ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUTE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

HEALTH INFORMATICS GRADUATE PROGRAM

MINING ROAD TRAFFIC ACCIDENT DATA FOR PREDICTING ACCIDENT SEVERITY TO IMPROVE PUBLIC HEALTH – ROLE OF DRIVER AND ROAD FACTORS IN THE CASE OF ADDIS ABABA

BY

ANTENEH FENTAHUN

JULY, 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUTE STUDIES
SCHOOL OF INFORMATION SCINCE AND SCHOOL OF PUBLIC HEALTH
HEALTH INFORMATICS GRADUATE PROGRAM

MINING ROAD TRAFFIC ACCIDENT DATA FOR PREDICTING ACCIDENT SEVERITY TO IMPROVE PUBLIC HEALTH – ROLE OF DRIVER AND ROAD FACTORS IN THE CASE OF ADDIS ABABA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABAB UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN HEALTH INFORMATICS

BY
ANTENEH FENTAHUN
JULY, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUTE STUDIES

SCHOOL OF INFORMATION SCINCE AND SCHOOL OF PUBLIC HEALTH

HEALTH INFORMATICS GRADUATE PROGRAM

MINING ROAD TRAFFIC ACCIDENT DATA FOR PREDICTING ACCIDENT SEVERITY TO IMPROVE PUBLIC HEALTH – ROLE OF DRIVER AND ROAD FACTORS IN THE CASE OF ADDIS ABABA

BY

ANTENEH FENTAHUN

JULY, 2011

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
Ato Mahder Alemayehu	Chairperson	_____	_____
Dr. Dereje Teferi	Advisor	_____	_____
Dr. Wakgari Deressa	Advisor	_____	_____
Dr.Million Meshesha	Examiner	_____	_____

Dedication

I dedicate this paper to mom, Asratie Biadegelegne for all the sacrifice she made for my education

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to my advisors, Dr. Dereje Teferi and Dr. Wakgari Deressa for their guidance, encouragement, and support throughout my research. I am also greatly indebted to Tibebe Beshah, who helped me in doing the proposal and providing me with the available resources.

My heartfelt thanks also go to Sajin Assefa Mezegebu and Abebe Asrat for their support in providing information on traffic accident situation in Addis Ababa.

My enormous thank goes to my friend Kedir Mohammed for his encouragement and support.
I am also indebted to Ejigu Tessema's family, Fate Awol, Yohannes Ejigu you were so special throughout my study thank you so much.

At last, but by no means the least, I would like to thank my friends, Geletaw Sahle, Biset Desalegne (our resource database), Haftom G/egeziabehere, Abel Dametew and Selam Assamnew for the constant assistance and the times we had during my study.

Table of Contents page

List of Figures and Tables

List of figures

List of tables

List of Acronyms

AARTCID: Addis Ababa Road Traffic Control and Investigation Department

AIDS: Acquired Immune Deficiency Syndrome

CRISP-DM: Cross Industry Standard Process for Data Mining

DALY: Disability-Adjusted Life Year

ERA: Ethiopian Road Authority

HIV: Human Immune Virus

MLP: Multi Layer Perception

RTA: Road Traffic Accident

WEKA: Waikato Environment for Knowledge Analysis

WHO: World Health Organization

Abstract

Road traffic accidents are among the top leading causes of deaths and injuries of various levels in Ethiopia. One of the solutions to reduce the problem of traffic accident is finding the causes through research, and data mining is one research tool in finding the causes of traffic accidents. The objective of this study is to identify and investigate drivers' and road factors that contribute to the cause of accident and to develop traffic accident prediction model.

In this research an attempt is made to apply the decision tree and rule induction predictive data mining techniques in major driver and road factors for car accidents and identify hidden patterns in the accident data set. To achieve this goal: the CRISP-DM 1.0 standard data mining methodology is adopted and the WEKA data mining tool is used to implement the ID3 , J48 and PART algorithms.

The data for this research is the RTA data of the years 2005-09 collected from the Addis Ababa Road Traffic Control and Investigation Department and local researchers. After preprocessing a total of 16,710 RTA records are used for building the models.

Various experiments are made iteratively by making adjustment of the parameters and using different number of attributes to come up with a meaningful output. Major factors of drivers and roads are identified and rules are generated using J48 decision trees and rule induction (PART algorithm). The comparison of the models using WEKA's experimenter showed that J48 slightly outperforms ID3 and PART algorithms.

In addition, the determinant factors of drivers and roads that cause road accidents are identified; these are LicenceGrade, subcity, RoadJunction, TypeofRoad, and LightCondition. In many data mining researches on traffic accidents, decision trees and neural networks are widely used but in this study rule induction and decision trees are used to built the different models that can solve the problem of public health in the society.

Chapter one

Introduction

1.1. Background

The encyclopedia defines road traffic accident as "any vehicle accident occurring on a public highway (i.e. originating on, terminating on, or involving a vehicle partially on the highway). These accidents therefore include collisions between vehicles and animals, vehicles and pedestrians, or vehicles and fixed obstacles. Single vehicle accidents, in which one vehicle alone (and no other road user) involved, are included. All fatality and injury totals include pedestrians, motorcyclists and bicyclists unless otherwise noted (Safecarguide, 2004).

A report by WHO (2004) estimated that worldwide over one million people are reportedly killed each year in road crashes, equivalent to three deaths every minute. Moreover, by the year 2020 road accidents will be the third leading cause of death. This puts road safety well ahead of wars, HIV/AIDS, malaria and (other) 'acts of violence' as world health problem. Among children aged 5-14years, and young people aged 15-29 years, road traffic injuries are the second-leading cause of death worldwide (WHO, 2004).

The cause of injuries worldwide is dominated by those incurred in road crashes. According to WHO (2004) deaths from road traffic injuries account for around 25% of all deaths. The annual number of road deaths varies from around 750,000 to 1,180,000- representing over 3,000 lives lost daily.

In low income countries and regions-in Africa, Asia, the Caribbean and Latin America-the majority of road deaths are among pedestrians, passengers, cyclists, users of motorized two wheelers, and occupants of buses and minibuses (WHO, 2004). Globally, the risk of dying in a road crash is far higher for vulnerable road users-pedestrians, cyclists and motorcyclists-than for car occupants. Africa has 4 percent of the world's cars but accounts for more than 11 percent of the world's traffic casualties. The WHO figures that road casualties in Africa are under reported by as much as twelve fold, and it predicts the death toll will rise an additional 80 percent by 2020, as the population grows and becomes more motorized.

1.1.1. Data Mining

The steady growth of computers and information technology helped the availability of data on different location with various formats. The abundance of data, together with the need for powerful data analysis tools in many countries has been described as data rich but information poor society (Han and Kamber, 2006).

This fast growth and tremendous amount of data, collected and stored in large and numerous databases need a powerful tool to elicit useful information. The tool helps to get benefit from the collected data, by identifying relevant and useful information. Data mining is one of the solutions to analyze huge amount of data and turn such data into useful information and knowledge (Han and Kamber, 2006).

Han and Kamber (2006) simply stated, "Data mining refers to extracting or "mining" knowledge from large amounts of data". There are some other terms which carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data or pattern analysis, and data archaeology.

In general data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, Han and Kamber (2006) classified data mining tasks into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database where as predictive mining tasks perform inference on the current data in order to make predictions.

1.1.2. Road Traffic Control System in Addis Ababa

Addis Ababa is the capital city of Ethiopia and home to the African Union, the Economic Commission for Africa and other international organizations. Hence, it is a city with a number of offices playing various roles in economic, social and political sector development of the country as well as the continent. Traffic Control and investigation Department of Addis Ababa Police

Commission which is located in Bole sub city is one of such offices with the following fundamental responsibilities (Zelalem, 2009):-

- ✓ Ensure the safety to the citizen by promoting safe and orderly flow of traffic on the city's street and highways.
- ✓ Enforce all laws and ordinances as they relate to each of the different forms of streets and highway traffic.
- ✓ Reduce the number of vehicular and pedestrian accidents.
- ✓ Develop and implement strategies that will improve the flow of traffic, remove obstacles to traffic movement and expedite motor vehicle traffic about the city.
- ✓ Reducing crime by improving the quality of life which is part of the goal of the police commission

To discharge these responsibilities the office has staffs mainly composed of traffic police officers and various equipments such as Motor Bicycles and different Automobiles.

1.2. Statement of the Problem

The underlying research problem that initiated this research is the fact that road traffic accidents are among the top leading causes of deaths and injuries of various levels in Ethiopia.

"Infectious diseases and, to a lesser extent, chronic conditions have been the focus of traditional public health efforts in low- and middle-income countries, while road traffic injury prevention has been considered the responsibility of the transport, police, and legal sectors" (Hyder, 2004). There are more than a million deaths each year from road traffic injuries around the world; millions more suffer morbidity and long-term disability. Moreover, road traffic injuries impact on the most productive members of a society and result in large-scale economic losses for a country. These are all strong reasons for traffic accident to be a public health issue.

Ethiopia has some of the most dangerous roads in the world and has pursued an ambitious road expansion policy in the past 10 years. The Ethiopian National Road Safety Coordination Office cites a road crash fatality rate of 114 deaths per 10,000 vehicles per year but the real figure may be higher due to underreporting. This compares to a road fatality rate of one death per 10,000 vehicles per year in the United Kingdom of Great Britain and Northern Ireland and an average fatality rate of 60 per 10,000 vehicles across 39 sub-Saharan African countries (WHO, 2004). In addition, the country is experiencing highest rate of such accidents resulting in fatalities and

various levels of injuries. The capital city, Addis Ababa shares 65% of the total accident in the country. Pedestrians are the most vulnerable ones in Addis Ababa; above 81% of accident fatalities are of accident type "car hit pedestrian". Moreover, 81% of crashes in Ethiopia are attributed to driver error (ERA, 2005).

In an attempt to prevent road accidents one role that can be played is researching the causes of traffic crashes and injuries and try to attack the problem from its root. To carry out such researches on voluminous, multi featured and historical accident data it requires some state of the art tool and technique. One such tool is data mining.

Research on road safety using data mining tools has been conducted for several years mainly in developed countries, and a few locally. Tibebe (2005) conducted a research on historical RTA data comprising a dataset of 4,658 accident records at Addis Ababa Traffic Office to investigate the application of data mining technology for the analysis of accident severity. He proved that data mining can be applied in road safety. Following Tibebe, Zelalem (2009) has also conducted a data mining research to classify drivers' responsibility on a given accident in Addis Ababa. In addition, Tibebe and Hill (2010) again did a research on road related factors on accident severity.

The previous researches have focused merely on single attributes that help to predict traffic accident in Addis Ababa, which shows there is a gap for further research that combines the drivers' information and road attributes to predict accident severity. Changes on traffic rules and regulations are made in the capital city, which has its own contribution in road safety after these researches have been done.

Thus this research answered the following questions:

- What are the main determinant factors (attributes) of drivers and road that cause traffic accident?
- What are the most interesting patterns or rules generated using the determinant factors of drivers and roads that can be used as a traffic rules and policies?
- Which data mining techniques perform well in developing a model that can identify and predict drivers and road determinant factors?

Moreover, although the existence of a large number of road accidents are shown by different studies and road traffic accident data are gathered periodically by the Addis Ababa traffic control and investigation department, due to lack of appropriate data analysis tools this historical and accumulated data has not been used for analysis.

The accumulated data is a major source of solution to analyze the determinant factors of the problem, more specifically drivers and road determinant risk factors that cause a great loss of life.

Furthermore, Tibebe Beshah and Shawandra Hill (2010) in their research on "*Mining Road Traffic Accident Data to Improve road Safety: Role of Road-related Factors on Accident Severity in Ethiopia*" have also recommended there is a need of further research on combining drivers information and road factors on prediction of traffic accident severity.

In this thesis, the researcher constructed a model that predicts the accident severity based on the drivers information and road characteristics, using a traffic accident data from Addis Ababa Road Traffic Control and Investigation Department (AARTCID).

1.3. Research Contribution

In this research an attempt is made to find out the applicability of data mining technology in identifying determinant factors of road and drivers which lead to traffic accident, it will have the following contributions in road safety and improving the public health:

- It will help health policy makers in planning health programs and to know main drivers' and road factors that contribute to traffic accident.
- It will also pave the way to develop better parameters in all aspects of traffic control system. Specifically it will support the Traffic Control Division of Addis Ababa in taking proper action, such as revising the existing traffic rules, against vehicle accidents.

1.4. Objectives

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

1.4.1. General objective

The general objective of the study is to identify and investigate drivers' and road factors that contribute to the cause of accident and to develop traffic accident prediction model that improves the public health.

1.4.2. Specific objectives

To accomplish the above stated general objective, the following specific objectives are carried out.

- Conduct a thorough review of literature on the existing data mining techniques and methods in general, and their application in road safety in particular.
- Review related literature in data mining as applied to road traffic accidents.
- Identify appropriate data mining algorithms and assess different data mining application software that are more appropriate to the problem domain, and select the best software.
- Select and extract the dataset required for analysis from the database of AARTCID.
- Prepare the data for analysis which includes adjusting inconsistent data encoding, accounting for missing values, and deriving other fields from existing ones;
- Compare and suggest the best model for prediction.
- Report the result and forward recommendation.

1.5. Methodology

In this research CRISP-DM standard data mining methodology is adopted. The important iterative activities that are undertaken in this research are: business understanding, data understanding, data preprocessing, selection of modeling technique, model building and model evaluation. Due to its documentation, inclusion of features to handle almost all activities performed in any data mining methods, and availability of the software; WEKA data mining tool is selected and used for this research.

Starting from the knowledge discovery processes used in early data mining projects, CRISP-DM defined and validated a data mining process that could be applicable in any industry sectors. This

methodology can make large data mining projects faster, cheaper, more reliable and more manageable (The CRISP-DM consortium, 2000).

This process model provides a simple *overview of the life cycle of a data mining project*. Corresponding phases of a data mining project are clearly identified throughout tasks and relationships between these tasks. Even if the model doesn't indicate it, there possibly exist relationships between all data mining tasks mainly depending on analysis goals and on the data to be analyzed. The sequence of the phases is not rigid; it is possible to move back and forth between different phases when ever required (The CRISP-DM consortium, August 2000). In this section, the researcher tries to briefly discuss the steps in CRISP-DM methodology.

Business/problem understanding: This initial phase focuses on understanding the data mining project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. Hence, in this research, an effort is first made to understand the problem domain, which is traffic accident and correct it into a data mining problem.

Data collection and understanding: In this phase, collection of original data is made. Activities are performed in order to get familiar with the data. Efforts to identify data quality problems are made, which helped the researcher to discover first insights into the data and to detect interesting subsets to form hypotheses for hidden information. Hence, in this research data is collected from previous researchers and AATCID database. A total of 18,419 traffic accident records from year 2005 to 2009 are used for the research. The dataset consists of a total of 40 attributes.

Data preparation and preprocessing: The data preparation phase covers all activities to construct the final dataset (data that are fed into the modeling tool(s)) need to be prepared from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools. Accordingly, missing values, outliers and noisy data are identified and handled, and data transformation/ reduction activities are also undertaken in this phase. Moreover feature/attribute construction or deriving new attributes through segmentation is also performed.

At this stage the collected data is arranged into a form that is suitable for the data-mining tool selected. Which means the data is prepared for analysis by collecting it to a new database. In addition, pre-processing tasks like handling noisy data, unknown values, missing values, deriving new fields from the existing ones, and summarization of data are done by taking into account the model building techniques. This helped a great deal in effectively applying the data mining tool and its algorithms on the accident data set.

Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there could exist several techniques for the same data mining problem. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

In this step of the data mining methodology, different algorithms supported by the selected data mining tool are examined by taking into consideration their application to the problem domain. Models are built and evaluations are carried out automatically to select the best model for prediction. Finally the appropriate predictive model is recommended for use by the decision makers, planners and health program developers.

Model Evaluation: At this stage in the research, a model is built that appear to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached (The CRISP-DM consortium, August 2000).

Deployment: Creation of the model is generally not the end of the research. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that it can be used. It often involves applying live models within an organization's decision making processes (The CRISP-DM consortium, August 2000).

1.5.1 Data Source Identification and Collection

The accident data for the research is collected from two sources. The first source of data is the AARTCID, where recent traffic accident data are collected. The second source is from local researchers, who have conducted data mining researches in traffic accident. The data set collected has been entered to the appropriate data mining tool and explored so that the researcher understood the data set properly.

1.5.2 Data Mining Tool Selection

It is important to assess tools and techniques early in the process since the selection of tools and techniques possibly influences the entire project.

Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. Han and Kamber (2006) suggest the following criteria to be used to assess the usefulness of data mining tools or software to the intended data mining task:

1. The goal of the data mining task in the research. The selected software should be able to provide the required data mining functions and methodologies. The data mining function carried out in this research is classification of the accident dataset, WEKA version 3.6.0 is used for this purpose.
2. Architecture and operating system. Some data mining software operate on specific types of architecture and operating systems, thus it is always wise to understand the computer architecture and the operating system on which the software is going to run. WEKA runs in Windows as well as Linux operating systems.
3. Data sources: specific data format on which the data mining software operate is also another important factor to consider. The accident data set is in MS-Excel format which is suitable for WEKA data mining software.
4. Scalability: the maximum number of columns and rows the software can efficiently handle. The target dataset has 40 columns and 18,420 records. The selected data mining software, WEKA version 3.6.0 supports up to 65,000 records with memory expansion.
5. Visualization capabilities: the variety, quality, and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of a data mining system. WEKA has a facility to visualize its outputs.

The above features of the software influenced the researcher to choose WEKA data mining software. In addition WEKA is easy to access and it provides a number of data mining functionalities such as classification, clustering, association, attribute selection, and visualization. WEKA is open source software where documents and supports are available freely. Familiarity of the researcher to the software is also another reason to select WEKA data mining tool. In the following section a description about WEKA software is presented. In addition to WEKA the following tools are used in the research:

- Ms- Excel is used for data preparation and pre-processing tasks for its filtering capability of attributes with different values.
- Ms-Word is used for documentation purposes.
- Ms- PowerPoint is used for preparing slides for presentation.

1.5.3 WEKA Data Mining Tool

WEKA is developed by the University of Waikato in New Zealand. "WEKA" stands for the Waikato Environment of Knowledge Analysis (Witten and Frank, 2005). The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms. WEKA has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. WEKA expects the data to be fed in ARFF format. It is necessary to have information about each attribute which cannot be automatically deduced from the attribute values (Witten and Frank, 2005).

WEKA includes a variety of tools for preprocessing a dataset, such as attribute selection, attribute filtering and transformation, feeding into a learning scheme, and tool to analyze the resulting classifier and its performance. WEKA is organized in packages that correspond to a directory hierarchy. The important packages of WEKA are association; attribute selection, classifiers, clusters, estimators, and filters packages (Whitten and Frank, 2005).

1.6. Scope and Limitation of the Study

This research is limited to the classification and analysis of traffic accident data from the year 2005 up to 2009 using road safety related drivers' information and road factors 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. In addition, intervention of the rules in the office is not included in this research report.

1.7. Thesis Organization

This research is organized into six chapters. The first chapter briefly discusses background to the problem area (i.e. Road traffic accident) and data mining technology, and states the problem, objective of the study, research methodology, scope and limitation of the research. Chapter two deals with literature review about data mining techniques and algorithms implemented in the study and its application in the area of road safety. Chapter three deals with business understanding about traffic accident situation in Addis Ababa and Ethiopia; where traffic accident is a core public health issue. Chapter four explains about data preparation for analysis. Chapter five presents the experimentation phase of the study. Results of decision tree and rule induction algorithm are discussed briefly. Finally, Chapter six provides conclusion, and offers recommendations for future work.

Chapter Two
Data Mining Technology

2.1 Introduction

In this chapter the essence of data mining, its application, how it is different from and similar to other related fields such as statistics, and the basic conceptual background of its methods and techniques is discussed.

The convenience and easy availability of various technologies in generating collecting, manipulating and storing data led to today's level of data explosion. Progress in database technology and other related fields in addition to the explosive growth in data collected from applications including business and management, government administration, science and engineering, and environmental control increased the demand for efficient and effective data analysis and data understanding tools (Han and Kamber, 2006). This in turn led to the increased availability of huge amounts of data, and its complex nature for simple statistical and manual analysis that necessitate the use of data mining technology.

Various definitions for data mining exist in literature. Han and Kamber (2001) defined data mining as the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. Data mining is the extraction of implicit, previously unknown, and potentially useful information from data (Whitten and Frank, 2005). While, according to Hand et al (2001) it is "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner". Two Crows Corporation (2005) also defines data mining as a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

2.2 Data Mining Technology and Statistics

In the past, statistical models have been widely used to analyze large amount of data. However, certain problems may arise when using classic statistical analysis on datasets with such large dimensions. For example, exponential increase happens in the number of parameters as the

number of variables increases and there will be invalidity of statistical tests as a consequence of sparse data in large contingency tables Chen and Jovanis (2002). This is where data mining comes to play.

Data Mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from large amounts of data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies. In data mining, data sets can be much larger than in statistics and data analyses are on a correspondingly larger scale.

Friedman (1997), on his paper "Data Mining and Statistics: What's the Connection?" have concluded that statistical models are more likely to be preferable when fairly simple models are adequate and important variables can be identified before modeling. However, when dealing with a large and complex data set, such as that of road accidents, the use of data mining methods seems particularly useful in identifying the relevant variables that make a strong contribution towards a better understanding of road traffic accident conditions.

Two Crows Corporation (2005) argues that Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. However, data mining is the application of different techniques like neural nets, decision trees and other Artificial Intelligence and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional.

2.3 Data Mining Methodology

In order to systematically conduct a data mining project, a general methodology is usually followed. There are some standard methodologies; One such methodology is CRISP, which is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study.

2.4 The CRISP-DM Methodology

CRISP-DM stands for CRoss Industry Standard Process for Data Mining. “It is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems” (Shearer, 2000). CRISP-DM came into existence in the early 90’s due to the growing interest for data mining and the absence of a "methodology" for knowledge discovery that would be neutral to any application, tool or industry. The CRISP-DM has six phases as depicted in the figure 2.1 (CRISP-DM 1.0, 2000).

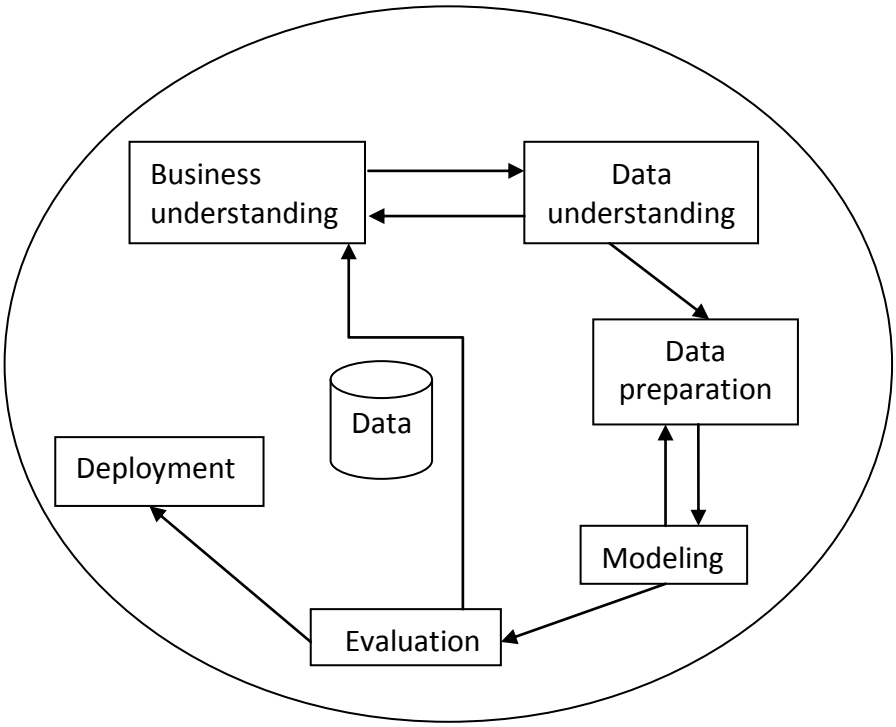


Figure 2.1: Phases of the CRISP-DM reference model (adapted from the CRISP-DM consortium, 2000)

The CRISP-DM data mining methodology is described in terms of a hierarchical process model as shown in the figure 2.2 below. It consists sets of tasks described at four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks and process instances (The CRISP-DM consortium, 2000).

At the top level, the data mining process is organized into a number of phases; each phase consists of several second-level generic tasks.

The second level is called generic, because it is intended to be general enough to cover all possible data mining situations. The generic tasks are intended to be as complete and stable as possible. Complete means covering both the whole process of data mining and all possible data mining applications. Stable means that the model should be valid for yet unforeseen developments like new modeling techniques.

The third level, the specialized task level, is the place to describe how actions in the generic tasks should be carried out in certain specific situations. For example, at the second level there might be a generic task called clean data. How the data should be cleaned is described in this level.

The fourth level, the process instance, is a record of the actions, decisions and results of an actual data mining engagement. A process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.

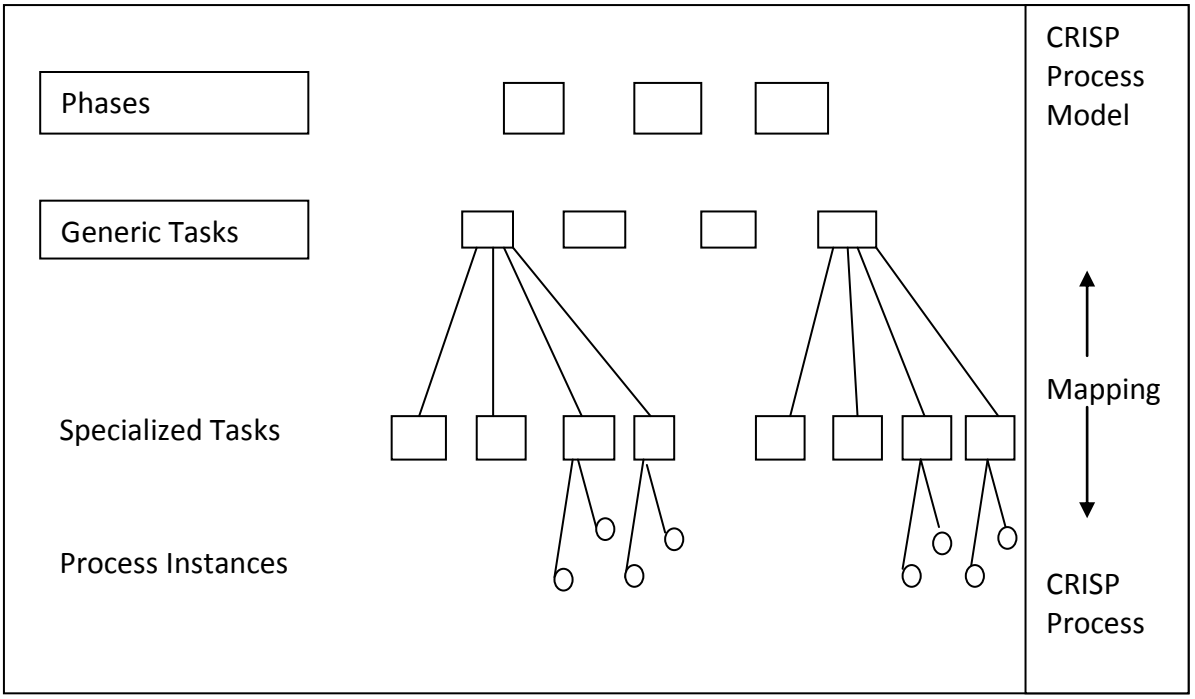


Figure 2.2: Four level breakdown of the CRISP-DM methodology

Source: The CRISP-DM consortium, 2000

As discussed in the above sections, in CRISP-DM a concept called data mining context drives mapping between the generic and the specialized level. The methodology is distinguished

between four different dimensions of data mining: application domain, data mining problem type, technical aspect and tool and technique (The CRISP-DM consortium, August 2000).

- The application domain is the specific area in which the data mining project takes place.
- The data mining problem type describes the specific class (es) of objective(s) that the data mining project deals with.
- The technical aspect covers specific issues in data mining that describe different (technical) challenges that usually occur during data mining.
- The tool and technique dimension specifies which data mining tool(s) and/or techniques are applied during the data mining project.

2.5 Data Mining Tasks

The cycle of data and knowledge mining comprises various analysis steps, each step focusing on a different aspect or task. Hand et. al (2001) propose the following categorization of data mining tasks.

2.6.1. Description and Summarization

This task of data mining aims to see general trends as well as extreme values quickly. Typically, getting the overview will at the same time point the analyst towards particular features, data quality problems, and additional required background information. Summary tables, simple univariate descriptive statistics, and simple graphics are extremely valuable tools to achieve this task.

2.6.2. Descriptive Modeling

Descriptive modeling tries to find models for the data. The aim of this model is to describe, not to predict models. As a consequence, descriptive models are used in the setting of unsupervised learning. Typical methods of descriptive modeling are density estimation, smoothing, data segmentation, and clustering.

Clustering is a well-studied and well-known technique in statistics. The most widely used method of clustering is k-means clustering. Although k-means is not particularly tailored for

large number of observations, it is currently the only clustering scheme that has gained positive reputation in both the computer science and the statistics community.

The reasoning behind cluster analysis is the assumption that the data set contains natural clusters which, when discovered, can be characterized and labeled. While for some cases it might be difficult to decide to which group they belong, we assume that the resulting groups are clear-cut and carry an intrinsic meaning. In segmentation analysis, in contrast, the user typically sets the number of groups in advance and tries to partition all cases in homogeneous subgroups.

2.6.3. Predictive Modeling

The aim of this task is to build a model that will permit the value of one variable to be predicted from the known values of other variables. In classification, the variable being predicted is categorical, while in regression the variable is quantitative. The term "prediction" is used here in a general sense, and no notion of a time continuum is implied. (Hand et. al, 2001).

Predictive modeling falls into the category of supervised learning; hence, one variable is clearly labeled as target variable Y and will be explained as a function of the other variables X. The nature of the target variable determines the type of model: classification model, if Y is a discrete variable, or regression model, if it is a continuous one. Many models are typically built to predict the behavior of new cases and to extend the knowledge to objects that are new or not yet as widely understood.

2.6.4. Discovering Patterns and Rules

The area of the previous tasks has been much within the statistical tradition in describing functional relationships between explanatory variables and target variables. There are situations where such a functional relationship is either not appropriate or too hard to achieve in a meaningful way. Nevertheless, there might be a pattern in the sense that certain items, values or measurements occur together more frequently. Association Rules are a method originating from market basket analysis to elicit patterns of common behavior.

2.6.5. Data Mining Techniques

A single data mining tool or technique is not equally applicable to all the above-mentioned tasks. Based on the nature of the problem under consideration and its proximity to the main divisions of data mining tasks, there is a need to choose the appropriate data mining techniques. Decision trees, Rule induction (Rule Learner), Neural Network, Clustering and Association Rule Mining are some of the data mining techniques that are used in most cases. Among these different techniques the researcher used Decision trees and Rule Induction technique in this research because their result is simple to explain for end user and these techniques support the selected data mining tasks for this research.

2.6.5.1. Decision Tree

A decision tree is a flowchart-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. Decision trees are most popular data mining techniques used for classification. Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules (Han and Kamber, 2006).

Since the construction of decision tree classifiers does not require any domain knowledge or parameter setting they are popular, and appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans (Han and Kamber, 2006).

The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology (Han and Kamber, 2006).

Classification using a decision tree is performed by routing from the root node until arriving at a leaf node (Bramer, 2007). Bramer (2007, page 41) extends his explanation of data type that can be handled by decision trees classification as:

> *Decision tree can represent diverse types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. Decision tree induction algorithms operate recursively. First, an attribute must be selected as the root node. In order to create the most efficient (i.e., smallest) tree, the root node must effectively split the data. Each split attempts to cut/trim down a set of instances (the actual data) until they all have the same classification.*

The best split is the one that provides the most information gain. Information in this context comes from the concept of entropy from information theory, as developed by Claude Shannon. Although "information" has many contexts, it has a very specific mathematical meaning relating to certainty in decision making. Ideally, each split in the decision tree should bring us closer to a classification. One way to conceptualize this is to see each step along the tree as removing randomness or entropy (Han and Kamber, 2006).

The process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class). The 'entropy method' of attribute selection is to choose to split on the attribute that gives the greatest reduction in (average) entropy, i.e. the one that maximizes the value of information gain. At any stage of this process, splitting on any attribute has the property that the average entropy of the resulting subsets will be less than (or occasionally equal to) that of the previous training set (Bramer, 2007).

The entropy of the training set is denoted by E. It is measured in 'bits' of information and is defined by the following formula as presented in Bramer (2007). The following explanation of using entropy based information gain for tree pruning in decision tree is extracted from Han and

Kamber (2006). Let node N represents or holds the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by equation 2.1.

$$\text{Info}(D) = -\sum_{i=1}^{k} \text{pi} \log_2(\text{pi}) \ldots\ldots\ldots\ldots\ldots \text{Equation 2.1}$$

Where pi is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_i, D| / |D|$. A log function to the base 2 is used, because the information is encoded in bits. Info (D) is just the average amount of information needed to identify the class label of a tuple in D. At this point, the information we have is based solely on the proportions of tuples of each class. Info (D) is also known as the entropy of D.

Suppose we were to partition the tuples in database D on some attribute A having V distinct values, {a1, a2. . . a_v} as observed from the training data. If A is discrete-valued, these values correspond directly to the V outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, {D_1, D_2… D_V}; where D_j contains those tuples in D that have outcome a_j of A. These partitions would correspond to the branches grown from node N. Ideally; we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). The amount of information we would still need (after the partitioning) in order to arrive at an exact classification is measured by equation 2.2.

$$\text{Info}_A{}^{(D)} = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * \text{Info}\ (D_j) \ldots\ldots\ldots \text{Equation 2.2}$$

The term D_j/D acts as the weight of the j^{th} partition. Info_A (D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected

information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$\text{Gain (A)} = \text{Info (D)} - \text{Info}_{A}^{(D)} \quad \text{.....Equation 2.3}$$

Gain (A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain (A)), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum InfoA(D)). So the researcher need to implement the entropy based attribute subset selection for tree pruning in this particular research. WEKA data mining tool selected for decision tree model building has J48 implementation which uses information gain method for tree pruning.

2.6.5.2. J48 algorithm

J48 is a version of an earlier algorithm developed by J. Ross Quinlan, C4.5. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. It is important to understand the variety of options available when using this algorithm, as they can make a significant difference in the quality of results. In many cases, the default settings will prove adequate, but in others, each choice may require some consideration.

The J48 algorithm gives several options related to tree pruning. Many algorithms attempt to "prune", or simplify, their results. Pruning produces fewer, more easily interpretable results. More importantly, pruning can be used as a tool to correct for potential over fitting. The basic algorithm described above recursively classifies until each leaf is pure, meaning that the data has been categorized as perfectly as possible (Han and Kamber, 2006).

This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular idiosyncrasies of that data. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data. This is

because pruning employs various means to relax the specificity of the decision tree, hopefully improving its performance on test data. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy (Witten and Frank, 2005).

J48 in WEKA3.6.0 employs two pruning methods. The first is known as subtree 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 raising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtree raising 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 a long time. This is due to the fact that subtree raising can be somewhat computationally complex (WEKA manual, 2008).

Error rates are used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest 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 is known as reduced-error pruning. Though the method is straight-forward, it also reduces the overall amount of data available for training the model. For particularly small datasets, it may be advisable to avoid using reduced error pruning.

Other error rate methods statistically analyze the training data and estimate the amount of error inherent in it. There are several other options that determine the specificity of the model. The minimum number of instances per leaf is one powerful option. This allows you to dictate the lowest number of instances that can constitute a leaf. The higher the number, the more general the tree is. Lowering the number will produce more specific trees, as the leaves become more granular. 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 encourages more generalized trees. There is also an option available for using Laplace smoothing for predicted probabilities. Laplace smoothing is used to prevent probabilities from ever being calculated as

zero. This is mainly to avoid possible complications that can arise from zero probabilities. The most basic parameter is the tree pruning option. If one decides to employ tree pruning, there may be a need to consider the options for pruning. It is important to know that depending on how the training and test data have been defined that the performance of an unpruned tree may superficially appear better than a pruned one (WEKA Manual, 2008).

It is also important to build models by intelligently adjusting these parameters. Often, only repeated experiments and familiarity with the data will tease out the best set of options as shown in Table 2.1 below which is taken from WEKA Manual (2008).

Table 2.1 J48 classifier descriptions of the parameters

Option	Description
ConfidenceFactor	The confidence factor used for pruning (smaller values incur more pruning).
BinarySplits	Whether to use binary splits on nominal attributes when building the trees.
debug	If set to true, classifier may output additional information to the console.
minNumObj	The minimum number of instances per leaf.
numFolds	Determines the amount of data used for reduced error pruning. One fold is used for pruning, the rest for growing the tree.
reducedErrorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning.
saveInstanceData	Whether to save the training data for visualization.
seed	The seed used for randomizing the data when reduced-error pruning is used.
subtreeRaising	Whether to consider the subtree raising operation when pruning.
unpruned	Whether pruning is performed.
useLaplace	Whether counts at leaves are smoothed based on Laplace.

2.6.5.3. ID3 algorithm

The ID3 algorithm starts with all the training samples at the root node of the tree. An attribute is selected to partition these samples. For each value of the attribute a branch is created, and the

corresponding subset of samples that have the attribute value specified by the branch is moved to the newly created child node. The algorithm is applied recursively to each child node until all samples at a node are of one class. Every path to the leaf in the decision tree represents a classification rule. Attribute selection at a node in ID3 and C4.5 algorithms are based on minimizing an information entropy measure applied to the examples at a node (Mehmed Kantardzic, 2003).

The original ID3 algorithm used a criterion called gain to select the attribute to be tested which is based on the information theory concept: entropy. The attribute-selection part of ID3 is based on the assumption that the complexity of the decision tree is strongly related to the amount of information conveyed by the value of the given attribute. An information-based heuristic selects the attribute providing the highest information gain, i.e., the attribute that minimizes the information needed in the resulting sub tree to classify the sample. An extension of ID3 is the C4.5 algorithm, which extends the domain of classification from categorical attributes to numeric ones. The measure favors attributes that result in partitioning the data into subsets that have low class entropy, i.e., when the majority of examples in it belong to a single class. The algorithm basically chooses the attribute that provides the maximum degree of discrimination between classes locally.

2.6.5.4. PART algorithm

According to Whitten and Frank (2005), many learning techniques look for structural descriptions of what is learned, descriptions that can become fairly complex and are typically expressed as sets of rules. Because they can be understood by people, these descriptions serve to explain what has been learned and explain the basis for new predictions. Classifications rules are a popular alternative to decision trees in representing the structures that learning methods produce. The antecedent, or precondition, of a rule is a series of tests just like the tests at nodes in decision trees, and the consequent, or conclusion, gives the class or classes that apply to instances covered by that rule, or perhaps gives a probability distribution over the classes. Generally, the preconditions are logically joined together by "AND", and all the tests must succeed if the rule is to work. It is also easy to read a set of rules directly off a decision tree. One rule is generated for each leaf. The antecedent of the rule includes a condition for every node on the path from the root to that leaf, and the consequent of the rule is the class assigned by the leaf.

One reason why rules are popular is that each rule seems to represent an independent "nugget" of knowledge.

PART is a class for generating decision list in WEKA. It builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule. Rules or decision lists which are generated using PART algorithm are more clear and understandable. As a result the researcher has also used this algorithm for modeling.

2.6 Data Mining Applications

Han and Kamber (2006) suggests that since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. Moreover they identified a few application domains which are summarized as follows.

- Data mining for biomedical and DNA data analysis has become a powerful tool by enabling semantic integration of heterogeneous and distributed genome databases, similarity search and comparison among DNA sequences, identification of co-occurring gene sequences (association analysis), linking genes to different stages of diseases (path analysis) and etc.
- Financial data analysis: Here data mining can be used to design and construct data warehouses for multidimensional data analysis, loan payment and customer credit policy analysis, classification and clustering of customers for target marketing and detection of money laundering and other financial crimes.
- Data mining for retail industry: Multidimensional analysis of sales, customers, products, time, and region, analysis of effectiveness of sales campaigns, analysis of customer loyalty and purchase recommendation and cross-reference of items.
- Data mining for Telecommunication industry: Multidimensional analysis of telecommunication data, fraudulent pattern analysis and the identification of unusual patterns, multidimensional association and sequential pattern analysis and use of visualization tools in telecommunication data analysis.

2.7 Related literature of Data mining for RTA

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. Thus applying data mining techniques to model traffic accident data records can help to

understand the problem and to find possible solutions. This can help the decision makers to formulate better traffic control policies.

Many data mining application researches on the analysis of RTA data have been done globally and a few locally. Reviewing this related works that were done using data mining tools and techniques at different place and time in the same problem domain gave the researcher an in-depth insight for this research. The researcher has summarized some of the most important works as follows.

Chong et.al (2002) in their paper on the analysis of the GES automobile accident data from 1995 to 2000 using machine learning paradigms investigated the performance of neural network, decision tree, support vector machines and a hybrid decision tree – neural network based approaches to predicting drivers' injury severity during traffic accidents in head on front impact point collisions. In their report they revealed that the classification accuracy obtained for the non-incapacitating injury, the incapacitating injury, and the fatal injury classes, the hybrid approach performed better than neural network, decision trees and support vector machines. For the no injury and the possible injury classes, the hybrid approach performed better than neural network. They also found out that the no injury and the possible injury classes could be best modeled directly by decision trees. They extended the research to possible injury, non-incapacitating injury, incapacitating injury, and fatal injury classes and showed that the model for fatal and non-fatal injury performed better than other classes.

They also underlined the importance of the ability of predicting fatal and non-fatal injury, since human fatality has the highest cost to society economically and socially. According to them the most important factor causing different injury level is the actual speed that the vehicle was going when the accident happened. Unfortunately, their dataset didn't provide enough information on the actual speed since speed for 67.68% of the data records' was unknown. They believed that if the speed was available, it was likely that it could have helped to improve the performance of models studied in their paper.

Tibebe (2005) conducted a research on historical RTA data comprising a dataset of 4,658 accident records at Addis Ababa Traffic Office to investigate the application of data mining technology for the analysis of accident severity. In his thesis he built various classification models using the decision tree technique by applying KnowledgeSEEKER algorithm of the KnowledgeSTUDIO data mining tool to help in decision-making process at the traffic office. The methodology he adopted had three basic steps namely data collection, data preparation, and model building and validation.

The model classifies accident severity into four classes, fatal injury, serious injury, slight injury and property-damage. He identified 'accident cause', 'accident type', 'road condition', 'vehicle type', 'light condition', 'road surface type' and 'driver age' as the basic determinant variables for injury severity level. Finally, he reported classification accuracy of the decision tree classifier to be 87.47%. He has also recommended that further research like identification of road, driver and vehicle factors that lead to road accidents should be undertaken.

Following Tibebe's work, Zelalem (2009) conducted a data mining research to classify drivers' responsibility on a given accident in Addis Ababa. The research uses decision tree and multilayer perception (MLP) neural network data mining techniques to analyze the accident data. The study focuses on predicting the degree of driver's responsibility for car accidents and identifying the important factors influencing the different levels of responsibility by using the RTA dataset of Addis Ababa Traffic Control and Investigation Department (AARTCID).

The researcher used WEKA data mining tool to build the decision tree (using the ID3 and J48 algorithms) and MLP (the back propagation algorithm) predictive models. Rules representing patterns in accident dataset have been extracted from the decision tree indicating important relationships between variables that influence driver's degree of responsibility such as; age, license grade, level of education, driving experience, and other environmental factors. According to the author, the accuracies of the models were 88.24% and 91.84% respectively. In addition the research reveals that, the decision tree model is found to be more appropriate for the problem type under consideration.

Finally Getnet (2009) investigated the potential application of data mining tools to develop models supporting the identification and prediction of major driver and vehicle risk factors that cause RTAs. The researcher used WEKA version 3-5-8 tool to build the decision tree (using the J48 algorithm) and rule induction (using PART algorithm) techniques. Performance of the J48 algorithm was slightly better than that of the PART algorithm. The license grade, vehicle service year, vehicle type, and experience were identified as the most important variables for predicting accident severity.

In this research determinant factors of drivers and road that cause traffic accident are identified Which will help health policy makers in planning health programs and it will also support the Traffic Control Division of Addis Ababa in taking proper action, such as revising the existing traffic rules, against vehicle accidents.

Chapter Three

Road Traffic Accidents

Road traffic accidents and their resulting fatalities may be regarded as a growing social and economic problem, especially in developing countries like Ethiopia, where the resources are limited.

Road traffic accident Kills more than 1.2 million and injures between 20 and 50 million people every year, as a results it became the ninth most common cause of death in 2004, and remains among the most central public health problems in the world (WHO, 2004). A tragic fact is that among the young people aged between 15 and 29 years, a road traffic injury is the most common cause of death worldwide. WHO reports that 90% of the road traffic deaths occur in low-income or middle-income countries.

This chapter deals with different literature reviews on the business domain, traffic accident. It is the first step of the CRISP-DM methodology. The concept of road traffic accident, the role of public health in reducing the problem and road traffic data analysis in Addis Ababa traffic office is discussed.

3.1 Traffic Accident as a Health Problem

Worldwide over one million people are reportedly killed each year in road crashes, equivalent to three deaths every minute. The World Health Organization report (2004) estimated that by the year 2020 road accidents will be the third leading cause of 'disability adjusted life years' (DALY) – putting road safety well ahead of wars, HIV/AIDS, malaria and (other) 'acts of violence' as a world health problem (depicted in Table 3.1 below). Currently, a number of hospital based surveys in developing countries have found that traffic accidents make up a very high proportion of the people being treated at accident and emergency departments and occupying hospital beds (WHO, 2004).

Table 3.1: Change in rank order of DALYs for the 10 leading causes of the global burden of disease Source: WHO, 2004 report

1990		2020	
Rank	Disease or injury	Rank	Disease or injury
1	Lower respiratory Infections	1	Ischaemic Heart Disease
2	Diarrheal diseases	2	Un polar Major depression
3	Prenatal conditions	3	Road traffic injuries
4	Un polar Major depression	4	Cerebovascular disease
5	Ischaemic Heart Disease	5	Chronic obstructive pulmonary disease
6	Cerebovascular disease	6	Lower respiratory Infections
7	Tuberculosis	7	Tuberculosis
8	Measles	8	War
9	Road traffic injuries	9	Diarrheal diseases
10	Congenital abnormalities	10	HIV

In the early times, road traffic safety has been assumed to be the responsibility of the transport sector, although the main focus within this sector has typically been limited to building infrastructure and managing traffic growth.

The world health organization in its report of 2004 described the important roles public health can play. This includes:

- Discovering, through injury surveillance and surveys, as much as possible about all aspects of road crash injury by systematically collecting data on the magnitude, scope, characteristics and consequences of road traffic crashes.
- Researching the causes of traffic crashes and injuries, and in doing so trying to determine: causes and correlates of road crash injury; factors that increase or decrease the risk; factors that might be modifiable through interventions.
- Exploring ways to prevent and reduce the severity of injuries in road crashes by designing, implementing, monitoring and evaluating appropriate interventions.

- Helping to implement, across a range of settings, interventions that appear promising, especially in the area of human behavior, disseminating information on the outcomes, and evaluating the cost-effectiveness of these programs.
- Working to persuade policy-makers and decision-makers of the necessity to address injuries in general as a major issue, and of the importance of adopting improved approaches to road traffic safety.
- Translating effective science-based information into policies and practices that protect pedestrians, cyclists and the occupants of vehicles.
- Promoting capacity building in all these areas, particularly in the gathering of information and in research.

The above discussion clearly indicates that one of the solutions to reduce traffic accident problem is researching, and data mining is one research tool in finding the causes of traffic accidents.

3.2 Road Traffic Accident Situation in Ethiopia

Road traffic accidents occur as a result of several factors associated with the traffic system, namely: road users, road environment, and vehicles. In Ethiopia, in 2002/3, 81% of all accidents are due to drivers error, 5% accounted for vehicle error, 4 % pedestrian error and 10% were associated with road environments (NRSCO, 2008). As it can be seen clearly from the figures accidents are highly attributed to drivers and road environments.

Road accident in Ethiopia is one of the worst accident records in the world. Moreover, road accidents are concentrated in Addis Ababa which is the capital city of Ethiopia accounting for 65% of the total accidents occurred in the country (NRSCO, 2008).

According to the report provided by the Interim National Road Safety Coordination Office, the reasons for the relatively high number of road traffic accident include:

- Lack of driving skills;
- Poor knowledge of traffic rules and regulations;
- Violation of speed Limit;
- Insufficient enforcement;

- Lack of vehicle maintenance;
- Animal drawn carts and animals frequently using in main highways;
- Lack of safety conscious design and planning of road network;
- Disrespect of traffic rules and regulations;
- Lack of general safety awareness by pedestrians; and
- Lack of medical facility in general, which increase the severity of accidents.

The apparent causes of accidents were identified by Traffic Police Officers. These causes were not accurate as identified on- the-spot of accident investigations. Usually, the report by the traffic police indicates a single cause of each accident but not multiple causes of the accident. However, the causes of road accidents are normally multi-factors always preceded by a situation in which one or more of road users have failed to cope with the road environment (Getu Segni, 2007).

In respect to the accident types, the report indicated that pedestrian hit by car is the leading accident type contributing 68% of fatalities while fell from car, roll over, collision with animals and others contribute 6%,13%,3% and 10% respectively. In the capital 82% of the road accident fatality is due to vehicle hitting pedestrian.

Table 3.2: Fatality percentage by RTA type across the country for years 2002-2007

Accident Type	**Fatality**
Hit pedestrian	68%
Fell from car	6%
Roll over	13%
Collision with animal	3%
Others	10%

3.3 Road Traffic Accident Recording System in Ethiopia

Regional departments of the Traffic Police are responsible for the recording of all traffic accidents under their jurisdictions. The Federal Police Commission is responsible for national accident data compilation and processing. In each Region's Woreda Police Station, accident data are reported manually. The traffic police accident data form contains accident classification, date, time, day of the week, year, age of the driver, sex, education of the driver, ownership of the vehicle, service year of the vehicle, defects of the vehicle, location of accident, road traffic condition, road surface condition, road junction type, weather and illumination condition, collision type, and property damage and parties injured (age, sex, physical fitness and the like).

Monthly reports are submitted to pertinent region Police Commissions. A yearly report from the Region Police Commission will then be submitted to the Federal Police Commission to generate national accident statistics. In Ethiopia, much of the information is needed for the traffic police's own activity, primarily, to enforce the law and carry out prosecutions. Some of the accident data are of no direct interest or use to the police, but are vital to the work of other organizations.

3.4 RTA Data Analysis in Addis Ababa Traffic office

For better implementation of road safety policies, it is essential to have suitable data sources. Many jurisdictions require the collection and reporting of road traffic incident statistics. Such data enables figures for deaths, personal injuries, and possibly property damage to be produced, and correlated against a range of circumstances. Analysis of this data help in identifying the causes and for taking counter measures.

RTA data is the interest of various stakeholders. The primary data source for the analysis of RTA's occurring at Addis Ababa is the accident data kept at the Traffic Office. The accident detail is recorded by the investigators at the place of the accident.

The data collected by the investigators mainly is used by the office for undertaking some simple and manual statistical analysis, such as analysis of accident severity rates: the rate of fatalities, serious injury, simple injury and property loss per week, per month and per-year. The analysis result is visualized in histograms. Moreover, traffic accident data in various sections of the

country and including those at the capital has been analyzed and reported by different researchers and stakeholders.

Tibebe Beshah and Shawndra Hill (2010) in their research on "*Mining Road Traffic Accident Data to Improve road Safety: Role of Road-related Factors on Accident Severity in Ethiopia*" have also recommended there is a need for further research on combining drivers information and road factors on prediction of traffic accident severity.

In general, data mining applications in the area of road safety reveals the rich potential of the field in the efforts against road traffic accident prevention. However, much of the attempts have not taken into account the combined drivers and road attributes to predict accident severity and to extract rules for traffic accident patterns. Thus, this work is unique in the sense that it employs the two main attributes in traffic accidents so as to come up with a model that could support the process of traffic accident reduction in the city.

Chapter Four

Data Preparation

One of the most important tasks in data mining is preparing the data in a way that is suitable for the specific data mining tool or software package to be used. Data preparation involves data selection, data cleaning, data construction, data integration and data formatting (The CRISP-DM consortium, August 2000).

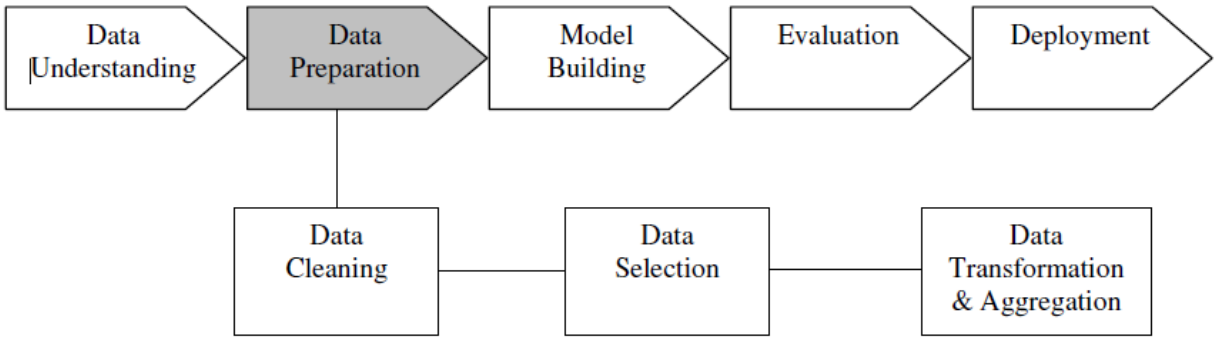


Figure 4.1 data preparation phase

In any data mining task the first step is clear understanding of the problem (i.e RTA in this study) to be solved and this has been already addressed in chapter three of this study which helped to know what data is required to perform the task.

In this chapter data understanding activities such as; data collection, data description and data formatting have been undertaken. Secondly, the data has been pre-processed by employing data cleaning and data selection techniques.

4.1. Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data in order to identify data quality problems, discover first insights into the data, or detect interesting subsets to form hypothesis for hidden information (The CRISP-DM consortium, August 2000).

Domain experts are consulted to have insight into the problem domain. The domain experts constitute two individuals from AATCID and National Road Safety and Coordination Office that are in charge of road traffic accident reduction and prevention and data collection on traffic accidents. On the basis of the insight gained from discussion with domain experts and review of relevant documents, a clear understanding of the data is achieved.

4.1.1. Business process understanding

In addition to consultancy of domain experts, knowing the business process also help the researcher to have a good understanding of the data.

Traffic Control and investigation Department of Addis Ababa Police Commission is the office, which is located in Bole sub city, is responsible for reduction and prevention of traffic accidents in the city. To discharge this responsibility the office has staffs mainly composed of traffic police officers and various equipments such as Motor Bicycles and different Automobiles.

In each sub city's Police Station, accident data are reported manually. The traffic accident data form contains accident classification, date, time, day of the week, year, age of the driver, sex, education of the driver, ownership of the vehicle, service year of the vehicle, defects of the vehicle, location of accident, road traffic condition, road surface condition, road junction type, weather and illumination condition, collision type, and property damage and parties injured (age, sex, physical fitness and the like).

When an accident occurs in a sub city the traffic officer during the accident fills the form and report to the sub city police station. Monthly these accident reports are compiled and reported to Addis Ababa police commission. The commission generates annual statistics about how many deaths, slight injury and sever injuries occurred.

Starting from 2005, in collaboration with WHO, the office has launched a project on automation of the traffic accident recording system in Addis Ababa and Dire Dawa.

4.1.2. Data Collection

It is a bare fact that the concept of data mining doesn't exist without data. There is some real benefit if the data is already part of a data warehouse. If the data has already been cleansed for a

data warehouse, then it most likely will not need further cleaning in order to be mined. Furthermore, many of the problems of data consolidation have already been addressed and maintenance procedures have been put in place. However, a data warehouse is not a requirement for data mining. Setting up a large data warehouse that consolidates data from multiple sources, resolves data integrity problems.

The data source for this research consumption is the traffic accident data kept at AARTCID and secondary data from previous researchers in the area. The data stored at AARTCID is partially automated in an Excel file format, and the rest in manual ledgers. It stores partial road accident records of years 2005-09 that occurred in the city. The total accident dataset obtained is around 18,419.

4.1.3. Formatting the Data

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last attribute being the outcome field the model to predict. It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute (The CRISP-DM consortium, 2000).

The WEKA data mining tool requires the dataset to be in a comma separated file format called the Attribute Relation File Format (ARFF). The ARFF file format is the standard way of representing datasets that consist of independent, unordered instances and does not involve relationships between instances (Whitten and Frank, 2005). Hence, the accident data set which was originally in an Excel file format is converted to an ARFF file format.

4.1.4. Data Description

After the initial data collection, the next step is describing the data set. The accident data set has forty attributes of text, number, and date & time formats. Among these attributes the car plate number and the driver's name hasn't been given by the office for the sake of privacy. Out of 40 attributes 17 are selected based on their relevance for the research objective by consulting domain experts. Table 4.1 shows the complete description of the whole attributes.

Table 4.1 Description of the whole attributes

S.No	Attribute Name	Data Type	Description
1	RegNo	Number	A key that identifies the accident uniquely.
2	Date	Date	The exact date on which the accident happened.
3	Time	Time	The time at which the accident happened
4	Day	Text	The week day on which the accident happened.
5	DriverSex	Text	The sex of the driver causing the accident.
6	DriverAge	Number	The Age of the driver causing the accident
7	LevelOfEducational	Text	The level of education of the driver causing the accident
8	DriversCarConnection	Text	Whether the driver is an owner or a professional driver or other.
9	DrivingExprience	Number	The driving experience of the driver causing the accident
10	VehicleType	Text	The type of vehicle in the accident
11	VehicleOwnership	Text	The owner of the vehicle (government, NGO, private,UN)
12	VehiclePeriodOfService	Number	The year of service of vehicle
13	VehcleStatus	Text	The status of the vehicle
14	Subcity	Text	The name of subcity where the accident occurred.
15	ParticularPlace	Text	Conventional name given to places in the city like kirkos, piazza.
16	Area	Text	Whether the accident occurred in school or market areas.
17	RoadSeparation	Text	How road segments are separated
18	RoadOrientation	Text	How the road is oriented
19	RoadJunction	Text	The type of road junction
20	RoadSurfaceType	Text	Whether the road surface is asphalt or ground.
21	RoadSurfCondition	Text	Is the condition of the surface of the road is dry, muddy or wet.
22	WeatherCondition	Text	The weather condition
23	LightCondition	Text	The light condition
25	AccidentType	Text	The type of the accident
26	AccidentSeverity	Text	The severity of the accident
27	NoOfVehiclesInvolved	Number	The total number of vehicles involved in the accident
28	AccidentFactor	Text	The cause for the accident
30	LicenceGrade	Number	The drivers license grade
31	Victim	Text	The victim of the accident
32	VictimAge	Number	The age of the victim due to the accident
33	VictimOccupation	Text	The occupation of the victim due to the accident
34	VictmsHealthCond	Text	The health condition of the victim
35	PedsMovment	Text	The pedestrian action
36	NoOfVictims	Number	Total number of victims in the accident
37	NoMaleVictims	Number	Total number of male victims
38	NoFemaleVictms	Number	Total number of male victims
39	Investigator	Text	Name of the investigator
40	Year	Number	Year the accident occurred

4.2. Data Preparation for Analysis

4.2.1. Data Cleaning

Data cleaning is a time-consuming and labor-intensive procedure in data preparation but one that is absolutely necessary for successful data mining (Witten and Frank, 2005). Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006). Thus, data cleaning has become a must in order to improve the quality of data so as to improve the accuracy and efficiency of the data mining techniques.

The data cleaning tasks to raise the data quality to the level required by the selected analysis techniques involves selection of clean subsets of the data and the insertion of suitable defaults. To this end, thorough discussion has been made with the office and it is found out that missing attribute values at the time of data entry are recorded as "unknown" and for those records the attribute is irrelevant they simply left it as a blank assuming that it would be obvious. To fix these problems 1,708 records with missing or unknown values for significant number of attributes are removed from the dataset which makes the final number of instances to be 16,710. In addition 43 records having no value for the dependent attribute are removed manually. Noisy values for attributes are also deleted and set to blank. For example "k?L" for vehicle ownership attribute, and driver experience with value "noliscence" are cleaned.

In replacing missing values, the ReplaceMissingValues data filtering method of WEKA is used. It replaces missing values with mean and modal values for numeric and nominal values respectively. Since all the attributes are nominal their corresponding missing values are replaced by modal values.

The percentage of records having missing values for each attribute is summarized in following table 4.2. Those records with significantly large missed values are deleted from the data set.

Table 4.2: Missing value statistics for the selected attributes (T=text, N=numerical).

S.No	Attribute Name	Data Type	Percentage of missing values
1.	DriverSex	T	11.6
2.	DriverAge	N	11.7
3.	DriversEducBackground	T	11.6
4.	DrivingExprience	N	11.4
5.	VehicleOwnership	T	0.1
6.	DriverVehicleRelationship	T	11.8
7.	LicenceGrade	N	11.5
8.	RoadJunction	T	0.007
9.	Subcity	T	0.008
10.	RarticularArea	T	0.008
11.	Roadseparation	T	0.007
12.	RoadOrientation	T	0.007
13.	Typeofroad	T	0.007
14.	Roadcondition	T	0.008
15.	Weathercondition	T	0.01
16.	Lightcondition	T	0.008

4.2.2. Data/ Attribute Selection

Records are evaluated and classified based on the values of their attributes. Of course, some of the attributes of a record may be irrelevant to the process of classification and thus should be excluded.

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 (Whitten and Frank, 2005).

Out of the 40 attributes of the original data set, 17 attributes (including the class attribute) which are believed by the domain experts to have significant contribution in assessing road and driver factors on accident severity, which is the focus of this research, have been selected.

4.2.3. Data Transformation

Data transformation can involve, smoothing or feature (attribute) construction, which works to remove noise from the data. Smoothing techniques include binning, regression, and clustering. Attribute construction on the other hand is a process where new attributes are constructed and added from the given set of attributes to help the mining process. Smoothing can also serve as data reduction, for example in the case of smoothing through binning; the number of the distinct values for a certain attribute is reduced (Han and Kamber, 2006).

Data transformation aims to manipulate the data so that its content and its format are most suitable for the data mining process. Accordingly, based on the office's classification of driver's experience, Driv_Exp_Cat is derived from the base attribute driver experience to categorize the input values as no experience, between 1 and 2, 3 and 5, 6and 10 and above 10 years. Driv_Age_Cat is also derived from driver's age attribute to classify the input values as less than 18, between 18 and 30, 31and 50 and above 50 years.

4.2.4. Data Set Format

WEKA needs data to be prepared in some formats and file types. The data sets provided to this software are prepared in a format that is acceptable for WEKA software. WEKA accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format.

In order to prepare the data in such format the records from the Microsoft excel database are saved as a Comma Delimited (CSV) file. Once all processing is completed and the file is converted to .**csv** format, WEKA either process the .**csv** format itself or a file in the form of Attribute Relation File Format (.arff). For this study the data is given to the software in .arff format. A screen shot of the output is depicted in figure 4.2.

WEKA file starts with the dataset's name followed by list of attributes. In fact, the dataset's name should be preceded by the symbol '@' and the word 'relation' (for example; @relation exp 2 unk replaced with blank where "exp 2 unk replaced with blank" is the name of the dataset) and each attribute name also starts with the same symbol and the word 'attribute' and following the name of the attribute including its possible values. If the variable or attribute is nominal, a list of possible values contained in a brace is required. By default the last attribute in the list of the attribute of the dataset designates the target class.

```
@relation 'exp 2 unk replaced with blank'

@attribute DriverSex {M,F}
@attribute DriverAge {Above_51,18_30,31_50,Below_18}
@attribute DriversEducationalBackground
{JuniorSchool,SeniorSecondarySchool,AboveSSS,PrimarySchool,BasicEdu
cation,Illiterate}
@attribute DriverVehicleRelationship {HiredDriver,Owner,Other}
@attribute DrivingExperience
{5_10_Years,2_5_Years,Above_10_Years,Below_1_Year,1_2_Years,Noexper
ience}
@attribute LicenceGrade
{Fifth,Third,Fourth,Second,NoLicense,Special,First}
@attribute VehicleOwnership
{Government,Private,InteOrg,PublicOrg,Defence,Police,Diplomatic,UN,
AU}
@attribute Subcity
{Kirkos,Addisketema,Arada,Lideta,Bole,Kolfe,Yeka,Lafto,Akaki,Gulele
}
@attribute ParticularArea
{Office,Residential,MarketArea,School,Churches,RecreationalArea,Hos
pital,Fabrika,Other}
@attribute RoadSeparation
{Island,BiDirectional,LineSeparation,SingleDirection}
@attribute RoadOrientation
{StraightPlain,Hill,DownHill,SlightlyBending,StraightSlightlyInclin
ed,StraightInclined,Straight}
@attribute RoadJunction {Y-Shape,NoJunction,T-
Shape,CrossRoad,Roundabout}
@attribute TypeOfRoad {GoodAsphalt,TornAsphalt,Rocky,NotAsphalted}
@attribute RoadConditions {Dry,Wet,Mud}
@attribute WeatherConditions
{GoodAir,Cold,HeavyRain,LightRain,Hot,Fog}
@attribute LightCondition
{DayLight,Dusk,Dawn,NightWithLight,NightWithWeakLight}

@attribute AccidentSeverity
{PropertyLoss,SlightInjury,SevereInjury,Fatal}

@data
M,Above_51,JuniorSchool,HiredDriver,5_10_Years,Fifth,Government,?,O
ffice,Island,StraightPlain,Y-
Shape,GoodAsphalt,Dry,GoodAir,DayLight,PropertyLoss
M,18_30,SeniorSecondarySchool,HiredDriver,2_5_Years,Third,Private,?
,Residential,BiDirectional,StraightPlain,NoJunction,GoodAsphalt,Wet
,GoodAir,DayLight,PropertyLoss
M,18_30,AboveSSS,HiredDriver,5_10_Years,Fourth,Private,Kirkos,Marke
tArea,Island,StraightPlain,T-
Shape,GoodAsphalt,Dry,GoodAir,DayLight,PropertyLoss
M,18_30,SeniorSecondarySchool,Owner,2_5_Years,Fifth,Private,Addiske
tema,Office,BiDirectional,StraightPlain,NoJunction,GoodAsphalt,Dry,
GoodAir,DayLight,SlightInjury
M,31_50,SeniorSecondarySchool,HiredDriver,Above_10_Years,Third,Priv
ate,Arada,Office,Island,StraightPlain,NoJunction,GoodAsphalt,Dry,Go
odAir,DayLight,PropertyLoss
M,18_30,JuniorSchool,HiredDriver,5_10_Years,Fourth,Government,Kirko
s,Office,LineSeparation,StraightPlain,NoJunction,GoodAsphalt,Dry,Go
odAir,DayLight,SlightInjury
F,31_50,AboveSSS,Owner,5_10_Years,Second,InteOrg,Arada,School,LineS
eparation,StraightPlain,T-
Shape,TornAsphalt,Dry,GoodAir,DayLight,PropertyLoss
M,18_30,SeniorSecondarySchool,HiredDriver,2_5_Years,Third,Private,L
ideta,MarketArea,Island,StraightPlain,NoJunction,GoodAsphalt,Dry,Go
odAir,Dusk,PropertyLoss
M,31_50,SeniorSecondarySchool,Other,Below_1_Year,Third,Private,Kirk
os,Office,LineSeparation,StraightPlain,NoJunction,TornAsphalt,Wet,G
oodAir,DayLight,SlightInjury
M,18_30,PrimarySchool,HiredDriver,5_10_Years,Third,Private,Bole,Res
idential,Island,StraightPlain,NoJunction,GoodAsphalt,Dry,GoodAir,Da
yLight,PropertyLoss
```

Figure 4.2 ARFF files for road traffic accident data set.

Once the list of attributes is completed, the word '@data' is used to indicate the beginning of the data. Finally each record is prepared by listing the values of each attribute separated by comma and missing values are represented by a "?". Then it is saved with a file name "exp 2 unk replaced with blank.arff"

Chapter Five

Experimentation

As presented in the previous chapter the data is well understood, explored, selected and clean enough to be used for model building. This chapter presents the detailed activities carried out in selecting a modeling technique, implementation of the technique selected using the most appropriate algorithms and evaluation of the models in order to select the best one for prediction.

The study focuses on identifying determinant factors of drivers and road that lead to traffic accidents and building a prediction model. Various classification models have been built by using decision tree and PART rules. The models have been tested on different number of the selected attributes and the significance of the outputs of the most important model is presented for analysis to the domain experts. Finally, the model with the best performance is selected.

5.1. Selection of Modeling Technique

According to the CRISP data mining standard methodology employed in this research, selecting the actual modeling technique to be used is the first step in modeling (The CRISP-DM consortium, 2000).

The most powerful predictive modeling methods include decision tree, Neural Networks, Support Vector Machine, Gene Expression Programming and Symbolic Regression, K-Means Clustering, Linear Discriminant Analysis, Linear Regression models and Logistic Regression models. In order to accomplish this research, the researcher used two data mining techniques. These are decision trees (using J48 and ID3 algorithm) and rule induction (using PART algorithm) for knowledge representation.

Decision trees are easy to build and understand. They can handle both continuous and categorical variables and can perform classification as well as regression. It automatically handles interactions between variables and identifies important variables.

The WEKA data mining tool implements decision tree using ADTree, DecisionStump, ID3 and J48 algorithms. The J48 algorithm is WEKA's version of the C4.5 decision tree algorithm developed by Whitten and Frank (Whitten and Frank, 2005).

Classification rules are a popular alternative to decision trees in representing the structures that learning methods produce. The antecedent, or precondition, of a rule is a series of tests just like the tests at nodes in decision trees, and the consequent, or conclusion, gives the class or classes that apply to instances covered by that rule, or perhaps gives a probability distribution over the classes. One reason why rules are popular is that each rule seems to represent an independent "nugget" of knowledge (Whitten and Frank, 2005).

Accordingly PART rule algorithm of WEKA used to represent knowledge/ pattern identified. PART is a class for generating decision list in WEKA. In an attempt to come up with significant rules, PART run on the accident dataset with different number of attributes.

Different models are built using these algorithms by changing the composition of the variables and parameters utilized so as to discover the most important model generating the most interesting rules.

5.2. Experiment one

The first experiment is conducted using 17 attributes that are selected during the data preparation phase. These are DriverSex, DriverAge, DriversEducationalBackground, DriverVehicleRelationship,DrivingExperience,LicenceGrade,VehicleOwnership,Subcity,ParticularArea,RoadSeparation,RoadOrientation,RoadJunction,TypeOfRoad,RoadConditions,WeatherConditions, LightCondition and AccidentSeverity. The accidentseverity attribute is the dependent attribute and the rest are independent variables or predictors.

Decision tree building using J48 algorithm

WEKA has implementation procedures of numerous classification and prediction algorithms to develop decision tree. J48 algorithm of decision tree technique is one of these algorithms which support both numeric and nominal predicators and nominal class attribute values.

J48 algorithm is an implementation of the C4.5 decision tree learner. The algorithm for induction of decision trees uses the greedy search technique to induce decision trees for classification. There are many parameters which can be adjusted in order to obtain better models with respect to the accuracy (or other parameters which can be used as measure for the quality of the model). These parameters allow greater control of the user in the process of learning the models (Witten and Frank, 2005).

J48 has a facility of generating outputs both in tree form and rule sets. The set of rules are generally easier to understand since each rule describes a specific context associated with a class. It also shows the hierarchy of the determinant factors or attributes. In section 2.6.5.2 there is a brief explanation of how J48 algorithm works and description of parameters.

The cleaned and preprocessed dataset of arff format is fed to WEKA software. All the selected attributes and other dataset are shown in figure 5.1.

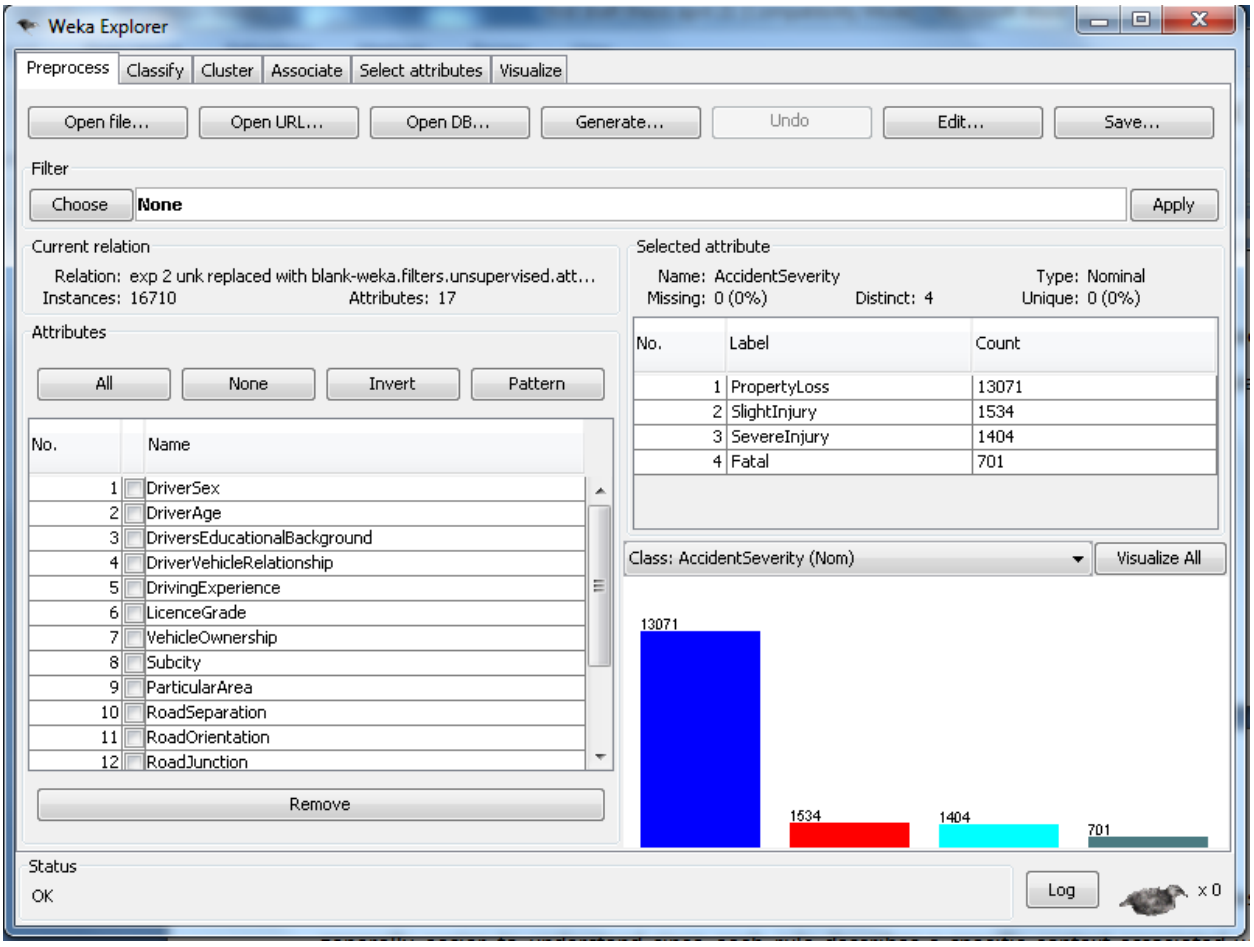


Figure 5.1 a screenshot that shows attributes prepared for experiment

In this first experiment all the 17 attributes related to drivers and road, which are believed by subject experts are used to build the model.

Since the explorer generally chooses sensible defaults (Whitten, 2005) the J48 decision tree algorithm with all its default parameters is run on the dataset. The default values for some of the parameters are: 0.25 for the confidence interval, pruning is allowed, the minimum number of objects for a leaf is 2. The training and testing is done using tenfold cross validation.

The k-fold (k=10) cross validation test options is used because the data set has unbalanced number of dependent class values; by doing so the partition and experiment could be more reliable. In this test option the accuracy estimate is the overall number of correct classifications from the k iteration divided by the total number of samples, which is k. After deciding the values of the parameters the algorithm is run to start building the model.

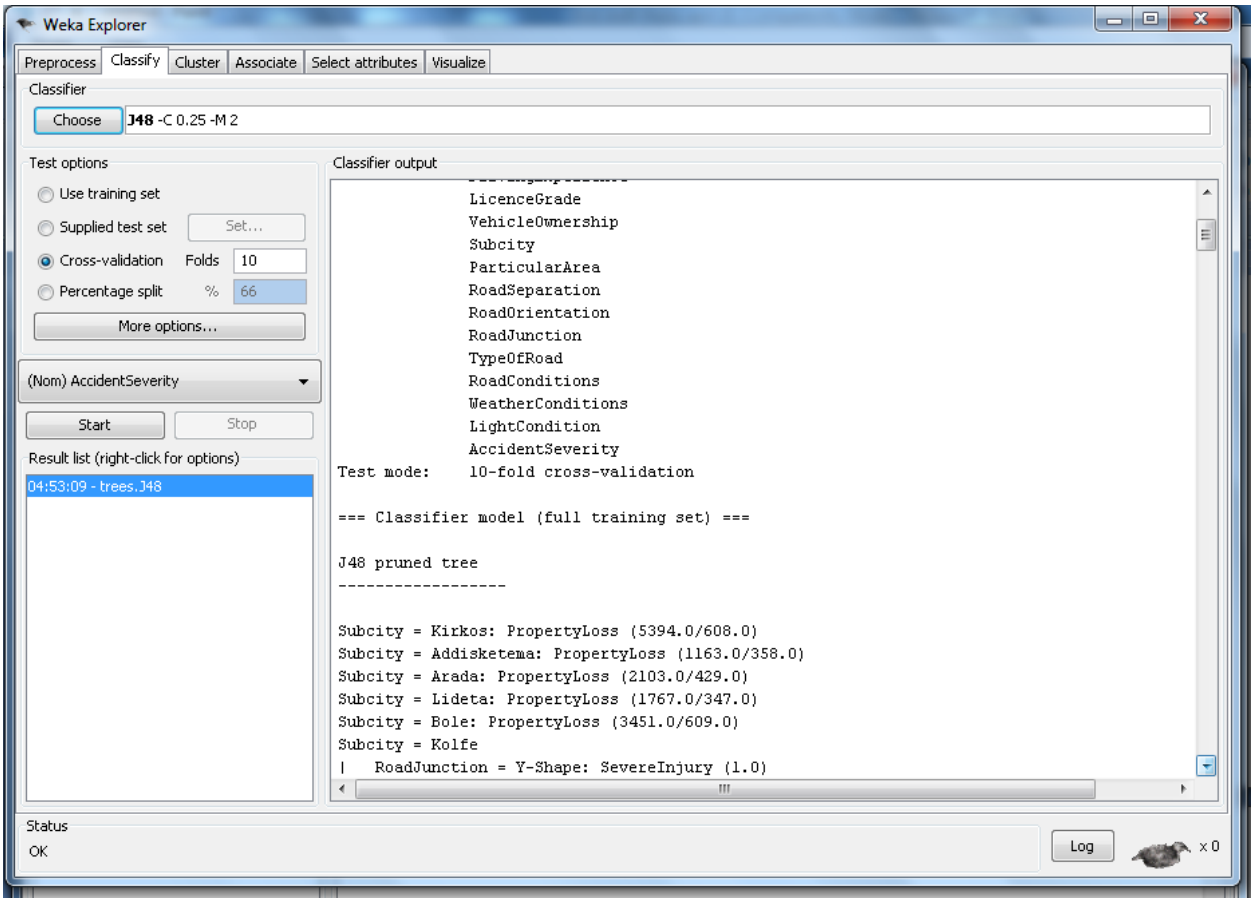


Figure 5.2 a screenshot of the J48 algorithm out put

As indicated in figure 5.3, out of the 16,710 records of the dataset 13,637 records are correctly classified and the model has an accuracy of 81.60%. The confusion matrix also shows that 13,046 out of 13,071 PropertyLoss, no SlightInjury are correctly classified, 424 out of 1404 SeverInjury and 167 out of 701 Fatal records are classified correctly.

In this experiment, the J48 algorithm used TypeofRoad and Driversex attributes in its pruned tree in few lines. This indicates that these attributes are considered as insignificant to discriminate records. Thus next experiment is done using 15 attributes including the dependent attribute to get interesting rules.

The decision tree generated in this learning scheme assured that as the size of the tree keeps increasing, it becomes difficult to analyze, interpret and generate rule sets. For this reason, the decision tree discovered in this training scheme has become complex to interpret. To address such difficulties an attempt is made to modify some of the parameters. The parameter minNumObj (minimum number of instances in a leaf) is set to 20, which are 2 in its default value and this figure is set after trying different values. The process of classifying of records proceeds as long as the number of records at each leaf node is reached 20.

```
=== Summary ===

Correctly Classified Instances          13637               81.6098%
Incorrectly Classified Instances         3073               18.3902 %
Kappa statistic                          0.2773
Mean absolute error                      0.1541
Root mean squared error                  0.28
Relative absolute error                 83.0564 %
Root relative squared error             91.9516 %
Total Number of Instances            16710

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 13046     1    16     8 |     a = PropertyLoss
  1520     0    12     2 |     b = SlightInjury
   900     2   424    78 |     c = SevereInjury
   381     2   151   167 |     d = Fatal
```

Figure 5.3: Statistical summary of experiment one

5.3. Experiment Two

This experiment is done using the 15 attributes, excluding Driversex and Typeofroad attributes. It is done using the default parameters values and 10 fold cross validation test option. The output shows that 13,632 records are correctly classified out of 16,710 with 81.57% accuracy which indicates that the DriverSex and TypeofRoad attributes have insignificance importance in building the model. The statistical summary is shown in Figure 5.4.

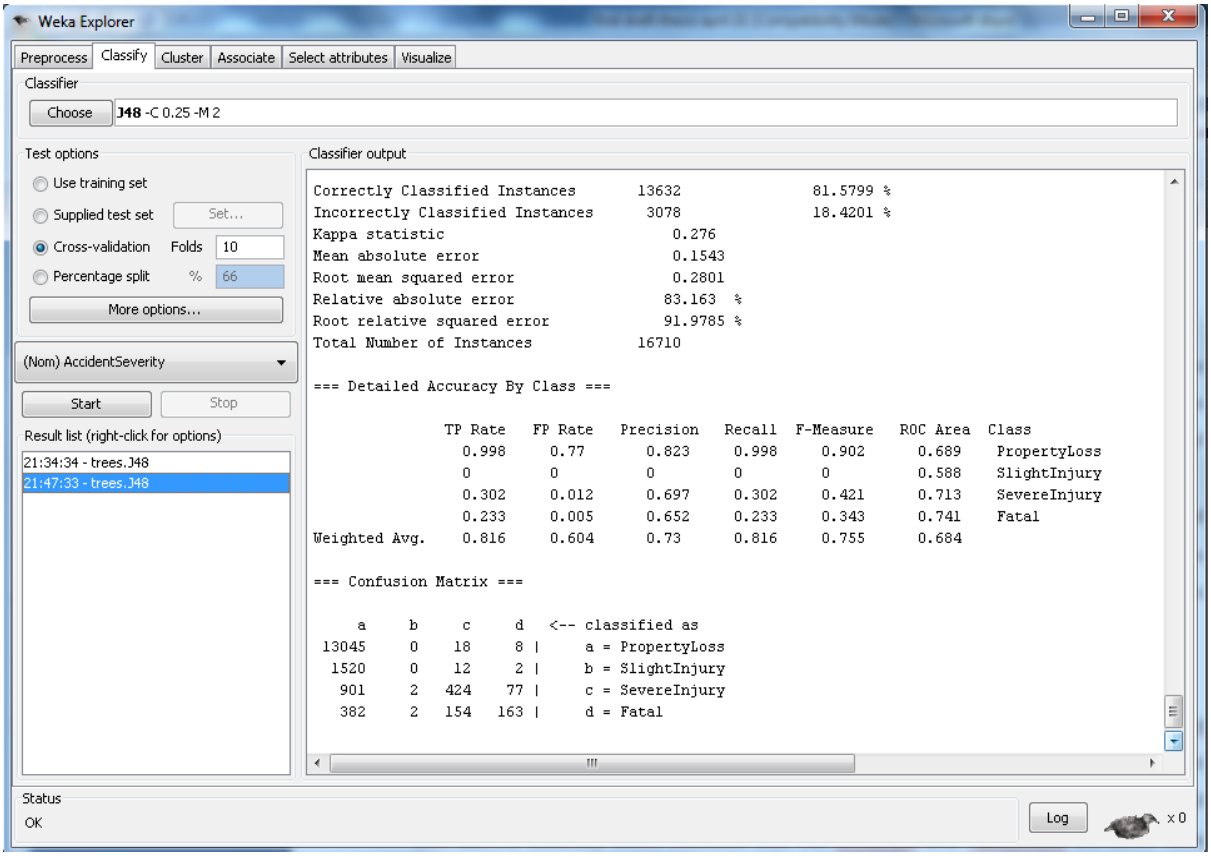


Figure 5.4 statistical summary of Experiment two

5.4. Experiment Three

In the process of building model, finding the best model measures like adjusting the values of the parameters are also taken. This experiment is carried out by varying the parameters of the ID3 and J48 algorithm, but changing the default parameters didn't result in a significant change to the performance of the model. Therefore the researcher decided to proceed the experiment by using the default parameters as the default parameters values work reasonably well in most case and which are strongly recommended by WEKA software. Besides this, changing the composition of input variables is done. That is, an attempt has been made to conduct the experiment by using different composition of variables by excluding and including some of the attributes to see if the accuracy of the learning scheme could be improved. Table 5.1, shows the result of the experiments using ID3 and J48 with 17, 15 and 14 different attribute subsets.

Table 5.1 performance of J48 and ID3 algorithm

S.No	Attribute subsets	Attribute removed	Algorithm	Accuracy (%)
1	17		J48	81.60
			ID3	83.02
2	15	Driversex and RoadType	J48	81.57
			ID3	82.84
3	14	RoadCondition	J48	81.50
			ID3	82.78

5.5. Experiment Four

In this scenario of the research, the researcher tried to build the model using 13 attributes, by removing the "Subcity" attribute, but the experiment resulted in generating no rules with a single leaf as shown in figure 5.5. It shows that this attribute is main attribute in generating interesting rules for classifying the accidents.

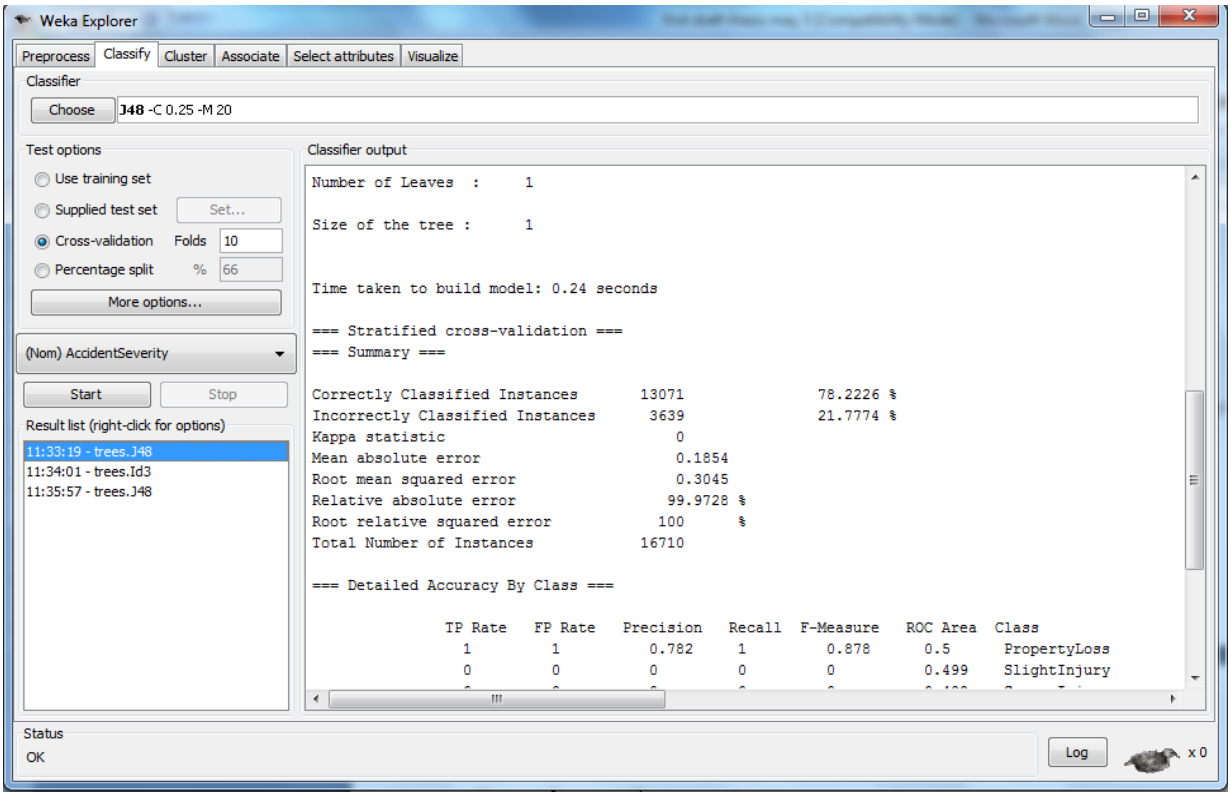


Figure 5.5 Output of Experiment 4 using J48 algorithm

5.6. Experiment Five

In this experiment, WEKA's attribute selection feature is used for performing the J48 and ID3 algorithms. As a result the following output is achieved, which rank attributes based on their information gain:

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

Search Method:
      Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 AccidentSeverity:(
      Information Gain Ranking Filter

Ranked attributes:
 8   0.16722 Subcity
 12  0.02162 RoadJunction
 6   0.01999 LicenceGrade
 10  0.01871RoadSeparation
 9   0.01863 ParticularArea
 5    0.0094 DrivingExperience
 3   0.00763 DriversEducationalBackground
 2   0.00669 DriverAge
 7   0.00594 VehicleOwnership
 4   0.00577 DriverVehicleRelationship
 16  0.00537 LightCondition
 1   0.00466 DriverSex
 11  0.00265 RoadOrientation
 15  0.00197 WeatherConditions
 13  0.00169 TypeOfRoad
 14  0.00115 RoadConditions

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

Based on the above ranking of attributes using the select attribute feature of WEKA, the fifth experiment is done by using six attributes including the class attribute, these are LiscenceGrade,subcity,RoadJunction,TypeofRoad,LightCondition and AccidnetSeverity. In this experiment the J48 algorithm has an accuracy of 81.15% and ID3 has 80.92% which shows a similar performance, however the ID3 algorithm has correctly classified 26 instances as SlightInjury out of 1530 where as the J48 algorithm classified none of the instances in the SlightInjury class. Another unique output from the J48 decision tree learner is the graphical output it produces for the tree it builds.

In decision tree the pruned tree has a hierarchy in that the most significant variable that used to discriminate the records is located at the top. The rules generated by J48 algorithm is presented in appendix I. Portion of the decision tree is depicted in figure 5.6

In decision tree, each rule is taken by reading the J48 pruned tree following the path from the root node to each leaves that contains the dependent class values. From the above experiments which are done using J48 and ID3 algorithm of decision tree technique, the LiscenceGrade attribute of drivers and subcity,RoadJunction,TypeofRoad, and LightCondition of the road are the most determinant factors to predict accident severity.

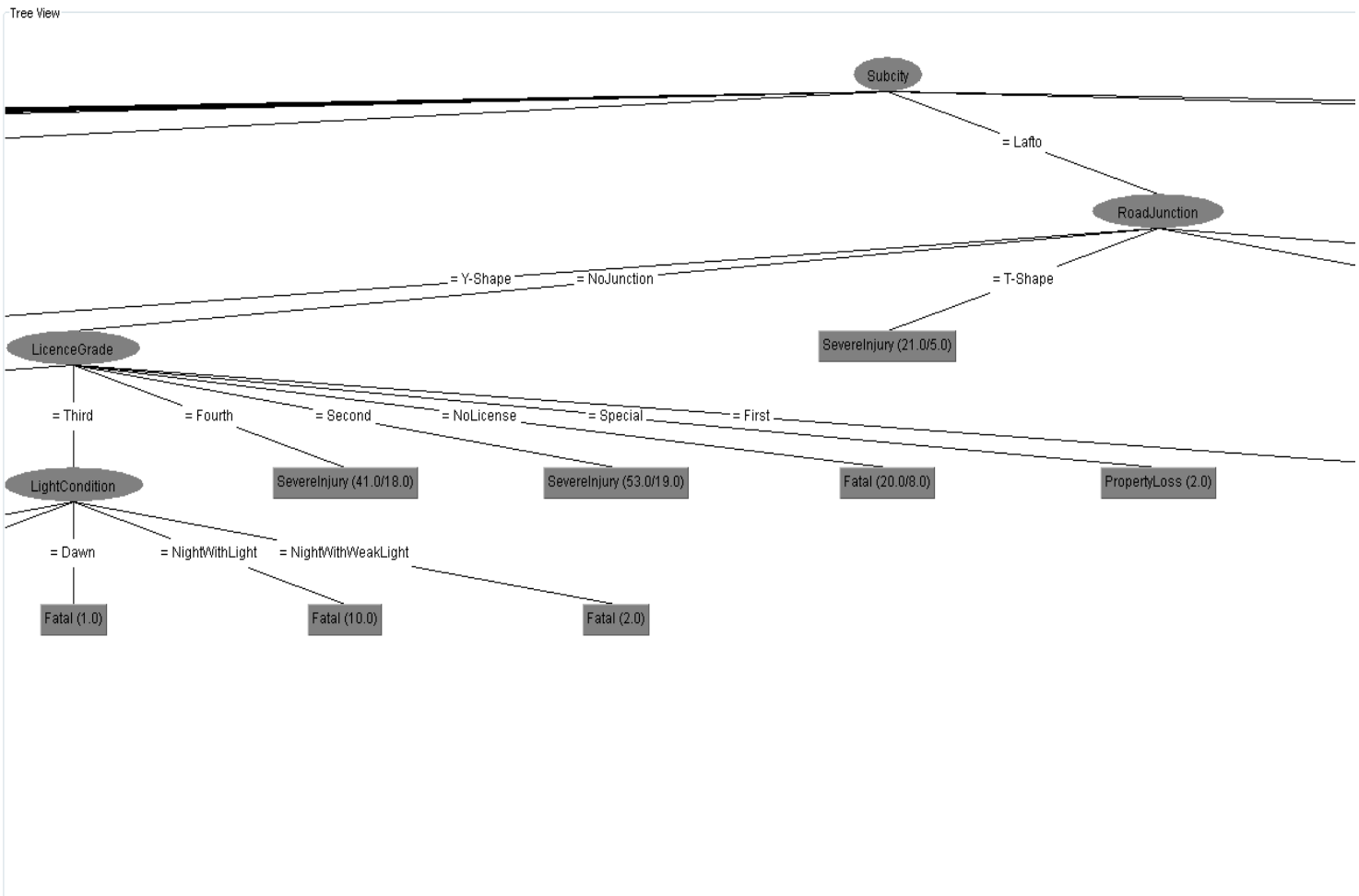


Figure 5.6 partial output of the J48 decision tree

5.7. Experiment Six

Model building using PART algorithm

In an attempt to come up with significant rules, PART is run on the accident dataset with different number of attributes. Ten fold cross validation has been used for testing.

Domain experts are consulted intensively in evaluating the significance of the rules. As a result the rules generated based on the six attributes: LiscenceGrade, subcity, RoadJunction, TypeofRoad, LightCondition and AccidnetSeverity are taken. The performance of the algorithm in generating the rules is 81.13% of accuracy. Figure 5.7 below, shows some of the one hundred rules generated by the PART algorithm. The rest of the rules can be found in Appendix II.

Using PART algorithm, the experiment showed that LiscenceGrade, Subcity and RoadJunction are the most important variables to classify records to their predefined class.

```
Subcity = Akaki AND
LicenceGrade = Fourth: Fatal (58.76/21.76)

Subcity = Akaki AND
RoadJunction = NoJunction AND
LightCondition = DayLight AND
LicenceGrade = Third: SevereInjury (48.44/14.14)

Subcity = Akaki AND
LicenceGrade = Fifth: SevereInjury (47.75/22.01)

Subcity = Akaki AND
LicenceGrade = Third: Fatal (25.46/6.46)

Subcity = Akaki AND
LightCondition = DayLight AND
LicenceGrade = Second: SevereInjury (19.1/2.01)

Subcity = Akaki AND
LightCondition = DayLight AND
LicenceGrade = NoLicense AND
TypeOfRoad = GoodAsphalt: SevereInjury (11.07/5.0)

Subcity = Akaki AND
LightCondition = DayLight AND
LicenceGrade = Special: SevereInjury (4.01/1.0)
```

Figure 5.7 partial output of the PART algorithm

5.8. Models Evaluation

Evaluation is one key point in any data mining process. It serves two purposes: the prediction of how well the final model will work in the future and an integral part of many learning methods, which help to find the model that best represents the training data.

In the series of experiments, evaluation of models is done based on performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rules generated. It is easy to learn that all the three classifiers are performing well and almost similarly with respect to the number of correctly classified instances.

But accuracy by itself doesn't tell everything about the efficiency of predictions, WEKA's experimenter has been utilized to implement ID3 and j48 decision tree learners and the PART algorithm for rule induction and to automatically analyze the models. The experiment type is a ten-fold cross-validation and model parameters have been set as follows.

Setting Modeling Parameters

In setting parameters for the J48 decision tree algorithm the default confidence factor of 0.25 is used, tree pruning is allowed; the number of folds is set to ten and the minimum number of instances to two. In the case of the ID3 algorithm there is no parameter to be set by the user. The parameters are customized by using WEKA's generic object editor window which can be referred at Appendix III for the J48 algorithm.

The three algorithms are executed at the same time, on the same data set, the same and equal number of attributes (6 attributes). Figure 5.8 below, shows the set up of WEKA's experiment environment window after the data is fed and the three algorithms ID3, J48 and PART have been selected.

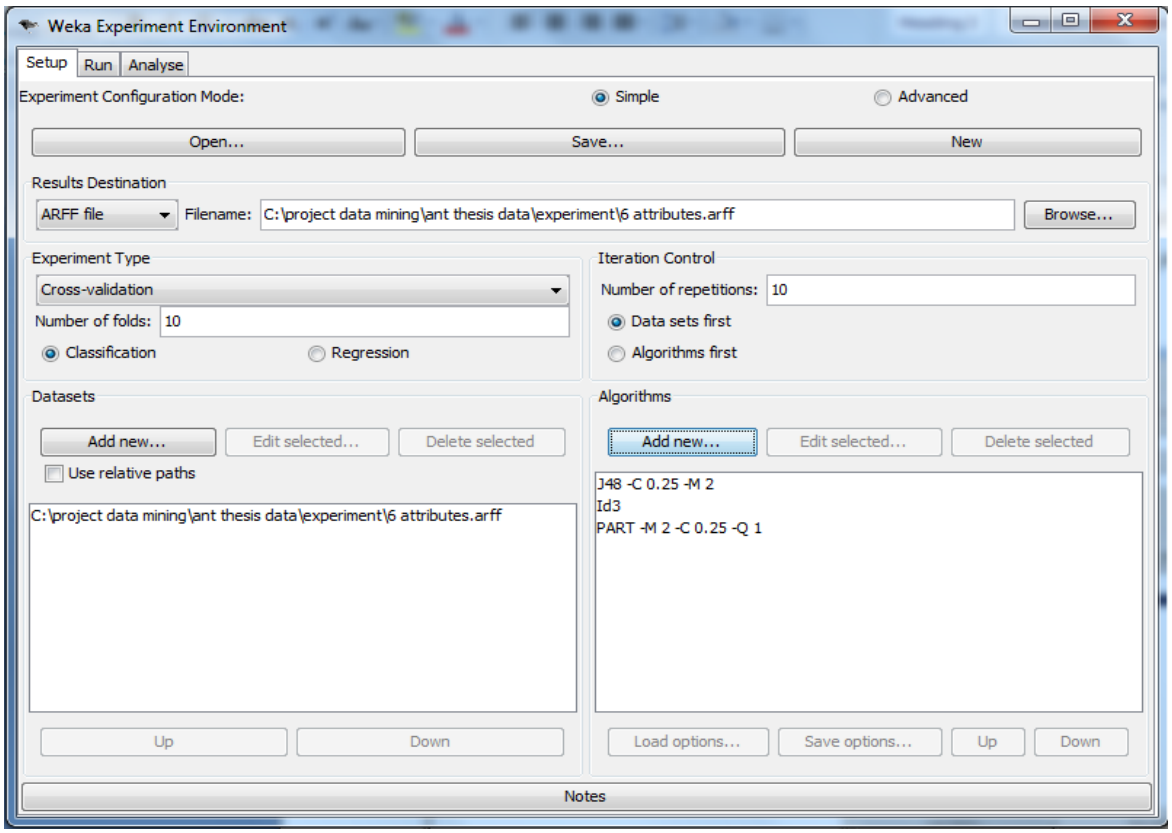


Figure 5.8 WEKA's Experiment Environment.

After finishing the necessary setup the algorithms are run to do the experiment. To analyze the experiment the percent-correct comparison measurement factor is selected from comparison field box as shown in figure 5.9, and then the three models have been analyzed by the experimenter automatically. The comparison result showed that the J48 algorithm slightly outperforms than ID3 and PART algorithms with an accuracy of 81.21%, 81.01% and 81.18% respectively.

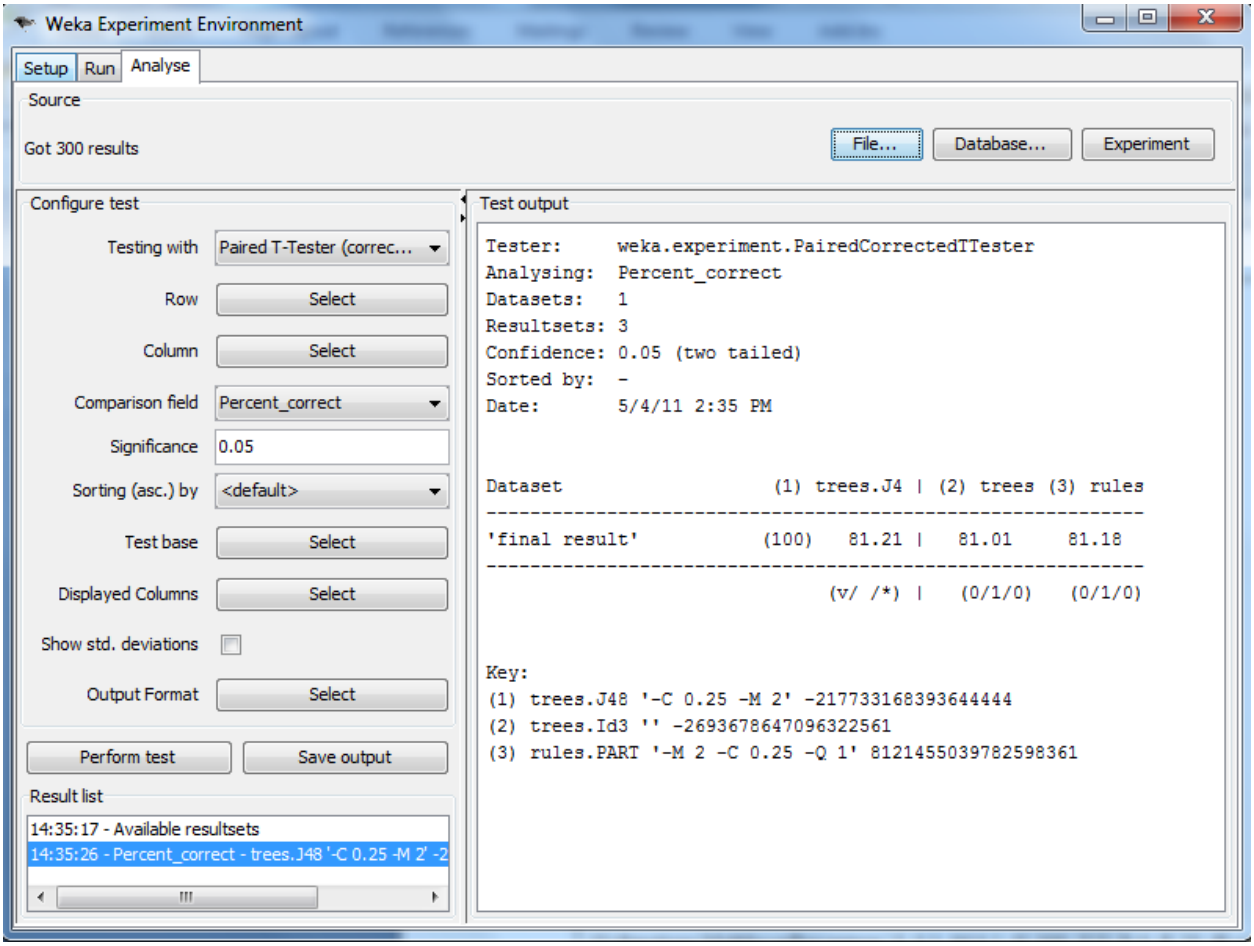


Figure 5.9 WEKA model comparison result

5.9. Discussion of Results

In the output of the J48 algorithm, at the beginning is a summary of the dataset, and the default 10 fold cross-validation which is used to evaluate it; then comes a pruned decision tree in textual form. The first split is on the "Subcity" attribute, and then, at the second level, the split is on "RoadJunction" and it goes on splitting on the most important attributes based on the measure of purity the algorithm uses until there is no more attribute left to split on or the data is completely classified. In the tree structure, a colon introduces the class label that has been assigned to a particular leaf, followed by the number of instances that reach that leaf, expressed as a decimal number because of the way the algorithm uses fractional instances to handle missing values. For example, an expression (1163.0/358.0) from the second leaf node in Appendix I imply a total of 1163 instances reached that leaf node out of which 358 are classified incorrectly. Beneath the tree structure the number of leaves is displayed; then the total number of nodes (Size of the tree).

The next part of the output gives estimates of the tree's predictive performance. In this case they are obtained using stratified cross validation with 10 folds. It can also be seen that out of the 16,710 instances, 13,632 (81.6%) are correctly classified and 3078(18.4%) are incorrectly classified in the cross-validation.

In addition to the classification error, the evaluation module also outputs the Kappa statistic, the mean absolute error, and the root mean-squared error of the class probability estimates assigned by the tree. The root mean squared error is the square root of the average quadratic loss. The mean absolute error is calculated in a similar way using the absolute instead of the squared difference. Finally, the confusion matrix at the bottom of the output shows that 13,045 out of 13,081 PropertyLoss, no SlightInjury are correctly classified, 424 out of 1404 SeverInjury and 163 out of 701 Fatal records are classified correctly.

Even though ID3 algorithm is slightly better than J48 and PART algorithms, some interesting rules from PART decision list and data visualization method are also considered to discuss the research result.

Rules from PART Algorithm

The classifier used some attributes to construct rules and provide the class predicted by the model. The numeric values which appeared in bracket next to the class label indicates the number of correctly and incorrectly classified records respectively.

For example Rule 2 can be interpreted as driver who has third level driving license and the accident happened in Akaki subcity and the road has no junction and day light condition which is classified as severe injury (48.44/14.14). This implies that, there are 48 records in the dataset that exactly satisfy this rule and 14 records are misclassified to this rule.

As the research indicates, Lafto, Akaki and Gulele are subcities with high number of sever accidents and fatal injuries. So the office has to improve the roads and perhaps assign more traffic police officers on these subcities by collaborating with Addis Ababa road authority.

In addition it is observed that those who don't have driving license create sever and fatal accident, as it is seen in rule 17 and 90. Therefore, rules and regulations should be strict on those who drive without driving license. The data visualization method also showed that drivers who

have second and third level driving license create most of the accident. Moreover, most of the accidents are caused by age group ranging from 18-30 and 31-50.

The data visualization showed that 95.81% (16,010) and 86.74% (14,495) of the accidents happened in good weather and day light conditions respectively, which indicates that weather and light conditions are not main factors for traffic accidents.

The following top ten rules are generated using the PART algorithm. The rules also showed that attributes such as LiscenceGrade, Subcity, RoadJunction, TypeofRoad, and LightCondition are found to be important attributes to classify accident types. This will help the traffic control and investigation department to focus their attention on these factors during the revision and construction of rules and policies.

1. Subcity = Akaki AND LicenceGrade = Fourth: Fatal (58.76/21.76)
2. Subcity = Akaki AND RoadJunction = NoJunction AND LightCondition = DayLight AND LicenceGrade = Third: SevereInjury (48.44/14.14)
3. Subcity = Akaki AND LicenceGrade = Fifth: SevereInjury (47.75/22.01)
4. Subcity = Gulele AND LightCondition = DayLight AND RoadJunction = NoJunction: SevereInjury (79.11/20.11)
5. Subcity = Gulele AND RoadJunction = CrossRoad: SevereInjury (5.0/1.0)
6. Subcity = Gulele AND LicenceGrade = Third: SevereInjury (4.01/1.01)
7. Subcity = Lafto AND LicenceGrade = Special AND LightCondition = NightWithLight: SevereInjury (2.05/0.05)
8. Subcity = Lafto AND LicenceGrade = Third AND LightCondition = DayLight: SevereInjury (73.88/19.18)
9. Subcity = Lafto AND LicenceGrade = Second: SevereInjury (56.51/20.09)
10. Subcity = Lafto AND
    LightCondition = NightWithLight AND
    LicenceGrade = Third: Fatal (7.96)

Chapter six

Conclusions and Recommendations

6.1. Conclusions

Managers do not have time to go through all records and data collections in their organizations. Moreover, the amount of data is bulky having several variables; it is extremely difficult to visualize patterns and relationships. They need filtered and simplified data from their large amount of records. Knowledge discovery systems could help them to pass correct decisions on their daily activity or improve their future plan. One tool is data mining technology which finds out hidden pattern from vast amount of data.

In this research an attempt has been made to apply the decision tree and Rule induction predictive data mining techniques in driver and road factors for car accidents and identify hidden patterns in the accident data set. To achieve this goal: the CRISP-DM 1.0 standard data mining methodology has been adopted and the WEKA data mining tool has been used to implement the ID3 , J48 and PART algorithms.

The data for this research is the RTA data of the years 2005-09 collected from the Addis Ababa Road Traffic Control and Investigation Department. After preprocessing out of 18419 records, 16,710 RTA records are remained and used for building the models.

Various experiments are made iteratively by making adjustment of the parameters and using different number of attributes to come up with a meaningful output. Major factors of drivers and roads are identified and rules are generated using J48 decision trees and rule induction (PART algorithm). The comparison of the models using WEKA's experimenter showed that J48 slightly outperforms ID3 and PART algorithms with an accuracy of 81.21%, 81.01% and 81.18% respectively.

In addition, the determinant factors of drivers and roads that cause road accidents are identified; these are LiscenceGrade, subcity, RoadJunction, TypeofRoad, and LightCondition.

In many data mining researches on traffic accidents, decision trees and neural networks are widely used but in this study rule induction and decision trees are used to build the different models that is believed to improve public health problem of the society.

In general, encouraging results are obtained by employing both decision tree and rule induction technique, and the rules generated by J48 and PART algorithm are easily understandable by subject experts in the department. Thus, the results obtained in this research have proved the applicability of data mining in road traffic accident preventing and controlling activities. More specifically it will provide valuable help in developing new methods to increase road safety, particularly in the phase of choosing the appropriate means and budget allocation of resources.

However, the research has also its own weakness, like accuracy of the models are not enough for efficient implementation of the models. To reduce this misclassification error, the office has to re-engineer the data collection system, train the data entry personnel, and fully automate the data recording system in collaboration with other organizations like WHO. This will enable the office to have a quality data which in turn increases the accuracy of the models to be built.

6.2. Recommendations

This research work is conducted mainly for academic purpose. However, the researcher highly believes that the findings of this research can be used by AATCID to further investigate the nature of traffic accident problem in Addis Ababa. The researcher makes the following recommendations based on the result of this study.

It is crystal clear that for an efficient data-mining task there is a need for an availability of electronic data. Although there is a good start, AATCID should take measures to store all its records with all the necessary attributes in an electronic format and to make all decisions based on collected records.

Data mining techniques could contribute a lot in identifying the most important factors of drivers and roads that could cause traffic accidents. Thus, it could be more important for the office to use the data mining technique as a tool for the decision making process. In other words, the AATCID could optimize its traffic accident prevention and control efforts by employing data mining technology.

Although both decision tree and rule induction reported promising results and hence could be applied in the area of traffic accident prevention, decision trees tend to perform better. Hence, it would be more optimal for the AATCID to employ the model developed with this technique.

This study focused mainly on methods that exploited pre-specified dependent variable and tested several models that work with these variables. Predictive models that do not require apriori output variable, such as association rules, should be further investigated.

To increase the classification accuracy of RTA types, data quality has to be improved. Re-engineering of a data collection system is left as further research topic to resolve many problems created in the gathering of data. Another future work is to test the applicability of other data mining techniques on other regions of the country.

References

1. Adnan A. Hyder, 2008 Road safety is no accident: a call for global action, Libyan journal of medicine, Issue 3, Vol.3, Pub Med publisher.
2. Bouckaert Remco, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, and Alex Seewald, 2008.WEKA Manual for Version 3-6-0. University of Waikato, New Zealand.
3. Bramer, M. 2007. Principles of Data Mining, Springer publisher, London, United Kingdom.
4. Chen, W. and Jovanis P., 2002, "Method for identifying factors contributing to driver-injury severity in traffic crashes", Transportation Research Record, unpublished research
5. Chong, M., Abraham A., and Paprzycki M., 2002. Traffic Accident Analysis Using Decision Trees and Neural Networks. Available at URL: http://falklands.globat.com/~softcomputing.net/informatica1.pdf (Accessed on February 20, 2011)
6. Ethiopian Road Authority; 2005 How safe are Ethiopian roads? Unpublished road safety report, Addis Ababa, Ethiopia
7. Friedman, J. H., 1997, "Data Mining and Statistics: What's the Connection?" Proceedings of the 29th Symposium on the Interface between Computer Science and Statistics, Texas.
8. Getnet M. (2009). Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling road traffic accidents: the case of Addis Ababa city, Msc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia
9. Getu Segni (2007). Causes of road traffic accidents and possible counter measures on Addis Ababa-shashemene roads, Msc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.
10. Han, Jiawei and Kamber, Micheline. (2006). Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.
11. Hand, D.J., Mannila, H., and Smyth, P. (2001). Principles of Data Mining. MIT Press.
12. Mehmed Kantardzic(2003).Data Mining: Concepts, Models, Methods, and Algorithms, ISBN13: 9780471228523, John Wiley & Sons Publisher
13. National Road Safety Coordination Office (2008). Unpublished Road Safety Report of years 2004-08. Addis Ababa, Ethiopia

14. National Road Safety Coordination Office, November 2006, Overview of the Road Safety Activities in Ethiopia, Federal Transport Authority, Addis Ababa, Ethiopia.
15. SafeCarGuide, 2004, International Injury & Fatality Statistics. Available at: www.safecarguide.com (accessed on February, 2011).
16. Shearer C., (2000), the CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing
17. The CRISP-DM consortium (August, 2000). Step-by-step data mining guide available at: URL: http://www.crisp-dm.org/CRISPWP-0800.pdf (Accessed on April 28, 2011).
18. Tibebe Beshah (2005). Application of data mining technology to support RTA severity analysis at Addis Ababa traffic office. M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.
19. Tibebe Beshah and Shawndra Hill (2010). Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia. Available at: www. ai-d.org/pdfs/Beshah.pdf (Accessed on: November 13, 2010).
20. Two Crows Corporation (2005), Introduction to Data Mining and Knowledge Discovery, 3rd Edition ISBN: 1-892095-02-5
21. Whitten I.H and Frank E. (2005). Data Mining: practical machine learning tools and techniques with java implementations. Morgan Kaufmann publishers. San Francisco.
22. WHO (2004). World report on road traffic injury prevention, Switzerland, Geneva.
23. Wikipedia the free encyclopedia (2010). Road Traffic Safety. Available at URL: http://en.wikipedia.org/wiki/Traffic_collision (Accessed on January 10, 2011).
24. Zelalem Regassa (2009). Determining the degree of driver's responsibility for car accident: the case of Addis Ababa traffic office. M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Appendices

Appendix I - Output of J48 Algorithm

=== Run information ===

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

Relation: exp 2 unk replaced with blank-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval-Sweka.attributeSelection.BestFirst -D 1 -N 5

Instances: 16710

Attributes: 6

LicenceGrade

Subcity

RoadJunction

TypeOfRoad

LightCondition

AccidentSeverity

Test mode: 10-fold cross-validation

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

J48 pruned tree

Subcity = Kirkos: PropertyLoss (5394.0/608.0)

Subcity = Addisketema: PropertyLoss (1163.0/358.0)

Subcity = Arada: PropertyLoss (2103.0/429.0)

Subcity = Lideta: PropertyLoss (1767.0/347.0)

Subcity = Bole: PropertyLoss (3451.0/609.0)

Subcity = Kolfe

| RoadJunction = Y-Shape: SevereInjury (1.0)

| RoadJunction = NoJunction: SevereInjury (248.0/107.0)

| RoadJunction = T-Shape: SevereInjury (32.0/17.0)

| RoadJunction = CrossRoad: PropertyLoss (10.0/1.0)

| RoadJunction = Roundabout

| | LicenceGrade = Fifth: Fatal (0.0)

| | LicenceGrade = Third: Fatal (0.0)

| | LicenceGrade = Fourth: Fatal (4.0)
| | LicenceGrade = Second
| | | LightCondition = DayLight: SevereInjury (2.0)
| | | LightCondition = Dusk: SevereInjury (0.0)
| | | LightCondition = Dawn: SevereInjury (0.0)
| | | LightCondition = NightWithLight: Fatal (2.0)
| | | LightCondition = NightWithWeakLight: SevereInjury (0.0)
| | LicenceGrade = NoLicense: Fatal (0.0)
| | LicenceGrade = Special: Fatal (0.0)
| | LicenceGrade = First: Fatal (0.0)
Subcity = Yeka: PropertyLoss (1920.0/429.0)
Subcity = Lafto
| RoadJunction = Y-Shape: SevereInjury (0.0)
| RoadJunction = NoJunction
| | LicenceGrade = Fifth
| | | LightCondition = DayLight: Fatal (27.0/11.0)
| | | LightCondition = Dusk: SevereInjury (3.0)
| | | LightCondition = Dawn: Fatal (0.0)
| | | LightCondition = NightWithLight: Fatal (0.0)
| | | LightCondition = NightWithWeakLight: Fatal (0.0)
| | LicenceGrade = Third
| | | LightCondition = DayLight: SevereInjury (67.0/15.0)
| | | LightCondition = Dusk: SevereInjury (2.0)
| | | LightCondition = Dawn: Fatal (1.0)
| | | LightCondition = NightWithLight: Fatal (10.0)
| | | LightCondition = NightWithWeakLight: Fatal (2.0)
| | LicenceGrade = Fourth: SevereInjury (41.0/18.0)
| | LicenceGrade = Second: SevereInjury (53.0/19.0)
| | LicenceGrade = NoLicense: Fatal (20.0/8.0)
| | LicenceGrade = Special: PropertyLoss (2.0)
| | LicenceGrade = First: SevereInjury (3.0/1.0)
| RoadJunction = T-Shape: SevereInjury (21.0/5.0)
| RoadJunction = CrossRoad: SevereInjury (10.0/2.0)
| RoadJunction = Roundabout
| | LicenceGrade = Fifth: SevereInjury (2.0)
| | LicenceGrade = Third: Fatal (2.0)

| | LicenceGrade = Fourth: Fatal (2.0)
| | LicenceGrade = Second: SevereInjury (2.0)
| | LicenceGrade = NoLicense: SevereInjury (1.0)
| | LicenceGrade = Special: SevereInjury (0.0)
| | LicenceGrade = First: SevereInjury (0.0)
Subcity = Akaki
| LicenceGrade = Fifth: SevereInjury (47.0/22.0)
| LicenceGrade = Third
| | LightCondition = DayLight: SevereInjury (50.0/15.0)
| | LightCondition = Dusk: SevereInjury (0.0)
| | LightCondition = Dawn: Fatal (2.0)
| | LightCondition = NightWithLight: Fatal (24.0/8.0)
| | LightCondition = NightWithWeakLight: SevereInjury (0.0)
| LicenceGrade = Fourth: Fatal (58.0/21.0)
| LicenceGrade = Second: SevereInjury (25.0/5.0)
| LicenceGrade = NoLicense: Fatal (19.0/9.0)
| LicenceGrade = Special: SevereInjury (4.0/1.0)
| LicenceGrade = First: SevereInjury (3.0)
Subcity = Gulele
| LightCondition = DayLight: SevereInjury (90.0/24.0)
| LightCondition = Dusk: SevereInjury (0.0)
| LightCondition = Dawn: Fatal (3.0)
| LightCondition = NightWithLight: Fatal (17.0/7.0)
| LightCondition = NightWithWeakLight: SevereInjury (0.0)

Number of Leaves : 62
Size of the tree : 74
Time taken to build model: 0.1 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	13561	81.155 %
Incorrectly Classified Instances	3149	18.845 %
Kappa statistic	0.262	
Mean absolute error	0.157	
Root mean squared error	0.2808	

Relative absolute error 84.6235 %

Root relative squared error 92.2086 %

Total Number of Instances 16710

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.997	0.767	0.824	0.997	0.902	0.689	PropertyLoss
0	0	0	0	0	0.587	SlightInjury
0.31	0.018	0.606	0.31	0.41	0.721	SevereInjury
0.138	0.005	0.567	0.138	0.222	0.763	Fatal
Weighted Avg.	0.812	0.602	0.719	0.812	0.749	0.685

=== Confusion Matrix ===

a	b	c	d	<-- classified as
13029	0	41	1	a = PropertyLoss
1519	0	15	0	b = SlightInjury
896	0	435	73	c = SevereInjury
377	0	227	97	d = Fatal

Appendix II- Output of PART Algorithm

=== Run information ===

Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1

Relation: exp 2 unk replaced with blank-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval-Sweka.attributeSelection.BestFirst -D 1 -N 5

Instances: 16710

Attributes: 6

LicenceGrade

Subcity

RoadJunction

TypeOfRoad

LightCondition

AccidentSeverity

Test mode: 10-fold cross-validation

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

PART decision list

Subcity = Kirkos AND

RoadJunction = T-Shape: PropertyLoss (1290.32/113.0)

Subcity = Kirkos AND

RoadJunction = CrossRoad AND

LightCondition = DayLight: PropertyLoss (491.0/40.0)

Subcity = Kirkos AND

RoadJunction = Roundabout: PropertyLoss (171.0/3.0)

Subcity = Kirkos: PropertyLoss (3424.26/451.23)

Subcity = Bole AND

RoadJunction = Roundabout AND

LicenceGrade = Third: PropertyLoss (85.3/2.0)

Subcity = Bole AND

RoadJunction = Roundabout AND

LicenceGrade = Second: PropertyLoss (37.45/2.0)

Subcity = Bole: PropertyLoss (3333.43/605.23)

Subcity = Lideta AND

RoadJunction = NoJunction AND

LightCondition = DayLight: PropertyLoss (1099.81/235.0)

Subcity = Lideta: PropertyLoss (670.15/112.05)

Subcity = Arada AND

LicenceGrade = Fifth AND

TypeOfRoad = GoodAsphalt: PropertyLoss (243.39/31.1)

Subcity = Akaki AND

LicenceGrade = Fourth: Fatal (58.76/21.76)

Subcity = Akaki AND

RoadJunction = NoJunction AND

LightCondition = DayLight AND

LicenceGrade = Third: SevereInjury (48.44/14.14)

Subcity = Akaki AND

LicenceGrade = Fifth: SevereInjury (47.75/22.01)

Subcity = Akaki AND

LicenceGrade = Third: Fatal (25.46/6.46)

Subcity = Akaki AND

LightCondition = DayLight AND

LicenceGrade = Second: SevereInjury (19.1/2.01)

Subcity = Akaki AND

LightCondition = DayLight AND

LicenceGrade = NoLicense AND

TypeOfRoad = GoodAsphalt: SevereInjury (11.07/5.0)

Subcity = Akaki AND

LightCondition = DayLight AND

LicenceGrade = Special: SevereInjury (4.01/1.0)

Subcity = Akaki AND

LightCondition = NightWithLight: SevereInjury (8.62/4.0)

Subcity = Akaki AND

LicenceGrade = First: SevereInjury (3.01)

Subcity = Akaki AND

LightCondition = DayLight: Fatal (2.02/0.02)

Subcity = Akaki AND

LightCondition = Dusk: Fatal (2.0/0.0)

Subcity = Gulele AND

LightCondition = DayLight AND

RoadJunction = NoJunction: SevereInjury (79.11/20.11)

Subcity = Arada AND

LicenceGrade = Fourth AND

RoadJunction = NoJunction: PropertyLoss (206.97/37.39)

Subcity = Arada AND

LicenceGrade = Fourth AND

RoadJunction = T-Shape AND

TypeOfRoad = GoodAsphalt: PropertyLoss (54.46/8.12)

Subcity = Arada AND

LicenceGrade = Third: PropertyLoss (965.35/191.08)

Subcity = Yeka AND

RoadJunction = T-Shape: PropertyLoss (244.1/38.0)

Subcity = Yeka AND

RoadJunction = Roundabout: PropertyLoss (106.0/6.0)

Subcity = Yeka AND

LightCondition = DayLight AND

LicenceGrade = Third: PropertyLoss (561.09/134.51)

Subcity = Arada AND

LicenceGrade = Second AND

LightCondition = DayLight: PropertyLoss (418.47/86.18)

Subcity = Yeka AND

LightCondition = DayLight AND

LicenceGrade = Second: PropertyLoss (282.71/62.78)

Subcity = Gulele AND

RoadJunction = NoJunction: Fatal (20.05/7.05)

Subcity = Yeka AND

LicenceGrade = Second: PropertyLoss (52.11/8.39)

Subcity = Yeka AND

LicenceGrade = Special AND

LightCondition = DayLight: PropertyLoss (18.54/2.19)

Subcity = Lafto AND

RoadJunction = T-Shape: SevereInjury (21.01/5.01)

Subcity = Yeka AND

LicenceGrade = Fourth: PropertyLoss (316.06/71.79)

Subcity = Lafto AND

LicenceGrade = NoLicense: Fatal (21.51/9.25)

Subcity = Lafto AND

LicenceGrade = First: SevereInjury (3.06/1.04)

Subcity = Lafto AND

LicenceGrade = Special AND

LightCondition = NightWithLight: SevereInjury (2.05/0.05)

Subcity = Lafto AND

LicenceGrade = Third AND

LightCondition = DayLight: SevereInjury (73.88/19.18)

Subcity = Gulele AND

RoadJunction = CrossRoad: SevereInjury (5.0/1.0)

Subcity = Gulele AND

LicenceGrade = Third: SevereInjury (4.01/1.01)

Subcity = Lafto AND

LicenceGrade = Second: SevereInjury (56.51/20.09)

Subcity = Lafto AND

LightCondition = NightWithLight AND

LicenceGrade = Third: Fatal (7.96)

Subcity = Yeka AND

LicenceGrade = Fifth: PropertyLoss (193.97/46.79)

Subcity = Lafto AND

LightCondition = DayLight AND

LicenceGrade = Fourth AND

RoadJunction = NoJunction: SevereInjury (35.37/16.0)

Subcity = Arada AND

LicenceGrade = Second AND

LightCondition = NightWithLight: PropertyLoss (74.11/15.04)

Subcity = Addisketema AND

RoadJunction = T-Shape AND

LicenceGrade = Third: PropertyLoss (88.47/17.86)

Subcity = Lafto AND

LightCondition = DayLight AND

LicenceGrade = Fifth AND

RoadJunction = NoJunction: Fatal (27.28/11.28)

Subcity = Addisketema AND

RoadJunction = CrossRoad: PropertyLoss (70.0/4.0)

Subcity = Lafto AND

LicenceGrade = Fifth: SevereInjury (7.24/0.23)

Subcity = Addisketema AND

LightCondition = DayLight AND

RoadJunction = T-Shape: PropertyLoss (106.44/22.0)

Subcity = Addisketema AND

RoadJunction = Roundabout: PropertyLoss (20.0)

Subcity = Addisketema AND

LightCondition = Dusk: PropertyLoss (8.19/2.0)

Subcity = Addisketema AND

LightCondition = Dawn: PropertyLoss (3.0)

Subcity = Arada AND

LicenceGrade = Fourth: PropertyLoss (53.25/8.42)

Subcity = Addisketema AND

LightCondition = DayLight AND

LicenceGrade = Special AND

TypeOfRoad = GoodAsphalt: PropertyLoss (2.19/0.07)

Subcity = Addisketema AND

LicenceGrade = Fifth: PropertyLoss (134.11/38.79)

Subcity = Addisketema AND

LightCondition = DayLight AND

LicenceGrade = Second AND

RoadJunction = NoJunction: PropertyLoss (113.44/35.86)

Subcity = Yeka AND

LicenceGrade = Third: PropertyLoss (86.63/18.14)

Subcity = Addisketema AND

LightCondition = DayLight AND

LicenceGrade = Third: PropertyLoss (341.39/117.48)

Subcity = Lafto AND

LightCondition = Dawn: SevereInjury (3.0/1.0)

Subcity = Kolfe AND

RoadJunction = CrossRoad: PropertyLoss (10.0/1.0)

Subcity = Kolfe AND

RoadJunction = Roundabout AND

LicenceGrade = Fourth: Fatal (4.0)

Subcity = Akaki: SevereInjury (2.01/0.01)

Subcity = Gulele: Fatal (2.01/0.01)

Subcity = Kolfe AND

RoadJunction = Roundabout AND

LightCondition = DayLight: SevereInjury (2.0)

Subcity = Kolfe AND

RoadJunction = NoJunction AND

LicenceGrade = NoLicense AND

LightCondition = DayLight: SevereInjury (9.66/2.25)

Subcity = Kolfe AND

RoadJunction = NoJunction AND

LicenceGrade = Third: SevereInjury (107.69/39.17)

Subcity = Lafto AND

LightCondition = DayLight AND

RoadJunction = NoJunction: PropertyLoss (2.04/0.03)

Subcity = Lafto AND

LightCondition = NightWithLight: PropertyLoss (4.49/2.48)

Subcity = Lafto AND

LightCondition = DayLight: Fatal (2.01/0.01)

Subcity = Lafto AND

LightCondition = Dusk: SevereInjury (2.0/0.0)

Subcity = Addisketema AND

LicenceGrade = Fourth: PropertyLoss (206.04/75.88)

Subcity = Arada AND

LightCondition = DayLight: SlightInjury (60.34/31.71)

Subcity = Kolfe AND

RoadJunction = T-Shape AND

LicenceGrade = Third AND

LightCondition = DayLight: SevereInjury (20.0/11.0)

Subcity = Kolfe AND

RoadJunction = T-Shape AND

LicenceGrade = Fourth: SevereInjury (6.0/2.0)

Subcity = Kolfe AND

LightCondition = DayLight AND

LicenceGrade = Fourth: SevereInjury (41.15/19.0)

Subcity = Kolfe AND

LightCondition = DayLight AND

LicenceGrade = Second AND

RoadJunction = NoJunction: SevereInjury (38.15/19.03)

Subcity = Kolfe AND

LightCondition = DayLight: SevereInjury (31.02/12.21)

Subcity = Kolfe AND

LicenceGrade = NoLicense: Fatal (7.3/2.03)

LicenceGrade = First AND

Subcity = Yeka: SevereInjury (10.24/2.22)

LicenceGrade = Third AND

LightCondition = NightWithLight AND

RoadJunction = NoJunction: SlightInjury (23.5/11.44)

LicenceGrade = Special: SlightInjury (5.25/1.21)

LicenceGrade = Third AND

LightCondition = NightWithLight: PropertyLoss (2.08/0.06)

LicenceGrade = Fourth: Fatal (12.61/5.49)

RoadJunction = T-Shape AND

Subcity = Arada: SevereInjury (5.0/2.0)

RoadJunction = CrossRoad: PropertyLoss (4.05/1.0)

Subcity = Kolfe AND

LightCondition = NightWithLight: Fatal (5.65/1.03)

LicenceGrade = First: PropertyLoss (4.21/2.1)

LightCondition = NightWithLight AND

Subcity = Addisketema AND

RoadJunction = NoJunction: PropertyLoss (6.14/2.68)

LicenceGrade = Third: Fatal (2.38/0.37)

TypeOfRoad = TornAsphalt: SevereInjury (4.0)

LicenceGrade = Second AND

LightCondition = DayLight: SevereInjury (2.33/0.3)

LightCondition = DayLight: PropertyLoss (65.36/36.1)

Subcity = Arada AND

LightCondition = NightWithLight: PropertyLoss (18.38/11.13)

Subcity = Addisketema: SlightInjury (3.24/1.24)

LicenceGrade = Second AND

LightCondition = Dusk AND

Subcity = Arada: PropertyLoss (2.0)

LicenceGrade = Second: SevereInjury (4.58/0.52)

LightCondition = NightWithLight: SevereInjury (4.38/1.38)

: Fatal (2.24/1.16)

Number of Rules : 100

Time taken to build model: 0.28 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	13570	81.2089 %
Incorrectly Classified Instances	3140	18.7911 %
Kappa statistic	0.2734	
Mean absolute error	0.155	
Root mean squared error	0.2797	
Relative absolute error	83.5534 %	
Root relative squared error	91.8557 %	
Total Number of Instances	16710	

=== Detailed Accuracy By Class ===

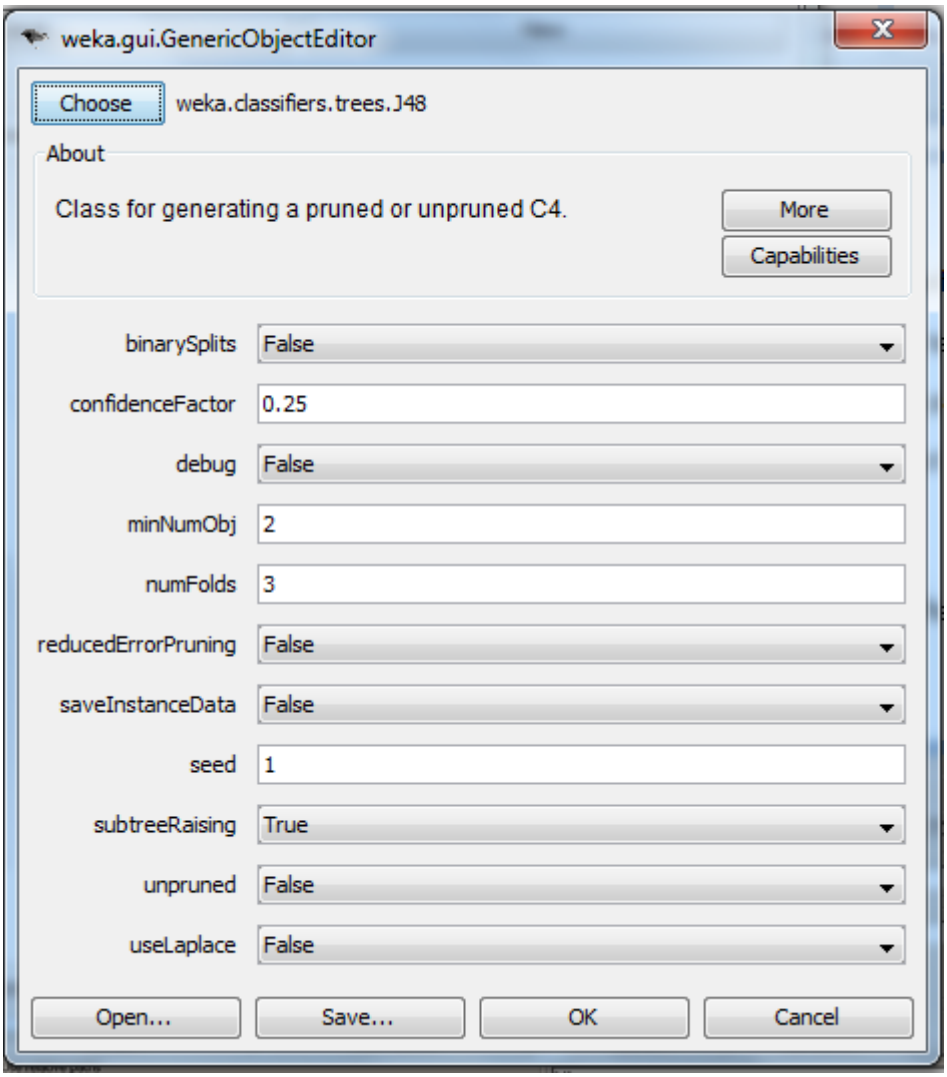
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.994	0.756	0.825	0.994	0.902	0.709	PropertyLoss
0.016	0.002	0.41	0.016	0.031	0.616	SlightInjury

0.311	0.017	0.621	0.311	0.414	0.745	SevereInjury
0.164	0.005	0.567	0.164	0.254	0.776	Fatal
Weighted Avg.	0.812	0.593	0.759	0.812	0.754	0.706

=== Confusion Matrix ===

a	b	c	d	<-- classified as
12994	30	41	6	a = PropertyLoss
1491	25	16	2	b = SlightInjury
884	4	436	80	c = SevereInjury
375	2	209	115	d = Fatal

Appendix III – J48 Algorithm Object Editor



Declaration

This thesis is my original work, has not been presented for a partial fulfillment of the requirement of a degree in any university and that all sources of material used for the thesis have been duly acknowledged.

_____

Anteneh Fentahun

July, 2011

This thesis has been submitted for examination with our approval as university advisors.

_____

Dr. Dereje Teferi

_____

Dr. Wakgari Deressa