ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

DETERMINING THE DEGREE OF DRIVER'S RESPONSIBILITY FOR CAR ACCIDENTS BY USING DATA MINING METHODS: THE CASE OF ADDIS ABABA TRAFFIC CONTROL AND INVESTIGATION DEPARTMENT

BY

Zelalem Regassa

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

Addis Ababa, Ethiopia

January, 2009

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

DETERMINING THE DEGREE OF DRIVER'S RESPONSIBILITY FOR CAR ACCIDENTS BY USING DATA MINING METHODS: THE CASE OF ADDIS ABABA TRAFFIC CONTROL AND INVESTIGATION DEPARTMENT

By

Zelalem Regassa

January, 2009

Name and Signature of Members of the Examining Board

_____

_____

_____

_____

Dedication

I dedicate this paper:

To my mother Degitu Fite whose love makes me strong.

To my late grand father Fite Hunde, I always miss him and I wish he didn't go so fast.

To those who paid the price so that we can also get proper education.

Table of Contents

List of Tables and Figures

List of Figures

List of Tables

Abstract

Road traffic accidents (RTAs) are now becoming one of the leading public health problems in Ethiopia. Due to accidents people are dying and getting disabled further resulting in property losses. It is more severe in the capital city of Ethiopia, Addis Ababa where most cars in the country are bustling. Most accidents in the capital are due to driver problems. And most of the fatalities are attributed to the accident type pedestrian hit by car.

In this research an attempt has been made to apply the decision tree and multilayer perceptron (MLP) neural network data mining techniques to analyse the accident data. The research 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).

In the research undertaking standard data mining methodology has been employed. Accordingly, the domain area; in this case the traffic control system is carefully studied, the accident dataset is carefully investigated to have a clear picture and verification of the data. Preprocessing the data by performing data cleaning, data selection, data transformation and replacing missing value activities are also among the crucial activities that have been carried out in this research.

Exploratory data analysis (EDA), descriptive modeling, predictive modeling (classification and regression), discovering patterns and rules and retrieval by content.

In accomplishing the data mining task a number of standard techniques and tools can be employed. The major data mining techniques according to Berry and Linoff (2004), are decision trees, neural networks, cluster Analysis and Statistical methods like, Bayesian inference, logistic regression, log-linear models, the common techniques of cluster analysis are :which are divisible algorithms, agglomerative algorithms, partitional clustering, and incremental clustering. Association Rules, genetic algorithms and fuzzy inference systems are also worth mentioning.

1.2 Road Traffic Control System

Addis Ababa is the capital of the Federal Democratic Republic of Ethiopia (FDRE). It is also 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. 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:-

- ✓ Ensure the safety to the citizen by promoting the 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 traffic police officers and various equipments such as Motor Bicycles and different Automobiles.

1.3 RTA

RTA is an accident that occurred on a way or street open to public traffic; resulted in one or more persons being killed or injured, and at least one moving vehicle was involved. Thus, RTA is collisions between vehicles; between vehicles and pedestrians; between vehicles and animals; or between vehicles and fixed obstacles (A. Persson, 2008). Every year about 1.2 million people are killed and more than 20 million are injured or disabled globally due to car accidents. 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 both 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 (2005), deaths from road traffic injuries account for around 25% of all deaths from injury. The annual number of road deaths varies from around 750,000 – 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 and that is probably conservative. 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.

1.4 Statement of the Problem and Justification of the study

The fundamental motivation for this research work is the prevalence of high road accident rate experienced by the nation and the capital, Addis Ababa, in particular. According to a report by WHO(2004), around 85% of all global road deaths and 96% of all children killed worldwide as a result of road traffic injuries occur in low-income and middle-income countries. In developing countries traffic accidents make up a very high proportion of the people being treated at accident and emergency departments and occupying hospital beds. Needless to mention the severe loss of property; people have been dying and injured due to car accidents. The capital 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, more than 80% of the accident factors are attributed to drivers' problem including, among other things, over speeding, disobeying traffic rules, driving with alcohol and selfish attitudes(NRSO,2008).The key determinants of the problem as identified by A. Persson (2008) are poor road network; absence of knowledge on road traffic safety; mixed traffic flow system; poor legislation and failure of enforcement; poor conditions of vehicles; poor emergency medical services; and absence of traffic accident compulsory insurance law.

The primary stakeholder in road safety problems of the city is AARTCID. As part of its daily activity the department collects and records road accidents occurring in the city to facilitate decision making.

In order to implement road safety policy effectively it is essential to have suitable data sources for monitoring and analyzing progress and evaluating the effects on safety of the measures taken. Analysis of this data may allow incident clusters and incident causes to be identified (www.wikipedia.com). According to WHO(2004) the loss due to road accidents is not affordable, especially in developing countries where resources are scarce and cannot be wasted on preventable 'accidents'. In an attempt to prevent road accidents one role that can be played is 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 risk and factors that might be modifiable through interventions. 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 technology is data mining.

A number of researches have been undertaken in analyzing accident data with different objectives using data mining technology at local level and internationally. One such research is that of Tibebe Besha (2005). He studied the accident data and proved the application of data mining in determining accident severity. He also recommended that other data mining techniques with better performance could be employed to come up with other important hidden patterns in the accident dataset.

Hence, in this research an attempt has been made to predict driver's degree of responsibility using decision tree and MLP techniques. It has also been tried to identify hidden patterns in the accident causing drivers behavior such as; the relationship between demographical, environmental and other attributes of the driver.

1.5 Objective of the study

The general and specific objectives of the proposed research are described as follows:

1.5.1 General Objectives

The general objective of the research is to build predictive data mining models to predict driver's degree of responsibility for an accident using RTA data and in doing so identify the hidden patterns in the data that influence the driver's degree of responsibility.

1.5.2 Specific objectives

The following are specific objective of the study so as to realize the above mentioned

general objective:-

- Investigate the traffic control system and identify the attributes that influence driver's responsibility for an accident.
- Predict the driver's degree of responsibility using the decision tree and MLP predictive techniques.
- Compare and suggest the best model for prediction.

1. 6 Research Methodology

In this research The WEKA data mining tool has been employed to implement most of the technical aspects of the CRISP-DM standard data mining methodology has been adopted. Based on the adopted methodology the important iterative activities that are undertaken in this research are: business understanding, data understanding, data preprocessing, and selection of modeling technique, model building and model evaluation. WEKA is the preferred tool for this research undertaking. The reason is that, it is extensively documented and it is equipped with multiple features to handle almost all activities performed in any data mining method. Last but not least it is public source software.

1.6.1 Literature Review

Different literatures on road safety and data mining have been reviewed. These include articles, magazines and other publications in the area of road safety to have a clear picture of the business and to understand the severity of road accidents. Data mining and machine learning literatures have also been extensively reviewed in an attempt to

discover the methods, tools and techniques that can be applied to achieve the data mining goal.

1.6.2 Data Collection/Creating Dataset

The accident data was obtained from the AARTICD. Some portion of the data is computerized in an excel file format in Amharic font and others are available in ledgers. The data was manually converted to English and completely described. The accident data was explored and re-explored using WEKA's data exploration facility. This was done to understand the accident data and to verify the quality of the data.

1.6.3 Data Preprocessing

At this stage different data preprocessing activities have been carried out based on the result of the data exploration result. Missing values were well attended by data cleaning process and some minor data transformation is also done. In addition data selection has been undertaken by selecting the appropriate records and attributes for the data mining.

The creation of the class label for the driver's degree of responsibility which is required for supervised learning is another major activity that has been carried out in this paper. Clustering technique which is an unsupervised learning technique is employed to determine the class label. As a result three classes 'Slight', 'Moderate' and 'Extreme' values are assigned based on expert judgment and used for prediction.

1.6.4 Experimentation and Result Analysis

In this research an attempt has been made to implement the most popular classification modeling techniques namely decision trees and neural networks (MLP).The models have

been built on different attribute subsets of the accident data. In all the experiments ten fold cross validation has been employed for training and testing the models. It is a model traing and testing method recommended by the WEKA developers. The performance of the different decision tree models has been studied and the one which is believed to be the best has been used for rule extraction and the rules where judged by the experts.

Evaluations of the models have been carried automatically to select the best model for prediction. In evaluating the models ten fold cross validation which is recommended by the WEKA developers has been used for training and testing. Three two models for decision tree and one for the MLP have been compared based on performance measures such as percent accuracy, true positive rate, precision , recall and others. And the one with the best score is selected.

1.7 Scope and Limitation of the Study

The scope of this research is limited to analysis of the accident data of the AARTCID of the years 2005-08 by employing decision tree and MLP techniques of classification for predictive data mining.

In conducting this research, the major limitations faced is that the data at the traffic office is not fully computerized and this research based on the partially computerized one. In addition, some important attributes are severely missing, inconsistent and incomplete. The absence of explicit information on level of responsibility of the driver for the accident can also be considered as a significant limitation.

1.8 Significance of the Research

It is the researcher's belief that the result of this research has a significant contribution in helping to improve the performance of the current traffic system to reduce car accidents especially by understanding important patterns in factors contributing to drivers' high level of responsibility for accidents. This could be achieved through enhancing the enforcement of traffic rules, awareness of the officials, the society and persuading policy makers and decision makers.

1.9 Research Organization

This research has 7 chapters. Chapter one is an introductory chapter. It presents the background of the study, statement of the problem, objective of the study and the methodology adopted in the research.

The second chapter is a discussion made on the data mining concept, importance, methods and techniques as described in different literatures.

In chapter three various literatures are summarized in an attempt understand the road traffic control business.

Chapter four is a report on the major activities performed during the accident data understanding and preprocessing phase of the data mining process. These are accident data collection, verification, data cleaning, data selection and data construction in data preprocessing.

The fifth chapter is a discussion made on the experimentation and result analysis phase of data mining process in which classification models are built and their performance is evaluated. The techniques used for modeling, the evaluation methods and the analysis of

the results are discussed in detail in this chapter.

Chapter Two

Data Mining Technology

2.1 Introduction

This chapter is about the essence of data mining, how it is important, how it is different from and similar to other related fields such as statistics, and the basic conceptual background of the methods and techniques of data mining as discussed in various data mining literatures by different authors.

2.1.1 Overview

We are living in a time where data is collected and stored in unprecedented volumes. Large and small enterprises collect data about their businesses, their customers, their human resources, their products, their manufacturing processes, their suppliers, their business partners, their local and international markets and their competitors. Turning this data into information and that information into knowledge has become a key component of the success of a business. Data contains valuable information that can support managers in their business decisions in effectively and efficiently running a business. Information is the basis for identifying new opportunities. Knowledge is the linchpin of society. (Graham Williams, 2004-08).

According Mehmed Kantardzic (2003), traditionally, analysts have performed the task of extracting useful information from recorded data. But, the increasing volume of data in modern business and science calls for computer-based approaches. As data sets have grown in size and complexity, there has been an inevitable shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and

sophisticated tools. The modern technologies of computers, networks, and sensors have made data collection and organization an almost effortless task. However, the captured data needs to be converted into information and knowledge from recorded data to become useful.

In recent years there has been an explosive growth of methods for discovering new knowledge from raw data (Mehmed Kantardzic, 2003). He adds that, in response to this, a new discipline of data mining has been specially developed to extract valuable information from such huge data sets. Given the proliferation of low-cost computers (for software implementation), low-cost sensors, communications, database technology (to collect and store data), and computer-literature application experts who can pose "interesting" and "useful" application problems, this is not surprising.

Different versions of definition for data mining exist in the data mining literatures. Some of them are: 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 D. 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".

2.2 Data Mining and Statistics

According to Berry and Linoff (2005), many of the data mining techniques were invented by statisticians or have now been integrated into statistical software; they are extensions

of standard statistics. They argue that although data miners and statisticians use similar techniques to solve similar problems, the data mining approach differs from the standard statistical approach in several areas:

- ❖ Data miners tend to ignore measurement error in raw data.
- ❖ Data miners assume that there is more than enough data and processing power.
- ❖ Data mining assumes dependency on time everywhere.
- ❖ It can be hard to design experiments in the business world.
- ❖ Data is truncated and censored.

They also generalize; the differences are differences of approach, rather than opposites. As such, they shed some light on how the business problems addressed by data miners differ from the scientific problems that spurred the development of statistics.

A. Ekhaus (2003) argued that data mining is useful for discovering relationships and Statistics is useful for analyzing relationships. The two disciplines need to coexist and methods that bridge the gap between the two are needed. Moreover, merging the two disciplines will develop approaches which discover "actionable decisions" with confidence that the decisions are actually "good".

The field of data mining, like statistics, concerns itself with "learning from data" or "turning data into information". Instead of trying to make a distinction between the two fields, it is rather important to note that data mining can learn from statistics- that, to a large extent, statistics is fundamental to what data mining is really trying to achieve. There is the opportunity for an immensely rewarding synergy between data miners and

statisticians. Data mining and statistics will inevitably grow toward each other in the near future because data mining will not become knowledge discovery without statistical thinking, statistics will not be able to succeed on massive and complex datasets without data mining approaches (Diego Kuonen, 2004).

2.3 Data Mining Methodology

In the business environment, complex data mining projects may require the coordinate efforts of various experts, stakeholders, or departments throughout an entire organization. In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements. One such model, CRISP (Cross-Industry Standard Process for data mining) was proposed in the mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining (StatSoft, Inc, 2008).

2.3.1 The CRISP-DM Methodology

The CRISP-DM data mining methodology is described in terms of a hierarchical process model as shown in figure 2.1. It consists sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task and process instance (The CRISP-DM consortium, August 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. This 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.

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.

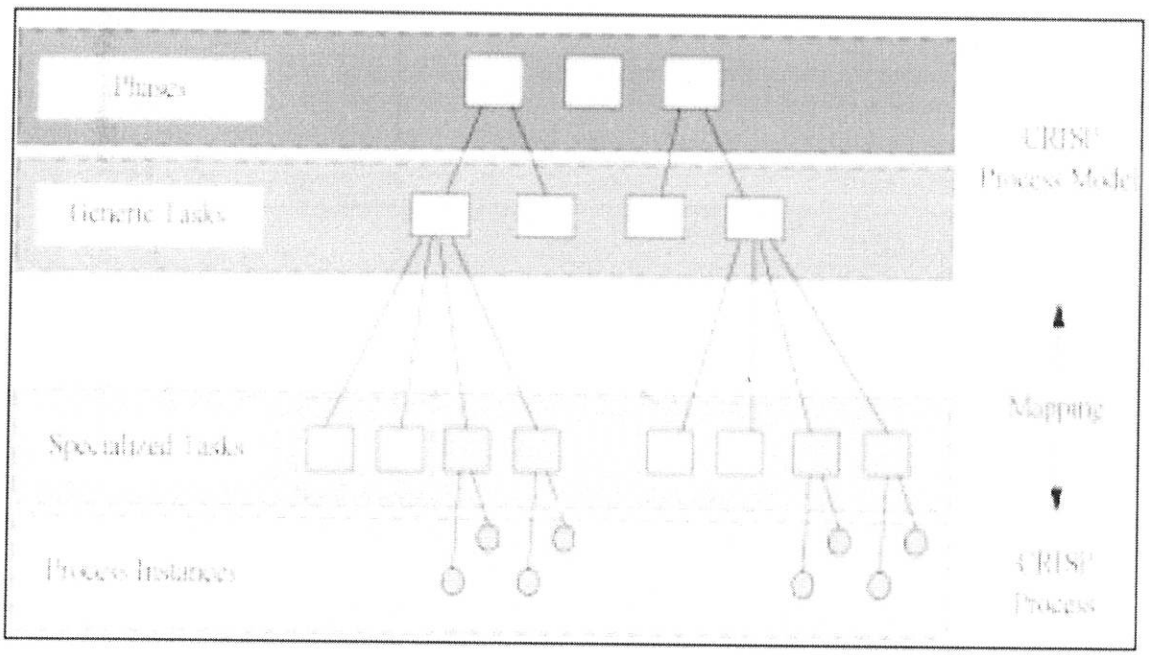


Figure 2.1: Four level breakdown of the CRISP-DM methodology (Adapted from: The CRISP-DM consortium, 2000)

In CRISP-DM a concept called data mining context drives mapping between the generic and the specialized level discussed above. The methodology 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.

2.3.2 The Data mining Process Model

A data mining model is concerned with the process of how to integrate data mining methodology into an organization, how to convert data into information, how to involve important stake-holders, and how to disseminate the information in a form that can easily be converted by stake-holders into resources for strategic decision making (StatSoft, Inc., 2008).

According to (The CRISP-DM consortium, August 2000), the life cycle of any data mining project consists of six phases (see Figure 2.2). They briefly outlined each phase as follows:

Business understanding: This initial phase focuses on understanding the 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.

Data understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data preparation: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

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

Evaluation: At this stage in the project you have built a model (or models) 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 project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying live models within an organization's decision making processes (The CRISP-DM consortium, August 2000).

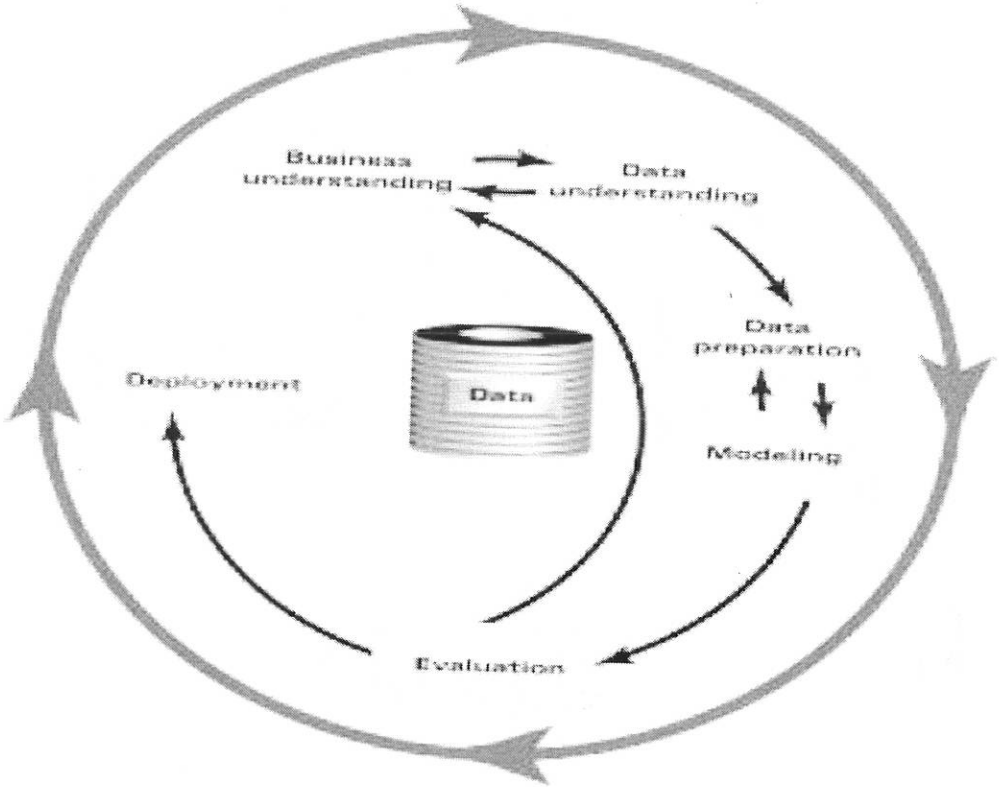


Figure 2.2: Phases of the CRISP-DM reference model (adapted from the CRISP-DM consortium, August 2000)

2.4 Data Mining Task

The two high-level primary goals of data mining in practice tend to be prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. Although the boundaries between prediction and description are not sharp (some of the predictive models can be descriptive, to the degree that they are understandable, and vice versa), the distinction is useful for understanding the overall discovery goal. The relative importance of prediction and description for particular data-mining applications can vary considerably. The goals of prediction and description can be achieved using a variety of particular data-mining methods (Usama Fayyad et al, 1996).

While, Berry and Linoff (2005) described data mining as having two general forms; directed and undirected data mining. Directed data mining involves searching through historical records to find patterns that explain a particular outcome. Directed data mining includes the tasks of classification, estimation, prediction, and profiling. Undirected data mining searches through the same records for interesting patterns. It includes the tasks of clustering, finding association rules, and description.

According to Mehmed Kantardzic (2003) the primary data-mining tasks are: Classification, Regression, Clustering, Summarization, Dependency Modeling and Change and Deviation Detection.

According to a publication by the CRISP-DM consortium (Augest, 2000), the data mining project involves a combination of different problem types, which together solve

the business problem. More over they have examined the basic concept of these methods and their inter connection as presented in the discussion that follows.

Data description and summarization: This aims at the concise description of characteristics of the data, typically in elementary and aggregated form. It provides an overview of the structure of the data. Sometimes, data description and summarization alone can be an objective of a data mining project.

Segmentation or clustering: This data mining problem type aims at the separation of the data into interesting and meaningful subgroups or classes. All members of a subgroup share common characteristics. Often, however, very often segmentation is a step towards solving other problem types. Then, the purpose can be to keep the size of the data manageable or to find homogeneous data subsets that are easier to analyze. Typically, in large datasets various influences overlay each other and obscure the interesting patterns. Then, appropriate segmentation makes the task easier.

Concept descriptions: It aims at an understandable description of concepts or classes. The purpose is not to develop complete models with high prediction accuracy, but to gain insights. Concept description has a close connection to both segmentation and classification. Segmentation may lead to an enumeration of objects belonging to a concept or class without any understandable description. Typically, there is segmentation before concept description is performed.

Classification: Classification assumes that there is a set of objects characterized by some attributes or features which belong to different classes. The class label is a discrete

(symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers), which assign the correct class label to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling. The class labels can be given in advance, for instance defined by the user or derived from segmentation. Classification is one of the most important data mining problem types that occur in a wide range of various applications. Many data mining problems can be transformed to classification problems.

Classification has connections to almost all other problem types. Prediction problems can be transformed to classification problems by binning continuous class labels, since binning techniques allow transforming continuous ranges into discrete intervals. These discrete intervals are then used as class labels rather than the exact numerical values and hence lead to a classification problem. Some classification techniques produce understandable class or concept descriptions. There is also a connection to dependency analysis because classification models typically exploit and elucidate dependencies between attributes.

Segmentation can either provide the class labels or restrict the dataset such that good classification models can be built. It is useful to analyze deviations before a classification model is built. Deviations and outliers can obscure the patterns that would allow a good classification model. On the other hand, a classification model can also be used to identify deviations and other problems with the data.

Prediction: Prediction is very similar to classification. The only difference is that in prediction the target attribute (class) is not a qualitative discrete attribute but a continuous

one. The aim of prediction is to find the numerical value of the target attribute for unseen objects. In the literature, this problem type is sometimes called regression. If prediction deals with time series data then it is often called forecasting.

Dependency analysis: Dependency analysis consists of finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of a data item given information on other data items.

Associations are a special case of dependencies, which have recently become very popular. Associations describe affinities of data items (i.e., data items or events which frequently occur together). A typical application scenario for associations is the analysis of shopping baskets.

Dependency analysis has close connections to prediction and classification, where dependencies are implicitly used for the formulation of predictive models. There is also a connection to concept descriptions, which often highlight dependencies. In applications, dependency analysis often co-occurs with segmentation.

2.5 Data Mining Technique

In the data mining literature several data mining techniques are discussed that can be employed to accomplish the various data mining tasks described in the previous section. Berry and Linoff (2005) argues that no single data mining tool or technique is equally applicable to all data mining tasks. The choice of a particular data mining technique depend on the business problems to be solved Graham Williams (2008). Some of the most important techniques as described in the literature are presented as follows.

In the machine learning literature, directed data mining is called supervised learning and undirected data mining is called unsupervised learning (Berry and Linoff, 2005).

2.5.1 Decision Tree

Various literatures described decision trees as a technique for classification in a number of ways. According to Mehmed Kantardzic (2003), decision trees and decision rules are data-mining methodologies applied in many real-world applications as a powerful solution to classification problems.

According to Han and Kamber (2001), a decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions (Hand). The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for that sample. A typical decision tree for classification of samples with two input attributes X and Y is given Figure 2.1. All samples with feature values X>0 and Y=B belong to Class2, while the samples with values X>0 belong to Class3, whatever the value for feature Y.

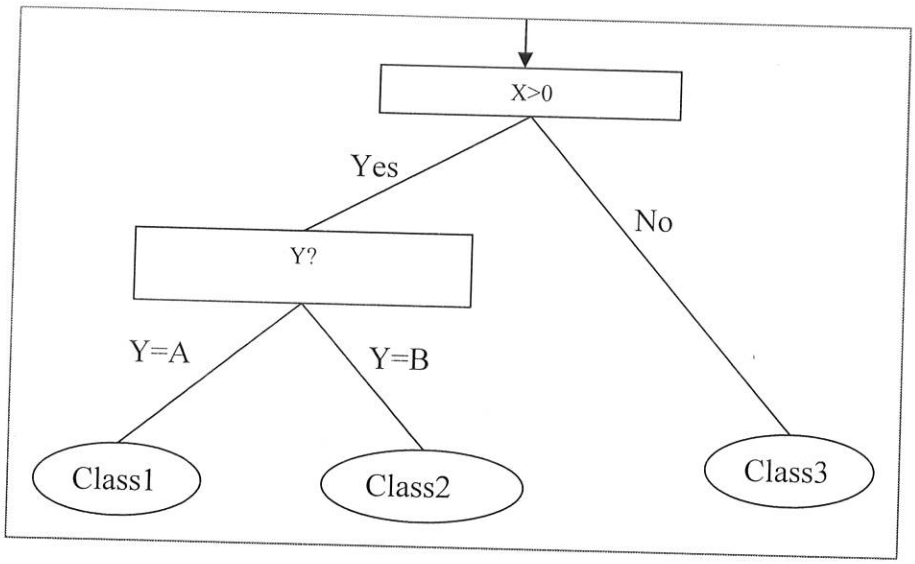


Figure 2.3: A simple decision tree with the tests on attributes X and Y

2.5.1.1 Decision Tree Algorithm

In the data mining and machine learning literatures decision trees are implemented using some common algorithms. A well-known tree-growing algorithm for generating decision trees based on univariate splits is Quinlan's ID3 with an extended version called C4.5. Greedy search methods, which involve growing and pruning decision-tree structures, are typically employed in these algorithms to explore the exponential space of possible models (Mehmed Kantardzic, 2003).

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 subtree 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. The general algorithm for a decision tree learner can be found at Appendix 1

2.5.3 Neural Network

An artificial neural network is an abstract computational model of the human brain. Similar to the brain, an ANN is composed of artificial neurons (or processing units) and interconnections. Although the term artificial neural network is most commonly used, other names include "neural network", parallel distributed-processing system (PDP), connectionist model, and distributed adaptive system. ANNs are also referred to in the literature as neurocomputers (Mehmed Kantardzic, 2003)

A neural network, as the name indicates, is a network structure consisting of a number of nodes connected through directional links. Each node represents a processing unit, and the links between nodes specify the causal relationship between connected nodes. All nodes are adaptive, which means that the outputs of these nodes depend on modifiable parameters pertaining to these nodes.

An artificial neural network is a massive parallel distributed processor made up of simple processing units. It has the ability to learn from experiential knowledge expressed through interunit connection strengths, and can make such knowledge available for use (Mehmed Kantardzic, 2003).

The use of artificial neural networks offers several useful properties and capabilities such as: Nonlinearity, Learning from examples, Adaptivity, Evidential Response, Fault Tolerance, Uniformity of Analysis and Design.

2.5.3.1 Model of an Artificial Neuron

An artificial neuron is an information-processing unit that is fundamental to the operation of an ANN (k). The block diagram (Figure 2.2), which is a model of an artificial neuron shows that it consists of three basic elements (Mehmed Kantardzic, 2003):

1. A set of connecting links from different inputs x_i (or synapses), each of which is characterized by a weight or strength w_{ki}. The first index refers to the neuron in question and the second index refers to the input of the synapse to which the weight refers. In general, the weights of an artificial neuron may lie in a range that includes negative as well as positive values.

2. An adder for summing the input signals x_i weighted by the respective synaptic strengths w_{ki}. The operation described here constitutes a linear combiner.
3. An activation function f for limiting the amplitude of the output y_k of a neuron.

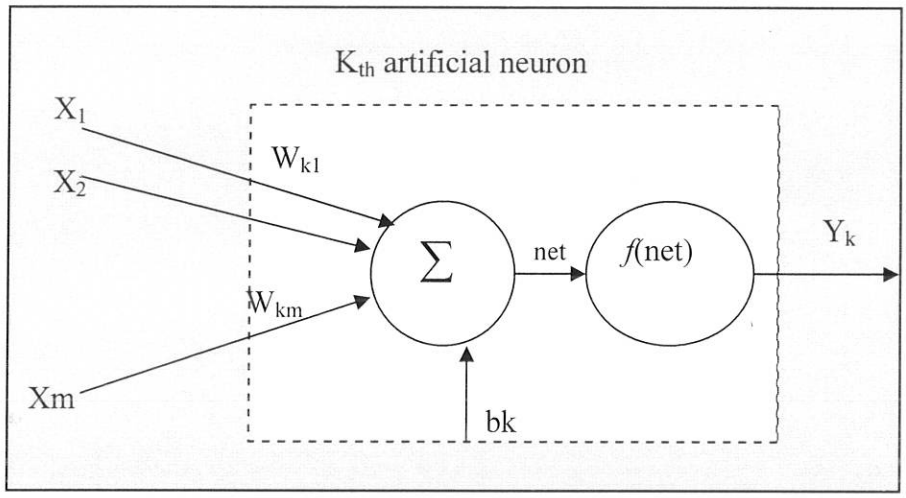


Figure 2.4: Model of an artificial neuron

The model of the neuron given in Figure 2.2 also includes an externally applied bias, denoted by b_k. The bias has the effect of increasing or lowering the net input of the activation function, depending on whether it is positive or negative.

A neuron's common activation functions are Hard Limit, Symmetrical Hard Limit,Linear, ,Sataurating Linear,Symmetric Saturating Linear, Log-Sigmoid, Hyperbolic Tangent Sigmoid. The description of the activation function is found at appendex2.

2.5.3.2 Architectures of Artificial Neural Networks

The architecture of an artificial neural network is defined by the characteristics of a node and the characteristics of the node's connectivity in the network (Mehmed Kantardzic, 2003). Typically, network architecture is specified. by the number of inputs to the

network, the number of outputs, the total number of elementary nodes that are usually equal processing elements for the entire network, and their organization and interconnections. Neural networks are generally classified into two categories on the basis of the type of interconnections: feedforward and recurrent.

The network is feedforward if the processing propagates from the input side to the output side unanimously, without any loops or feedbacks. In a layered representation of the feedforward neural network, there are no links between nodes in the same layer; outputs of nodes in a specific layer are always connected as inputs to nodes in succeeding layers. If there is a feedback link that forms a circular path in a network then the network is recurrent.

There are different types of Neural Networks discussed in the literature. According to DERG (2008) when used without qualification, the terms "Neural Network" (NN) and "Artificial Neural Network" (ANN) usually refer to a Multilayer Perceptron Network. However, there are many other types of neural networks including Probabilistic Neural Networks, General Regression Neural Networks, Radial Basis Function Networks, Cascade Correlation, Functional Link Networks, Kohonen networks, Gram-Charlier networks, Learning Vector Quantization, Hebb networks, Adaline networks, Heteroassociative networks, Recurrent Networks and Hybrid Networks.

2.5.3.3 The MLP Model

Multilayer feedforward networks are one of the most important and most popular classes of ANNs in real-world applications (Mehmed Kantardzic, 2003). Typically, the network consists of a set of inputs that constitute the input layer of the network, one or more

hidden layers of computational nodes, and finally an output layer of computational nodes. The processing is in a forward direction on a layer-by-layer basis. This type of artificial neural networks is commonly referred to as multilayer perceptrons (MLPs).

A multilayer perceptron has three distinctive characteristics:

1. The model of each neuron in the network includes usually a nonlinear activation function, sigmoidal or hyperbolic.
2. The network contains one or more layers of hidden neurons that are not a part of the input or output of the network. These hidden nodes enable the network to learn complex and highly nonlinear tasks by extracting progressively more meaningful features from the input patterns.
3. The network exhibits a high degree of connectivity from one layer to the next one.

The following diagram illustrates a perceptron network with three layers (DTREG, 2003):

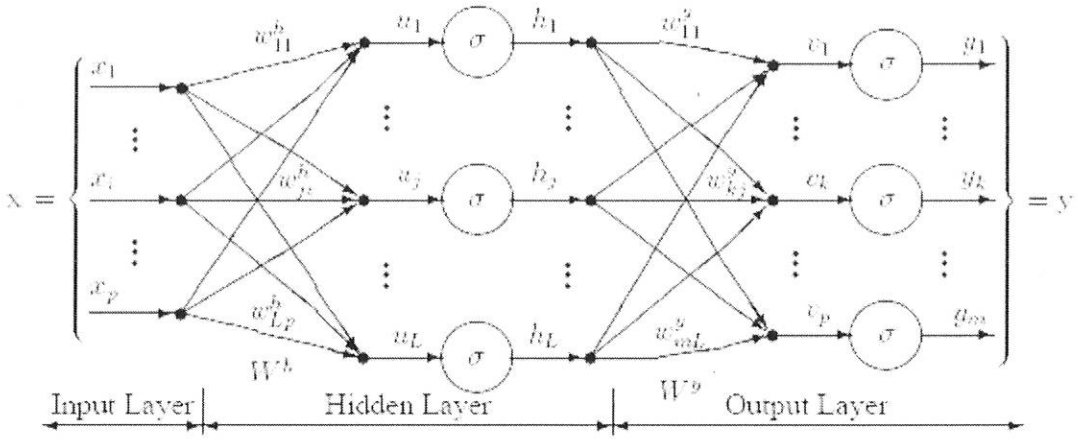


Figure 2.5: A model for MLP

This network has an input layer (on the left) with three neurons, one hidden layer (in the middle) with three neurons and an output layer (on the right) with three neurons. There is one neuron in the input layer for each predictor variable.

MLP has been applied successfully to solve some difficult and diverse problems by training them in supervised or unsupervised manner with a highly popular algorithm known as error back-propagation algorithm.

2.5.3.3 Back-propagation Algorithm

Training a neural network is the process of setting the weight on the inputs of each of the units in such a way that the network best approximates the underlying function. Back-propagation is the most commonly used method for training multilayer feed forward neural network. This training scheme is used for adjusting the connection weight of each unit in such a way that the error between the desired output and the actual output is

reduced. The role of a training algorithm is to set the network's weight and threshold so as to minimize predictive error by the network. The error of a particular configuration of the network can be determined by running all the training cases through network, comparing the actual output of generated with the desired or targeted output. The difference is combined together by an error function to give the network error. The most common error function is the sum squared error, where individual error of output on each case squared and summed together. The back-propagation algorithm can be referred at Appendix 2.

2.5.4 Other Data Mining Techniques

There are also other learning techniques described in the data mining and machine learning literatures. Some of them are sited by (K) as: Statistical Methods where the typical techniques are Bayesian inference, logistic regression, ANOVA analysis, and log-linear models. Cluster Analysis, the common techniques of which are divisible algorithms, agglomerative algorithms, partitional clustering, and incremental clustering. Association Rules represent a set of relatively new methodologies that include algorithms such as market basket analysis, apriori algorithm, and WWW path-traversal patterns. Genetic Algorithms are very useful as a methodology for solving hard optimization problems. Fuzzy Inference Systems are based on the theory of fuzzy sets and fuzzy logic. Fuzzy modeling and fuzzy decision making are steps very often included in the data-mining process. N-dimensional Visualization Method: Typical data-mining visualization techniques are geometric, icon-based, pixel-oriented, and hierarchical techniques.

These techniques have been proved to be applicable to solving different types of real world problems, some of which are described in the section that follows.

2.6 Data Mining Applications

Han and Kamber (2001) 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 examined 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.

According to Usama Fayyad et al (1996), the main data mining application in business areas includes marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents while in science, one of the primary application areas is astronomy.

A survey by Two Crows Consulting (Two Crows, 2001) found these applications of data mining: Ad revenue forecasting , Churn (turnover) management, Claims processing, Credit risk analysis, Cross-marketing, Customer profiling, Customer retention, Electronic commerce, Exception reports, Food-service menu analysis, Fraud detection, Government policy setting, Hiring profiles, Market basket analysis, Medical management, Member enrollment, New product development, Pharmaceutical research, Process control, Quality control, Shelf management/store management, Student recruiting and retention, Targeted marketing and Warranty analysis.

2.6.1 Data Mining Application for RTA Data analysis

A number of data mining application researches has been undertaken for the analysis of RTA data at locally and globally. Some of the most important works are summarized and presented as follows.

In their paper on the analysis of the GES automobile accident data from 1995 to 2000 using machine learning paradigms, Miao Chong et al (2001) 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. In their past research they focused mainly on distinguishing between no-injury and injury (including fatality) classes. 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 drivers' fatality has the highest cost to society economically and socially. According to them the very important factors 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 extremely likely that it could have helped to improve the performance of models studied in their paper.

Tibebe Beshah (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 accident severity. In this 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%.

In this chapter the basic concept of data mining, the problems it handles and how it is important, the techniques that can be employed and the methodology it adopts have been discussed as described by different data mining literatures. Some important related literatures in the application of data mining have also been reviewed.

Chapter Three

Road Traffic Accidents

3.1Introduction

Throughout the world, roads are bustling with cars, buses, trucks, motorcycles, mopeds and other types of two- and three-wheelers. By making the transportation of goods and people faster and more efficient, these vehicles support economic and social development in many countries. But while motorized travel provides many benefits, it can also do serious harm unless safety is made a priority. Pedestrians and cyclists using roads are particularly at risk. Crashes are frequent. Deaths and injuries are common (WHO, 2004).

Traffic on roads may consist of pedestrians, ridden or herded animals, vehicles, streetcars and other conveyances, either singly or together, while using the public way for purposes of travel. Traffic laws are the laws which govern traffic and regulate vehicles, while rules of the road are both the laws and the informal rules that may have developed over time to facilitate the orderly and timely flow of traffic.

Traffic is formally organized in many jurisdictions, with marked lanes, junctions, intersections, interchanges, traffic signals, or signs. Traffic is often classified by type: heavy motor vehicle (e.g., car, truck); other vehicle (e.g., moped, bicycle); and pedestrian (www.wikipedia.com).

A car accident is a road traffic incident which usually involves at least one road vehicle being in collision with, another vehicle, another road user, or a stationary roadside object, and which may result in injury or property damage. Phrases used to describe accidents

include: auto accident, car crash, car smash, car wreck, fender bender, motor vehicle accident (MVA), personal injury collision (PIC), road accident, RTA, road traffic collision (RTC), road traffic incident (RTI), and traffic collision (www.wikipedia.com).

3.1.2 Road Traffic Accidents

Road crashes, causing death, injury, and damage have always happened. History tells of many notable historic personalities who were the victim of such incidents. Louis IV of France died in 954 after falling from his horse, as did at least two kings of England: William I in 1087 and William III in 1702(www.wikipedia.com).

RTAs are a major public health concern causing thousands of injuries and premature deaths each year. Globally, RTAs result in an estimated 1.2 million deaths and a further 50 million injuries per year (World Health Organization, 2004). In 1990, RTAs were the ninth leading cause of the global burden of disease1 and they are predicted to rise to the third leading cause by 2020.This figure is set to increase by 83% in low and middle-income countries which already shoulder 90% of the global road traffic death toll (World Health Organization, 2004), warns a new report launched in Paris on World Health Day, 7 April 2004. Road traffic crashes cost the world US$ 518 billion every year, says the World report on road traffic injury prevention, the first ever joint report on the subject released by WHO and the World Bank. It also adds that the bulk of the global burden of road traffic-related deaths occur in low and middle-income countries, with countries in South-East Asia and the Western Pacific regions accounting for more than half of all road traffic deaths. The report says the economic implications of this for developing countries are grave since more than half of all road traffic deaths occur among young adults

between 15 and 44 years of age. According to this report, the annual bill footed by such countries for road traffic crashes stands at an estimated US$ 65 billion more than the total amount received by these same countries in development assistance and representing between 1% and 2% of their gross national product.

The global burden of disease report by (Harvard University, 2005) establishes RTAs as a health problem, especially in terms of years of life lost. The report states that for men aged 15 -44, RTAs are the biggest cause of ill-health and premature death worldwide, and the second biggest in the developing regions, surpassed only by depression". The authors also suggest that "the high toll of RTAs in developing regions has received little attention from public health specialists in the past". In projections made for 2020, they expect RTAs to rise accordingly and could rise to third place from ninth worldwide (in terms of lost years of life).

In Ethiopia over the last ten years RTAs claimed over 3,065 lives every year and the damage to property caused by these accidents is estimated to be more than Birr 500 million and road accidents in the capital account for 65% of the total accidents occurred in the country (National Road Safety Coordination Office, 2008).

3.1.2 Road Traffic Safety, Accident Types and Accident Factors

Road traffic safety aims to reduce the harm (deaths, injuries, and property damage) resulting from crashes of road vehicles. Harm from road traffic crashes is greater than that from all other transportation modes (air, sea, space, off-terrain, etc.) combined. It deals exclusively with road traffic crashes, how to reduce their number and their consequences.

Strictly speaking, most accidents are not accidents at all: they are collisions that could and should have been avoided (smartmotorist.com).They identified four factors that contribute to the vast majority of collisions. In ascending order they are: equipment failure, roadway design, poor roadway maintenance, and driver behavior. Their research also indicated that, over 95% of motor vehicle accidents (MVAs, in the USA, or RTAs,, in Europe) involve some degree of driver behavior combined with one of the other three factors. Most are caused by excessive speed or aggressive driver behavior. Driving behavior is aggressive if it is deliberate, likely to increase the risk of collision and is motivated by impatience, annoyance, hostility and/or an attempt to save time. The New York State Police characterize aggressive driving by different traffic violations such as, excessive speed, frequent or unsafe lane changes, failure to signal, tailgating, and failure to yield the right of way, disregarding traffic controls and impaired driving.

The report by the National Road Accident Bureau of Ireland (2000) noted, amongst other things, that:

- Compared to older drivers, young drivers were deemed to be to a large extent responsible for a higher proportion of the fatal / injury accidents in which they were involved.
- Compared to young female drivers, young male drivers were deemed to be to a large extent responsible for a higher proportion of accidents in which they were involved.
- The responsibility rate was higher for young drivers with provisional licenses than for young drivers with full licenses.

- ❖ Young Driver Accidents (YDAs) tend to be more severe than accidents not involving young drivers.

3.2 RTA data Analysis at Addis Ababa

In order to implement road safety policy effectively it is essential to have suitable data sources for monitoring and analyzing progress and evaluating the effects on safety of the measures taken. 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 may allow incident clusters and incident causes to be identified (www.wikipedia.com).

RTA information is the interest 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 office has been undertaking some simple and manual statistical analysis on the accident data. It is mainly analysis of accident severity rates such as 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. More over 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 Besha (2005), studied the 2000-04 accident data with 4,658 records at AARTCID and classified accident severity in to four classes fatal injury, serious injury, simple injury

and property damage using decision tree. In his research he identified 'accident cause', 'accident type', 'road condition', 'vehicle type', 'light condition', 'road surface type' and 'driver age' as the most important determinant variables for level of injury severity of an accident. More over he recommended the possibility to consider other accident variables and apply different techniques like neural networks in order to come up with better model and dig out other important patterns hidden in the accident dataset. He also added that more number of experiments and testing techniques seem appropriate.

The National Road Safety office has also analyzed the accident data throughout the country for the years 2002-2007. The analysis was made to find out the most important accident factors based on accident information in order to take safety measures to reduce the accident severity in the country. To this end the office has carried out the following major activities:

- Extensively studied the road accident data and came up with important accident factors.
- Identified actions to be taken before, during and after an accident occurs.
- Identified important safety measures that are believed to be fruitful and feasible and planed a safety project for their implementation.

According to the office the basic accident factors are road and its environment, pedestrian, vehicle, driver and combination of two or more of them (NRSCO, 2008). Driver behavior, vehicle problem, pedestrian fault, road design and others contribute 81%, 5%, 4%, 1% and 9% for the occurrence of RTA's across the country. As it can be seen clearly from the figures accidents are highly attributed to driver cause. The major

driver problems are denying priority for pedestrian, over speeding, over loading, breaking traffic rules and other aggressive and selfish behaviors. Hence the driver's level of education, age, experience skill, and their connection to the vehicle needs a critical investigation.

Table 3.1: Percentage fatality by accident factor across the nation for years 2002-2007

S.No	Accident Factor	Fatality and injury
1	Driver behavior	81%
2	Equipment failure	5%
3	Pedestrian failure	4%
4	Road design	1%
5	Others	9%

Considering the accident types, the report indicted that pedestrian collision is the leading accident type contributing 68% of fatalities while roll over, felling from car, collision with animals and others contribute 13%, 6%, 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

S.no	Accident Type	Fatality
1	Hit pedestrian	68%
2	Roll over	13%
3	Fell from car	65%
4	Collision with animal	3%
5	Others	10%

In general, the above discussions show that various important analyses can be performed on the acci dent data and the high risk on roads and the causes are also identified. .Accordingly, the focus of this study is to investigate driver's behaviors and other related variables of the accident data that influence the different level of driver's responsibility for RTA's in Addis Ababa. The researcher believes that it helps to bring attention and to spotlight one of the most dangerous dilemmas threatening the traveling public which is well recognized by the public.

Chapter Four

Analysis and Design

4.1 Introduction

In chapter three, the data mining task has been determined by carefully investigating the business area which is the RTA. This chapter explains the important activities carried out in achieving the classification task by applying a standard data mining methodology.

In this research the iterative CRISP data mining methodology is adopted. Firstly, the traffic control system and road safety is studied which is reported in chapter three of this paper. Next, data understanding activities such as; data collection, data description and data verification have been undertaken. Secondly, the data has been pre-processed by employing data cleaning and data selection techniques. At the third step, modeling techniques have been selected. Next, different models have been built on the dataset with the selected attributes. Finally, the models are evaluated based on standard data mining model evaluation techniques. The method is schematically shown in Figure 4.1 below.

4.2 Data understanding

4.2.1 Overview

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

4.2.2 Data Collection

The fist step in defining a data mining task is the specification of the task relevant data, that is, the data on which mining is to be performed (Han and Kamber, 2005). Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart (The CRISP-DM consortium, August 2000). 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, and loads the data into a query database can be an enormous task, sometimes taking years and costing millions of dollars. Data mining can be performed on data from one or more operational or transactional databases by simply extracting it into a read-only database (The CRISP-DM consortium, August 2000).

The data source for this research consumption is the traffic accident data kept at AARTCID. The data storage is partially automated in an Excel file format. It stores partial road accident records of years 2005-07 that occurred in the city. The total accident dataset obtained is around 8,345. The painstaking conversion of the Amharic data to English happened to be time consuming because of the large data volume and the search for equivalent and well expressive English term.

4.2.3 Formatting the Data

Formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool. 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 field being the outcome field the model is 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, August 2000).

The WEKA data mining tool requires the dataset to be in a comma separated file format called the Attribute Relation File Format (ARFF) file format. 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 was converted to an ARFF file format.

4.2.4 Data Description

The accident record has forty columns or attributes of text, number, and date and time format. Among these attributes the car plate number and the driver's name have been kept secrete by the office for the sake of keeping the privacy of the accused. Table 4.1 shows the complete description of the attribute names and data types or format and what they store.

Table 4.1: Description of the accident dataset

S.No	Attribute Name	Data Type	Description
1	RegNo	Number	A key that identifies an accident uniquely.
2	Date	Date	The exact date on which an accident happened.
3	Time	Time	The time at which an accident happened
4	Day	Text	The week day on which an accident happened.
5	DriverSex	Text	The sex of the accident causing driver.
6	DriverAge	Number	The Age of the accident causing driver
7	LevelOfEducational	Text	The level of education of the accident causing driver
8	DriversCarConnection	Text	Whether the driver is an owner or a professional driver. or other.
9	DrivingExprience	Number	The driving experience of the accident causing driver
10	VehicleType	Text	The type vehicle
11	VehicleOwnership	Text	The owner of the vehicle
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 accident occurred.
15	ParticularPlace	Text	The convensional name given to places in the city like Olompia.
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

attribute values at the time of data entry are recorded as "not known" 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 some 50 records with missing or unknown values for significant number of attributes were removed from the datasets. Around 157 of the records with huge amount attributes with missing values have been deleted from the datasets while some of the crucial ones have been replaced by the appropriate values. 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 were replaced by modal values.

4.3.2 Construct data

This task includes constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes (The CRISP-DM consortium, August 2000).

The categorical attributes, 'Driver Age', 'Driving Experience', 'Vehicle Period Of Service', 'No Of Vehicles Involved', 'License Grade' and 'Victim Age' have been transformed to letter codes to enhance clarity as indicated in the Table 4.3.

Table 4.3: Attribute Construction.

S.NO.	Attribute	Value	Code
1	DriverAge	Below-18, 18-30, 31-50, Above-50 years	A, B,C,D & E respectively
2	DrivingExperience	Below-1, 1-2,2-5, 5-10 & Above 10 years	A, B, C, D, & E respectively
3	VehiclePeriodOfService	Below 2, 2-5, 5-10 and Above 10 years	A, B, C & D respectively
4	NoOfVehiclesInvolved	1, 2, & 3	One, Two and Three
5	VictmAge	Below 7, 7-18, 18-30, 30-50 & Above 50 years	A, B, C, D & E respectively

4.3.4 Clustering for Classification

The WEKA data mining tool implements various supervised and unsupervised filtering algorithms that transform the input dataset in some way. The AddCluster filtering algorithm is one such algorithm which is an unsupervised filtering algorithm that assigns a new nominal attribute (a cluster number like cluster1, cluster2 ...) representing the cluster assigned to each instance by a selected clustering algorithm. Certain attributes can be ignored when clustering (Whitten, 2005).

Hence, the accident dataset was fed to the WEKA data mining tool. The three columns or attributes accident type, accident severity and accident cause which are used by domain experts to identify the level of driver's responsibility were selected. Figure 4.1 shows the WEKA's explorer window after the data is fed and the 'AddCluster' with the k means clustering algorithm with k=5 is selected.

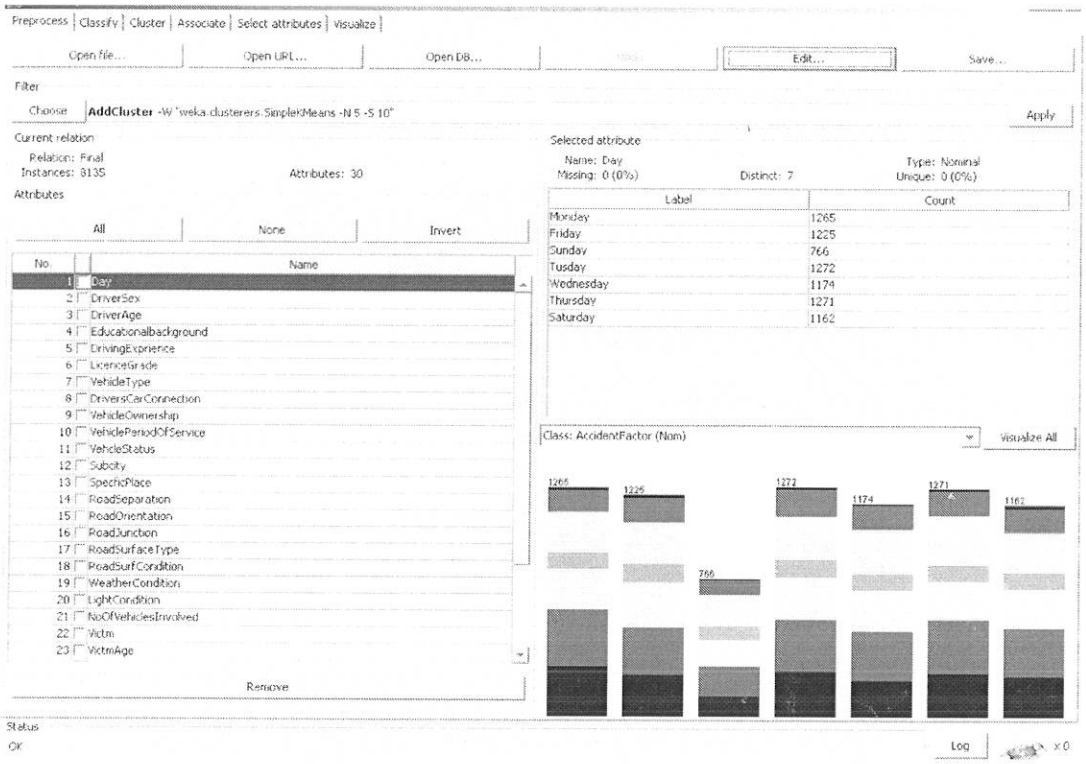


Figure 4.1: WEKA's explorer window.

The Addcluster filtering algorithm with the k-means clustering algorithm was run on the whole dataset for different number of clusters until the difference between clusters became insignificant. The histogram below shows the distribution of the dataset into the five clusters.

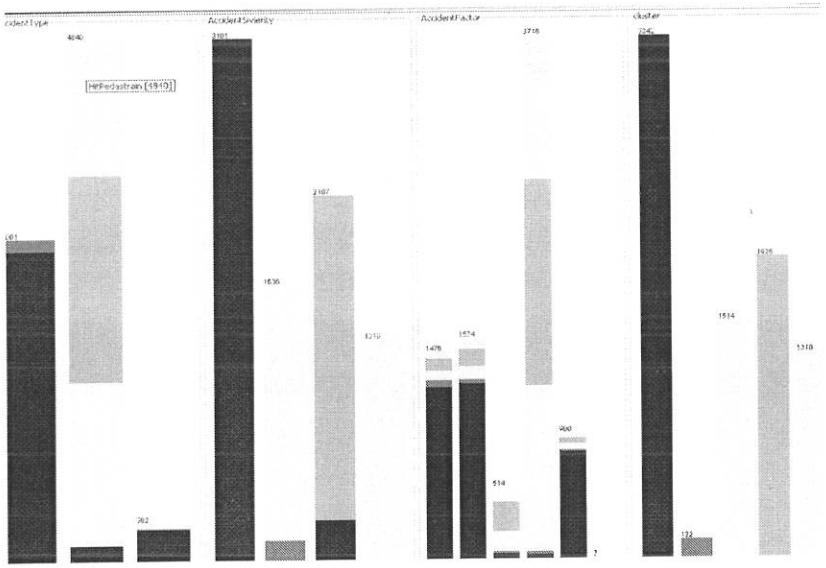


Figure 4.2: WEKA's data visualization output.

The output was forwarded to the domain experts to assign driver's responsibility based on how the clusters are related to the three attributes. They suggested that assigning responsibility to those accidents involving the 'car to car collision' is so complex and subjective. Moreover, they give special attention to accidents involving pedestrians. In addition, the fatality rate of RTA's highly sever for the accident type 'hit pedestrian', contributing more that 82%, (National Road Safety Coordination Office, 2008). As a result, only 3,373 accidents with the 'hit pedestrian' accident types which are driver caused are selected and others are considered as outliers and omitted. The new dataset now has a nominal attribute named cluster which replaces the accident type, accident severity and accident cause columns. It has three values namely; cluster1, cluster2 and cluster3. The class label cluster is renamed to 'DriverDegreeOfResponsibility' and based on domain expert opinion the cluster values clusrer1, cluster2 and cluster3 are renamed as "moderate", "extreme" and "slight" respectively. Hence the attributes 'accident type',

'accident severity' and 'accident factor' are now replaced by the new class label attribute 'DriverDegreeOfResponsibility'.

5.3.5 Data selection

In selecting data which covers selection of attributes as well as selection of records to be used for analysis the researcher applied different criteria including, relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types, data mining principles for data selection, domain knowledge expertise and results from the exploration task reported in the previous section.

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). However, automatic methods are also useful. There are two fundamentally different approaches for the automatic method, the filter method and the wrapper method. The filter method makes an independent assessment based on the general characteristics of the data; while the wrapper methods evaluates the subset using the machine learning algorithm that will ultimately be employed for learning. In WEKA ClassifierSubsetEval is a method used to select attribute subset by using a classifier to evaluate attribute set while, WrapperSubsetEval uses a classifier plus cross-validation to evaluate attribute subset, (Whitten and Frank, 2005). WEKA's Single attribute evaluators are ChiSquaredAttributeEval, GainRatioAttributeEval and InfoGainAttributeEval. The worth of the attribute subset is determined using the full set of training data or a process of

cross-validation. In this research both methods, manual and automatic methods for attribute selection have been employed.Manually, 16 attributes which are believed by the domain experts to have significant contribution in the assessing driver's level of responsibility, which is the focus of this research, have been selected. And they are: DriverAge, EducationalLevel, DrivingExprience, LicenceGrade, VehicleType, VehiclePeriodOfService, VehcleStatus, RoadSeparation, RoadOrientation, RoadJunction, RoadSurfaceType, RoadSurfCondition, WeatherCondition, LightCondition, VehicleMovemet and DegreeOfResponsibility. However, the significance of these attributes is to be tested during experimentation and different attribute selection mechanism would be employed due to the iterative nature of data mining process.

The automatic attribute selection methods such as; WEKA's attribute subset evaluators ClassifierSubsetEval, WrapperSubsetEval and single attribute evaluators GainRatioAttributeEval and InfoGainAttributeEval can be employed during experimentation. In this case, the attribute subset evaluator "WrapperSubsetEval" selected the 13 attributes as shown in the Figure 4.3

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

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 168
        Merit of best subset found:    0.514

Attribute Subset Evaluator (supervised, Class (nominal): 16 AccidentFactor):
        Classifier Subset Evaluator
        Learning scheme: .classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Hold out/test set: Training data
        Accuracy estimation: classification error

Selected attributes: 1,3,4,5,6,7,8,9,10,11,12,14,16,: 13
              DriverAge
              Educationalbackground
              DrivingExprience
              LicenceGrade
              VehicleType
              VehiclePeriodOfService
              RoadSeparation
              RoadOrientation
              RoadJunction
              RoadSurfaceType
              RoadSurfCondition
              LightCondition
              VehicleMovemet
```

Figure 4.3: WEKA's attribute selector output.

In this chapter all the important works undertaken in understanding and preprocessing the RTA dataset has been reported. The accident dataset is now well qualified to be mined to identify factors influencing driver's different level of responsibility for car accident and predict the degree of responsibility.

Chapter Five

Experimentation and Analysis of Results

As clearly presented in the previous chapter the data is now well understood, explored, selected and is 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 one for prediction.

5.1 Introduction

The study focuses on determining the degree of driver's responsibility for car accidents. In doing so, an attempt has also been made to identify the behavioral and environmental factors contributing to the different drivers degrees of responsibility.

In this research undertaking, various classification models have been built by using decision tree and neural network techniques. The models have been tested on different number of the selected attributes and the significance of the outputs of the most important model was presented for analysis to the domain experts. Finally, the model with the best performance is selected.

5.2 Selecting a Modeling Technique

Data mining is about building models from data. We build models to gain insights into the world and how the world works. A data miner, in building models, deploys many different data analysis and model building techniques. Our choices depend on the business problems to be solved (Graham Williams, 2008). 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, August 2000). Many techniques have been developed for classification or predictive modeling, and there is an art to selecting and applying the best method for a particular situation (DETERG, 2008). 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 conducting this research, decision tree and neural network techniques have been used. Each of these techniques has many attractive features.

Decision trees a re easy to build and are easy to understand. They can handle both continuous and categorical variables and can perform classification as well as regression. They automatically handle interactions between variables and they identify important variables. Neural networks on the other hand, have wide applicability. They have been successfully applied to a wide variety of classification and regression problems. Neural networks have the theoretical capability of modeling any type of function and they have tremendous accuracy (DETERG, 2008).

The WEKA data mining tool implements decision trees by 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. While the MLP is implemented using the backprpagation algorithm (Whitten and Frank, 2005).

The back propagation algorithm measures the overall error of the network by comparing the values produced on each training example to the actual value. It then adjusts the weights of the output layer to reduce, but not eliminate, the error. It then assigns the error to earlier nodes of the network and adjusts the weights by connecting those nodes, further reducing overall error. This technique for adjusting the weights is called the generalized delta rule. There are two important parameters associated with using the generalized delta rule. The first is momentum, which refers to the tendency of the weights inside each unit to change the "direction" they are heading in. That is, each weight remembers if it has been getting bigger or smaller, and momentum tries to keep it going in the same direction. The learning rate controls how quickly the weights change. The best approach for the learning rate is to start big and decrease it slowly as the network is being trained (Berry and Linoff, 2005)

5.3 Techniques for Data Mining Model Evaluation

Evaluation is the key to making real progress in data mining (Whitten and Frank, 2005). According to Han and Kamber (2000), estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classier will correctly label future data, i.e., data on which the classifier has not been trained. Accuracy estimates also help in the comparison of different classifiers. Holdout and cross-validation are two common techniques for assessing classifier accuracy, based on randomly-sampled partitions of the given data. In the holdout method, the given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set.

The standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use stratified 10-fold cross-validation (Whitten and Frank, 2005). The data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning procedure is executed a total of 10 times on different training sets. Finally, the 10 error estimates are averaged to yield an overall error estimate.

5.4 Experimentation

As a result of the extensive data pre-processing activities that have been done the 3,373 accident dataset chosen for this purpose is clean, it has no missing value, there is a well identified class label which is "DriverDegreeOfResponsibility", with three nominal values: "Extreme", "Moderate" and "Slight" and it is in an ARFF file format . During data exploration, different number of attributes has been selected by different selection techniques. Now the data can be fed to the modeling tool WEKA to build and test the various models using the methods that have been chosen.

One way of using WEKA is to apply a learning method to a dataset and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances. A third is to apply several different learners and compare their performance in order to choose one for prediction. Many classifiers have tunable parameters, which can be accessed through a property sheet or object editor. A common evaluation module is used to measure the performance of all classifiers (Whitten, 2005).

5.4.1 Building a decision tree

In building the various decision tree models the ID3 and J48 algorithms have been utilized with different parameters and number of attributes. As mentioned earlier J48 is WEKA's version of the popular C4.5 decision tree learner.

5.4.2 The Experiment

In the first experiment, the 3,373 accident dataset having 14 attributes, 13 of them independent variables and the 14th one the "DriverDegreeOfresponsibility" being the dependent or the class label attribute has been fed to the WEKA's explorer. Since the explorer generally chooses sensible defaults (Whitten, 2000) the J48 decision tree algorithm with all its default parameters was 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 3. The training and testing is done using ten fold cross validation. Figure 5.1, shows the WEKA explorer window after the classifier is run.

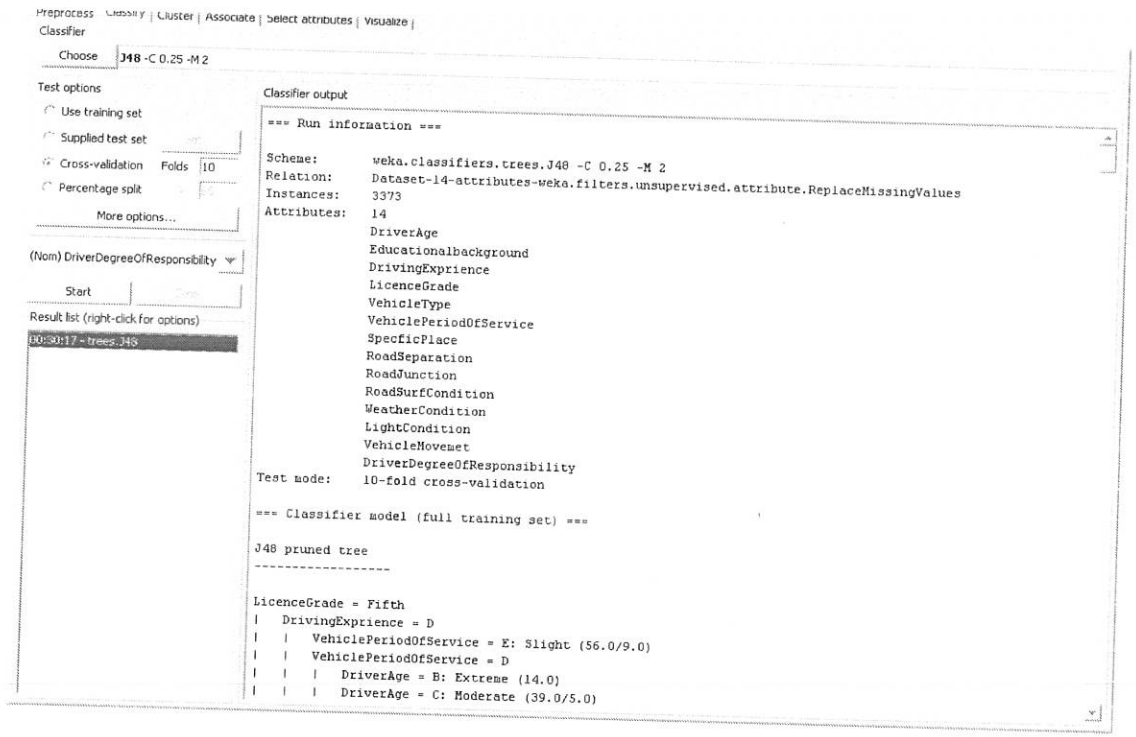


Figure 5.1: The explorer window with the J48 output.

The detailed out put of the J48 decision tree learner shows the decision tree graphically and textually in addition the detailed accuracy and the confusion matrix. Figure 5.2 shows the detailed output consisting part of the generated decision tree, the accuracy and the confusion matrix.

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

LicenceGrade = Fifth
|   DrivingExprience = D
|   |   VehiclePeriodOfService = E: Slight (56.0/9.0)
|   |   VehiclePeriodOfService = D
|   |   |   DriverAge = B: Extreme (14.0)
|   |   |   DriverAge = C: Moderate (39.0/5.0)
|   |   |   DriverAge = A
|   |   |   |   VehicleMovemet = MovingStraight: Moderate (5.0)
|   |   |   |   VehicleMovemet = Turning: Extreme (5.0/1.0)
|   |   |   |   VehicleMovemet = EnteringJunction: Moderate (2.0)
|   |   |   DriverAge = D
|   |   |   |   Educationalbackground = SSS
|   |   |   |   |   RoadJunction = No_Junction: Extreme (3.0/1.0)
|   |   |   |   |   RoadJunction = CrossRoad: Slight (0.0)
|   |   |   |   Educationalbackground = Elementary: Moderate (7.0/1.0)
```

```
|  |  |  |  Educationalbackground = Junior: Moderate (3.0)
|  |  VehiclePeriodOfService = C
|  |  |  DriverAge = B: Slight (2.0)
|  |  |  DriverAge = C: Moderate (8.0)
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2930               86.8663 %
Incorrectly Classified Instances       443               13.1337 %
Kappa statistic                          0.8017
Relative absolute error                 25.6598 %
Root relative squared error             56.9004 %
Total Number of Instances             3373
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   Class
                 0.863    0.063      0.858     0.863     0.861     Slight
                 0.907    0.08       0.874     0.907     0.89      Moderate
                 0.827    0.055      0.873     0.827     0.849     Extreme
=== Confusion Matrix ===

   a   b   c   <-- classified as
 889  68  73 |   a = Slight
  64 1165  55 |   b = Moderate
  83 100 876 |   c = Extreme
```

Figure 5.2: Output from the J4.8 decision tree learner.

5.4.3 Result Analysis

Figure 5.2 above, shows the full output of the J48 learner. At the beginning is a summary of the dataset, and the fact that the default tenfold cross-validation was used to evaluate it. Then comes a pruned decision tree in textual form. The first split is on the "LicenseGrade" attribute, and then, at the second level, the split is on "RoadSeparation" 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, the expression (56.0/9.0) from the first leaf node in

figure...means that a total of 56 instances reached that leaf node out of which 9 are classified incorrectly. Beneath the tree structure the number of leaves is printed; 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 3,373 instances, 2,930 (86.871%) are correctly classified and 443(13.139%) are incorrectly classified in the cross-validation. As well as 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 indicates that out of the 2,930 correctly classified instances, 889 are correctly classified as "Slight", 1165 as "Moderate" and 876 as "Extreme" Driver's degree of responsibility. Regarding the misclassified instances; of the 141 class "slight" miss classified instances 68 are misclassified as class "moderate" and 73 as class "Extreme" , of the 119 misclassified class "Moderate" instances, 64 are gone to class "Slight" and 55 to class "Extreme" . While 83 and 100 instances of class "Extreme" instances are wrongly classified as "Slight" & "Moderate" respectively. Another unique output from the J4.8 decision tree learner is the graphical output it produces for the tree it builds.

```
  a   b   c  <-- classified as
939  38  53 |  a = Slight
 34 1201  49 |  b = Moderate
 41  58 960 |  c = Extreme
```

Figure 5.3: Detailed accuracy of the MLP model.

As can be seen from the out put in figure the accuracy of the MLP model is 91.91%. The confusion matrix at the bottom of the figure details the accuracy.

Other MLP models have also been trained using 10 fold cross validation and the result is summarized in Table 5.2.

Table 5.2: Performance of the MLP model on different attribute subsets

S.No	attributes	Number of epoch	Momentum	Learning rate	Accuracy
1	14	300	0.25	0.3	91.00
2	12	300	0.25	0.3	91.40
3	10	300	0.25	0.3	91.84

As it can be seen from the table the MLP model performed almost the same. It must be noted that the parameters are all the same all the three cases and the recommended default parameters are used for the momentum and training time.

5.5 Knowledge Representation

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. 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. Generally, the preconditions are logically ANDed together, 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. In an attempt to come up with significant rules, PART was run on the accident dataset with different number of attributes. Ten fold cross validation has been used for testing and the minimum number of objects in a leaf is set to twenty. Domain experts have been consulted intensively in evaluating the significance of the rules. As a result the rules generated based on the ten attributes: DriverAge, Educationalbackground, DrivingExprience, LicenceGrade, VehicleType, VehiclePeriodOfService, WeatherCondition, LightCondition, VehicleMovemet and DriverDegreeOfResponsibility(dependent). The performance of the algorithm in generating the rules is 88.45% of accuracy. Figure 5.4 below, shows some of the one hundred seventy rules generated by the PART algorithm. The rest can be found at Appendix 3.

```
PART decision list
------------------

LicenceGrade = Third AND
DriverAge = B AND
VehicleType = Taxi&MiniBus: Extreme (221.0)

LicenceGrade = Second AND
VehiclePeriodOfService = E AND
DriverAge = D: Slight (89.0/1.0)

LicenceGrade = Fourth AND
DriverAge = C: Moderate (253.0/16.0)

DrivingExprience = E AND
VehiclePeriodOfService = D: Moderate (327.0/29.0)

DriverAge = B AND
VehiclePeriodOfService = D AND
LicenceGrade = Third: Extreme (42.0)

VehiclePeriodOfService = D AND
DriverAge = C AND
LicenceGrade = Second: Moderate (98.0/9.0)

DrivingExprience = E AND
DriverAge = C: Moderate (213.0/29.0)

VehicleType = Automobil AND
VehiclePeriodOfService = E: Slight (161.0/15.0)
```

Figure 5.4: Partial output from PART rule generator.

The rules above indicate that driver's degree of responsibility varies with the different possible combination of their experience, age, educational level, license grade, and environmental conditions. For instance, it can be seen from the rules that there are more scenarios for most taxi drivers to be extremely responsible for accidents than other drivers in the same age and experience categories.

In the previous research by Tibebe Beshah (2005), RTA severity was classified as 'Property Damage', 'Fatal Injury', 'Serious Injury' and 'Simple Injury' using the data mining software KnowledeStudio and its decision tree algorithm KnowledgeSeeker. He obtained an accuracy of around 87%. In doing so he also extracted hidden patterns in the

RTA data which are mainly composed of different accident causes and environmental factors that influence accident severity for the whole accident type. The work did not attempt to significantly investigate the driver properties. However, as it was well explained in Chapter 1, section 1.4 of this paper, drivers contribute to above 80% RTA in the capital Addis Ababa and around 81% of which is 'Car-Hit-Pedestrain' accident type. Hence, the current research further investigates the RTA dataset with the objective to predict driver's responsibility for an accident in which pedestrians are involved. As it can be seen from the extracted rules more driver attributes are incorporated in the current study than the previous one. It is the researcher's belief that the various stake-holders and experts in the area can make significant use of the rules in making appropriate decisions to reduce accidents that occur due to driver's fault. To achieve this intensive investigation of the rules by the domain experts is mandatory.

5.6 Model Evaluation

So far, different models have been built to determine the driver's degree of responsibility and identify important factors influencing the different degrees of responsibility. Significant rules have also been extracted from the best decision tree as discussed above. Evaluating the different models in order to select one for prediction is also another crucial step in data mining methodology. As it has been described in the previous section evaluation of models can be done by applying different measure of performance.

As indicated at the beginning of this chapter one of the ways to use the WEKA data mining tool is to apply several different learners and compare their performance in order to choose one for prediction. To this end, WEKA's experimenter has been utilized to

implement ID3 and j48 decision tree learners and the back propagation algorithm for the MLP and to automatically analyze the models. The experiment type is a ten fold cross-validation and model parameters have been set as follows.

5.6.1 Setting modeling parameters

In setting parameters for the J48 decision tree algorithm the default confidence factor of 0.25 was used, tree pruning was allowed; the number of folds was set to ten and the minimum number of instances to twenty. In training the MLP the auto build parameter has been set to add and connect up hidden layers in the network. The number of hidden layer nodes is set according to the formula: (number of attributes + number of classes)/2, where the number of classes is 3 in all the experiments and the number of attributes varies from 10 to 13.. The learning rate and momentum are the recommended defaults 0.3 and 0.2 respectively. The "nominaltobinary" parameter is set to automatically convert the nominal attributes to binary. The training which is the number of epochs to train through is set 300. Other less important default parameters are left unchanged. 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 to at Appendix 4 and Appendix 5 for the J48 and MLP respectively.

Figure 5.5 below, shows the WEKA's experiment environment window after the data is fed and the three algorithms ID3, j48 and MLP have been selected.

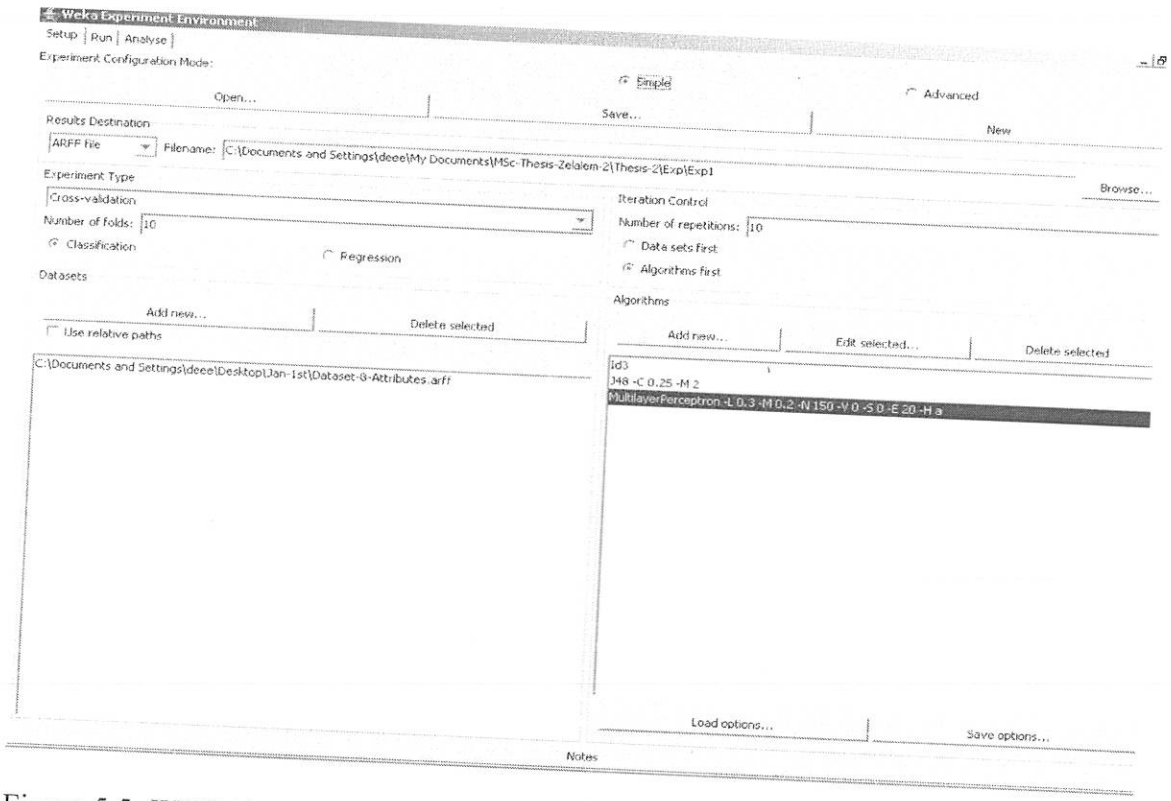


Figure 5.5: WEKA's Experiment Environment.

After choosing the algorithms and setting parameters the accident data set with ten attributes (used already for rule extraction) has been fed to the experimenter. The experiment has been run on the data to build the three models all at once.

Then the three models have been analyzed by the experimenter automatically. The MLP has been used as a base model to compare the other two against. The performance tests have been based on different performance measures and are discussed as follows:

The first analysis was performed by testing the models based on the percent-correct statistic and the following result output has been obtained.

```
Analysing:  Percent_correct
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Date:       1/8/09 8:10 AM


Dataset                 (3) function | (1) trees (2) trees
                        ----------------------------------
Dataset-10-attributes-wek(100)   91.84 |   87.11 *   88.24 *
                        ----------------------------------
                             (v/ /*) |   (0/0/1)   (0/0/1)
Skipped:

Key:

(1) trees.J48 '-C 0.25 -M 2' -217733168393644444
(2) trees.Id3 '' -2693678647096322561
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 300 -V 0 -S 0 -E 20 -H a' 572250905027665169
```

Figure 5.6: WEKA's percent-correct test output

The three models are compared based on the percent correct statistic: this is selected by default. The three methods are displayed horizontally, numbered (1), (2), and (3), as the heading of the table. The labels for the columns are repeated at the bottom, ID3, J48 and function for the MLP. The value in parentheses at the beginning of the dataset row (100) is the number of experimental runs: 10 times 10-fold cross-validation. The percentage correct for the three schemes is shown in Figure 87.11% for method 1, 88.24% for method 2, and 91.84% for method 3(the base method). The symbol placed beside a result indicates that it is statistically better (v) or worse (*) than the baseline scheme; in this case the function (MLP) at the specified significance level (0.05, or 5%). Here, method 1 and method 2 are significantly worse than method 1, because its success rate is followed

20. Usama Fayyad et al (1996). From Data Mining to knowledgeDiscovery in Databases. Available URL:http://citeseer.nj.nec.com/fayyad96from.html

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 (2008). Road Traffic Safety.Available at URL: http://en.wikipedia.org/wiki/Road_traffic_safety

Appendices

Appendix 1: Decision Tree Algorithm

Algorithm : (Generate_decision_tree) Generate a decision tree from the given training data.

Input: The training samples, *samples*, represented by discrete valued attributes; the set of candidate attributes, *attribute-list*.

Output: A decision tree.

Method:

```
1) create a node N;
2) if samples are all of the same class, C then
3)     return N as a leaf node labeled with the class C;
4) if attribute-list is empty then
5)     return N as a leaf node labeled with the most common class in samples; // majority voting
6) select test-attribute, the attribute among attribute-list with the highest information gain;
7) label node N with test-attribute;
8) for each known value a_i of test-attribute // partition the samples
9)     grow a branch from node N for the condition test-attribute=a_i;
10)    let s_i be the set of samples in samples for which test-attribute=a_i; // a partition
11)    if s_i is empty then
12)        attach a leaf labeled with the most common class in samples;
13)    else attach the node returned by Generate_decision_tree(s_i, attribute-list - test-attribute);
```

Appendix 2: The Backpropagation Algorithm

Algorithm : (Backpropagation) Neural network learning for classification, using the backpropagation algorithm.

Input: The training samples, *samples*; the learning rate, *l*; a multilayer feed-forward network, *network*.

Output: A neural network trained to classify the samples.

Method:

```
1) Initialize all weights and biases in network;
2) while terminating condition is not satisfied {
3)     for each training sample X in samples {
4)         // Propagate the inputs forward:
5)         for each hidden or output layer unit j
6)             I_j = Σ_i w_ij O_i + θ_j; //compute the net input of unit j
7)         for each hidden or output layer unit j
8)             O_j = 1/(1+e^{-I_j}); // compute the output of each unit j
9)         // Backpropagate the errors:
10)        for each unit j in the output layer
11)            Err_j = O_j(1 − O_j)(T_j − O_j); // compute the error
12)        for each unit j in the hidden layers
13)            Err_j = O_j(1 − O_j) Σ_k Err_k w_jk; // compute the error
14)        for each weight w_ij in network {
15)            Δw_ij = (l) Err_j O_i; // weight increment
16)            w_ij = w_ij + Δw_ij; } // weight update
17)        for each bias θ_j in network {
18)            Δθ_j = (l) Err_j; // bias increment
19)            θ_j = θ_j − Δθ_j; } // bias update
20)    } }
```

Appendix 3: PART Extracted Rules

VehiclePeriodOfService = D AND
DriverAge = C AND
VehicleType = Automobil AND
LicenceGrade = Second: Moderate (96.0/3.0)

VehiclePeriodOfService = E AND
LicenceGrade = Second AND
DriverAge = D: Slight (89.0/1.0)

DriverAge = C AND
VehicleType = Taxi&MiniBus AND
LicenceGrade = Third AND
DrivingExprience = D: Extreme (28.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
DrivingExprience = E AND
VehicleType = StationW: Moderate (24.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
VehicleMovemet = MovingStraight AND
LicenceGrade = Fourth: Moderate (39.0/4.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
DrivingExprience = E AND
VehicleType = Automobil: Moderate (16.0)

DriverAge = B AND
VehiclePeriodOfService = D AND
DrivingExprience = C: Extreme (52.0)

LicenceGrade = Special AND
VehicleMovemet = MovingStraight: Moderate (9.0/1.0)

LicenceGrade = Second AND
DrivingExprience = D AND
VehicleType = Automobil: Slight (101.0)

VehiclePeriodOfService = D AND
LicenceGrade = Third AND
DrivingExprience = B: Extreme (26.0)

VehiclePeriodOfService = D AND
DrivingExprience = E AND
LicenceGrade = Fifth AND
WeatherCondition = GoodAir: Moderate (43.0)

VehiclePeriodOfService = D AND

LicenceGrade = Third AND
DrivingExprience = D AND
DriverAge = B: Extreme (24.0)

VehiclePeriodOfService = D AND
LicenceGrade = Third AND
DriverAge = A: Extreme (31.0/6.0)

VehiclePeriodOfService = D AND
DrivingExprience = E AND
VehicleType = Automobil: Moderate (60.0/1.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
VehicleMovemet = Turning: Moderate (27.0/1.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
LicenceGrade = Second AND
DrivingExprience = D: Moderate (14.0/2.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
LicenceGrade = Second AND
DrivingExprience = E: Moderate (7.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
WeatherCondition = GoodAir AND
LicenceGrade = Third AND
Educationalbackground = SSS: Extreme (17.0/3.0)

DriverAge = C AND
VehicleType = Automobil AND
VehiclePeriodOfService = C AND
DrivingExprience = E: Moderate (27.0)

DriverAge = C AND
VehicleType = Automobil AND
VehiclePeriodOfService = A: Moderate (24.0/1.0)

VehiclePeriodOfService = E AND
DriverAge = B AND
VehicleType = Automobil: Slight (33.0/1.0)

LicenceGrade = Second AND
DrivingExprience = D AND
VehicleType = Pickup10Qu: Slight (18.0)

VehiclePeriodOfService = D AND
VehicleType = Taxi&MiniBus AND

Educationalbackground = SSS: Extreme (18.0/5.0)

VehiclePeriodOfService = D AND
VehicleType = Taxi&MiniBus AND
Educationalbackground = Junior: Extreme (13.0)

LicenceGrade = Second AND
DrivingExprience = D AND
VehicleMovemet = MovingStraight AND
VehiclePeriodOfService = E: Slight (10.0)

LicenceGrade = Second AND
DrivingExprience = D AND
VehicleMovemet = Turning: Slight (7.0)

VehiclePeriodOfService = D AND
VehicleType = Taxi&MiniBus AND
Educationalbackground = Elementary: Extreme (12.0)

DriverAge = C AND
VehiclePeriodOfService = D AND
DrivingExprience = D: Moderate (23.0/4.0)

LicenceGrade = Second AND
Educationalbackground = SSS AND
VehiclePeriodOfService = E: Slight (74.0/3.0)

DriverAge = C AND
DrivingExprience = E AND
LicenceGrade = Fifth: Moderate (41.0)

DriverAge = C AND
VehicleType = Automobil AND
VehiclePeriodOfService = C AND
VehicleMovemet = MovingStraight AND
Educationalbackground = Above_SSS: Moderate (10.0)

DriverAge = C AND
VehicleType = Automobil AND
DrivingExprience = E: Moderate (38.0/3.0)

LicenceGrade = Fourth AND
DriverAge = C AND
VehicleType = Bus: Moderate (23.0/1.0)

VehiclePeriodOfService = D AND
VehicleType = Truck AND
LicenceGrade = Fourth: Moderate (34.0)

DriverAge = B AND
VehiclePeriodOfService = B AND

VehicleMovemet = MovingStraight: Extreme (10.0/2.0)

DriverAge = B AND
VehiclePeriodOfService = D AND
VehicleType = Truck: Extreme (8.0)

DriverAge = B AND
VehiclePeriodOfService = E AND
VehicleType = Bus: Slight (23.0)

VehicleType = Taxi&MiniBus AND
DriverAge = A AND
LicenceGrade = Third: Extreme (16.0)

DriverAge = B AND
VehiclePeriodOfService = D AND
VehicleType = Pickup10Qu: Extreme (7.0)

LicenceGrade = Second AND
DrivingExprience = D AND
VehiclePeriodOfService = C: Slight (7.0)

VehiclePeriodOfService = D AND
VehicleType = Bus AND
WeatherCondition = Hot: Moderate (10.0)

VehiclePeriodOfService = E AND
VehicleType = Automobil AND
LicenceGrade = Third AND
DrivingExprience = D: Slight (30.0)

LicenceGrade = Third AND
VehiclePeriodOfService = D AND
DrivingExprience = D: Extreme (11.0)

DriverAge = B AND
VehiclePeriodOfService = E AND
VehicleMovemet = Turning AND
DrivingExprience = D: Slight (6.0)

DriverAge = B AND
LicenceGrade = Third AND
VehicleMovemet = MovingStraight AND
VehiclePeriodOfService = C: Extreme (20.0/1.0)

DriverAge = B AND
VehiclePeriodOfService = E AND
VehicleMovemet = Turning AND
DrivingExprience = C: Slight (5.0)

DriverAge = B AND
DrivingExprience = NoLicence AND
Educationalbackground = SSS: Slight (10.0/2.0)

VehicleType = Taxi&MiniBus AND
DrivingExprience = C: Extreme (22.0)

LicenceGrade = Fourth AND
DriverAge = C: Moderate (59.0/3.0)

DrivingExprience = C AND
VehicleType = Automobil AND
Educationalbackground = SSS: Slight (27.0/3.0)

VehiclePeriodOfService = D AND
VehicleType = Truck: Moderate (21.0/4.0)

VehiclePeriodOfService = D AND
VehicleType = Bus AND
DrivingExprience = D: Moderate (11.0/1.0)

VehiclePeriodOfService = D AND
VehicleType = Bus AND
WeatherCondition = GoodAir AND
DrivingExprience = E: Moderate (4.0/1.0)

VehiclePeriodOfService = D AND
VehicleType = Bus: Extreme (16.0/5.0)

DrivingExprience = C AND
VehiclePeriodOfService = E AND
DriverAge = C: Slight (13.0/1.0)

DrivingExprience = C AND
VehiclePeriodOfService = E AND
DriverAge = A AND
Educationalbackground = Above_SSS: Slight (7.0)

DrivingExprience = C AND
VehiclePeriodOfService = A AND
VehicleType = Truck: Extreme (5.0)

DrivingExprience = C AND
VehiclePeriodOfService = A AND
VehicleType = Automobil: Slight (4.0)

DrivingExprience = C AND
VehiclePeriodOfService = A AND
VehicleType = Pickup10Qu: Slight (3.0/1.0)

DrivingExprience = C AND

VehiclePeriodOfService = E AND
DriverAge = B AND
LicenceGrade = Fourth: Slight (18.0/2.0)

DrivingExprience = E AND
LicenceGrade = Fourth AND
LightCondition = DayLight AND
DriverAge = A: Moderate (30.0)

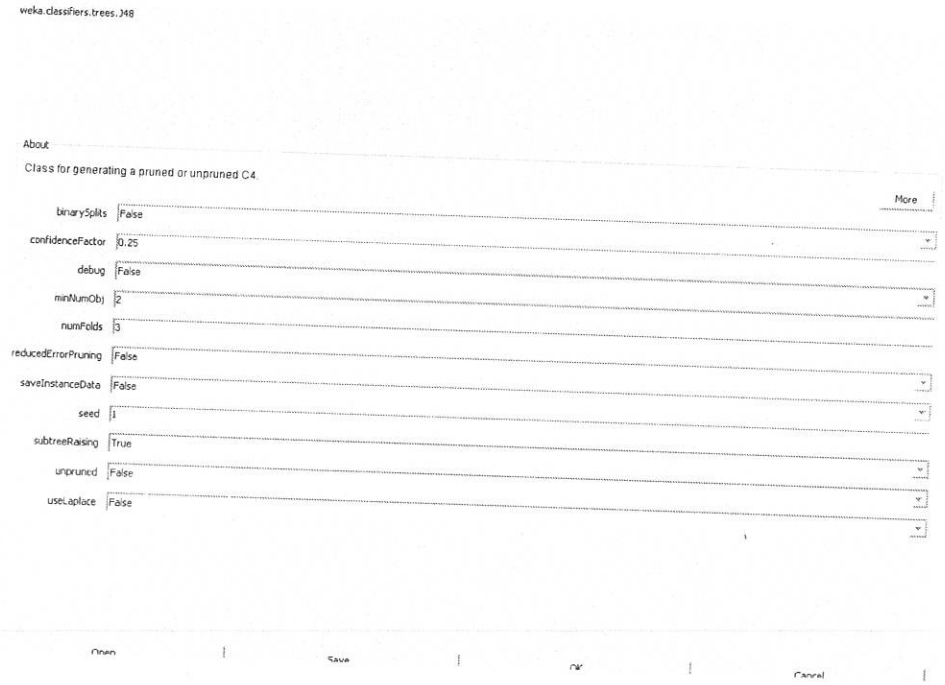
VehiclePeriodOfService = E AND
VehicleType = Automobil AND
DrivingExprience = E AND
WeatherCondition = GoodAir: Slight (28.0/1.0)

DrivingExprience = E AND
LicenceGrade = Fifth AND
VehicleType = Truck: Moderate (31.0/3.0)

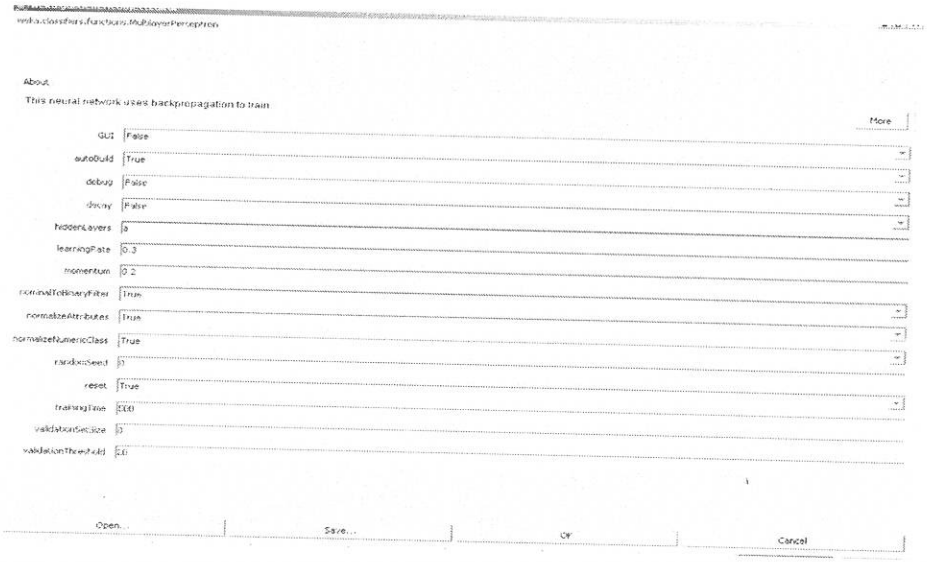
VehiclePeriodOfService = E AND
LicenceGrade = Second AND
DrivingExprience = B: Slight (18.0/1.0)

DriverAge = B AND
VehiclePeriodOfService = D AND
VehicleMovemet = MovingStraight AND
DrivingExprience = D AND
LicenceGrade = Fifth: Extreme (5.0)

Appendix 4: WEKA's object editor window (J48)

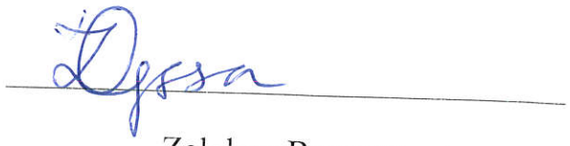


Appendix 5: WEKA's object editor window (MLP)



Declaration

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



Zelalem Regassa

January 2009

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

Dr. Kumudha Raimond

January 2009