKA attribute selection. From the WEKA attribute selection criteria the most common ones which are used for decision tree building are information gain attribute evolution and gain ratio attribute evaluation. However, in real world it is common to find a number of irrelevant attribute that could be easily known their irrelevance before adopting any complicated techniques. In addition to that attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean [53]. Thus, it is important to exclude those attribute that are not important for analysis in order to simplify the task of decision tree. For this research the researchers prefers two selection strategies in order to select attributes. The first way is identifying attributes by consulting and discussing with domain experts. The second mechanism is consulting other similar literatures. For the second way the researcher refer and use some attributes used by Leul (2003).

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

Crime Profile: “Crime Code”, “Crime Type”, “Date”, “Time”, “Particular Place”, “Crime Level”, “Crime Registration Date”, “Crime day-situation number”, “Crime Record Number”, “Sub city’, “Kebele” where the crime was committed and “crime Category”.

Victim’s Profile: “Victim Sex”, “Victim Age”, “Victim Job”, “Victims Religion”, “Victim Marriage Status”, and “Victim Nationality”.

Offenders profile: “Offender Sex”, “Offender Age”, “Offender Job”, “Offender Education Level”, “Offender Religion”, “Offender marital Status”, “number of offender involved”, “Offender seized Condition”, “Date of offender identified lately”, “number of offenders late identified”, “number of Offenders lately arrested”, “number of Offenders lately not arrested”, “Date of Offenders arrested lately”, Offenders Nationality”, Offenders current Condition”.

Others: “Judgment”, “Body Arrested Offender”, “Price of stolen property”, “Price of stolen property arrested”, “price of stolen property returned to owner”, “Date the cases sent to court”, “Name of legal institution”, “Date of Decision”, “Decision Body”, “Police Station”.

As mentioned earlier the crime data set has 43 attributes of text, number, date and time formats. Among these attributes the name of victims and offenders hasn’t been given by the commission for the sake of privacy. Out of 43 attributes 19 are selected based on their relevance for the research objective and opinion of domain expert. Table 3.2 shows general description of all attributes.

Table 3.2Description of the whole attributes of the study.

Attribute Name	Data Type	Description
CrimeCode	Number	A code assigned to different crime types to differ them one crime type from the others.
CrimeType	Nominal	Types of crime occurred.
Date	Date	The exact date on which the crime occurred.
Time	Time	The time at which the crime occurred.
ParticularPlace	Nominal	The place where the crime occurred
VictimSex	Nominal	The sex of the victim who was affected by the crime.
VictimAge	Number	The Age of the victim, who was affected by the crime.
Victim Job	Nominal	The occupation of the victim, who is affected by the crime.
VictimsReligion	Nominal	The religion of the victim, who is affected by the crime.
VictimMarrtialStatus	Nominal	The marital status of the victims.
OffenderSex	Nominal	The sex of the criminal who cause crime.
OffenderAge	Nominal	The sex of the criminal who cause crime.
OffenderJob	Nominal	The occupation of the offender, who cause crime.
OffenderEduL	Nominal	The level of education of the offender causing the crime.
OffenderReligion	Nominal	The religion of the offender, who caused crime.
OffendermarrtialStatus	Nominal	The marital status of the victims.
Judgmentpassed	Nominal	The decision passed by the court and prosecution.
CrimeLevel	Nominal	The level or the degree of crime.
CrimeRegistrationDate	Date	The date, when the crime registered.
Crimeday-situationnumber	Number	The record number of the document, which narrate how the crime happened.
CrimeRecordNumber	Number	The serial number of the crime record.
Subcity	Nominal	The name of sub city where the crime occurred.

KebeleCrimeOccured	Number	The kebele, where the crime occurred.
VictimNationality	Nominal	The Nationality of the victims.
Noofoffenderinvolved	Number	The numbers of criminals who involved on causing crime.
OffenderseizedCondition	Nominal	The conditions of the criminal/offender i.e. arrested or not.
Dateofoffenderidentifiedlately	Date	The date when the criminals identified lately.
Noofoffenderslateidentified	Date	The number of criminals identified lately.
NoofOffenderslatelyarrested	Number	The number of criminals, who was arrested lately.
NoofOffenderslatelynotarrested	Number	The number of criminals, who was identified lately but not arrested.
DateofOffendersarrestedlately	Date	The date of the criminal arrested lately.
OffendersNationality	Nominal	The nationality of offender.
BodyArrestedOffender	Nominal	The body that arrest the offender (police, society, other legal body etc)
Priceofstolenproperty	Number	Estimation price of stolen property.
Priceofstolenpropertyarrested	Number	Estimation Price of stolen property which is arrested by the society or legal body.
Priceofstolenpropertyreturnedto owner	Number	Estimation price of stolen property which is returned to the owner.
OffenderscurrentCondition	Nominal	The current conditions of the criminal/offender i.e. guarantee or imprison.
Datethecasessenttocourt	Nominal	Date of the case (investigation result) sent to the court from the police.
Nameoflegalinstitution	Nominal	The name of the legal institution. e.g. Guelele court, Arada court, Ledeta court etc.
DateofJudgmentPassed	Nominal	The Date of the final decision passed by the court.
DecisionBody	Nominal	The legal body that pass the final decision. E.g. court or prosecution.
PoliceStation	Nominal	The police station which investigates the crime occurred.
CrimeCatagory	Nominal	The general category of the specific crime.

From eleven crimes related profile attributes “Crime Code”, “Crime Type”, ”Year”, “Time”, “Particular Place”, “Crime Level” and “Crime Category” are selected. The others were not

considered as relevant for the purpose of this research. For instance crime registration date is a date, when the victims reported the case to the police. This attribute has not direct relevance to the research rather the specific date the crime occurred have directly importance to the research.

From six Victim's related Profile attributes "Victim Sex", "Victim Age", "Victim Job", "Victims Religion" and "Victim Marriage Status" are selected. The other attribute which was Victim Nationality has not selected because it is not relevant in showing patterns.

From fifteen Offenders profile attributes "Offender Sex", "Offender Age, "Offender Job", "Offender Educational Level", "Offender Religion", and "Offender marital Status" are selected, because all of them were directly related to the problem to be solved to attain the research objective. The rest, "number of offender involved", "Offender seized Condition", "Date of offender identified lately", "number of offenders late identified", "number of Offenders lately arrested", "number of Offenders lately not arrested", "Date of Offenders arrested lately", "Offenders Nationality", and "Offenders current Condition" are not. For instance, "Offenders nationality" has been removed because it contains single value and has no role for prediction to the subject under study. From the rest attributes like "Judgment passed", "Body Arrested Offender", "Price of stolen property", "Price of stolen property arrested", "price of stolen property returned to owner", "Date the cases sent to court", "Name of legal institution", "Date of Decision", "Decision Body", and "Police Station" the attribute "police station" was selected since it is the place where the crime was investigated. But as the experts stated that the rest have no any direct contribution to predict the crime. Table 3.3 shows the crime code which stands for the specific crime type and their main category.

Table 3.3 Descriptions of crime codes and their category.

Crime code	Specific crime type occurred	Crime Category
10	Forged Money usage and possessing	Crime against society
13	Deforestation	Crime against property
16	Murder	Crime Against person
17	Trying to commit Murder	Crime Against person
21	Simple assault, beat, body injury, committing body crippled	Crime Against person
22	Arson	Crime against property
24	Snatching	Crime against property
25	Trying to snatch vehicle, Trying to theft vehicle	Crime against property
26	Theft with house break	Crime against property
28	Purse-snatching or pocket-picking	Crime against property
29	Theft of vehicle part or accessory	Crime against property
30	Property theft from vehicle	Crime against property
31	Theft (miscellaneous thefts)	Crime against property
32	Hiding, buying, possessing and selling theft or missed property	Crime against property
33	Miscellaneous cheating including non-sufficient fund (cheque),Preparing forged document, using forged document etc.	Crime against property
34	Behave disloyal/unfaithful act/Breach of trust	Crime against society
35	Drugs possession and use	Crime against society
37	Taking bribes	Crime against property
38	Raping virginity forcefully or Immature girl	Crime Against person
39	Rape forcefully non-virgin women	Crime Against person
40	Homosexual, Trying to commit Rape	Crime Against person
41	Holding/possession forbidden gun	Crime against society
42	Beat and Threat, Threat, Insult, Insult and Threat, Owning and using others property, Character assignation, False Witness, Disrespectful others resident, Kick off Disturbance, Creating False Accuse, Authority misuse, Theft of Copy right/patent, Illegal commerce, Abuse of government resource, let alone infant, Hiding document/evidence etc .	Other miscellaneous crimes against society.
43	Violation of rules and regulations, Get lost from imprison, Disturbing workplace, Disturbing	Other offences related to violation of rules and

	Environment, Property damage, Illegal withdrawal from military service, Human trafficking, Damaging infrastructure, Disturbing Conference, Trying to Theft of vehicle part, Absent from workplace without permission, Trying to Theft,	regulation
44	Theft supported by gun.	Crime against property
45	Theft of vehicle	Crime against property

Crime Code attributes: the data type of crime code attribute is number. It has 27 distinct values. Nationally, 46 crime types (codes) were identified and recognized; from which 27 crime types were committed at Guelele sub city from 2003-2006 E.C. Table 3.4 shows the frequency of each crime code in the study.

Table 3.4 Summary of Crime Code attributes

CrimeCode: **Number**		
Distinct Value	Frequency	Percentage
10	12	0.02%
13	22	0.43%
16	61	1.19%
17	109	2.1%
21	1086	21%
22	17	0.33%
24	262	5.1%
25	3	0.05%
26	179	3.5%
28	100	1.9%
29	301	5.8%
30	33	0.6%
31	1284	25%
32	9	0.17%
33	257	5%
34	161	3.15%
35	51	0.99%
37	7	0.13%
38	40	0.8%
39	46	0.9%
40	34	0.6%
41	13	0.25%
42	581	11%
43	396	7.7%
44	5	0.09
45	36	0.7%
Total	5106	100

Year attribute: it is numeric attribute having 4 distinct values from 2003 -2006 E.C. Table 3.5 shows the year attribute with the number of crimes which had got the final decision.

Table 3.5: Statistical summary of Year attribute.

Year: **Numeric**	Frequency (No of Crime which had got final decision)	Percentage
2003	1533	30%
2004	1583	31%
2005	1509	29.5%
2006	481	9.4%
Total	5106	100%

Time attribute: it is numeric attribute having 24 distinct values from 0100 to 2400 Military times. But for this particular research the researcher convert and use world standard time because of its convenience and familiarity. Table 3.6 shows time attribute with the number of crimes which was committed within a given time. The crime which was committed from 9:00:00AM-12:00:00AM accounts 23.2%.It means 23.2 % of the crime was committed from 9:00:00AM-12:00:00AM. In addition, 20 % of the crime was committed from1:00:00PM-4:00:00PM.

Table 3.6: Statistical summary of time attribute.

Time: **Numeric/time**	Frequency	Percentage	World Standard Time
0100-0400	500	9.7%	1:00:00AM-4:00:00AM
0500-0800	636	12.4%	5:00:00AM-8:00:00AM
0900-1200	1186	23.2%	9:00:00AM-12:00:00PM
1300-1600	1024	20%	1:00:00PM-4:00:00PM
1700-2000	931	18.2%	5:00:00PM -8:00:00PM
2100-2400	648	12.6%	9:00:00PM-12:00:00AM
Unknown	167	3.2%	
Night unknown	7	0.13%	
Missed Value	7	0.13%	
Total	5106	100%	

Victim's sex attributes: it contains two valid values that are either male or female. The detail is presented in Table 3.7.As shows in the tables 58.53% were male victims and which was more than female victims (34.1%).This implies that more number of male than female was victims.

Table 3.7: Statistical summary of Victim's sex attributes.

VictimSex: **Nominal**		
Distinct Value	Frequency	Percentage
Male	2987	58.53%
Female	1742	34.1%
Undefined	364	7.12%
Missed values	13	0.21%
Total	5106	100%

Victim's age attribute: the age attribute contain a numerical valid value in the range from 5 to 100 years old. Victims have various ages and the details of the age of the victims are presented in table 3.8.The figure indicates 18.5% of the victims lie in the age group 26-30.Since Ethiopia youth policy defines youth as to include part of the society who are between 15-29 years [65].Which is clearly indicates that most of the victims are young.

Table 3.8: Statistical Summary of Victim age attribute

VictimAge: **Numeric**	Frequency	Percentage
0-15	14	0.27%
16-20	513	10.04%
21-25	827	16.1%
26-30	948	18.5%
31-35	640	12.5%
36-40	587	11.4%
41-45	258	5.05%
46-50	344	6.7%
51-55	168	3.2%
56-60	159	3.11%
Above 60	279	5.4%
Undefined	364	7.1%
Missed Value	5	0.09%
Total	5106	100%

Victim's Job attribute: the job attribute contain a nominal value and refers to the job of the victim the time she or he was affected by the crime. As shows Table 3.9 the victims who engaged on their own private work accounts (55.6%).In addition, others account (0.37%) and which encompass or refers victims who was engaged on begging, farming, driving and guarding activities.

Table 3.9: Statistical Summary of Victim Job attribute.

VictimJob: **Nominal**	Frequency	Percentage
Private	2842	55.6%
Government	749	14.6%
Unemployed	535	10.4%
Daily worker	97	1.8%
Housewife	347	6.7%
Student	138	2.7%
NGO	5	0.09%
Undefined	364	7.1%
Others	19	0.37%
Missed value	10	0.19%
Total	5106	100%

Victim's Religious Attribute: The religious attribute contain a nominal value and refers to the religion of the victims. Table 3.10shows that the victims who follows orthodox religion accounts (77.8%).Others, refers to government which was consider itself as a victim and its religion status is undefined. The number of missed value is very few and it accounts 0.11%.

Table 3.10: Statistical Summary of Victim's religion Attribute.

VictimsReligion: **Nominal**	Frequency	Percentage
Orthodox	3977	77.8%
Muslim	489	9.5%
Protestant	254	4.9%
Catholic	16	0.31%
Undefined	364	7.14%
Missed Values	6	0.11%
Total	5106	100%

Victim Marriage Status attributes: the attribute of marital status contains two distinct values namely married and single. The frequency of the attribute with possible nominal value is

described in Table 3.11.As the table shows below, modal (the highest frequency) value for marital status was married. The missing value of the attribute was insignificant and it accounts 0.09%. The Undefined refer to the government (Law), which is considered as victim.

Table 3.11: Statistical summary of Victim Marriage Status attribute.

VictimMarrtialStatus: **Nominal**	Frequency	Percentage
Married	2730	53.4%
Single	2007	39.3%
Undefined	364	7.1%
Missed value	5	0.09%
Total	5106	100%

Offender sex attributes: Offender sex attributes: it contains two valid values that are either male or female. The detail is presented in table 3.12. Male offender accounts (88.2%).It shows that most of the offender was Males.

Table 3.12: Statistical summary of offender's sex attribute.

OffenderSex: **Nominal**	Frequency	Percentage
Female	557	10.9%
Male	4504	88.2%
Missing Value	45	0.88%
Total	5106	100%

Offender age attribute: the age attribute contain a numerical value in the range from 18 to 89 years old offender. Based on the scope of the study the offender's age includes 18 years and above, as a result the age range of victims and offenders was vary. Offenders have various ages and the details of the age of them are presented in table 3.13. The figure indicates 33.2% of the offenders lie in the age group 18-22. As mentioned before Ethiopia youth policy defines youth as

to include part of the society who are between 15-29 years [65].It is clearly indicates that most of the offenders was also young.

Table 3.13: Statistical summary of offender's age attribute.

OffenderAge: **Numerical**	Frequency	Percentage
18-22	1698	33.2%
23-27	1637	32.06%
28-32	665	13.02%
33-37	457	8.9%
38-42	327	6.4%
43-47	126	2.4%
48-52	73	1.4%
53-57	23	0.4%
58-62	23	0.4%
63 and above	32	0.62%
Missed value	45	0.8%
Total	5106	100%

Offender Job attributes: the job attribute contain a nominal value and refers to the job of the offender during the time she or he committed crime. Table 3.14 shows that the criminal who engaged on their own private work during the time she or he committed crime accounts (57.5%).In addition, others account (1.4%) and which encompass offenders who was engaged on begging, farming, driving, brokering, church servant, nursing, and guarding activities during the time she or he committed crime.

Table 3.14: Statistical summary of Offender's Job attribute.

OffenderJob:**Nominal**	Frequency	Percentage
Private	2936	57.5%
Government	239	4.68%
Unemployed	1118	21.8%
Daily worker	342	6.6%
Housewife	59	1.15%
Student	286	5.6%
NGO	4	0.07%
Others	72	1.4%
Missed value	50	0.9%
Total	5106	100%

Offender educational level attributes: the educational level attribute refers to the level of education the offender hold during the time she or he committed crime. Originally the data was coded as numeric data but the researcher converts it to nominal data type for the sake of convenience, consistency and uniformity. The data contains 53 missing values, which accounts 1.03% of the total data. Table 3.15 describes statistical summary of offender educational level attributes.

Table 3.15: Statistical summary of offender educational level attribute.

OffenderEduL: **Nominal**	Frequency	Percentage
Illiterate	635	12.4%
Able to read	8	0.15%
Primary	1055	20.6%
Junior	915	17.9%
Secondary	1965	38.4%
Technical and vocational	203	3.9%
Diploma	161	3.13%
First Degree	98	1.9%
Master	4	0.07%
Church education	9	0.17%
Missing value	53	1.03%
Total	5106	100%

Offender religion attribute: The religion attribute contain a nominal value and refers to the religion of the offender during the time she or he committed crime. Table 3.16shows that the offender who follows orthodox religion accounts (87.4%) and it is followed by Muslim (9.3%). The number of missed value is very few and it accounts 0.9%.

Table 3.16: Statistical summary of offender religion attributes.

OffenderReligion:**Nominal**	Frequency	Percentage
Adventist	4	0.07%
Catholic	3	0.05%
Muslim	476	9.3%
Orthodox	4465	87.4%
Protestant	110	2.15%
Missing Value	48	0.9%
Total	5106	100%

Offender marital status attribute: the attribute of marital status contains two distinct values of married and single. The frequency of the attribute with possible nominal value is described in Table 3.17.

As Table 3.17 shows below, modal (the highest frequency) value for marital status was single and accounts for (72.9%). The number of missed value is very few and accounts 0.03%. It is very little and the researcher replaced with its modal value.

Table 3.17: Statistical summary of offenders marital status attribute.

OffendermarrtialStatus: **Nominal**	Frequency	Percentage
Married	1328	26%
Single	3725	72.9%
Missed Value	53	1.03%
Total	5106	100%

Crime Level attributes: Table 3.18 depicts the distribution of records based on the crime level attribute. This consists of three classes, i.e. high, medium and low levels. Therefore out of the 5106 records 1056 (20.6%) records belongs to the class of ‘High level’, 2160 (42.3%) records in the class of ‘medium’ and the remaining 1890 (37.01%) records are in the class of ‘low’.

Table 3.18: Statistical summary of Crime Level attributes.

CrimeLevel: **Nominal**	Frequency	Percentage
High Level	1056	20.6%
Middle Level	2160	42.3%
Low Level	1890	37.01%
Total	5106	100%

Police station attribute: The attribute of police station contains seven nominal values. It refers to the police stations where the study was conducted. Table 3.19 shows the number of crimes

investigated under each police station from 2003-2006 E.C. From the total crimes committed 32.3% of them were investigated under Paster police station and it is followed by Addisu Gebeya and accounts16.8%.

Table 3.19: Statistical summary of police station attribute

PoliceStation: Nominal	Frequency	percentage
Addis Ababa police Commission (Head Quarter)	158	3.09%
Guele Command Post	620	12.14%
Paster	1,650	32.3%
Menene	808	15.8%
Sheromeda	631	12.3%
Addisu Gebeya	858	16.8%
Kechene	381	7.46%
Total	5,106	100%

3.4 Data preparation and preprocessing

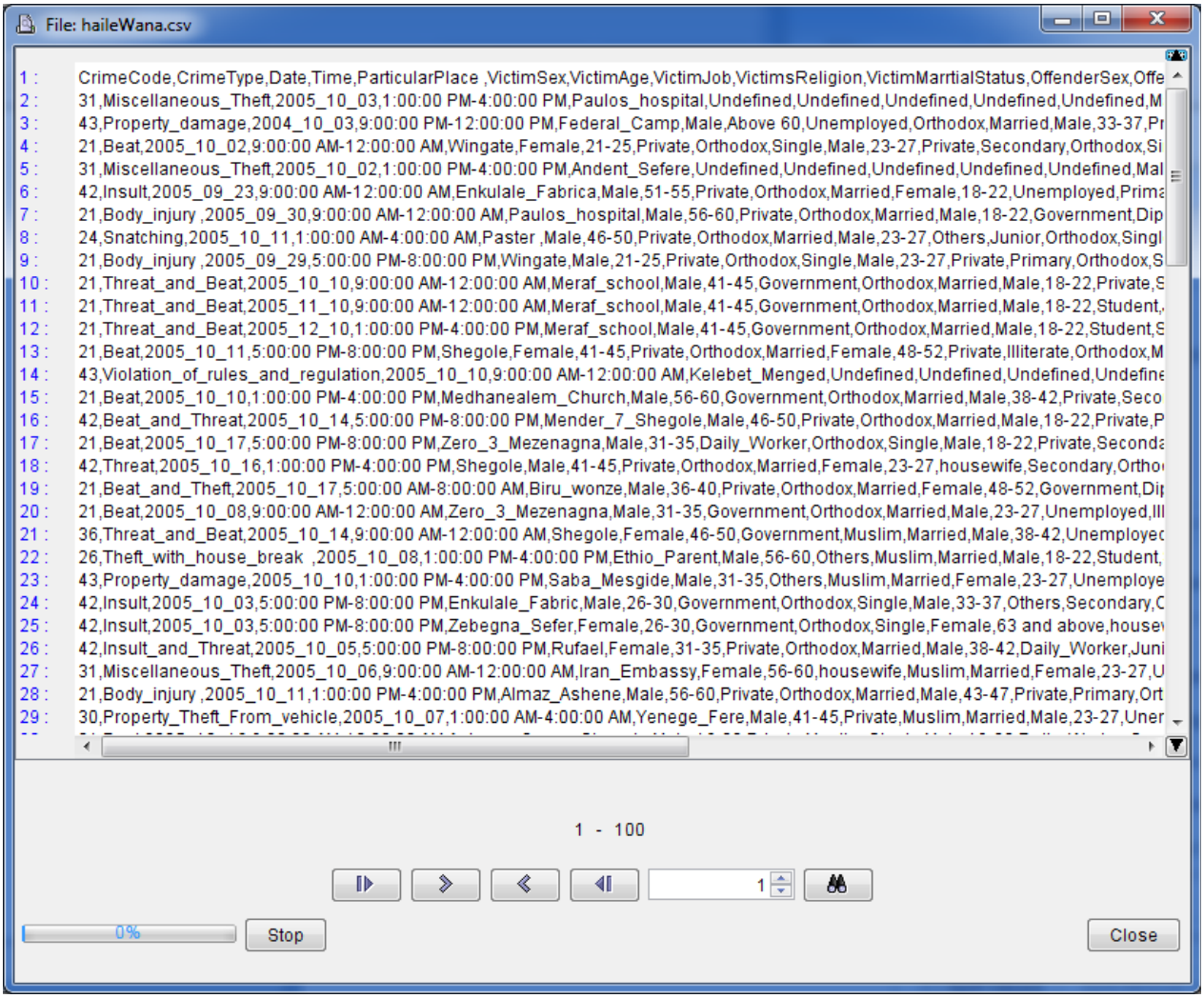
Proceeding to data mining step with low-quality data will lead to low-quality result [14].Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to improve the quality of the data and, consequently, of the mining result raw data is preprocessed so as to improve the efficiency and ease of the mining process. Data preprocessing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. Therefore to enhance the performance of algorithms, the researcher tried to assess such problem which requires due emphasis by applying different preprocessing techniques. Such methods are data cleaning, data reduction, and data integration and transformation. Each of these methods discussed as follows.

3.4.1 Data Cleaning

Data that is to be analyzed by data mining techniques can be incomplete (lacking attribute value or containing only aggregate data), noisy (containing errors, or outlier values which deviate from the expected), and inconsistent (discrepancy).Even if most of the data was collected manually from unpublished documents, some of the existing data are not clean; they need the effort of the researcher in the area to get rid of bad data, noise (invalid) data and filling missing values under each column. Removing of such record was done as the records with this nature are few and their removal does not affect the entire dataset.

As a result, the researcher makes use of MS-excel 2007 built-in functions to search and replace, auto fill mechanisms and DataPreparator-1.7 and WEKA open source software to identifies, fill errors and missing values. Figure 3.1 shows the representation of the data in comma-separated value (csv) format.

Figure 3.1 Representation of the data in CSV (comma separated value) format

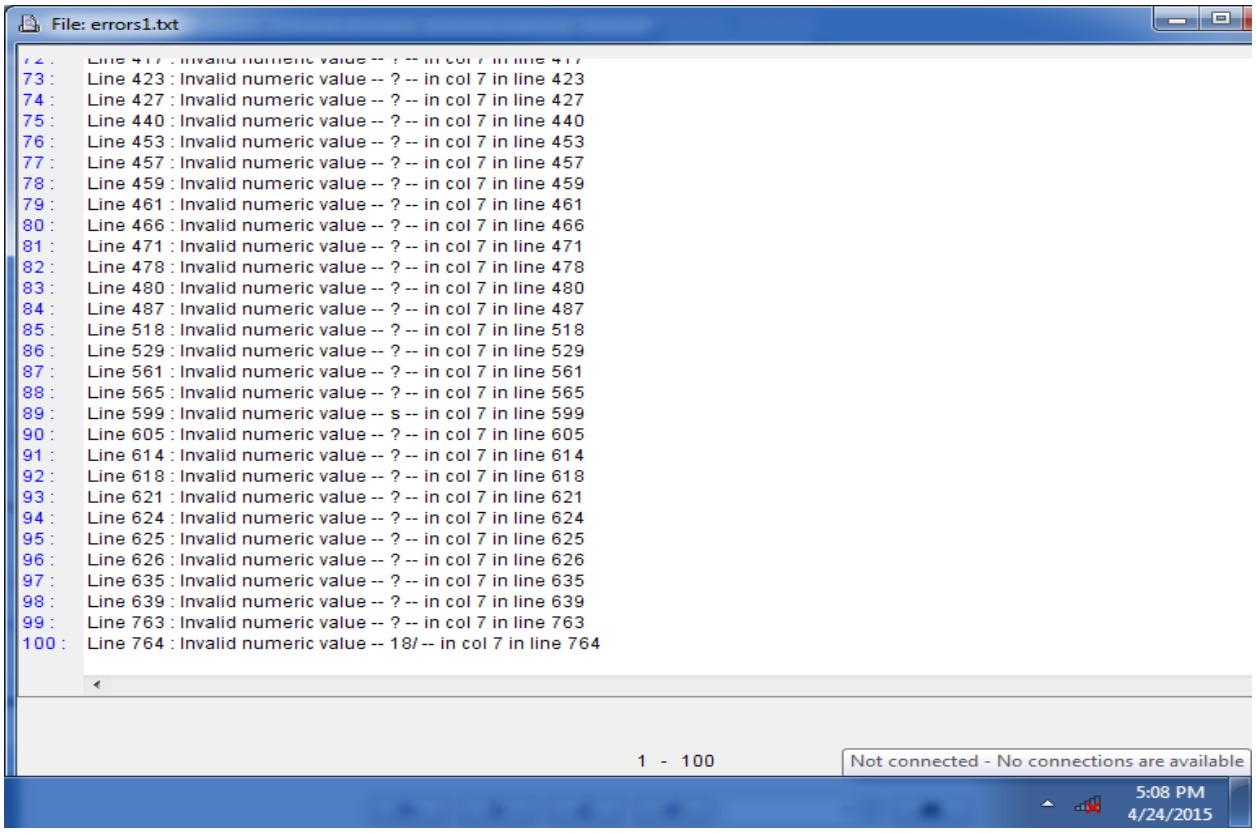


3.4.1.1 Missing Value Handling

Missing attribute values in the data is most likely associated with unavailability of interesting information, lack of knowhow on the importance of data at the time of entry, misunderstanding of the data, the respondents him/her self may refuse to answer certain questions or they may not know the answer exactly or may answer in unexpected manner. Figure 3.2 shows how the researcher identified missing value, outliers and errors in the data. Accordingly, the researcher

has been analyzing the crime dataset and identifying missing values and taken measures to solve the problem as follows.

Figure 3.2 Representations of missing value and errors in the data



As the researcher depicted in section 3.3.3; percentage of missing values for each attribute were calculated. As a result, the missing value of most attributes were less than one percent (1%) except attributes of offender educational level and marital status and both accounts 1.03%. To handle the problem of missing value for nominal data type attributes, replacing with modal value is recommended in [16]. Some common strategies for calculating missing values include using the modal value (for nominal variables), the median (for ordinal variables), or the mean (for continuous variables).Therefore, based on the above principle the researcher handles the missing value and WEKA preprocessing and MS-excel find and replace missing value techniques also used. WEKA fills using the most frequent (modal) value methods which is same as the above

principle. Additionally, manually tracing and fixing the missed value is other techniques used by researcher. Summary of handled missing value is illustrated in Table 3.20.

Table 3.20 Summary of handling missing value

Attribute name and their data type	Percentage (%) of missing value	Replaced with	Justification/Techniques applied
Time: Time	0.13%	[11:00-12:00)	Most frequent value
VictimSex: Nominal	0.21%	Male	Most frequent value
VictimAge: Numeric	0.09%	32.2	Mean
VictimJob: Nominal	0.19%	Private	Most frequent value
VictimsReligion: Nominal	0.11%	Orthodox	Most frequent value
VictimMarrtialStatus: Nominal	0.09%	Married	Most frequent value
OffenderSex: Nominal	0.88%	Male	Most frequent value
OffenderAge: Numeric	0.8%	26.6	Mean
OffenderJob: Nominal	0.9%	Private	Most frequent value
OffenderEduL: Nominal	1.03%	Secondary	Most frequent value
OffenderReligion: Nominal	0.9%	Orthodox	Most frequent value
OffendermarrtialStatus: Nominal	1.03%	Single	Most frequent value

3.4.1.2 Handling outlier value

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers [14]. The data stored in a database may reflect outlier – noise, exceptional case, or incomplete data object and random error in a measure of variable. These incorrect attribute values may be due to data entry problems, faulty data collection, inconsistence in naming the value (miss spelt) and technology limitation. According to Han et al [14] there are four methods for handling of noise data. These are binning method, clustering, regression and combined computer and human inspection.

Since most of the data have collected manually from unpublished record and they also have more of nominal values, outliers are not a problem of the data used in this research. The attribute age for both the victim and the offender can have the outlier values but both are written in interval format. For instance for victims 0-15,16-20,21-25,26-30,31-35,36-40,41-45,46-50,51-55,56-60 and above 60. Regarding offender age, it is started from 18 based on the scope of this research and categorized as 18-22, 23-27, 28-32, 33-37, 38-42,43-47, 48-52, 53-57, 58-62 and above 63.

3.5 Data Transformation and Reduction

Data integration is a preprocessing task of combining data from multiple sources, database, data warehouse, or different file structure. As a result of this combination process data that is fine on its own can become problematic because of different formats and structures, conflicting and redundant data and data at different level of detail. As mentioned before in the data used in this study there is no significant problem, because the commission used relatively uniform crime record format. But in terms of field understanding there was a difference. For instance ‘Menene’ police station record the value of Crime day-situation number for Crime record number and vice versa. But it was solved by discussion with senior experts.

According to Han et al [14] data reduction techniques include discretization and concept hierarchy generation where raw data values for attribute are replaced by ranges or higher conceptual level. Concept hierarchy allows the mining of data at multiple level of abstraction, and is a powerful tool for data mining.

In this research some attributes were discretized to obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. As a result, the researcher was performing some activities while we translate the data from

source language (Amharic) to English. In addition, the following sections depicted the details of them.

3.5.1 Discretization and concept hierarchy generation

As Han et al [14] mentioned, raw data values for attributes are replaced by range or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining. That is, mining on the reduced data set should be more efficient yet produce the same or almost the same analytical result.

Discretize the educational level attribute value: based on the current and previous curriculum of different level of education, the researcher tried to associate and classify educational level of offenders. For this reason educational level category was arranged and categorized grade as follows: 1-6 as “primary”, grade 7-8 as “junior”, grade 9-12 including former named as grade 12 complete as “secondary”, and the others were used as it is, since it was understandable and comfortable.

Discretize the job attribute value: Similarly the researcher tried to associate and classify the different distinct value of victim’s and offenders job attributes to high level of concept. The detail has been shows below in Table 3.21.

Table 3.21 Victim and offender's Job attributes value discretization

AttributeName	Old value	Transformed value
Victim Job	Commerce, retired, taxi assistant, Musician and private	Private
	Police, military, government	Government
	Jobless	Unemployed
	Daily worker	Daily worker
	Jobless married women	Housewife
	Student	Student
	NGO	NGO
	Nurse, teacher, farmer, guard	Others

3.6 WEKA understandable format

Once the data sets are pre-processed and cleaned, the next steps is changing Ms-Excel format into CSV (comma separated value) WEKA understandable format for experimentation. Table 3.22 shows comparison of the original dataset and the refined target dataset.

Table 3.22 Summary of original and Target Dataset

Parameters	Originaldataset	Target dataset	
Fields	42+1	19	
Total number of record from 2003-2006	10,483	5106	
File Format	Manual unpublished crime record and Ms-Excel	.xls	.csv

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS OF RESULTS

In this chapter, the researcher describes the techniques that have been used in developing data mining predictive model that can predict crime levels, analyze the relation of the demographic profile of offenders and victims and extract crime patterns. This research incorporated the typical stage that characterizes a data mining process. The study has been organized according to hybrid processing model, which is described and discussed in sub section 1.5.1.1 of Chapter One and sub section 2.2.4 and figure 2.1 of Chapter Two. Here, the researcher discusses the experimentation process, results obtained, compared the model and results, and finally presented it in a way that the organization can easily understand and use it.

4.1 Model building

Modeling is one of the major tasks which are undertaken under the phase of data mining in hybrid methodology. Model implementation and tool selection are tasks that should be discussed in this chapter. Here, the researcher is interested in explaining the models selected to achieve the mining goal. As mentioned in the methodology section in Chapter One, the problems or the classification model was addressed. Some of the tasks include: selecting the modeling technique, experimental setup or design a model and compare the model.

4.1.1 Classification model building

Selecting data mining technique depends on data mining goals. Consequently, to attain the objective of this study, two classification techniques have been selected for model building. In classifications model building, the researcher intended to use and build decision tree (J48 algorithm), and PART rule induction algorithm. For experimentation, those algorithms are

employed by considering different parameters for model building such as pruning, unpruning and testing model performance with selected attributes and all attributes in both training and testing dataset. The experiment and its scenario are depicted in Table 4.2. The researcher selected the above algorithms since the task is a classification one. Those classification algorithms are easy to understand, to interpret the result of the model and they are convenient to the expected result set.

4.1.2 Experimental Setup

In any data mining process, before building a model, we need to generate a procedure or mechanism to test the model's quality and validity. For instance, in supervised data mining tasks such as classification, it is common to use classification accuracy measures or error rates as quality measures for data mining models. Besides, other standard measures including precision, recall, and ROC are available. Therefore, the test design specifies that the dataset should be separated into training and tests set, and builds the model on the training set and estimates its quality on a separate test set. With regards to this, Two Crows Corporation [16] reported that the process of building predictive models requires a well-defined training and validation protocol in order to ensure most accurate and robust prediction.

In this research, 5,106 dataset are used for training and testing. The researcher believed that these numbers of data fulfill the minimum requirement for data mining techniques and methods. WEKA 3.7.10 software has been used to set up and measure the quality, validity and test of the selected model. For the purpose of this study,K-fold (10-folds) cross validation and percentage split test options are used because of their relatively low bias and variations [53].Accordingly, the datasets are partitioned into 80-20 percentage split option, meaning 80% of the dataset are used for training and remaining for testing. To build the model of this research, the researcher

used different independent and dependent variables or attributes. ***CrimeLevel, Time, victimMartialStatus, OffenderMartialStatus, victimJob and offenderJob*** attributes have been used as a class attribute in different experiments interchangeably as needed to address the objectives of this study. Those target classes were selected based on the objective of the study.

4.1.3Attribute ordering

Since attribute selection is important in decision tree models, the researcher tried to rank the attribute based on information gained. It was calculated based on entropy value of the attribute. As Witten and Frank [53] explain information gain is calculated from sum of entropy for every attribute. The formula for calculating intermediate value is:

$$\textbf{Info (D)} = -\sum m_{i=1} p_i \log_2 p_i$$

Where, P_i is the probability that an arbitrary tuples in sets of training D belongs to certain class. Info (D) is also known as the entropy of D. After one calculates information gain for each attribute, he/she should select the one with the highest information gain as the root node, and continue the calculation recursively until the data is completely classified by J48 algorithm.

For the purpose of this research, the WEKA software is used to compute the information gain and rank according to the importance of the attribute for the classifier. The following figure 4.1 shows the rank of attributes that have been selected by domain experts. From the total of 43 attributes, 19 of them were subjected to experiments. Attributes, such as crime type, crime code, date, particular place, police station, crime category, victim's job, offender marital status, offender job, victim's age, victim's religion, time, offender educational level, offender's sex, offender's age, victim's sex, victim marital status and offender religion have been used. Figure 4.1 shows results of ranked attributes.

Figure 4.1: Results of ranked attributes

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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 19 CrimeLevel):
        Information Gain Ranking Filter

Ranked attributes:
 1.20788    3 CrimeType
 1.0231     1 CrimeCode
 0.81519    4 Date
 0.56269    6 ParticularPlace
 0.117     18 PoliceStation
 0.04708    2 CrimeCategory
 0.03621    9 VictimJob
 0.03474   17 OffendermarrtialStatus
 0.03405   14 OffenderJob
 0.02885    8 VictimAge
 0.02751   10 VictimsReligion
 0.0261     5 Time
 0.02319   15 OffenderEduL
 0.01977   12 OffenderSex
 0.01657   13 OffenderAge
 0.01426    7 VictimSex
 0.01373   11 VictimMarrtialStatus
 0.00659   16 OffenderReligion

Selected attributes: 3,1,4,6,18,2,9,17,14,8,10,5,15,12,13,7,11,16 : 18
```

4.1.4. Running Experiments

As it has been shows in experimental setup of 4.1.2 for training and testing the classification model the researcher used two methods. Method one consists percent split method (holdout), where 80% of the data have been used for training and the remaining 20% for testing. In method two, k-fold cross validation method, the data was divided into 10 folds, some fold are used as testing and the remaining ones are for training.

Based on the above methods, establishing scenario for model to be developed is very important to see the model result and analysis of each result, the method is also useful to compare the result of one model with the next one and finally to find out the outperforming model based on criteria of evaluation.

4.1.4.1 Model building using J48 decision tree

J48 is one of the most common decision tree algorithms that are used today to implement classification techniques using WEKA [16]. This research used J48 algorithm to identify and investigate the relation and demography factors of victims and offenders who were exposed to crime and to develop crime prediction model. This algorithm is implemented by modifying parameters such as confidence factor, pruning, unpruning and other parameters available in Table 4.1.Therefore, it is very crucial to understand the available parameters to implement the algorithms, as it can make a significant difference in the quality of the result.

Table 4.1 Parameters for building J48 tree

Name	Possible value	Default value	Description
Number of instance	1,2…	2	Minimum number of instances in leaves. High values results in smaller trees)
unpruned tree	Yes/no	no	Use unpruned tree (the default value 'no' means that the tree is pruned).
confidence factor	10^{-7}—0.5	0.25	Confidence factor used in post pruning (smaller values incur more pruning)
subtreeraising	Yes/no	Yes	Whether to consider the sub tree raising operation in post pruning
use binary splits	Yes/no	No	Whether to use binary splits on nominal attributes when building the tree

As indicated in the above Table, the commonly used parameter of J48 algorithm was illustrated with various options related to tree pruning. Pruning produces fewer, more easily interpreted results. In other words, pruning can be used as a tool to correct potential over fitting. The algorithm recursively classifies until each leaf is pure, meaning until dataset has been categorized as close to their possible respective classes. Hence, the process ensures the maximum accuracy on the training data, but creates excessive rules.

In General, Witten and Frank [53] show the various options of tree pruning in J48 algorithm as follows: Basically, J48 employs two pruning methods. The first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf -- basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub tree rising. In this case, a node may be moved from leaves upwards the tree, replacing other nodes along the way. Sub tree rising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking for a long time. This is due to the fact that sub tree rising can be somewhat computationally complex.

Reduce-error pruning option- Error rates are used to make actual decision about which part of the tree to be replaced. There are multiple ways to do so. The simplest method is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over fitting. This approach reduces the overall amount of data available for training the model.

Confidence factor option: confidence factor is the approach that seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires confidence threshold, which by default is set to 25%. This option is important for determining how specific or general the model should be. If the value of the confidence factor is lowered on the selected model, the training data is expected to fit in closely to the data we would like to test. As a result of this, a more pruned or more generalized tree will be produced.

Minimum number of instances per leaf option: this is the lowest number of instances that can constitute a leaf. The higher the number, the more general the tree will be. Lowering the number will produce more specific tree as the leaves become more granular.

Binary split option: The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using an inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourage more generalized trees.

Laplace smoothing option: this option is used to prevent probabilities from ever being calculated as zero. This is mainly done to avoid possible complication that can arise from zero probabilities.

If the data mining researcher decides to employ tree pruning, it is advisable to consider the above options. Depending on how the training and test data have been defined, the performance of an

unpruned tree may specifically appear better than a pruned one. This can be a result of over fitting. Hence, it is important to repeatedly experiment with models by intelligently adjusting these parameters to obtain the best set of options. According to Witten and Frank [53], five basic familiar parameters usually used for data mining research are depicted in Table 4.1. But for this particular research, the default parameters with the following experimental scenario have been applied.

Table 4.2 Experiments and Scenario

Experiments	**Scenarios**
J48 Unpruned Tree Model Generation	1. J48 Unpruned with all attributes
	2. J48 Unpruned with selected attributes
J48 Pruned Tree Model Generation	3. J48 pruned tree model with all attributes
	4. J48 pruned tree model with selected attributes
PART unpruned decision list	5. PART unpruned with all attributes
	6. PART unpruned with selected attributes
PART pruned decision list	7. PART pruned tree model with all attributes
	8. PART pruned tree model with selected attributes

To sum up, to build predictive model, 5106 instances and 19 attributes were used through both PART and decision tree algorithm. The models generated with all attributes were compared with models with selected attributes. As mentioned in Table 4.1, J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Accordingly, based on the two methods selected (pruned and unpruned), the following series of experiments have been conducted. Finally, the classification accuracy and interesting patterns generated from both models help to analyze the relation between/ among demographic features of offender, victims

and type of crimes. Summary of the result of the experimentation for both are depicted in tables relating to each experiment.

As mentioned in methodology, sub section 1.5.1.1, one of the compulsory steps of hybrid methodology next to model building is evolution of the model. Accordingly, the performance of the model has been evaluated based on the criteria presented in 1.5.1.1 of Chapter One which include performance accuracy, confusion matrix value, and TP and FN rate, number of leaves and size of the tree generated and ROC and execution time.

Experiment 1: Classification of records using the *CrimeLevel* Class

This experiment uses the attribute crime Level to classify records. The attribute consist of three classes namely high, medium and low level.

Table 4.3 shows that the experimentation has been performed based on two model validation techniques. The first method 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 second 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 purpose of using those parameters was to assess the performance of the learning scheme by switching the proportion of testing dataset. The first experiment has been tested with four scenarios mentioned in Table 4.2 with 80-20 percentage split criteria.

When we compare the result of method I, scenario No 2 and 4 have been found to be the best models. Out of 1021 records both models correctly classify 1006 instances. Besides, in terms of

number of leaves and size of tree, scenario No3 performs better. But regarding times span taken to build a model, scenario No2 registered minimum time. The result of those learning schemes are summarized and presented in Table 4.3.

Table 4.3 Experimentation results of J48 Algorithm based on the two methods

Model Characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test option	I- (80/20)				II- 10-fold			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	**All**	All	Reduced	Reduced
Accuracy (%)	94.5	98.5	94.3	98.5	**95.3**	99	94.9	98.9
Time taken	0.27	0.23	0.3	0.25	**0.27**	0.05	0.27	0.22
Noof leaves	3034	11109	2702	14758	**3034**	11109	2702	14758
Size of trees	3057	11165	2713	14806	**3057**	11165	2713	14806
AV.TPR (%)	94.5%	98.5%	94.3%	98.5%	**95.3%**	99%	95%	98.9%
AV.FPR (%)	3.3%	1%	3.5%	1%	**2.8%**	0.5%	3%	0.6%
AV.PR	0.94	0.986	0.943	0.985	**0.953**	0.99	0.95	0.989
AV.RR	0.94	0.985	0.943	0.985	**0.953**	0.99	0.95	0.989
AV.ROC	0.97	0.996	0.948	0.996	**0.977**	0.99	0.976	0.999
CCI	965	1006	963	1006	**4867**	5061	4850	5052
ICI	56	15	58	15	**239**	45	256	54

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, I: 80-20 percentage split option (Holdout), II: 10-fold cross validation.

The last comparison in method one was made on average ROC. It has registered a performance of 99.6% in scenario No 2 and 4. Generally from method I, scenario No2 is selected since the performance of the model is better. But the complexity of the generated tree from stated scenario is high.

The next experiment has been tested with 10-fold cross validation. The result shows that scenario No2 (building decision tree unpruned with all attribute) registered best accuracy and accounts 99%. It correctly classifies 5061 instances out of 5106. In addition, scenario No2 (unpruned with all attribute) registered better time to build the model.

Analogously, concerning the number and size of tree, scenario No3 is relatively better (less complex and understandable). Regarding the average ROC performance measure indicates #4 (99.9%), which performs better than the rest of the three scenarios. Finally, when we compare the actual correctly classified instance based on their label class, i.e. classifying crime levels as high, medium and low level for all scenarios, irrespective of number of attributes, unpruned one registered better than pruned.

However, when we compare the size and leaves of tree of unpruned J48 model, the number is enormous and complex compared to the pruned one. As a result of this, the algorithm might not reach optimality and generate more generalized decision tree rules. Han et al [14] also reported the reason for this to be the problem of over fitting. This is a fundamental problem in learning

algorithms. Besides, such situation has its own impact in classifying performances, particularly classifying unseen or new instances. Subsequently, to solve the problem, the researcher selected pruned scenario that performs with better accuracy. Accordingly, scenario No1 (Building pruned decision tree with all attribute) of 10-fold cross validation (method II) has been selected as the best J48 decision tree model.

Consequently, Table 4.3 also shows that irrespective of reduced or all attribute, the performance of unpruned J48 algorithm has registered better than pruned in terms of accuracy. However, in addition to the above performance metrics used, in terms of accuracy, tree size and number of leaves, scenario No1(building decision tree pruned with all attribute) of 10 fold cross validation is relatively more understandable and less complex than others models generated. Therefore, the performance of J48 pruned tree classifier with all attribute gives valuable information in predicting crime level as compared to other models. The detailed confusion matrix as a result of pruned J48 algorithms with all attribute has been presented below in table.

Table 4.4 Confusion Matrix output of the J48 algorithm with 10-fold cross validation.

Actual	Medium Level	Low Level	High Level	Total	Correctly Classified (accuracy rate)
Medium Level	**2058**	66	36	2160	95.2%
Low Level	84	**1801**	5	1890	95.2%
High Level	16	32	**1008**	1056	95.4%
Total	2158	1899	1049	5106	95.3%

As shown in the confusion matrix above, the J48 learning algorithm scored an accuracy of 95.3 %, which indicates that ,out of the total number of records supplied, around 4867(95.3%) records are classified correctly and 239(4.7%)misclassified or incorrectly classified. Furthermore, the resulting confusion matrix of this experiment has shown that, 95.2% of the records are correctly classified at the medium and low levels. This shows that out of the 2160 medium level crime records, 95.2% of them classified in their respective class, while 66(3%) of them are misclassified as low level, and 36 (1.6%) of them misclassified as high level. In addition to this, out of the 1890 low level crime record,1801(95.2%) of them are classified correctly in their designated class, i.e. low level, while 84(4.4%) of them are misclassified as medium level crime and only 5(0.2%) low level record misclassified as high level crime.

Most classification algorithms seek models that attain the highest accuracy, or equivalently, the lowest error rate when applied to the test set. According to Han et al [14] hypothesis testing, error can be categorized into two based on the classification of being true as false and false as true. These errors are called Type I (false positive error) and Type II (false negative error). Thus, in the experiment in this research context, classifying High level crimes as medium or low and low level crime as medium and high, etc. may mislead to make a wrong decision. As a result, police and law enforcement bodies may allocate and utilize unequal, unfair, ineffective and inefficient resources to prevent and control crimes.

As discussed in Table 4.3, the size of the tree and the number of leaves produced from method two of scenario No1 model training are 3057 and 3034 respectively. This seems that it is difficult to navigate through all the nodes of the tree in order to derive out with valid sets of rules.

Therefore, to ease the process of generating rules or to make it more understandable, the researcher attempts to modify the default values of the parameter so as to minimize the size of the tree and the number of leaves.

In this regard, the minNumObj (minimum number of instances in a leaf) parameter was tried with 10, 20, 30, 50 100 and 200.However, the minNumObj set to 200 gives a better tree size and minimum accuracy compared with the other trials. With this value of theminObj, the process of classifying records proceeds until the number of records at each leaf reached 200.The confusion matrix with minNumObj=200 is shown in Table 4.5.

The experiment has shown an improvement in the number of leaves and tree size. The size of the tree is dropped from 3057 to 65 and the number of leaves decreased from 3034 to 64.As we can see from Table 4.5, the resulting confusion matrix shows that the J48 decision tree algorithm scored 93.8 % accuracy. This means out of the total 5106 records, 4791 (93.8%) are correctly classified and the remaining are misclassified.

Table 4.5 Summary of pruned J48 algorithm with minNumObj=200

Actual	Predicted			Total	Correctly Classified (accuracy rate)
	Medium Level	Low Level	High Level		
Medium Level	2055	71	34	2160	95.1%
Low Level	107	1778	5	1890	94%
High Level	65	33	958	1056	90.7%
Total	2227	1882	997	5106	93.8%

In other words, the confusion matrix of this experiment has shown that 2055(95.1%) of the 2160 total medium level crime are correctly classified as medium level crime. But 71(3.2%) and 34(1.5%) of them are misclassified as low and high level crime respectively. Besides the result shows that out of 1890 total low level crime, 1778(94%) of the records are correctly classified as low level crime while 107(5.6%)and 5(0.2%)are misclassified as medium and high crime levels respectively. When we compare the result of this experiment, classifying crimes at their correct level outweighs from those misclassifying crimes levels. Similarly, the execution time has reduced from 0.27 to 0.06 seconds and the complexity of the tree has reduced. Besides, the accuracy of the model has lowered from 95.3% to 93.8% as minNumObj getting large showing a considerable change. From this experiment, one can conclude that minNumObj and accuracy of model performance have shown inversely related pattern. The generated possible decision tree with reduced dataset is shown in Appendix A.

Therefore, the result of table 4.3 shows that even if the decision tree developed in the learning scheme is relatively complex. It seems more accurate to classify records.

Experiment 2: Classification of records using time target class

This experiment uses the attribute time to classify records. The attribute consist of eight classes; 1:00:00PM-4:00:00PM, 9:00:00PM- 12:00:00PM, 9:00:00AM-12:00:00AM,1:00:00AM-4:00:00AM, 5:00:00PM-8:00:00 PM, 5:00:00AM-8:00:00AM, Night Unknown and Unknown. Table 4.6 shows that experimentation results of J48 Algorithm based on the two methods.

Table 4.6 Experimentation results of J48 Algorithm based on the two methods.

Model Characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test option	I (80-20)				II 10-fold			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy (%)	78.3	80.4	79.4	81.6	**84**	86	84.5	86.3
Time taken	0.44	0.14	0.19	0.13	**0.14**	0.34	0.25	0.34
No of leaves	28109	29533	34523	36642	**28109**	29533	34523	36642
Size of trees	28564	30017	34958	37099	**28564**	30017	34958	37099
AV.TPR (%)	78.4%	80.4%	79.4%	60%	**74.3%**	86%	84.6%	77.2%
AV.FPR (%)	4.7%	4.3%	4.6%	0.3%	**0.6%**	2.9%	3.2%	0.6%
AV.PR	0.783	0.804	0.796	0.875	**0.805**	0.861	0.846	0.822
AV.RR	0.784	0.804	0.794	0.60	**0.743**	0.861	0.846	0.772
AV.ROC	0.94	0.942	0.945	0.938	**0.954**	0.960	0.961	0.958
CCI	800	821	811	834	4292	4394	4319	4408
ICI	221	200	210	187	814	712	787	698

Similarly, when we compare the results of method I. scenario No4 registered better accuracy. Out of 1021 records 834(81.6%) of them have been correctly classified. Besides, in terms of number of leaves and size of tree, scenario No1 perform better and relatively easy to understand/interpret. But regarding duration taken to build a model scenario No4 registered minimum time.

The next experiment has been tested with based on the four scenario mentioned in Table 4.2 with 10-fold cross validation. The result shows that scenario No4 (building decision tree unpruned with all attribute) registered best accuracy and accounts 86.3%. It correctly classifies 4408 instances out of 5106. In addition, scenario No1(pruned with all attribute) registered better time to build the model.

Analogously, regarding the number of leaves and size of tree, scenario No1 is relatively better and considered as the best one because it reduces the complexity of generated tree. The average ROC performance measure indicates that scenario No3 accounts 96.1%, which performs better than the rest of the three scenarios. Finally, when we compare the actual correctly classified instance, irrespective of number of attributes, unpruned registered better than pruned.

Accordingly, scenario No1 (Building pruned decision tree with all attribute) of 10-fold cross validation (method II) has been selected as the best J48 decision tree model. Table 4.6 above shows output from the J48 decision tree learner using the default value of the parameters.

The accuracy of this learning scheme was 84% this indicates that out of the total number of records supplied 4286 (84%) records were classified correctly while the remaining 820(16%) records have been classified incorrectly. The summary of the output is presented in Table 4.6. Moreover, the results of the experiment have shown that about 86.4 %of the records in the class of 1:00:00PM-4:00:00PM were correctly classified, while also about 86.5% and 88% percent of

the records in the 9:00:00PM-12:00:00PM and 9:00:00AM-12:00:00AM classes respectively have been classified correctly.

Figure 4.2 Confusion Matrix output of the J48 algorithm for time target class

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 838  24  54  37  32  28   0  11 |   a = 1:00:00 PM-4:00:00 PM
  24 513  42  23  18  15   1  12 |   b = 9:00:00 PM-12:00:00 PM
  59  37 977  30  58  21   0  11 |   c = 9:00:00 AM-12:00:00 AM
  38  32  32 365  27   3   0   3 |   d = 1:00:00 AM-4:00:00 AM
  58  25  68  13 729  29   0   9 |   e = 5:00:00 PM-8:00:00 PM
  43  20  35  26  43 464   0   5 |   f = 5:00:00 AM-8:00:00 AM
   2   2   0   0   0   0   3   0 |   g = Night Unknown
   6  10  13   3  13   7   0 115 |   h = Unknown
```

An attempt has been also made to conduct the experiment by excluding some of the attributes to see if the accuracy of learning scheme could be improved. For instance, by dropping some attributes, such as crime code, victims' religion and offenders' religion. But the accuracy of learning scheme was not possible to improve to above 84%.This revealed that about 16 records are wrongly classified and this may account to the error made by the crime analysis experts in classifying the data employed in this study.

Experiment 3: Classification of records using *VictimMartialStatus* target class

This experiment uses the attribute Victim Martial Status to classify records. The attribute consists of three classes, namely single, married and undefined marital status. Table 4.7 shows experimentation results of J48 Algorithm with two methods. The third experiment as well has been tested with the two methods as mentioned above. The first was 80-20 percentage split method with four scenario mentioned earlier.

Table 4.7 Experimentation results of J48 Algorithm based on the two methods.

Model Characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test option	I (80-20)				II 10-fold			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy (%)	86.7	96.4	86.5	96.3	**87.2**	97.3	86.7	97.1
Time taken	0.3	0.08	0.31	0.08	**0.08**	0.23	0.16	0.05
No of leaves	770	7891	1568	9756	**770**	7892	1568	9756
Size of trees	797	7980	1593	9840	**797**	7980	1598	9840
AV.TPR (%)	86.8%	96.5%	86.6%	96.4%	**87.2%**	97.3%	86.8%	97.1%
AV.FPR (%)	12.2%	3.1%	12.4%	3.3%	**12.2%**	2.4%	12.6%	2.7%
AV.PR	0.868	0.965	0.866	0.964	**0.872**	0.973	0.868	0.971
AV.RR	0.868	0.965	0.866	0.964	**0.872**	0.973	0.868	0.971
AV.ROC	0.926	0.989	0.926	0.990	**0.927**	0.995	0.923	0.996
CCI	886	985	884	984	4453	4970	4430	4959
ICI	135	36	137	37	653	136	676	147

When we compare the results of method I. In terms of accuracy, scenario No2 has been found to be better model. Correctly classified instances are 96.4%. Moreover, in terms of number of leaves and size of tree, scenario No1 registered better performance. This scenario shows a less complicated tree than the other three. But regarding times taken to build a model, scenario No2 have less time. When we compare the model in terms of ROC area, scenario No 4 has been

perform better and accounts 99%. To sum up, even though the complexity of the generated tree is high, scenario No 2 registered better classification performance. As a result scenario No2 is selected from method I.

Similarly, the next experiment has been tested with four scenarios with 10-fold cross validation. The result depicts that scenario No 2 registered the best accuracy and accounts 97.3%. It correctly classifies 4970 instances out of 5106. Nevertheless, concerning the number of leaves and size of trees, scenario No 1 was comparatively better and it generated slightly understandable and less complex tree. Regarding on ROC performance, scenario No 4 registered better than the rest three scenario and accounts 99.6%. Finally, when we compare the actual correctly classified instance based on their label class i.e. classifying victims marital status as married, single and undefined for all scenarios, irrespective of number of attribute, unpruned registered better than the pruned one.

To sum up, when one compare the size and leaves of trees generated from pruned and unpruned J48 model, the tree generated from unpruned has been enormous and complex. As a result of this, the algorithm might not reach optimality and may not generate more generalized decision tree rules. Moreover, such situation has its own impact on classification performance particularly classifying unseen or new instances. Consequently, to solve the problem the researcher selected pruned scenario that perform better accuracy. Accordingly, scenario No 1 from method II selected as the best J48 decision tree model based on the aggregate performance of accuracy, computation time and visibility of the tree generated. Table 4.8 shows the confusion matrix for selected model.

Table 4.8 Confusion Matrix output of the J48 algorithm with 10-fold cross validation.

Actual	Predicted			Total	Correctly Classified (accuracy rate)
	Undefined	Married	Single		
Undefined	364	0	0	364	100%
Married	0	2480	255	2735	90.6%
Single	0	398	1609	2007	80.1%
Total	364	2878	1864	5106	87.2%

The result of the experiment has shown that about 100% of the records in the class of undefined marriage status have been classified correctly; also about 90.6% and 80.1% of the records of the married and single marital status classes respectively have been classified correctly. In addition to this the resulting confusion matrix has shown that out of 5106 crime records 4453(87.2) of them are correctly classified. Thus, this also indicates that about 12.8% records are wrongly classified.

Experiment 4: Classification of records using *victimJob* target class

This experiment uses the attribute victim's job to classify records. The attribute consists of nine classes, namely undefined, unemployment, private, government, daily-worker, housewife, student, NGO and others. Table 4.9 shows experimentation result of J48 Algorithm based on the two methods.

Table 4.9 Experimentation results of J48 Algorithm based on the two methods.

Model Characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test option	I (80-20)				II 10-fold			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy (%)	93.5	94	90.4	93.8	**93.8**	95.1	91.7	95.2
Time taken	0.34	0.09	0.38	0.3	**0.09**	0.31	0.39	0.06
No of leaves	8339	9713	10707	14890	**8339**	9713	10707	14890
Size of trees	8541	9928	10918	15123	**8541**	9928	10918	15123
AV.TPR (%)	93.5%	94%	90.4%	93.8%	**93.9%**	95.1%	91.8%	95.3%
AV.FPR (%)	5.4%	5%	8.8%	4.5%	**4.5%**	3.1%	7.5%	3.2%
AV.PR	0.935	0.940	0.905	0.939	**0.938**	0.951	0.919	0.953
AV.RR	0.935	0.940	0.904	0.938	**0.939**	0.951	0.918	0.953
AV.ROC	0.985	0.986	0.973	0.985	**0.989**	0.990	0.979	0.989
CCI	955	960	923	958	**4794**	4858	4687	4865
ICI	66	61	98	63	**312**	248	419	241

As the above table shows, from eight different experiments, scenario No 1 (from method II) is the best model based on the aggregate result of number of leaves and size of trees generated, accuracy, time taken to build the model and ROC value. Correctly and incorrectly classified instances at this accuracy are 4794(93.8%) and 312(6.2%) respectively from 5,106 instances. Figure 4.3 shows the confusion matrix for the selected model. In addition to this the resulting

confusion matrix has shown that out of 364 undefined job records 364 (100%) of them are correctly classified. This shows that the model have correctly classified those crime data in their respective classes. Also, out of 2852 private records, 2759 (96.7%) of them are correctly classified and 44(1.5%) of them incorrectly classified as government. In addition, out of 347 housewives 321(92.5%) of them are correctly classified and 18(5.1%) are misclassified in private. Figure 4.3 shows the confusion matrix for scenario No 1 (method II).

Figure 4.3 Confusion Matrix output of the J48 algorithm for victim job target class

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i   <-- classified as
 364    0    0    0    0    0    0    0    0 |    a = Undefined
   0  477   33   14    0    0   11    0    0 |    b = Unemployed
   0   14 2759   44    9    2    9   15    0 |    c = Private
   0   14   90  644    0    0    0    1    0 |    d = Government
   0    0    7    0   90    0    0    0    0 |    e = Daily_Worker
   0    1    8    0    0   10    0    0    0 |    f = Others
   0    6   18    2    0    0  321    0    0 |    g = housewife
   0    2   10    1    0    0    0  125    0 |    h = Student
   0    0    1    0    0    0    0    0    4 |    i = NGO
```

Experiment 5: Classification of records using *offenderMartialStatus* target class

This experiment uses the attribute offender Martial Status to classify records. The attribute consists of two classes, namely single and married. The fifth experiment as well has been tested with the two methods as mentioned above. The first was 80-20 percentage split method with four scenarios mentioned earlier. Table 4.10 shows experimentation result of J48 Algorithm with two methods.

Table 4.10 shows that even if the accuracy is somehow low, the second experiment (method II scenario No1)still perform better when we see the aggregate result based on number and size of

tree, accuracy, time taken to build model and ROC value. The scenario shows that 4020 instances were correctly predicted as their respected classes. In addition, when we compare the actual correctly classified instance based on their label class i.e. classifying offenders marital status as married and single for all scenarios, irrespective of number of attribute, still unpruned registered better than pruned one.

Table 4.10 Experimentation results of J48 Algorithm based on the two methods.

Model Characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test option	I (80-20)				II 10-fold			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy (%)	78.8	88.1	94.8	96.1	**78.7**	88.1	95.6	97.2
Time taken	0.27	0.23	0.06	0.23	**0.08**	0.27	0.28	0.05
# of leaves	762	9662	8438	9083	**762**	9667	8438	9083
Size of trees	779	9793	8634	9288	**779**	9793	8634	9288
AV.TPR (%)	78.8%	88.1%	94.8%	96.2%	**78.7%**	88.2%	95.6%	97.2%
AV.FPR (%)	57.5%	21.4%	12.9%	9.6%	48.5%	21.3%	8.4%	4.9%
AV.PR	0.765	0.881	0.948	0.962	**0.771**	0.880	0.956	0.972
AV.RR	0.788	0.881	0.948	0.962	**0.787**	0.882	0.956	0.972
AV.ROC	0.741	0.906	0.966	0.971	**0.747**	0.917	0.985	0.988
CCI	805	900	968	982	**4020**	4503	4882	4965
ICI	216	121	53	39	**1086**	603	424	141

Similarly, when we compare the model in terms of size and leaves of tree, unpruned J48 models are still enormous and complex relative to pruned one. As a result of this, the algorithm might not reach optimality and generate more generalized decision tree rules. This is a fundamental problem in learning algorithms. Besides, such situation has its own impact in classification performance particularly classifying unseen or new instance. Subsequently, to solve the problem the researcher also selected pruned scenario that perform better accuracy. Accordingly, scenario No1 (method II) selected as the best J48 decision tree model. Table 4.11 shows the detail of confusion matrix for the selected model.

Table 4.11 Confusion Matrix output of the J48 algorithm with 10-fold cross validation.

Actual	Single	Married	Total	Correctly Classified (accuracy rate)
Single	**3532**	246	3778	93.4%
Married	840	**488**	1328	36.7%
Total	4372	734	5106	78.7%

From this experiment, 4020 instances have been classified correctly and 1086 instances have moved to the wrong class. Furthermore, out of 3778 single offender records, 3532 (93.4%) of them are correctly classified and 246(6.6%) of them are incorrectly classified as married. In addition, out of 1328married offender records, 488(36.7%) of them are correctly classified and 840(63.3%) are misclassified as single. The offenders' marital status is significantly misled and unreliable information would be provided for decision makers and crime analysts.

Experiment 6: Classification of records using *offenderJob* target class

This experiment uses the attribute offender job to classify records. The attribute consists of nine classes, namely unemployed, private, government, daily-worker, housewife, student, NGO and others. Table 4.12 shows experimentation result of J48 Algorithm based on the two methods.

Table 4.12 Experimentation result of J48 Algorithm based on the two methods.

Model Characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test option	I (80-20)				II 10-fold			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy (%)	89.9	92.2	89.9	92.3	**91.8**	94.2	91.8	94.1
Time taken	0.33	0.23	0.34	0.3	**0.27**	0.27	0.31	0.08
No of leaves	10619	11778	10619	15408	**10619**	11778	14228	15408
Size of trees	10916	12092	10916	15708	**10916**	12092	14508	15708
AV.TPR (%)	89	92.3	89.9	92.4	**91.9**	94.2	91.9	94.2
AV.FPR (%)	9	6.3	9	5.8	**6.7**	4.2	7	4.4
AV.PR	0.899	0.923	0.899	0.924	**0.918**	0.942	0.918	0.942
AV.RR	0.898	0.923	0.899	0.924	**0.919**	0.942	0.919	0.942
AV.ROC	0.968	0.973	0.968	0.977	**0.983**	0.986	0.983	0.987
CCI	818	942	818	943	**4691**	4812	4690	4809
ICI	103	79	103	78	**415**	294	416	297

As the above table shows that relatively scenario No1with 10-fold cross validation test option relatively perform better when we see the aggregate results in terms of number and size of trees, accuracy, time taken to build model and ROC value. It correctly classifies 4691 instances out of 5106. In addition, when we compare the actual correctly classified instance based on their label class i.e. classifying offenders' job as unemployed, private, government , daily-worker, housewife, student, NGO and others of all scenarios, irrespective of number of attributes unpruned registered better than pruned one. Figure 4.4 shows the detail of confusion matrix for the selected model.

Figure 4.4 Confusion Matrix output of the J48 algorithm for offender job target class

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h   <-- classified as
 986  123    6    0    0    0    3    0 |    a = Unemployed
  79 2837   15    6   25    2   22    0 |    b = Private
   4   31  201    0    0    0    3    0 |    c = Government
   2   19    1   46    1    1    2    0 |    d = Others
  10   25    1    1  247    0    2    0 |    e = Student
   0    1    0    2    0   55    1    0 |    f = housewife
   2   20    4    0    0    0  316    0 |    g = Daily_Worker
   0    0    0    0    1    0    0    3 |    h = NGO
```

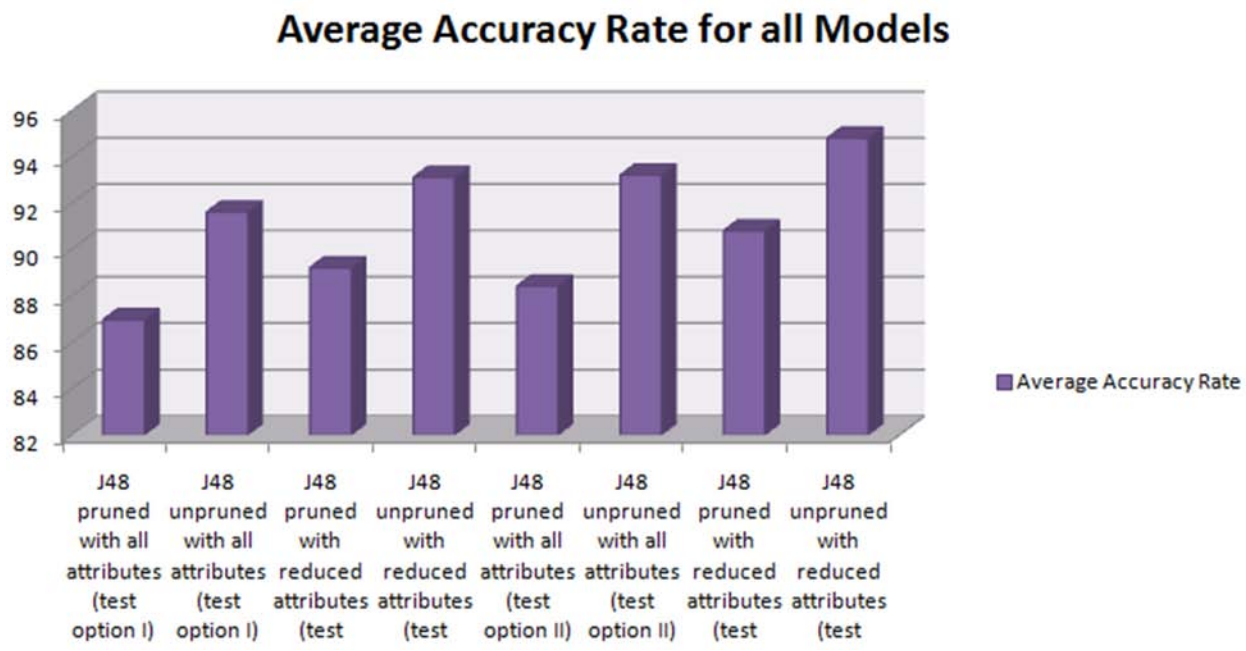
The confusion matrix for J48 decision tree presented above in Figure depicts that out of the total records provided to the program, 986 (88.1%), 201(84.1%) and 247(86.1%) records have been classified correctly in the class of unemployed, government and student respectively. On the other hand, 123 (11%) records have been incorrectly classified as private while actually they were supposed to be in the unemployed class and 31(12.9%) records have been classified incorrectly as private while actually they are in the government. This portrays that from the total

records, 4691 (91.8%) records have been classified correctly while the remaining 415(8.2%) records were classified incorrectly.

4.1.4.1.1 Comparison of J48 Algorithm

In general when we compare both methods in terms of classification model accuracy, 10 fold cross validation (method II) registered better than 80/20 percentage split method. Their minimum and maximum accuracy results are 78.3% and 98.5% for 80/20 percentage split method and 78.7% and 99%, for 10 fold cross validation method. Additionally, the following figure 4.5 shows variations of accuracy among models.

Figure 4.5 Average accuracy of all models in the J48 experiment



4.1.4.1.2 Generating Rules from Decision Trees

To make a decision tree model which is more human-readable, each path from root to leaf can be transformed into an IF-THEN rule. If condition is satisfied, the conclusion follows. The algorithm decision tree is the best known method for deriving rules from classification trees.

This is simply by traversing any given path from the root node to any leaf. The numbers in parentheses at the end of each leaf indicate the number of examples in the leaves. The number of misclassified examples would also be given after a slash, and hence it is possible to compute the success fraction (ratio) to estimate the level of confidence or likelihood of predictability of the class that tells how much the rule is strong. From the entire models that were generated, J48 pruned tree model with all attributes is selected as the best model for rule generation. Rules provided by decision tree models can be easily assimilated by human without any difficulty.J48 pruned tree model with all attributes have produced different rules. However, the researcher selected few rules that cover most of the data points in the study. After the rule extraction, the researcher show to domain experts to discuss upon the generated rules. Some of the rules generated by J48 pruned tree models with all attributes are:

Rule 1: If CrimeType = Deforestation AND VictimJob= undefined AND OffenderEduL= Primary: THEN Medium Level (26/7).

Rule 2: If crime type = Snatching AND PoliceStation =MPS AND OffenderJob =Private AND OffenderMartialStatus = Married AND VictimJob Unemployed: THEN High_Level (6/0).

Rule 3: If Crime type = Theft_of_Vehicle part AND offenderMartialStatus =Single AND OffenderEduL = Secondary: THEN High_Level (120.0/0).

Rule 4: If Crme type = Trying_to_Commit_Murder AND VictimJob = Private ANDVictimAge = 26-30: THEN Medium Level (20.0/4.0).

Rule 5: If Victim age = 16-20 AND OffenderMartialStatus =Married AND CrimeCatagory = Crimes_Against_Person: THEN VictimMartialStatus will be Single (50.0/6.0).

Rule 6: If VictimAge = 16-20 AND OffenderMartialStatus = Single: THEN Victim Martial Status will be Single (432.0/12.0).

Rule 7: If VictimsReligion = Orthodox AND VictimAge =26-30 AND ParticularPlace = Mender_7_Shegole: THEN VictimMartialStatus will be Single (15/0).

Rule 8: If VictimReligion = Muslim AND VictimAge = 21-25 AND ParticularPlace = Medhanealem_School AND VictimSex = Male: THEN Victims marital Status will be Single (12/0).

Rule 9: If VictimAge = 26-30 AND ParticularPlace = 'Meketiya' AND PoliceStation – SPS AND Time =9:00:00 PM -12:00:00 PM: THEN victim Martial Status will be Married (10.0/1.0).

Rule 10: If particularplace = Enkulale_Fabrica AND CrimeCatagory = Crime_Against_Property AND offenderReligion = Orthodox: THEN 1:00:00 PM - 4:00:00 PM (12.0/3.0).

Rule 11: If victimSex = Male AND ParticularPlace =Rufael AND PoliceStation = PPS AND OffenderMartialStatus = Married AND CrimeCatagory = Crime_Against_Property: THEN Victim job will be Private (16.0/0).

Rule 12: If VictimSex = Male AND ParticularPlace = Medhanealem_School AND OffenderJob = Private: THEN Victim job will be private (35.04/4.02).

Rule 13: If VictimSex = Male AND ParticularPlace = Menene AND CrimeLevel = Medium_Level AND VictimReligion = Orthodox: THEN Victim job will be government (14.0/0).

Rule14: If VictimSex = Male AND ParticularPlace = Meketiya AND VictimMartialStatus = Single AND OffenderJob = Private AND VictimAge = 26-30: THEN Victim job will be Private (8.0/0).

Rule 15: If ParticularPlace = Rufael AND CrimeType = Miscellaneous_Theft AND OffenderJob = Unemployed AND OffenderEduL = Secondary: 1:00:00 AM – 4:00:00 AM (5.0/0).

Rule 16: If ParticularPlace = Iran_Embassy AND VictimJob = Private : 1:00:00 PM – 4:00:00 PM (15.02/0.02).

Rule 17: If ParticularPlace = Arat_Menta AND OffenderJob = Private AND PoliceStation = KPS AND OffenderMartialStatus = Single: 1:00:00 PM – 4:00:00 PM (12/0)

Rule 18: If ParticularPlace = Margeja AND VictimJob = Private AND CrimeCatagory = Crime_Against_Person AND OffenderEduL = Junior: THEN offender job status will be Unemployed (19.0/0).

Rule 19: If ParticularPlace = Zero_3_Kebele_Mezenagna AND OffenderMartialStatus = Single AND OffenderReligion = Orthodox AND VictimSex =Male: THEN Offender job status will be Private (16/01).

Rule 20: If particularPlace = Wingate_industry_Sefer AND PoliceStation =PPS AND VictimJob = Private AND OffenderEduL = Secondary: THEN Offender job status will be Private (14.0/0).

Rule 21: ParticularPlace = Mender_7_Shegole AND CrimeType = Miscellaneous_Theft_Act AND VictimsReligion = Orthodox AND VictimSex = Female: THEN Offender job status will be Private (17.0/3.0).

Rule 22: If PoliceStation = SPS AND CrimeType = Miscellaneous_Theft_Act AND VictimAge = Above 60: THEN the particular place will be Sheromeda (5.0/0).

Rule 23: If PoliceStation = AGPS AND CrimeType = Miscellaneous_Theft_Act AND Time = 1:00:00 PM – 4:00:00 PM AND VictimAge = 21-25: THEN the particular place will be Oil_Libya (4.0/0).

Rule 24: If PoliceStation = AGPS AND CrimeType = Deforestation AND VictimJob = Undefined: THEN the particular place will be Deleber_Forest (9.0/0).

Rule 25: If policeStation = KPS AND OffenderEduL = Junior AND OffenderJob = Unemployed: THEN the particular place will be Margeja (23.0/0).

Rule 26: If Policestation = PPS AND CrimeType = Threat AND OffenderEduL = Second AND Time = 9:00:00 PM – 12:00:00 PM AND VictimSex = Male: THEN the particular place will be Filance (6.0/0).

Rule 27: If ParticularPlace = Antari_Sefere AND VictimJob = Private AND CrimeCatagory = Crime_Against_Property: THEN Offender Martial status will be Single (21.0/3.0).

Rule 28: If ParticularPlace = Meketiya AND VictimJob = Unemployed: THEN Offender Martial status will be single (23/0).

Rule 29: If ParticularPlace = Biru_Wonze AND VictimJob = Private AND CrimeType = Miscellaneous_Theft_Act AND Time = 1:00:00 PM – 4:00:00 PM: THEN Offender Martial status will be Married (9.0/0).

Rule 30: If OffenderAge = 33-37 AND ParticularPlace =Medhanealem_School AND VictimSex = Male: THEN Offender Martial status will be Married (10.0/0).

As depicted in the above rules generated from the tree built, the experimentation helps to get demographic features which characterize offenders and victims as well as features which characterized crimes types, locations and times committed. These rules will also help to predict demographic features of new offenders and victims. For example, if a new victim is between 16 and 20, and the criminal marital status is married and the crime committed is categorized as crime against person (like, Simple assault, beating, body injury, crippling part of a body, etc.), then her/his marital status is regarded as single (as defined in rule 5).Similarly, if a new crime is committed at 'EnkulaleFabrica', a place located around 'Paster', and if the crime is categorized as crime against property(like, snatching, different type of theft, arson, miscellaneous cheating, taking bribes, etc.) and the offender is the follower of orthodox religion then the crime is committed from the time 1:00:00 PM-4:00:00PM (as defined rule 10).

All attributes being constant in the first four rules, the type of crimes committed matters the level of crime (high, medium and low). The domain expert accepted this by saying 'nationally, those crime types like, deforestation and trying to commit murder labeled as middle level crimes and

snatching and theft of vehicle accessories as high'. According to those expertise, there are 46 crime titles identified nationally.

Rule 5 and 6 depicted that whatever the offender's marital status is, if the victim's age is between 16 and 20, his/her marital status will become single.

4.1.4.2 Model Building Using PART Rule Induction Algorithm

The second data mining classification technique applied in this research was PART rules induction algorithm. As mentioned in chapter two of literature review section 2.3.1.5, there are many rule induction algorithms but the researcher selected PART for the reason that PART has the ability and potential to produce accurate and easily interpretable patterns/rules that helps to achieve the research objectives. As mentioned in chapter two PART is a separate-and-conquer rule learner and proposed by Witten and Frank [53]. It works by generating a rule that covers a subset of the training examples and then removing all examples covered by the rules from the training set. The final rule set is the collection of the rules discovered at every iteration of the process. The rules are in standard form of IF-THEN rules.

To build the Rule induction model, WEKA software package and the same 5,106 crime data set were used as an input. The experiments were performed analogously as the researcher did in former model. The parameters are partially adjusted and default values were used with reduced and all attributes. Accordingly, the experiment of all scenarios with the two model validation methods (80/20 percentage split and 10-fold cross validation) was conducted.

Experiment I: Classification of records using *CrimeLevel*target class

This experiment uses the attribute Crime Level to classify records. The attribute consists of three classes, namely High, Medium and Low Level. The first experiment as well has been tested based on the two methods as mentioned above. Accordingly, the experiment of all scenarios based on both methods is illustrated below in Table 4.13.

Table 4.13 Experiment result of PART algorithm based on the two methods

Model characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test Option	I				II			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy	94.6	97.2	**94.7**	97.6	**95.3**	97.7	94.7	97.7
Time taken	1.25	1.78	**0.86**	1.89	**0.78**	1.82	0.92	1.79
No of rules	391	654	**377**	786	**391**	654	377	654
AV.TPR	94.6	97.3	94.7	97.6	95.3	97.7	94.8	0.977
AV.FPR	3.6	1.8	3.1	1.2	2.6	1.3	3.1	1.3
AV.PR	0.947	0.973	0.947	0.977	0.953	0.978	0.948	0.978
AV. RR	0.946	0.973	0.947	0.976	0.953	0.977	0.948	0.977
AV.ROC	0.990	0.987	0**.987**	0.993	**0.992**	0.995	0.989	0.995
CCI	966	993	967	997	4868	4991	4840	4991
ICI	55	23	54	24	238	115	266	115

As the above table shows, the registered performance (accuracy) of PART with unpruned decision list is better than the pruned one. Among the four scenarios experimented with 80-20 percentage split (method I), scenario No3 relatively registered better aggregate performance in terms of accuracy, time taken to build the model, rules generated and ROC value. It registered the accuracy of 94.7%.This shows that out of 1021 records, 967(94.7%) of them correctly classified, while 54(5.2%) are misclassified.

The next experiment has been tested with four scenarios as mentioned above with 10-fold cross validation. The result shows that scenario No 2 and 4 registered the best accuracy and accounts to 97.7%. It correctly classifies 4991 instances out of 5106. In addition, scenario No1 registered better time to build the model.

Analogously, in terms of the number of rules generated scenario No 3 is relatively minimum, hence it is considered best because it reduces the complexity of the generated rules. The average ROC performance measure indicates that scenario No 4 perform better than the rest of the three and accounts 99.5%.Finally, when we compare the actual correctly classified instance based on their label class i.e. classifying crime levels as high, medium and low level for all scenarios, irrespective of number of attribute unpruned registered better than pruned.

However, the rules generated from unpruned, PART models are found to be relatively enormous and complex to the pruned one. This is a fundamental problem in learning algorithms. Besides, such situation has its own impact on classification performance particularly classifying unseen or new instances. Subsequently, to solve the problem the researcher selected pruned scenario that performs in a better accuracy. Accordingly, scenario No1 (method II) with 10-fold cross validation test option selected as the best PART rule induction model.

Consequently, Table 4.13,irrespective of reduced or all attribute the performance of unpruned PART algorithm registered better than pruned one in terms of accuracy. However, based on rules generated, from scenario No3 with 10-fold cross validation test option is more understandable

and less complex than others models generated. The detailed confusion matrix for selected model has been discussed below in Table 4.14.

Table 4.14 Confusion Matrix output of the PART algorithm based on 10-fold cross validation.

Actual	Predicted			Total	Correctly Classified (accuracy rate)
	Medium Level	Low Level	High Level		
Medium Level	**2054**	63	43	2160	95%
Low Level	78	**1800**	12	1890	95.2%
High Level	11	31	**1014**	1056	96
Total	2143	1894	1069	5106	**95.3%**

Table 4.14 summarizes the correct classification and incorrect classifications in each class labels for the PART model built in experiment one. According to the confusion matrix, among the 2160 instances which are actually defined to be medium crime level, 2054 of them are correctly classified while the rest instances were misclassified, 63 in low level class and the rest 43 into a high class crime level.

Among the 1890 instances which are actually regarded as low level crime, 1800 of them are correctly classified while the rest are misclassified, 78 in to medium level crime and only 12 in to high crime level. The overall accuracy is obtained from the percentage of the total correct classifications (2054+1800+1014) to the total dataset provided to the software (5106).

Experiment II: Classification of records using time target class

This experiment uses the attribute time to classify records. The attribute consists of eight classes; 1:00:00 PM-4:00:00PM, 9:00:00PM-12:00:00 PM, 9:00:00AM-12:00:00AM, 1:00:00AM-4:00:00AM, 5:00:00PM-8:00:00 PM, 5:00:00AM-8:00:00AM, Night Unknown and Unknown.

Table 4.15presented below shows that experimentation result of PART Algorithm for two methods.

Table 4.15 Experimentation results of PART Algorithm based on the two methods.

Model characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test Option	I				II			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy	72.4	73.2	71.5	71.8	**78.4**	78.8	77.9	78
Time taken	7.11	6.15	7.1	6.3	**7.24**	6.3	7.02	6.4
No of rules	1343	1473	1345	1450	**1343**	1473	1345	1450
AV.TPR	72.5	73.3	71.6	71.9	78.4	78.8	78	78
AV.FPR	5.8	5.6	6.1	5.9	4.4	4.4	4.5	4.5
AV.PR	0.725	0.733	0.715	0.719	0.785	0.789	78	0.781
AV. RR	0.725	0.733	0.716	0.719	0.784	0.788	78	0.780
AV.ROC	0.931	0.922	0.933	0.926	0.952	0.948	0.952	0.952
CCI	740	748	731	734	4004	4026	3981	3985
ICI	281	273	290	287	1102	1080	1125	1121

Table 4.15 shows that the experiment which has been tested with 10-fold cross validation test option performing better. Accordingly, scenario No2 registered the best accuracy and accounts to 78.8%. It correctly classifies 4026 instances out of 5106.

Similarly, among the four scenarios experimented with 10-fold cross validation (method II) with varying parameters, scenario No1 relatively registered better aggregate performance in terms of accuracy, time taken to build the model, tree complexity (rules generated) and ROC value. It registered the accuracy of 78.4%.This shows that out of the training set of 5106 records, 4004(78.4%) of them are correctly classified, while 1102(5.2%) misclassified.

Accordingly, scenario No1 (Building pruned decision tree with all attribute) of 10-fold cross validation (method II) selected as the best PART rule induction model for this particular experiment. Figure 4.8 below shows output from the PART rule induction learner for the selected model.

Figure 4.6 Confusion Matrix output of the PART algorithm for time target class

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 838  24  54  37  32  28   0  11 |   a = 1:00:00 PM-4:00:00 PM
  24 513  42  23  18  15   1  12 |   b = 9:00:00 PM-12:00:00 PM
  59  37 977  30  58  21   0  11 |   c = 9:00:00 AM-12:00:00 AM
  38  32  32 365  27   3   0   3 |   d = 1:00:00 AM-4:00:00 AM
  58  25  68  13 729  29   0   9 |   e = 5:00:00 PM-8:00:00 PM
  43  20  35  26  43 464   0   5 |   f = 5:00:00 AM-8:00:00 AM
   2   2   0   0   0   0   3   0 |   g = Night Unknown
   6  10  13   3  13   7   0 115 |   h = Unknown
```

The accuracy of this learning scheme was 78.4%, and this indicates that out of the total number of records supplied 4004 (78.4%) records were classified correctly while the remaining 1102(21.6%) classified incorrectly. Moreover, the results of the experiment has shown that about 81.8 %of the records in the class of 1:00:00PM-4:00:00PM were correctly classified while about 79.1% and 81.8% of the records in the 9:00:00PM-12:00:00PM and 9:00:00AM-12:00:00AM classes respectively were classified correctly.

Experiment III: Classification of records using *VictimMartialStatus* target class

This experiment uses the attribute Victim Martial Status to classify records. The attribute consists of three classes', namely single, married and undefined marital status. The third experiment as well has been tested with the two methods as mentioned above. Table 4.16 shows experimentation result of PART Algorithm with two methods.

Table 4.16 Experimentation results of PART Algorithm based on the two methods.

Model characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test Option	I				II			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy	91.3	94.8	90.2	94.4	**91.5**	95.7	91.2	95
Time taken	2.11	2.59	2.15	2.9	**2.11**	2.2	2.15	2.76
No of rules	743	906	751	926	**743**	906	751	926
AV.TPR	91.4	94.8	90.2	94.4	91.6	95.8	91.3	95.1
AV.FPR	8.1	4.6	9.3	5.3	8.1	3.7	8.5	4.4
AV.PR	0.914	0.948	90.2	0.944	0.916	0.958	0.913	0.951
AV. RR	0.914	0.948	90.2	0.944	0.916	0.958	0.913	0.951
AV.ROC	0.967	0.979	0.966	0.984	**0.977**	0.984	0.976	0.987
CCI	933	968	921	964	4677	4891	4660	4854
ICI	88	53	100	57	429	215	446	252

Table 4.16 also shows that relatively method II (10-fold cross validation test option) scenario No1 still performs better when we see the combined results of number and size of tree, accuracy, time taken to build model and ROC value. It correctly classifies 4677 instances out of 5106. In addition, when we compare the actual correctly classified instance based on their label class i.e. classifying victims' marital status as single, married and undefined for all scenarios, irrespective of number of attribute still unpruned registered better than pruned one. Subsequently to solve the problem, the researcher also selected pruned scenario that performs in better accuracy. Accordingly, scenario No1 of 10-fold cross validation (method II) is selected as the best PART rule induction model. Table 4.17 shows the detail of confusion matrix for the selected model.

Table 4.17 Confusion Matrix output of the PART algorithm with 10-fold cross validation.

Actual	Predicted			Total	Correctly Classified (accuracy rate)
	Undefined	Married	Single		
Undefined	**364**	0	0	364	100%
Married	0	**2572**	163	2735	94
Single	0	266	**1741**	2007	86.7
Total	364	2838	1904	5106	**91.5%**

The confusion matrix depicted that out of the total of records provided to the program, 364(100%), 2572(94%) and 1741(91.5%) records were classified correctly in the class of undefined, married and single respectively. On the other hand, 163 (5.9%) records were incorrectly classified as single while actually they were supposed to be in the married class also 266 (13.2%) records were classified incorrectly as married while actually they are in the single class. Hence, this indicates that records whose class is married were classified with a minimum error compared to the record in the class single.

Experiment IV: Classification of records using *victimJob* target class

This experiment uses the attribute victim job to classify records. The attribute consists of nine classes, namely undefined, unemployment, private, government, daily-worker, housewife, student, NGO and others. Table 4.18 below shows experimentation result of PART rule induction Algorithm with two methods.

Table 4.18Experimentation results of PART Algorithm based on the two methods.

Model characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test Option	I				II			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy	86.7	88.8	85.9	88.2	**88.7**	90.9	88.5	89.5
Time taken	5.02	4.63	5.05	4.58	**5.01**	4.71	5.18	4.85
No of rules	938	1110	958	1112	**938**	1110	958	1106
AV.TPR %	86.8	88.8	86	88.2	88.8	91	88.5	89.5
AV.FPR	9.6	6.2	9.5	6.5	7.4	4.8	7.6	5.4
AV.PR	0.865	0.887	0.856	0.879	0.886	0.910	0.884	0.896
AV. RR	0.868	0.888	0.860	0.882	0.888	0.910	0.885	0.895
AV.ROC	0.967	0.969	0.965	0.967	**0.980**	0.981	0.979	0.981
CCI	886	907	978	901	4534	4645	4520	4572
ICI	135	114	143	120	572	461	586	534

Similarly, when we compare the model in terms of rules generated, unpruned PART models are still relatively enormous and complex to the pruned one. As a result of this, the algorithm might not reach optimality and generate more generalized rules. This is a fundamental problem in learning of algorithms. Besides, such situation has its own impact on classification performance particularly classifying unseen or new instance. Subsequently, to solve the problem the researcher selected pruned scenario that performs in a better accuracy. Accordingly, scenario No1 (Building pruned PART decision list with all attribute) of 10-fold cross validation (method II) is selected as the best PART model. Figure 4.7 below shows the detail of confusion matrix for the selected model.

Figure 4.7 Confusion Matrix output of the PART algorithm for victim job target class

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i   <-- classified as
 364    0    0    0    0    0    0    0    0 |   a = Undefined
   0  408   76   35    1    0   12    3    0 |   b = Unemployed
   0   30 2681   87   13    4   26   11    0 |   c = Private
   0   30  112  586    5    0    8    8    0 |   d = Government
   0    1   14    3   79    0    0    0    0 |   e = Daily_Worker
   0    0    9    1    0    9    0    0    0 |   f = Others
   0    8   39    4    2    0  292    2    0 |   g = housewife
   0    5   18    4    0    0    0  111    0 |   h = Student
   0    0    1    0    0    0    0    0    4 |   i = NGO
```

The result from this experiment shows that out of the 5106 total records, 4534 (88.7%) of them are correctly classified. And 572 (11.2%) records are incorrectly classified. In addition to this, the resulting confusion matrix has shown that out of 364 undefined job records, 364 (100%) of them are correctly classified. This shows that the model has correctly classified those crime data in their respective class. And out of 2852 private records, 2681 (94%) of them are correctly classified and 87 (3%) of them are incorrectly classified as government. In addition, out of 347 housewives 292 (84.1%) of them are correctly classified and 39 (11.2%) misclassified in private.

Experiment V: Classification of records using *offenderMartialStatus* target class

This experiment uses the attribute offender Martial Status to classify records. The attribute consists of two classes, namely single and married. This experiment as well has been tested based on the two methods as mentioned above. Table 4.19below shows experimentation result of PART Algorithm with two methods.

As Table 4.19 shows, even if the accuracy is somehow low, the second experiment (method II scenario No1) still performs better when we see the aggregate result based on number and size of tree, accuracy, time taken to build model and ROC value. It correctly classified4632 instances out of 5106. In addition, when we compare the actual correctly classified instances based on their label class, i.e. classifying offenders' marital status as married and single for all scenarios, irrespective of number of attributes still unpruned registered better than the pruned one.

Table 4.19 Experimentation results of PART Algorithm based on the two methods.

Model characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test Option	I				II			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute %	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy	91.1	91.2	91.6	94.9	**90.7**	93.6	91.5	95.1
Time taken	4.52	3.95	4.19	3.88	**4.01**	4.24	4.23	3.82
No of rules	789	1057	794	981	**789**	1057	794	981
AV.TPR %	91.2	91.3	91.7	94.9	90.7	93.7	91.6	95.1
AV.FPR	21.6	14.3	20	11.4	18	10.5	16.8	8.3
AV.PR	0.910	0.914	0.915	0.949	0.905	0.937	0.914	0.951
AV. RR	0.912	0.913	0.917	0.949	0.907	0.937	0.916	0.951
AV.ROC	0.936	0.912	0.947	0.973	**0.953**	0.956	0.965	0.982
CCI	931	932	936	969	4632	4783	4676	4858
ICI	90	89	85	52	474	323	430	248

In the same way, while we compare the model in terms of generated rules unpruned PART models still enormous and complex relative to pruned one. As a result of this, the algorithm might not reach optimality and generate more generalized rules. This is a fundamental problem in learning of algorithms. Besides, such situation has its own impact in classification performance particularly classifying unseen or new instance. Subsequently, to solve the problem the researcher also selected pruned scenario that performs in a better accuracy. In view of that,

scenario No1 (Building PART pruned decision list with all attribute) with 10-fold cross validation (method II) selected as the best PART rule induction model. Table 4.20below shows the detail of confusion matrix for the selected model.

Table 4.20Confusion Matrix output of the PART algorithm with 10-fold cross validation

Actual	Medium Level	Low Level	High Level	Total	Correctly Classified (accuracy rate)
Medium Level	**817**	14	6	837	97.6
Low Level	23	**700**	2	725	96.5
High Level	3	7	**428**	438	97.7
Total	843	721	436	2000	97.2

From this experiment, 4632 instances were correctly classified whereas 474 instances moved to the wrong class. And out of 3778 single offender records, 3606 (95.4%) of them are correctly classified and 172 (4.5%) of them are incorrectly classified as married. In addition, out of 1328 married offender records, 1026 (77.2%) of them are correctly classified and 302 (22.8%) are misclassified as single. This significantly misleads offender's marital status and provides unreliable information for decision makers and crime analysts.

Experiment VI: Classification of records using *offenderJob* target class

The last experiment uses the attribute offender job to classify records. The attribute consist of nine classes, namely unemployed, private, government, daily-worker, housewife, student, NGO and others. Table 4.21 shows experimentation result of PART Algorithm with two methods.

Table 4.21 Experimentation results of PART Algorithm based on the two methods.

Model characteristics	Experiment (Scenario)							
	1	2	3	4	1	2	3	4
Test Option	I				II			
Pruned	Yes	No	Yes	No	Yes	No	Yes	No
Attribute %	All	All	Reduced	Reduced	All	All	Reduced	Reduced
Accuracy	85.1	86.6	84.1	87.1	**87.3**	89	87.2	88.6
Time taken	5.88	5.49	5.46	5.35	**5.68**	5.1	5.73	5.27
No of rules	1004	1185	1051	1157	**1004**	1185	1051	1157
AV.TPR %	85.1	86.7	84.1	87.2	87.4	89.1	87.2	88.7
AV.FPR	10.7	8.3	11.9	7.6	8.7	6.5	8.9	6.7
AV.PR	0.852	0.870	0.842	0.874	0.873	0.892	0.872	0.888
AV. RR	0.861	0.867	0.841	0.872	0.874	0.891	0.872	0.887
AV.ROC	0.961	0.963	0.958	0.966	**0.973**	0.976	0.972	0.978
CCI	869	885	859	890	4461	4549	4453	4529
ICI	152	136	162	131	645	557	653	577

As mentioned earlier, table 4.21 also shows that method II (10-fold cross validation test option) scenario No**1** relatively perform better when we see the aggregate results of number and size of tree, accuracy, time taken to build model and ROC value. It correctly classified 4461 instances out of 5106. In addition, when we compare the actual correctly classified instance based on their label class, i.e. classifying offenders job as unemployed, private, government , daily-worker, housewife student, NGO and others for all scenarios, irrespective of number of attribute, still unpruned registered better than the pruned one. Subsequently to solve the problem, the researcher also selected pruned scenario that performs in a better accuracy. Accordingly, scenario No1 (Building pruned decision list with all attribute) of 10-fold cross validation (method II) selected as the best PART model. Figure 4.8 shows the detail of confusion matrix for the selected model.

Figure 4.8 Confusion Matrix output of the PART algorithm for offender job target class

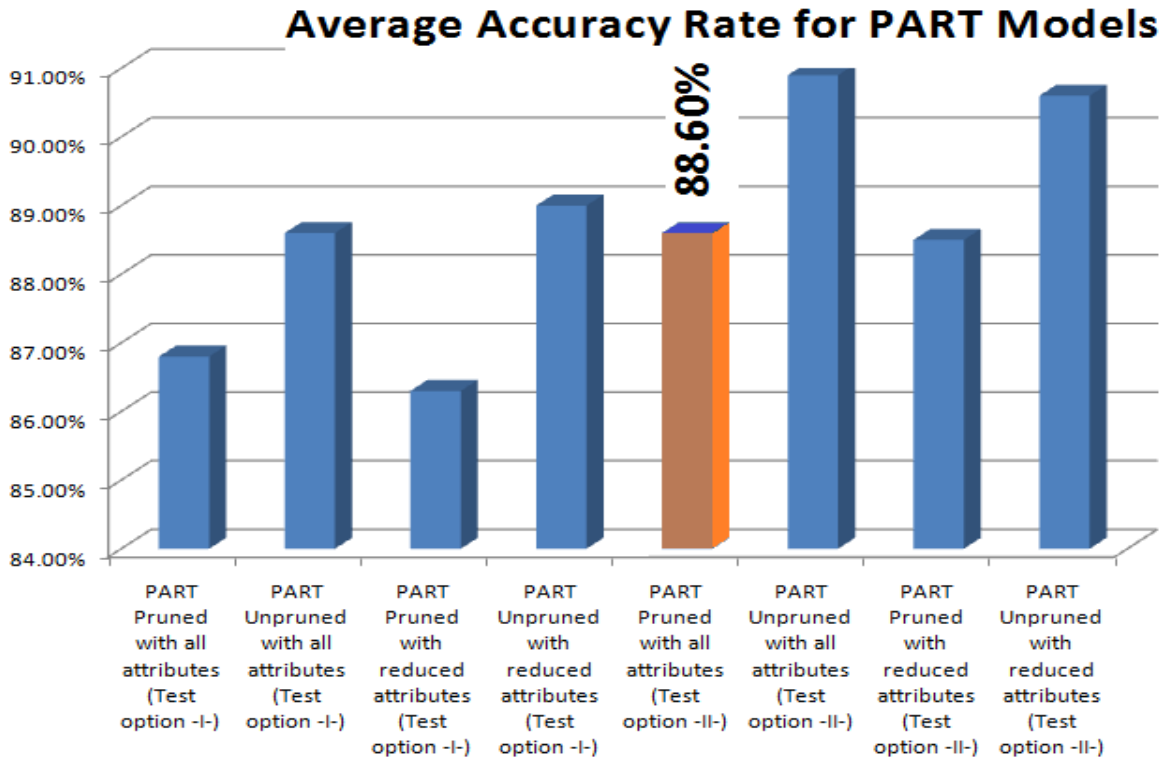
```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h   <-- classified as
 938  143   14    4   11    1    7    0 |    a = Unemployed
 132 2726   33   12   39    3   41    0 |    b = Private
   8   39  177    3    4    0    8    0 |    c = Government
   3   19    1   39    4    0    6    0 |    d = Others
  15   38    1    1  229    0    2    0 |    e = Student
   1    6    0    0    2   50    0    0 |    f = housewife
   5   30    5    2    1    0  299    0 |    g = Daily_Worker
   0    0    0    0    1    0    0    3 |    h = NGO
```

Figure 4.10 depicts that out of the total records provided to the program, 938 (83.8%), 177(74%) and 229(80%) records were classified correctly in the class of unemployed, government and student respectively. On the other hand, 143 (12.7%) records were incorrectly classified as private while actually they were supposed to be in the unemployed class also, 39 (16.3%) records were classified incorrectly as private while actually they are in the government. This portrays that from the total records, 4461 (87.3%) were classified correctly while the remaining 645(12.7%) records were classified incorrectly.

4.1.4.2.1 Comparison of PART Algorithm

As the researcher did for J48 algorithm, when both methods are compared in terms of classification model accuracy and 10-fold cross validation, (method II) still registered better than 80/20 percentage split method. Their minimum and maximum accuracy results are 71.5% and 97.6% for 80/20 percentage split method and 77.9 % and 97.7 % for 10 fold cross validation method. Additionally, the following Figure 4.9shows average accuracy rate for all models of PART rule induction algorithm.

Figure 4.9 Average accuracy of all models in the PART algorithms.



4.1.4.2.2 Analyzing interesting Rules from PART algorithms

PART rule induction algorithm is the second method in the study for generating rule. But, listing all the rules here will be quite cumbersome, thus rules which are highly predictive are selected and discussed at the finding of this study based on success ratio. The success ratio of rules is found in parenthesis just at the end of the predictive rules. The numbers in parenthesis at the end of each rule tells the number of instances in the rule. If one or more rules were not pure (that is all in the same class), the number of misclassified cases also are given after slash (/). The researcher has converted these numbers into percent to compare the chance of the rule to be correct with that of its chance to be incorrect. The greater the number before the parenthesis the greater the chance of the rule to predict the class indicated by that particular rule. After the rule

extraction, the researcher also went back to domain experts to discuss upon the generated rules. Therefore, some of the rules generated by PART pruned tree models with all attributes are:

Rule: 1 If VictimJob = Undefined AND OffenderJob = Private AND OffenderEduL = Secondary AND OffenderReligion = Orthodox AND Time = 9:00:00 AM-12:00:00 AM AND OffenderAge = 23-27: THEN crime level will be low_Level and nifty one records of the crime correctly satisfying this rule. (91.0/2.0)

Rule 2: If CrimeType = "Miscellaneous_Theft" AND OffenderSex = "Male" AND PoliceStation = "AGPS" then crime level will be Medium Level and night six records of the crime correctly satisfying this rule. (96.0)

Rule 3: If CrimeType = "Snatching" AND PoliceStation = "MPS" AND OffenderJob = "Private"ANDOffendermarrtialStatus = "Single" then crime level will be High_Level and thirty records of the crime correctly satisfying this rule. (30.0)

Rule 4: If CrimeCode> "28" AND OffenderSex = "Male" AND OffenderJob = "Private" then crime level will be Medium _Level and fifty five records of the crime correctly satisfying this rule. (55.0)

Rule 5: If CrimeCode> "41" AND CrimeCode<= "43" then crime level will be Low_Level and three hundred eighteen records of the crime correctly satisfying this rule. (318.0/24.0)

Rule No 5 implies that crimes like beating, threatening, insulting, violation of rules and regulations, breaking out of imprison, disturbing workplace etc. are classified as low level crime. There are 342 records having this property, from which 318 records are correctly classified and

the rest 24 records are misclassified. This implies that if the crime has a code greater than 41 and less than or equal to 43, the likely hood of crime level being predicted as low is 92.9%.

Rule 6: If CrimeType = “Miscellaneous_Cheating_Act” AND VictimJob = “Private” AND Time =“1:00:00 PM-4:00:00 PM” THEN crime level will be Medium_Level and forty three records of the crime correctly satisfying this rule. (43.0)

Rule 7: ParticularPlace =Mender_7_Shegole AND VictimReligion = Orthodox AND CrimeCode>25 AND OffenderSex = Male AND VictimSex = Female AND VictimMartialStatus = Single: THEN 1:00:00 PM – 4:00:00 PM (8.0/1.0)

The above rule implies that 8 records have the following characteristics: if the place where the crime is committed is at Mender 7 Shegole, and the victim is female, single and follower of orthodox religion and the offender is male, then the time she is affected or exposed to crime is between 1:00:00 PM-4:00:00PM. The likelihood that the crime being committed between the stated times gaps is 88.8%.

Rule 8: PoliceStation = SPS AND VictimJob = Unemployed AND OffenderEduL = Primary AND OffenderJob = Private: 9:00:00 AM – 12:00:00 (10.0).

Rule 9: VictimAge 26-30 AND VictimJob = Private AND OffenderJob = Private AND OffenderReligion = Orthodox: 5:00:00 AM = 8:00:00 AM (12.0).

Rule 10: CrimeCode< = 25 AND PoliceStation =PPS AND OffenderSex = Male AND VictimJob = Private AND OffenderEduL = Secondary: THEN 5:00:00 PM - 8:00:00 PM (8.0).

Rule 11: VictimSex = Male AND ParticularPlace =Meketiya AND VictimMartialStatus = Married: THEN Victims job will be Private (46.05/0.02).

Rule 12: VictimSex = Male AND ParticularPlace = Rufael AND PoliceStation = PPS AND OffendeMartialStatus = Single AND OffenderAge=23-27: THEN Victim job will be Private. (19.0)

Rule 13: VictimSex = Male AND VictimAge = 31-35 AND OffenderJob = Private AND VictimReligion = Orthodox: THEN Victim job will be Private (48.0/4.0).

Rule 14: VictimSex = Male AND OffenderJob = Governoment AND VictimAge = 21-25 AND OffenderEduL = Secondary: THEN Victim job will be Government (11.0).

Rule 15: VictimAge = 51-55 AND OffenderJob = Private AND VictimJob = Private: THEN victims marital status will be Married (45.0).

Rule 16: VictimAge = 56-60 AND CrimeCatagory = Crime_against_Property AND OffenderSex = Male: THEN Victim marital status will be married (69.0).

Rule 17: VictimAge = 16-20 AND OffenderMartialStatus = Single AND CrimeType = Beat: THEN Victim marital status will be Single (142.0).

Rule 17, selected from the rules generated by PART gives correct result of 142 out of 142 instances that it covers. From this, the likelihood of predictability of victim's marital status by the above predictors is 100 %.

Rule 18: ParticularPlace = Mender_7_Shegole AND CrimeType = Miscellaneous_Theft AND VictimReligion = Muslim: THEN Offender job will be Private (21.0).

Rule 19: ParticularPlace = Enkulale_Fabric AND VictimMarrtialStatus = Married AND CrimeCode< = 33: THEN Offender job will be Private (20.01).

Rule 20: OffenderEduL = Diploma AND VictimJob = Government AND OffenderMartialStatus = Married:THEN Offender job will be Private (8.0)

Rule 21: ParticularPlace = Filance AND VictimJob = Private AND VictimMartialStatus = Single: THEN Offender job will be Private (19.0/1.0)

Rule 22: ParticularPlace = Iran_Embassy AND OffenderMartialStatu = Married AND VictimsReligion = Muslim: THEN Offender job will be Private (7.0).

Rule 23: OffenderAge = 18-22 AND ParticularPlace = Rufael : THEN Offender marital status will be Single (41.02/6.0).

Rule 24: OffenderAge = 18-22 AND ParticularPlace = Rufael AND Policestation = PPS AND Time = 5:00:00 PM – 8:00:00 PM: THEN Offender marital status will be Single (11.0).

Rule 25: ParticularPlace = Rufael AND PoliceStation =PPS AND VictimJob = Private AND VictimMartialStatus = Married AND CrimeCatagory = Crime_against_Property AND VictimsReligion = Orthodox:THEN Offender marital status will be Married (12.0).

As can be observed from the above partial list of rules, the classifier has used selected attributes to construct rules and provide the class predicted by the model. As discussed before the numeric values which appeared in the bracket next to the class label indicate the number of records having that property (i.e. property stated in the rule) and incorrectly classified records respectively. For instance, rule 23 could be interpreted as offender age is between 18 and 22, the place where the crime is committed is at Rufael (Addisu Gebeya), and the offender marital status is single. There are 47 records having this property. From which 41 records are correctly

classified and the rest 6 records are misclassified. From this, the likelihood of predictability of offender marital status by the stated predictors is 87.2 %.

These rules have indicated that attribute such as crime type, age, educational level, job, marital status; sex and particular place are the most important basis for classification.

For instance in the case of rule1:

If there are both male and female offenders and

If none of the victim are engaged on Private, government, daily-Worker, housewife,

Student, unemployed, NGO and

If none of the offender have educational level of illiterate, junior, primary, Diploma, first

Degree and Master and

If none of the offender are followers of Muslim, protestant, catholic and undefined religion and

If none of the crime type is committed from 5 AM-8AM, 1PM-4PM, 5PM-8PM, 3PM-12PM and

If none of offender have age from 18-22, 28-32, 33-32, 33-37, 38-42…63 and above

Then the crime committed is ***Low level*** *crime.*

4.2 Comparison of J48 and PART

Selecting a better classification technique to identify and investigate the relation between demographic factors of victims and offenders who were exposed to crime and to develop crime prediction model is the aim of this study. For this reason, the two classification models with their respective best performance accuracy are listed in the table 4.22 below.

Table 4.22 Comparison of the result of the J48 and PART models

Algorithms	Average Performance registered (%)	Average Time Taken	Average Correctly classified instances	Average Misclassified instances
J48	88.4%	0.15	4520	586
PART	88.6%	4.13	4529	577
10-fold Cross validation test option				

The results in Table 4.22 showed that PART rule induction model perform better compared to J48 decision tree. In terms of tree size and number of leaves, J48 pruned with all attribute is relatively more understandable and less complex to humans. Therefore, the performance of J48 pruned tree classifier with all attribute gives valuable information in predicting the target classes as PART model.

Moreover, PART rule induction is more self-explanatory, simple and easy to understand because it produces in the forms of “if then” condition. It generates rules that can be presented in simple human language. Based on rules generated the researcher found that crime type, age, education level, job, marital status, sex and particular place are the most prevalent demographical factors to commit crimes and to be a victim of crime.

Therefore, it is reasonable to conclude that the PART is more appropriate to this particular case than the J48 decision tree. Thus, the model that is developed with the PART rule induction classification technique is taken as the final working classification model to identify and investigate the relation of demographic factors of victims and offenders who were exposed to

crime and to develop crime prediction model that could help the police commission of Addis Ababa in decision making for crime detection and prevention.

4.3 Result of re-evaluation PART Models

Evaluation is one the key points in any data mining process. It serves the prediction of how well the final model will work in the future. Based on this concept the researcher tried to re-evaluate and validate PART models by using 2000 randomly selected datasets. But the researcher select one target class since to avoid confusion and complexity. Table 4.23 shows the detail of confusion matrix for crime level target class. Crime level target class is preferred because of its better accuracy registered.

Table 4.23 Result of re-evaluation of PART model with 2000 dataset

Actual	Medium Level	Low Level	High Level	Total	Correctly Classified (accuracy rate)
Medium Level	**817**	14	6	837	97.6
Low Level	23	**700**	2	725	96.5
High Level	3	7	**428**	438	97.7
Total	843	721	436	2000	97.2

The performance of the model with this testing dataset was 97.2%. This result implies that out of the 2000 testing datasets, the developed PART classification model predicted 1945 (97.2%) records correctly. Which means the performance of the model is validated using randomly selected 2000 dataset and registered even better performance accuracy 97.2% this indicates the model is pretty good. Based on its performance some interesting rules generated from PART algorithm discussed as follows.

Table 4.24. Sample classification rules for offender and offence relation

Rule Number	"IF" Part	"THEN" Part	Success ratio	%
1	OffenderEduL = Secondary AND OffenderAge = 18-22 AND OffenderReligion = Orthodox AND OffenderSex = Male AND OffendermarrtialStatus = Single:	Miscellaneous_Theft	(301.0/80.0)	79%
2	OffendermarrtialStatus = Single AND OffenderAge = 23-27 AND OffenderReligion = Orthodox AND OffenderEduL = Secondary AND OffenderSex = Male:	Beat	(266.0/47.0)	84.9%

Table 4.24 depicts that, to come up with the above significant rules, PART is run on the crime data set with seven attributes including crime type, offenders' age, sex, marital status, educational level, job and religion. As a result, the first rules showed that offender who have the stated demographic features commit miscellaneous theft with the accuracy of 79 %. For instance offenders, who commit beat and offender who commit miscellaneous theft, have different profiles of age. Table 4.25 depicts some rules that can show the connection of crime types with some demographic features of victims.

Table 4.25 Sample classification rules for Victims and Offence relation

Rule Number	"IF" Part	"THEN" Part	Success ratio	%
1	VictimAge = 21-25 AND VictimsReligion = Muslim AND VictimSex = Female AND VictimMarrtialStatus = Single AND VictimJob = Private:	Snatching	(130.0/2.0)	98.4 %
2	VictimJob = Private AND VictimSex = Male AND VictimAge = 26-30 AND VictimsReligion = Orthodox AND VictimMarrtialStatus = Single :	Theft_with_house_break	(107.0/2.0)	98.1 %
3	VictimSex = Male AND VictimAge = 31-35 AND VictimJob = Private:	Theft_of_vehicle_part:	(77.0/22.0)	77.7 %

Similarly, the classifier used six attributes to construct the rules stated in Table 4.25. Attributes like crime type, victim's sex, age, marital status, job and religion were used to generate the rules. As the rule depicts that, there is connection between the crime committed and the demographic feature of the victims'. For instance, the crime type 'theft of vehicle part' affect victims whose age is between 31-35 while crime type 'theft with house break' affects victims whose age is between 26-30. Table 4.26 depicts some rules that can show connection of offender's profile, the time offender prefers to commit crime and the particular place where the crimes were committed.

4.26 Sample classification rules for offender, time and place relation

Rule Number	"IF" Part	"THEN" Part	Success ratio	%
1	CrimeType = Murder AND Time = 1:00:00 AM-4:00:00 AM AND OffenderEduL = Secondary:	Kolash_Sefere	(8.0/4.0)	66.6%
2	CrimeType = Theft_with_house_break AND OffenderEduL = Primary AND OffenderJob = Unemployed AND OffendermarrtialStatus = Married AND ParticularPlace = Mender_7_Shegole:	1:00:00AM-4:00:00AM	(8.0)	100%
3	OffenderEduL = Secondary AND Time = 9:00:00 AM-12:00:00 AM AND OffenderAge = 18-22 AND OffenderJob = Private AND OffenderSex = Male CrimeType = Miscellaneous_Theft:	Mender_7	(173.0/16.0)	91.5%

Similarly to generate the above rules predicted in Table 4.26, the classifier used nine attributes, namely crime type, time, particular place, offenders' sex, age, marital status, educational level, job and religion. As a result, PART also generate interesting rules that can depict the connections that exist among offenders demography, the time when the crime committed and the particular place where the crime was committed. For example, rule No 2 indicates that eight instances in the crime dataset have such features. In addition, as the domain expert approved that the time from 1:00:00 AM-4:00AM is the common time most of the offender prefers to commit theft with house break crime types.

To sum up, Based on the rules generated from PART algorithm the most frequent crime types that were committed in the sub city are miscellaneous theft, beat or assault, theft of vehicle part, snatching and threat.

Most of the thefts of vehicle accessories are committed by male and they have secondary education level, their age is between 23 and 27, job status is private, unmarried and follower of orthodox religion. Accordingly, most of the victims are males, whose age are between 31 and 35, and engaged on their private work. This type of crime is committed around ‘Rufael’, ‘Paulos’ and ‘Markos Megazen’ (‘Paster’). This crime mostly committed from 1:00:00PM to 4:00:00PM. Based on nationally identified crime level it is categorized as High Level Crime.

The other crime types mostly committed is snatching, most of the offenders is males, whose age is between 18 and 22, engaged on private works, and has secondary education level and single and follower of orthodox religion. But, most of the victims by this crime type is female, whose age is between 21 and 25, and their job status is private worker. Most of this crime types are committed around Lazarist_School (Addisu Gebaye) ‘Adem_Bedaane’ (‘Sheromeda’), ‘Ethio-Parent’ and ‘Kelem_Ammba’ around ‘paster’. This crime mostly committed from 5:00:00PM-8:00:00PM. Based on nationally identified crime level it is categorized as High Level Crime.

Most of the miscellaneous theft crimes are committed by male offenders whose age was between 18 and 22; they have secondary educational level, engaged on private work, unmarried and follower of orthodox religion. Accordingly, most of the victims are males (except snatching crime type), whose age is between 21 and 30, and engaged on their own private work. Most of those miscellaneous thefts were committed at ‘Mender_7’ (‘Shegole’), ‘Rufael’ and ‘Medhanealem School’ (‘Paster’). Most of the time, this crime is committed from 9:00:00AM-12:00:00AM. Based on nationally identified crime level it is categorized as medium Level Crime.

Whereas, beat or assault crimes are mostly committed by male offenders who have secondary education level, whose age is between 23 and 27, job status is private, single and follower of orthodox religion. Similarly, most of the victims are males, whose age are between 21 and 25, and engaged on private work. Most of those crimes were committed at 'Meketiya' ('sheromeda'), '19_kebele' ('Menene') and 'Shero meda taxi tera'. This crime mostly committed from 5:00:00 PM to 8:00:00 PM. Based on nationally identified crime level it is categorized as low Level Crime.

Threat crime type is also one of the most and frequently committed crime types. Most of the offenders are males, whose age is between 23 and 27, engaged on private work, unmarried and have primary educational level and follower of orthodox religion. Accordingly, most of the victims are male, whose age is between 21-25, and engaged on private work. Most of this kind of crime is committed around 'Tsedu sefere' ('Kechene'), 'Filance' ('Paster'), 'Medhanealem school' ('paster'), and 'Meketiya' ('Sheromeda'). This crime type mostly committed from 9:00:00 PM to 12:00:00PM. Based on nationally identified crime level it is categorized as low Level Crime.

The strength of this study compared with other similar research output is that, in this research the most demographic relationship between offenders and victims were identified and investigated. That means it try to answer questions like, who is the offender and what type of crime they commit? In addition, the research also tried to identify offenders most preferred time and place to commit offences were mark out. Similarly, the study also point out the most frequent crime types committed in Addis Ababa, particularly at "Gulele sub city" with their crime level (how much it is serious). Those results are very crucial for law enforcement body like police; for instance to allocate and efficiently utilized their limited resource, to identify the patterns as well as trend of

crime in advance and to be proactive in detecting and preventing crime and to understand the victims and offenders behaviors.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

For this research, the data was collected from seven police centers namely, Addis Ababa police Commission (Head Quarter), ‘Guele Command Post’, ‘Paster’, ‘Menene’, ‘Sheromeda’, ‘Addisu Gebeya’ and ‘Kechene’. The data was stored in manual unpublished crime record and computers. Data preprocessing and preparing the data for model building was conducted in both MS-Excel filtering and WEKA filters tools. The attribute selection was made with the help of domain experts and WEKA information gain attribute evaluation. The modeling Methodology applied was Hybrid Data Mining Model. Different literatures were reviewed with the aim to bring the problem into Data Mining problem.

The selected data mining task to be implemented for this research are classification. For classification model J48 decision tree and PART (Projective Adaptive Resonance Theory) rule induction algorithms were used. Twelve experiments were undergone for the two techniques. These experiments were performed as six with J48 and six with PART. For each techniques eight experimental scenarios with two data mining algorithm validation techniques (10-fold cross validation and 80-20 percentage split validation) were used.

As a result PART rule induction algorithms registered better average performance with 88.6 % accuracy running with 10-fold cross validation with default parameter using 19 attributes than any experimentation done for this research purpose. The performance of the model is validated using randomly selected 2000 datasets and registered performance accuracy of 97.2% which is pretty good model

The classification task is performed to help the classifier to predict the class of a particular record. The dependent variables for the classifier were crime level, time, victim marital status, offender marital status, victim job and offender Job which serves as the class values. The prediction is done to show relation between/among the demography factors (profile) of victims and offenders who were exposed to crime and to identify factors to predict crime level. To this end, with extensive discussion of domain experts, 'educated guesses' and pattern (rules) generated out of 19 attributes 7 attributes are selected as a determinant variables to predict the target classes stated above. Namely, crime type, age, educational level, job, marital status, sex and particular place.

Based on the analysis, the researcher identifies the most frequent crime types that were committed in the sub city. Offences like, miscellaneous theft, beat or assault, theft of vehicle part, snatching and threat are the most and the common types of crimes observed in 'Guelele' sub city. These crimes were mostly committed by male offender, whose age is between 18 and 27, and they have secondary education level, engaged on private work, their marital status is single and the follower of orthodox religion.

Similarly, those crime types mostly committed in 'Gulele sub city' affects male victims whose age is between 16 and 35 and engaged on private work. Specifically, miscellaneous theft crime type affect victims whose age is between 26 and 30 while snatching crime type affect those victims whose age is between 21 and 25. Analogously, theft of vehicle accessories affect those victims whose age was between 31 and 35.

The most preferable time to commit miscellaneous theft was the time between 9:00:00AM-12:00:00PM. Most of this crime type concentrated around Mender_7 ('Shegole'), 'Rufael', and 'Medhanealem_school' ('Paster').Similarly, the crime type beat and simple attack mostly

committed around ‘Meketiya’ (‘sheromeda’), ‘19_kebele’ (‘Menene’) and ‘Shero meda taxi tera’. This crime mostly committed from 5:00:00 PM to 8:00:00 PM. Most theft of vehicle accessories offences committed around ‘Rufael’, ‘Paulos’ and ‘Markos Megazen’ (‘Paster’). The most preferable time to commit theft of vehicle part crime was from 1:00:00PM to 4:00:00PM. Analogously, most offenders prefer the time from 5:00:00 PM – 8:00:00PM to commit snatching crime. This crime type mostly committed around Lazarist_School (Addisu Gebaye), ‘Adem_Bedaane’ (‘Sheromeda’), ‘Ethio-Parent’ and ‘Kelem_Ammba’ (‘Paster’). Lastly, the other crime type which was frequently committed in the sub city was threating. Which is mostly committed around ‘Tsedu sefere’ (‘Kechene’), ‘Filance’ (‘Paster’), ‘Medhanealem school’ (‘paster’), and ‘Meketiya’ (‘Sheromeda’). Mostly offenders preferred the time between 9:00:00 PM and 12:00:00PM to commit this threating offence.

In general, as commented by domain experts the result from this researche are encouraging. It is important to determine attributes and their values to understand the profile of both victim and offenders. In addition, determining attributes and their values also help to identify which victim are exposed to which crime and which offenders are exposed to commit crime. The classification techniques also shows which attribute are common in a given target class (class label).This helps to generate rules to identify the potential victim and offender for a specified crime category. Having this knowledge Addis Ababa Police commission can design appropriate training programs and crime prevention and investigation strategies.

5.2 Recommendations

In this study an attempt has been made to identify and investigate the demographic relation of victims and offenders who were exposed to crime and develop crime prediction model that could

help the police commission of Addis Ababa in decision making, crime detection and prevention. Even though this research is done for academic purpose; its output would help law enforcement body to identify potential victims and offenders proactively. This research has identified the profile of victims and offenders by generating rules like If (particular place = 'Rufael' and police station = PPS and Crime type = Miscellaneous theft and time = 1:00:00 PM – 4:00: 00 PM then offender marital status = single) which is to mean: A theft crime which is committed at 'Rufael' (the area i.e. belongs to PPS) from 1:00:00 PM up to 4:00:00 PM are committed by offender whose marital status is single. Additionally it predicts which time is preferable for which type of crime. It is helpful for the commission to use such kind of rules to design programs to protect victims as well as potential offenders, to identify target areas and to allocate resources efficiently.

Although the research output is encouraging and usable, the following recommendations are given for future consideration by the concerned bodies:

- To perform these tasks more efficiently and effectively, more emphasis should be given to data collection and organization. Addis Ababa police commission better to work on national standard database format for recording crime related data with a sustainable maintenance facility to render adequate and instant decision. For this the commission better to give adequate training for its staff to equip them with the current data management technology.

- Even the available attributes are not enough to generate different analysis results. For example educational status of victim, previous crime history of offender, the time span (period) the offender/victim stay in Addis Ababa, physical description of offender and

where they were born and grew up (geographic information) are equally important to describe the profile of offenders and victims.

- 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. So that experts can consult the system in their problem solving and decision making process.

- Although both the Decision tree and Rule induction approaches resulted in an encouraging output, still performance improvement is expected. Hence, other classification techniques such as Neural networks, Naïve bayes classifier, Time series analysis and summarization which have also been proved to be important techniques can be tested in order to investigate their applicability to the problem domain in the program by using the entire crime dataset.

REFERENCES

[1] Dubey Nikhil and S. K. Chaturvedi, "A Survey Paper on Crime Prediction Techniques Using Data Mining," *Journal of Engineering Research and Application,* vol. 4, no. 3, pp. 396-400, 2014.

[2] A. Sharma and R. Kumar, "The Obligatory of an algorithm for matching and predicting crime-using Data Mining techniques," *International Journal of Computer Science and Technology (IJCST),* vol. 4, no. 2, pp. 289-292, 2013.

[3] Merriam-Webster, Webster's Dictionary for Students., 4th ed., New York: Federal Street, 2011.

[4] 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.

[5] H. Chen, W.Chung, J.Xu, G.Wang, Y. Qin and M. Chau, "Crime data mining: A general framework and some examples," *IEEE Computer Society,* pp. 50-56., 2004.

[6] Z. S. Zubi and A. A. Mahmmud, "Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention.," *International Journal of Computer,* vol. 8, pp. 39-45, 2014.

[7] P. Gera and R. Vohra, "City Crime Profiling Using Cluster Analysis," *International Journal of ComputerScience and Information Technologies,* vol. 5, pp. 5145-5148., 2014.

[8] E. R. Groff and N. G. L. Vigne, " Forecasting the Future of predictive Crime Mapping.," *Journal of Crime prevention Studies,* no. 13, pp. 29-57, 2002.

[9] K. Z. Hussain, M. Durairaj and G. R. J. Farzana, "Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model," *International Journal of Computer Science and Information Technology & Security (IJCSITS),* vol. 2 , pp. 25-29, 2012.

[10] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland], Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data, Trento: University of Trento, 2014, pp. 1-10.

[11] S.Yamuna and N. Bhuvaneswari, "Datamining Techniques to Analyze and Predict Crimes," *The International Journal of Engineering And Science (IJES),* vol. 1, no. 2, pp. 243-247, 2012.

[12] P. Bergeron and C. A. Hiller, Competitive intelligence: Annual review of information science, 2002.

[13] K.Cios and L. Kurgan, "Trends in data mining and knowledge discovery," in *Advanced Techniques*

in Knowledge Discovery and Data Mining, London, Springer Verlag, 2005, pp. 1-26.

[14] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques (3rd ed.)., Amsterdam: Elsevier Inc., 2012.

[15] L. Sharma and N. Mehta, "Data Mining Techniques: A Tool For Knowledge Management System In Agriculture.," *International Journal of Scientific and Technology research,* vol. 1 , no. 5, pp. 67-73., 2012.

[16] Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Potomac: Two Crows, 1999.

[17] O. R. Zaiane, Principles of Knowledge Discovery in Databases., Alberta: University of Alberta.: University of Alberta., 199.

[18] T. Silwattananusarn and KulthidaTuamsuk, *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 2 , no. 5, pp. 1-12, 2012.

[19] H. Sahu, S. Shrma and S. Gondhalakar, "A Brief Overview on Data Mining Survey," *International Journal of Computer Technology and Electronics Engineering (IJCTEE),* vol. 1, no. 3, pp. 114-121., 2011.

[20] Girma Aweke, *Predicting HIV infection risk factor using voluntary counseling and testing data: A case of African AIDS initiative international. (Unpublished M.Sc thesis),* Addis Ababa: Addis Ababa University, 2012.

[21] A. D. M. &. K. D. Process, "A Data Mining & Knowledge Discovery Process," in *Data Mining and Knowledge Discovery in Real Life Applications*, Vienna, Austria, I-Tech, 1999, pp. 1-16.

[22] U. Fayyad, G. P. -Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *American Association for Artificial Intelligence,* vol. 39, no. 11, pp. 37-54., 1996.

[23] G. K. Kahlon and G. Kaur, "Methodology for dynamic and knowledge discovery static data mining.," *A Journal of Multidisciplinary Research,* vol. 2, no. 5, pp. 1-13, 2013.

[24] O. Maimon and L. Rokach, "Introduction to Knowledge discovery," in *Data Mining and Knowledge Discovery Handbook*, New York, Springer, 2005, pp. 1-17.

[25] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing,* vol. 5, pp. 13-22., 2000.

[26] A. Azevedo and M. F. Santos, "KDD, SEMMA AND CRISP-DM: A parallel overview.," in *IADIS European Conference Data Mining 2008 (pp. 182-185). l: IADIS.*, de Infesta Portugal, 2008.

[27] SAS Institute Inc., Data Mining and the Case for Sampling: Solving Business Problems Using SAS

Enterprise Miner Software, NC: SAS Institute Inc., 1998.

[28] K. J. Cios, W. Pedrycz, R. W. Swiniarski and L. A. Kurgan, Data Mining: A Knowledge Discovery Approach, New York: Springer., 2007.

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

[30] Y. Singh and A. S. Chauhan, "Neural Networks in Data Mining.," *Journal of Theoretical and Applied Information Technology* , pp. 37-42, 2005.

[31] J. M. David and K. Balakrishnan, "Significance of classification techniques in prediction of learning disabilities," *International Journal of Artificial Intelligence and Application.,* vol. 1, pp. 111-120, 2010.

[32] J.-s. Li, H.-y. Yu and X.-g. Zhang, "Data Mining in Hospital Information System," in *New Fundamental Technologies in Data Mining*, Zhejiang, Zhejiang University, 2011, pp. 143-172.

[33] S. Chauhan, S. chauhan and N. Kumar, "A Survey of Data Mining Techniques," in *1st International Conference on Research in Science, Engineering & Management (IOCRSEM 2014)*, 2014.

[34] V. Kumar, Understanding Complex Datasets: Data Mining with Matrix Decompositions, Minnesota: Taylor and Francis Group, LLC., 2007.

[35] H. M. Moshkovich and A. I. M. D. L.Olson, "Rule induction in data mining: effect of ordinal scales," *Journals of expert systems with application,* vol. 22, pp. 303-311, 2002.

[36] B. M. Ramageri, "Data Mining Techniques and Applications.," *Indian Journal of Computer Science and Engineering*, vol. 1, no. 4, pp. 301-305, 2010.

[37] S. Taneja and R. Sapra, "Hybrid Approach for Classification Tree Generation," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 3, no. 1, pp. 240-242, 2014.

[38] R. Kumar, A. K. Kapil and A. Bhatia, "Modified tree classification in Data Mining," *Global journal of computer Science* , vol. 12, no. 2, pp. 1-6., 2012.

[39] 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.

[40] L. Rokach and O. Maimon, "Decision Tree. In Data (pp.).," in *Mining and Knowledge Discovery Handbook*, New York, Springer, 2010, pp. 166-192.

[41] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," *International Journal of Computer Applications,* vol. 98, no. 22, pp. 13-17, 2014.

[42] V. Kumar and N. Rathee, "Knowledge discovery from database Using an integration of clustering and classification," *International Journal of Advanced Computer Science and Applications,* vol. 2, no. 3, pp. 29-33, 2011.

[43] D. Matyja, *Applications of data mining algorithms to analysis of medical data (MSc Thesis),* Sweden: Blekinge Institute of Technology., 2007.

[44] S. O. Danso, *An Exploration of Classificatiion Predictiion Techniques in Data Mining:The Insurance Domain. Dissertation Presented to the School of Design, Engineering, and Computing.,* Bournemouth: Bournemouth University, 2006.

[45] S.Neelamegam and E.Ramaraj, "Classification algorithm in Data mining: An Overview," *International Journal of P2P Network Trends and Technology (IJPTT),* vol. 4, no. 8, pp. 369-374, 2013.

[46] F. Voznika and L. Viana, "Data Mining Classification," *Journal of computer science and Technology,,* vol. 3, no. 2, pp. 1-6, 2015.

[47] 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.

[48] J. Furnkranz, D. Gamberger and N. Lavrac, Foundation of rule learning, Heidelberg: Springer, 2012.

[49] K. Thearling, "Data Mining," Online, n.d.

[50] A. A. Freitas, A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery, United Kingdom: Springer-verlag, 2002.

[51] G. P. a. D. Haussler, "Boolean Feature Discovery in Empirical Learning," vol. 5, pp. 71-99, 1990.

[52] R. C. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," *Journal of Machine Learning,* vol. 11, pp. 63-91, 1993.

[53] I. H. Witten and E. Frank, Data mining : practical machine learning tools and techniques, San Francisco : Morgan Kaufmann, 2005.

[54] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, Canada: John Wiley and Son, 2011.

[55] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," *Journal of Bioinformatics,* pp. 1-12, 2010.

[56] Gebremedhin Gebreyohanis, *Application of Data Mining techniques to predict children dataset: the case of love for children organization.(unpublished M.Sc thesis),* Addis Ababa: Addis Ababa

University, 2012.

[57] F. Siraj and M. A. Abdoulha, "Mining Enrolment Data Using Predictive and Descriptive Approaches," in *Knowledge-Oriented Applications in Data Mining*, Utara, Universiti Utara Malaysia, 2011, pp. 53-72.

[58] C.-H. Yu, M. W. Ward, M. Morabito and W. Ding, "Crime Forecasting Using Data Mining Techniques," in *IEEE 11 th international conference on data mining workshops*, Boston.

[59] S. V. Nath, "Crime pattern detection using data mining," in *Proceeding of the 2006 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, Washington, 2006.

[60] Z. S. Zubi and A. A. Mahmmud, "Crime Data Analysis using Data Mining techniques to improve crimes prevention," *International Journal of Computer*, vol. 8, pp. 39-45, 2014.

[61] L. W. Asegehgn, *The application of Data Mining in crime prevention: the case of Oromia Police Commission (unpublished M.Sc thesis)*, Addis Ababa: Addis Ababa University, 2003.

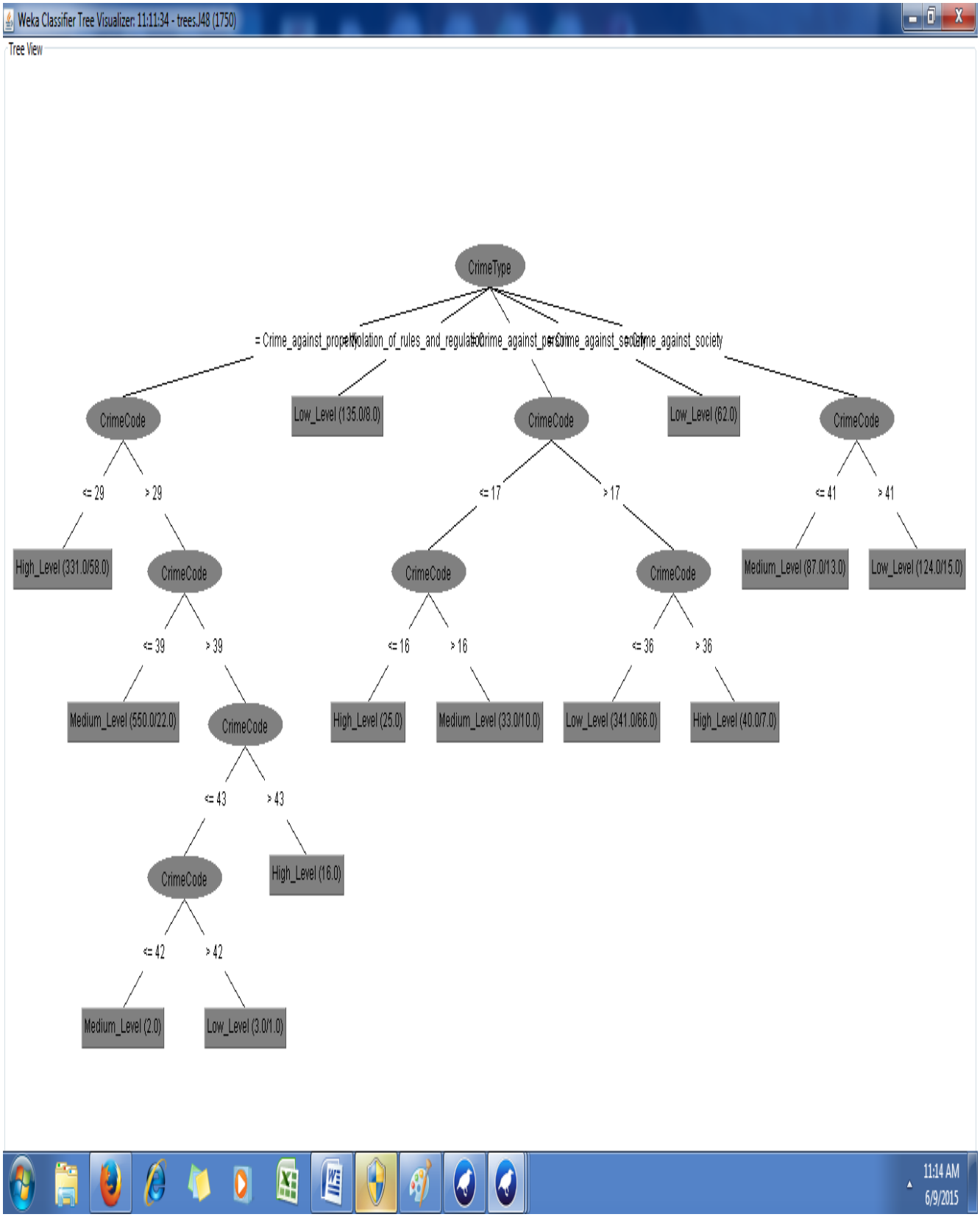
[62] J. E. Douglas, A. W. Burgess, B. A. G. and R. K. Robert, Crime Classification Manual: A Standard system for investigating and classifying violent crimes, 2nd ed., San Francisco: Jossey-Bass, 2006.

[63] R. L. Akers, "Rational Choice, Deterrence, and Social Learning Theory in Criminology: The Path Not Taken," *Journal of Criminal Law and Criminology*, vol. 81, no. 3, pp. 653-676, 1990.

[64] G. S. Linoff and M. J. Berry, Data Mining techniques for Marketing, sales, and customer relationship management, 2nd ed., Indiane: Wiley, 2004.

[65] Ministry of Youth, Sport and Culture of Ethiopia (MYSC), "Official web site of Ministry of Youth, Sport and Culture of Ethiopia (MYSC)," 2013. [Online]. Available: http://www.mysc.gov.et/youth.html. [Accessed Tuesday August 2015].

Appendix A. The generated Sample decision tree (J48 algorithm)



Appendix B. Sample PART decision rule list with 10-fold Cross Validation

```
ParticularPlace  = Paster AND
OffenderReligion = Orthodox AND
PoliceStation = PPS AND
VictimsReligion = Orthodox: 5:00:00 AM-8:00:00 AM (10.01/2.01)


CrimeType = Theft_with_house_break   AND
VictimJob = Private AND
OffenderEduL = Primary AND
OffenderSex = Male AND
OffenderReligion = Orthodox: 1:00:00 PM-4:00:00 PM (18.0/5.0)


CrimeType = Pocket_Theft AND
OffenderAge = 23-27: 5:00:00 PM-8:00:00 PM (14.0/3.0)


CrimeType = Raping_virginity_forcefully_Immature girl: 1:00:00 PM-4:00:00 PM (17.0/6.0)


CrimeType = Pocket_Theft AND
OffenderJob = Unemployed AND
OffenderReligion = Orthodox AND
ParticularPlace  = Medhanealem_Church : 5:00:00 AM-8:00:00 AM (8.0)


ParticularPlace  = Paulos_hospital AND
OffenderJob = Unemployed AND
OffenderReligion = Orthodox: 1:00:00 PM-4:00:00 PM (12.0)


ParticularPlace  = Kidanemehret  AND
CrimeCatagory = Crime_against_person AND
OffendermarrtialStatus = Single: 5:00:00 PM-8:00:00 PM (9.01/0.01)


ParticularPlace  = Zero_3_Mezenagna AND
OffenderJob = Private AND
OffendermarrtialStatus = Single AND
OffenderSex = Male AND
CrimeCode > 26: 5:00:00 AM-8:00:00 AM (6.01/0.01)
```

Time target class

```
VictimSex = Male AND
ParticularPlace  = Kebet_Beret AND
OffenderJob = Private: Private (19.02/2.01)


VictimSex = Male AND
ParticularPlace  = Dera_Sefere AND
CrimeCode <= 26: Private (15.01)


VictimSex = Male AND
ParticularPlace  = Menene AND
VictimsReligion = Orthodox: Government (14.02/0.01)


VictimSex = Male AND
VictimAge = 31-35 AND
OffenderJob = Unemployed AND
OffenderReligion = Orthodox AND
ParticularPlace  = Mender_7_Shegole: Private (8.0)
```

Victim Job target class

```
VictimAge = 16-20 AND
ParticularPlace  = Meketiya AND
PoliceStation = SPS: Student (10.0)


VictimAge = 16-20 AND
OffendermarrtialStatus = Single: Single (174.0/8.0)


ParticularPlace  = Shegole AND
OffenderReligion = Orthodox AND
CrimeLevel = Medium_Level: Single (13.0)


ParticularPlace  = Filance AND
OffenderJob = Private AND
VictimSex = Male: Single (19.01/0.01)
```

Victim Marital Status target class

```
ParticularPlace  = Zero_3_Kebele Mezenagna AND
OffendermarrtialStatus = Single AND
OffenderReligion = Orthodox AND
VictimSex = Male: Private (16.01)

ParticularPlace  = Rufael AND
VictimSex = Male AND
VictimAge = 26-30 AND
VictimMarrtialStatus = Single: Private (27.0)
CrimeType = Breach_of_trust AND
VictimJob = Private AND
OffenderEduL = Secondary: Private (18.0/3.0)

CrimeType = Drugs_possession_and_use AND
ParticularPlace  = Alemtsaye_Deldeye AND
Time = 1:00:00 PM-4:00:00 PM: Unemployed (4.0)

CrimeType = Pocket_Theft AND
OffenderEduL = Junior: Unemployed (8.0)

ParticularPlace  = Rufael AND
VictimJob = Private AND
PoliceStation = PPS AND
VictimAge = 16-20: Private (10.0)
```

Offender Job

target class

```
ParticularPlace  = Rufael AND
PoliceStation = PPS AND
CrimeType = Miscellaneous_Theft AND
Time = 1:00:00 PM-4:00:00 PM AND
CrimeCode > 26: Single (6.0)
```

Offender Martial Status target class

```
OffenderAge = 23-27 AND
CrimeType = Theft_with_house_break   AND
VictimsReligion = Orthodox AND
VictimJob = Private: Single (16.0)


ParticularPlace  = Tsedu_Sefere AND
VictimMarrtialStatus = Married AND
VictimJob = Private AND
PoliceStation = KPS AND
OffenderEduL = Secondary AND
Time = 9:00:00 AM-12:00:00 AM: Married (8.0)
```

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources used for the thesis have been duly acknowledged.

HAILEMARIAM NEGUSSIE

2015

The thesis has been submitted for examination with my approval as University Advisor

---

Dr. GashawKebede

2015