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

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT THE PRIORITIZATION OF DANGEROUS CRASH LOCATIONS
THE CASE OF ADDIS ABABA TRAFFIC OFFICE

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

BY

HALELUYA KIFLU ADANE

JANUARY, 2009

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

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT THE PRIORITIZATION OF DANGEROUS CRASH LOCATIONS THE CASE OF ADDIS ABABA TRAFFIC OFFICE

BY

HALELUYA KIFLU ADANE

JANUARY, 2009

NAME AND SIGNATURE OF MEMBERS OF THE EXAMINING BOARD

_____ _____

_____ _____

_____ _____

DEDICATION

I WOULD LIKE TO DEDICATE THIS PAPER TO W/RO WUDINESH YIIRDAW, WHO I HAVE ALWAYS BEEN AND WILL ALSO, BE THINKING ABOUT.

ACKNOWLEDGMENT

I would like to extend my gratitude to my advisor, Dr. Rahel Bekele who gave me constructive ideas on my work. I owe special thanks to Ato Mesfin Getachew, for his comments on the research document.

My especial thanks goes to my beloved family for their support and encouragement, particularly to Amini, I am also grateful to my sisters and brothers who always stands by me. My father your unfailing support and belief in me have been a constant source of strength and enthusiasm, God Bless your spirit.

I would also like to express my appreciation to the wonderful friends I have had at the Department of Information Science most of whose my group members. Special thanks to Teshome Alemu who always devoted his time on my favorite. Unforgettable old friends, Amlake Admasu and Fasil Tsegaye, I appreciate your true friendship.

Last but not least, I would like to thank the New Abyssinia College staff; the Dean, Ato Nega Kebedom, in particular who encouraged me, without whose good will I would not have accomplished this research.

Thanks to the Omnipotent, Omnipresence and Almighty, who is with me on every step.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AATO	Addis Ababa Traffic Office
ARFF	Attribute-Relation File Format (ARFF)
CSV	Comma Separated Value
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
DMKD	Data Mining and Knowledge Discovery
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
OLAP	On-Line Analytical Processing
RTA	Road Traffic Accident

ABSTRACT

The development of automotive industry, the slowly improvement of the roadways and the behavior of the traffic participants increased the number of the road accidents. Traffic accident results in loss of life, human injury and financial prejudices. Road Traffic Safety which is currently one of the highest priorities may be affected by a number of factors. One important group of bottlenecks in traffic safety are dangerous accident locations. Addis Ababa is a city where the number of traffic accident is increasing from time to time. Identification of high crash locations in the city will either protect the accident occurrences or minimize the rate of damage to be caused.

This paper reports on the findings of a research that had the objective to prioritize high crash locations and predict exposure of the society on different crash locations. The study used data obtained from the Addis Ababa Traffic Office. In order to prioritize high crash locations different data mining tools and techniques were used.

The data mining process in this research is divided into two major phases. During the first phase data was prepared and formatted into the appropriate format for the respective data mining software to be used (Weka 3.5.8). The second phase contains model building for prioritization using decision tree classification. In the classification phase J4.8 algorithm were employed to generate rules.

Traffic accident locations were prioritized based on their degree and number of fatality occurrence. The patterns obtained from the J-48 algorithm separated these locations as: death, severe injury, and light injury.

The outcome of the study is highly useful for the Traffic police office on developing traffic management system; for the society, drivers and pedestrians, on pre-informing the accident occurrences on those black spots. It also provides valuable information for making decisions effectively for road safety investment projects.

CHAPTER ONE

Introduction

1.1 Background

Nowadays data can be stored in many different types. The steady and amazing growth of computers and information technology even hampered the availability of data on different location with various formats. The abundance of data, coupled with the need for powerful data analysis tools has been described as data rich but information poor society (Han & Kamber, 2001).

This fast growth and tremendous amount of data, collected and stored in large and numerous databases need a powerful tool to elicit useful information. The tool helps to get benefit from the collected data, by identifying relevant and useful information. Data mining is one of the solutions to analyze huge amount of data and turn such data in to useful information and knowledge (Han & Kamber, 2001).

Defining a scientific discipline is always a controversial task; researchers often disagree about the precise range and limits of their field of study. Due to this, there exists variety of definition for the term data mining. For the sake of this research, the researcher selected the following working definitions:

Data mining 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 (Hand and Mannila, 2001).

Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database (Giudici, 2003).

With this respect one of the areas where vast amount of data stored is traffic offices in which a traffic accident record is stored. So a work on the field of data mining is needed to convert such data to useful information.

Overview of Traffic accident

The traffic accident is usually caused by the failure of one or more multitude of factors. Variety of the factors are mentioned by the New York personal injury lawyers (New York Personal injury Lawyers, 2007). Some of them are: the safety condition of the vehicle, the safety condition of the road and its environment, the safe behavior of the driver, and the road design and layout etc.

Reducing the number of accidents therefore need an integrated approach on the above factors. This can be carried out by improving the active and passive safety of cars, by sensitizing and enforcing car drivers to be more careful and by reducing the hazardousness of roads etc. The latter involves identifying sites with large accident risk so as to make the necessary infrastructure and traffic management system changes for reducing the riskiness of the site.

Traffic accident is one of the main causes of the fatalities in Addis Ababa. As such, it has become one of the causes that aggravate the death of many people. Too many people are really dying due to car accidents. It's an unacceptable situation in Ethiopia besides hunger and all kinds of terrible diseases (Capua, 2006).

According to the traffic office report of 1999 E. C, there were 2224 accidents in relation to human beings out of which about 347 were dead and others were injured. The report on the traffic office also showed 19, 573 cars were damaged on that year. This resulted in enormous amount of property loss. The estimated loss of property according to the office was 23,094,667 birr. (Addis Ababa Traffic Office, 1999).

This shows that, there is significant amount of loss on resources of the country, specially the loss of human resource.

1.2 Statement of problem & its importance

The Addis Ababa city traffic office has several branches based on the 10 Sub cities ("Kifle Ketema") of the city. The central traffic office performs various activities; some of them are improving the flow of traffic, remove obstacles on traffic movement and speed up motor vehicle traffic of the city etc. One of the tasks in the main traffic office is keeping record of accidents in the city. In order to do this, there is a database built using Microsoft Access. There are 47 fields on a table, and currently the database contains more than 12,000 records. The traffic office uses this data for report generation, traffic police distribution and research purpose; the latter one belongs to the function of this thesis.

The Addis Ababa Traffic Office (AATO) has such vast amount of data with variety of fields. On the other hand it assigns traffic polices to different locations of the city based on simple report generated using SQL queries, such as maximum crash of last month.

It is difficult to assign traffic polices or traffic management system on all cites or part of a city. This is because traffic accidents do not occur equally on all cites or part of a city (World Bank group, nd.). In addition to this, there is insufficient number of human power in the office. In order to efficiently mange limited human resource, or develop a new traffic system such as speed camera on certain places there should be detailed investigation of past records on traffic accidents which select and rank dangerous traffic locations.

An attempt was done on investigating severity of accident but not pointing hazardous locations. Because of this the office can not plan and implement allocation of its staff with greater efficiency on those locations, and distribution of human resource (traffic polices) too.

Lack of traffic system worsened the problem observed, and aggravated the existence and wide spread of traffic accident. Drivers cause accidents specially while there is no traffic system or a person to watch over them. Due to this the number of criminal drivers and the accident rate is increasing from time to time. The alarming increase on accident records in the office is an evidence for this.

Not only increasing in number of criminal drivers, but also no means to pre-inform the drivers/pedestrians to take care of their movement around those black spots. The drivers can not

have detailed understanding of those areas and can not take more care while they are driving on such type of locations than locations which are less hazardous.

Besides this the society does not have awareness to keep himself/herself safe from an accident on his/her regular movement. Different groups of the society should know their exposure on various areas of the city, especially on places where they regularly move. Such as Students around their school, adults around their institution, children and old ages around their residence and churches respectively. By studying and identifying danger crash locations or black spots, provision of information will be facilitated.

In addition to this, if a traffic police is assigned to look after the traffic movement on right location, the number of criminal drivers who escapes from that location after committing an accident on pedestrians as well as other properties will be eliminated if not minimized with certain ratio.

Thus it is worth to prioritize dangerous crash locations and design a predictive model that supports the traffic office on providing clear picture of dangerous traffic locations on the city, and factors that make these locations more sever than the others. Pin pointing these areas will help the traffic office to give emphasis on human resource management and to look over safe traffic flow too.

So this study gave background information for the abovementioned problems. Especially it has threefold advantage at the end. It will help the traffic office, the drivers and the pedestrians. Having identified those dangerous locations, the traffic office can apply it on its traffic

management system and decision making. The driver can predict the nature of pedestrians and their exposure to accident. The pedestrians can get a background information on which age group was exposed on the given location and their probability of safety or exposure to danger at a location.

At the end, the result of this study with other researchers' work will actually strengthen alleviation of problems on traffic management system, particularly at the central office and generally all over Addis Ababa city.

1.2.1 Research questions

Based on the foregoing, the research attempts to answer the following questions.

- **i.** What are the mechanisms that the traffic police use to identify dangerous traffic locations?
- **ii.** What are the factors contribute to the hazardousness of a site based on the road and its environment?
- **iii.** Which data mining algorithm helps to select and rank dangerous crash locations?
- **iv.** Which specific sites need more concentration of traffic police assignment and traffic management systems?

1.3 Objectives of the study

The general objective of the research is to investigate the potential applicability of data mining technology in developing a model that can prioritize dangerous traffic locations at the city of Addis Ababa.

In order to achieve the general objective indicated above, the research will have the following specific objectives:

- To develop an understanding of the application domain through review of relevant documents and interview with domain experts from AATO.
- To recognize the occurrence rate of traffic accident location as a whole and particularly the Ethiopian context.
- To understand the possible opportunities of data mining application on selecting and prioritizing Traffic Accident Locations.
- To assess and choose among the various data mining software which are more appropriate to the selecting, ranking and predicting accident location.
- To build and validate models using the selected tools and technologies.
- To show the results and make recommendations on what should be done next.

1.4 Methodology

The following methods were employed in conducting the study.

1.4.1 Data Source

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

The data used for the study was obtained from the central office of the Addis Ababa Traffic Office. The office keeps records of accident occurrences on a centralized manner using MS Access database. The records are collected from 1998 to 2000 E.C. and stored on the database using Amharic Language on a single table which has 47 fields. At the time the study started the table consisted of 12,441 records from seven sub cities.

The database contains data on different perspectives, such as data in relation to the road where the accident occurs, the driver who owns the car, the condition of the car, the environment on specific accident occurrence such as weather and air condition, and data on the injured person such as his age, work place, and others are included (See Annex one for detail).

While understanding the data source, the researcher learned that there were more attributes in the traffic accident dataset than actually required for this analysis. The attributes can be characterized as redundant or non-variant.

1.4.2 Tool selection

There are various tools available for data mining, such as Knowledge Studio, Weka, Xlminer, SA, SPSS, STATA and others. Among those tools, Weka is selected as a tool. In addition Visual Basic 6 is also used on the study. These are discussed as follows.

Weka was adopted for undertaking the experiment for prioritizing dangerous crash locations based on Decision Tree algorithm. Weka is Collection of machine-learning algorithms with an open-source Java package that supports Numeric, nominal, string, and date format files for processing data on several methods.

The algorithms of WEKA can either be applied directly to a dataset or called from one's own Java code. WEKA is also well suited for developing new Machine Learning schemes. WEKA is open source software issued under the GNU General Public License. It incorporates an association rule learner. In addition to the learning schemes, WEKA also comprises several tools that can be used for datasets preprocessing (Palous, nd).

While Weka was used for Decision Tree model building, Visual Basic 6 is used for interface design on prediction model building, which was based on the patterns obtained from the result of the decision tree model.

The reason why Weka and Visual Basic 6 were selected for the study was: familiarity of the researcher in earlier applications, their appropriateness for the problem domain, and Weka’s free availability.

1.4.3 Data collection and preprocessing

For data mining application, the appropriate data should be made available first, and provided to the data mining tool. The data collection and preprocesses for dangerous crash location identification, prioritization are discussed as follows.

The availability of the database by itself doesn't fulfill everything for the study, it needs further processes. Some of these include data analysis, feature selection, preprocessing, and training of the model through decision tree.

As mentioned earlier all the data was in Amharic language, hence it needed language transformation. The data was extracted and put from MS Access to MS excel, and important fields and their values were translated to English language using subject experts on the area. After it was translated to English language, the basic activities performed were attribute selection, data cleaning and preprocessing.

The necessary attributes were selected assisted by domain experts in the area, and then the dataset was cleaned by filling missed values manually; removing tuples from the dataset, and aggregation of data values were performed. Then the cleaned data was converted to .csv format and finally to .arff format which was an input to the tool. By doing this the appropriate input for developing the decision tree model on Weka was achieved.

1.4.4 Training and Model Building

Training the decision tree model helped to get a pattern on accident locations that were classified as hazardous sites based on certain variable inputs. The study was done by successive training and model building until the final decision tree model was achieved. First it was trained with the entire data set using J48 classifier for decision tree modeling technique; however the model obtained was not satisfactory. Then several attempts were done using 10 fold cross validation and partitioning the data set 66% for training and 34%testing.

While evaluating a model, its accuracy and the number of leaves generated were taken into consideration. The one with higher number of accuracy and least number of leaves was selected.

Finally based on assessment and evaluation of decision tree models the best model that was found with an interesting pattern was selected. The rules generated from the decision tree model were supported by domain experts too. Then by using the leaves/classes of the decision tree, a predictive model was built for those locations based on the given age group. And that provided a good ground for result interpretation and recommendation.

1.5 Scope and limitation of the study

The scope of the research was limited to assessing the possible application of data mining technology at Addis Ababa Traffic Office; it was limited to identifying and ranking crash locations associated with road accidents on human.

The study was also limited to examine the potentials of data mining techniques in developing classification and prediction model by the use of decision tree, and developed a prototype, in support of traffic control activities. The identification of its cause was left to the subject expert.

Traffic road accident is caused due to several factors, one of them is road location, other cases of Road Traffic Accident were not considered in this study, such as driver quality and quality of the vehicle.

1.6 Thesis organization

This thesis report is organized under five chapters. The first chapter deals with the general overview of the study including background on data mining and traffic accident, statement of the problem & its importance, the general and specific objectives, methodology of the research and finally the limitation.

The second chapter is devoted to literature review of data mining technology and Knowledge Discovery in Databases. Emphasis is given on different data mining models and modeling methods. Assessment of the application of data mining on Road Traffic Accident is also discussed in detail.

Under chapter three attempts had been made to review literatures to know about trends in road transport, road types and junctions, ranking methods of black spots on traffic system and road safety. Accidents on road junctions and exposed age groups are also assessed.

Chapter four, deals with experimentation. It includes source identification, data collection and different preprocessing steps. Tasks like data collection, cleaning, attribute selection, and others are reported in detail. The next tasks on the chapter are experimentations; this basically comprises training, building and validation of the models in addition to analysis and interpretation of the results.

Finally the last chapter presents conclusions of the study. Furthermore recommendations of the research on AATO and other possible study areas related to Road Traffic Accident are presented in the final chapter.

Chapter Two

Data Mining Technology

2.1 Introduction

This chapter presents review of related literatures on data mining. This includes what data mining is, different phases evolved on data mining, variety of models, decision trees and various activities involved in it. The chapter also discussed studies done on data mining and other research works specially concentrated on traffic accidents. This enables to easily understand the focus of the study and application of data mining for the problem domain.

2.2 Data Mining

Nowadays organizations are collecting larger and larger amounts of data. The capabilities of both generating and collecting data have been increasing in the last several decades. Contributing factors include the wide spread use of bar codes, the computerization of many business, scientific and governmental transactions, advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data.

As the collected data grows in organizations, there was a need for new automated methods that can enable them to convert the collected data into useful information and knowledge. The fast growing, tremendous amount of data, collected and stored in large and numerous databases, has far

exceeded our human capability for comprehension with out powerful tools. To get benefit from the collected data, there should be a way to identify relevant and useful information (Han and Kamber, 2001).

According to Witten and Frank (2000), there is a gap between the generation of data and our understanding of it. As the volume of data increases, the proportion of it that people understand decreases. This is because people cannot elicit certain patterns with out a tool to help them. Kamber and Han (2001) proposes data mining as an appropriate solution for the above problem which is the automated extraction of patterns representing knowledge stored in large databases and data warehouses.

So data mining which was evolved from the need for the extraction of useful information and knowledge is an application specific issue, and various techniques have been developed to solve different application problems. Some examples are mining association rules, classification, clustering and sequential patterns. Data mining tools are used for performing data analysis on the databases or data warehouses to find data patterns, this enables institutions to improve their business decision. As (Two Crows Corporation, 1999) stated, Data mining gives managers a powerful new tool to improve the job they are doing.

Therefore the purpose of data mining is to elicit patterns in retrospective data, not substituting specialists, so this knowledge can be applied to problem solving and decision making by analysts and mangers. The data mining system can automatically find and show new patterns that will lead us to fresh insight. Examples of this might be discriminating among subsets of the data with differing characteristics, and inferring probabilities of future events from historical data.

Data mining is not a one time activity, instead an evolutionary process that incorporates several steps. Thearling (2003) describes data mining techniques as a result of a long process of development in research areas such as Statistics, Artificial Intelligence, and Machine Learning. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to forthcoming and proactive information delivery. Data mining is used in organizations where massive data collection, powerful multiprocessor computers, and data mining algorithms are available.

2.3 Data Mining and knowledge discovery in databases (KDD)

Finding useful information from a database had been given several names serving synonymously. Some of the names include, knowledge transaction, information discovery, information harvesting, data archeology, data pattern processing, data dredging, and data mining (Han & Kamber, 2001)

In some data mining literatures, the term data mining is used as a synonym for Knowledge Discovery in Databases (KDD). For example, Seidman (2001) uses the term data mining interchangeably with KDD. On the other hand, Han and Kamber (2001) and Two Crows Corporation (1999) view KDD as the overall process of discovering useful knowledge from data, and data mining as a particular but essential step in the process of knowledge discovery in databases.

Treating Data Mining as a phase in knowledge discovery is also supported by Pal and Jain (2005). According to them Knowledge discovery (KD) is a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from large collections of

data. Where as Data Mining is one of the crucial KD steps. DM is concerned with the actual extraction of knowledge from data; in contrast to the KD process is concerned with many other activities.

However People often use the terms DM, KD and DMKD as synonymous. it is sometimes difficult to distinguish the usage of these two terms due to this reason some give emphasis on DM others on KD the rest treat the two together as Data mining and knowledge discovery (DMKD), Pal and Jain (2005).

Due to the popularity of the term data mining than the longer term Knowledge Discovery in Databases, Han and Kamber (2001) favored to adapt the broader view of data mining functionality, and use the term data mining.

In this study, however, the terms data mining and knowledge discovery process both refer to the entire process from data collection through pattern identification and reporting. So these terms, data mining and knowledge discovery process will be used interchangeably. The reasons behind this adoption are, being consistent with major data mining projects, use the corresponding experiences, and avoid any confusion between the two phrases, ‘data mining’ and ‘knowledge discovery in databases’.

2.3.1 Knowledge discovery process

To see certain patterns on the KDD process one has to pass several steps. The KDD process consists of an iterative sequence of many steps. The steps are essential to ensure that useful knowledge is derived from the data. Several professionals discussed these steps varying from

five to nine phases. According to Pal and Jain (2005) the steps of DMKD consists six major phases, these are discussed briefly as follows.

The first step of DMKM is **understanding the problem domain.** In this step one works closely with domain experts to define the problem and determine the project goals, identify key people, and learn about current solutions to the problem. It involves learning domain-specific terminology. This step may also include initial selection of potential DM tools.

After understanding the domain the second step follows, which is **Understanding the data.** This step includes collection of sample data and deciding which data will be needed, including its format and size. If background knowledge exists, some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DMKD goals. At this phase data need to be checked for completeness, redundancy, missing values, plausibility of attribute values, and the like.

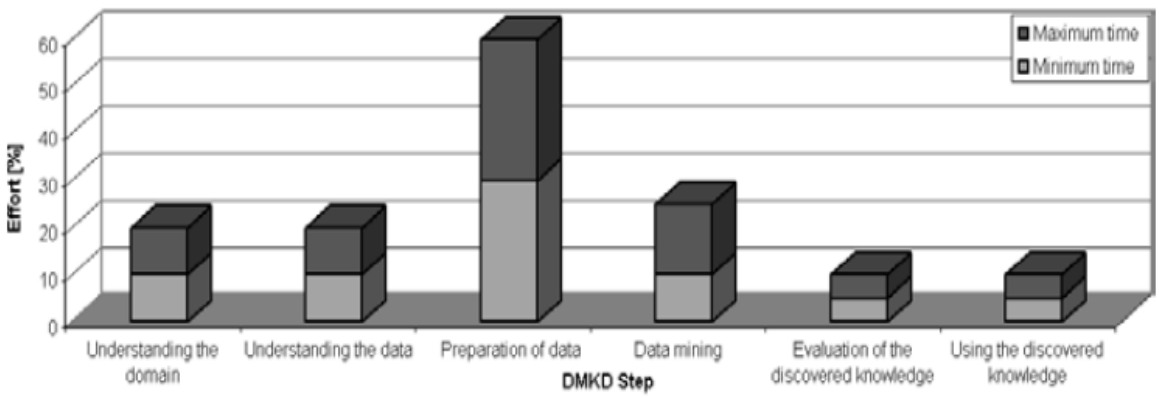
The third step is **Preparation of the data.** This is the key step on which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire project effort. In this step, we decide which data will be used as input to the data mining tools in the next step. It may involve sampling of data, running correlation and significance tests, cleaning data like checking for completeness of data records and correcting for noise. The cleaned data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say by means of discretization), and by summarization of data (data granularization). The result is new data records, meeting specific input requirements for the planned, to-be-used DM tools.

The fourth step is **performing data mining.** Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools, and selection of the new ones if needed. Data mining tools include many types of algorithms, such as rough and fuzzy sets, Bayesian methods, evolutionary computing, Machine Learning, neural networks, clustering, and preprocessing techniques. This step involves the use of several DM tools on data prepared in the third step.

Based on the result of data mining, **Evaluation of the discovered knowledge** is done which belongs to the fifth step. This step includes understanding the results by owners of the data that check whether the new information is truly novel and interesting, and checking the impact of the discovered knowledge. The entire DMKD process may be revisited to identify which alternative actions could be taken to improve the results.

The final step evolve on DMKM is **using the discovered knowledge.** This step is entirely in the hands of the owners of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains within an organization. A plan to monitor the implementation of the discovered knowledge should be created and the entire project documented.

The Time elapsed on the above steps may vary from one phase to the other. According to Pal and Jain (2005), it is the preprocessing step that uses maximum time on KDD processes; the comparison of time between these phases is presented on figure 2.1.



Pal and Jain (2005)

Figure 2.1: A comparative Efforts and time spent on KMDM process

The KDD process can involve significant iteration and can contain loops between any two steps. The feedback loops are necessary since any changes and decisions made in one of the steps can result in changes in later steps Pal and Jain (2005).

2.3.2 Tasks of data mining

The above phases of data mining can be encountered in variety of tasks. Data mining as a filed of study contains several tasks to be performed in it. In order to have a clear picture it is convenient to categorize it into different types of *tasks*, corresponding to different objectives for the person who is analyzing the data.

Different authors categorize data mining tasks to various divisions; Berry and Linoff (2004) divide them in to six tasks: Classification, Estimation, Prediction, Affinity grouping, Clustering, and Description and profiling".

The first three are all examples of directed data mining, where the goal is to find the value of a particular target variable. Affinity grouping and clustering are undirected tasks where the goal is to uncover structure in data without respect to a particular target variable. Profiling is a descriptive task that may be either directed or undirected.

2.4 Data mining and data warehousing

Data in the real world generally reside in dozens of disparate systems. In response to integrating the disparate systems, data warehousing is needed which is the process of bringing diverse data together from throughout an organization for decision-support purposes (Berry and Lineoff, 1997).

Nowadays, many businesses are trying to transform their data from various data sources by consolidating it from different systems into a single accessible source of information called a data warehouse. The single repository, from distributed information is suitable for data analysis (Anoop Singhal, 2007).

There is some real benefit if your data is already part of a data warehouse. The data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart. The problems of cleansing data for a data warehouse and for data mining are very similar. 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. (Two Crows, 2005)

However a data warehouse is not a requirement for data mining. Setting up a large data warehouse can be an enormous task, sometimes taking years and costing millions of dollars. It is possible to mine data from one or more operational or transactional databases by simply extracting it into a read-only database. This new database functions as a type of data mart.

Figure 2.2: Data mining data mart extracted from operational databases

The other concept that makes Data warehouse different is, it's isolated from other databases. It is maintained and stored separate from other databases; this makes it special and selective for data mining. According to Anoop Singhal (2007) three of the key features that make data warehouse selective are its Subject Oriented, integrated and time variant nature.

2.5 Data mining models

A model according to Berry and Linoff (2004) is an explanation or description of how something works that reflects reality well enough that it can be used to make inferences about the real world. Data mining is all about creating models. As shown in Figure 2.3, models take a set of inputs and produce an output. The data used to create the model is called a *model set*. The model set has three components training, validation and tst sets: The *training set* is used to build a set of models, the *validation set* is used to choose the best model of these, and the *test set* is used to determine how the model performs on unseen data.

Data mining techniques can be used to make three kinds of models for three kinds of tasks: descriptive profiling, directed profiling, and prediction. Descriptive profiling produces descriptive models. On the other hand, both *directed profiling* and *prediction* have a goal in mind when the model is being built. The difference between them has to do with time frames. In profiling models, the target is from the same time frame as the input. In predictive models, the target is from a later time frame. *Prediction* means finding patterns in data from one period that are capable of explaining outcomes in a later period.

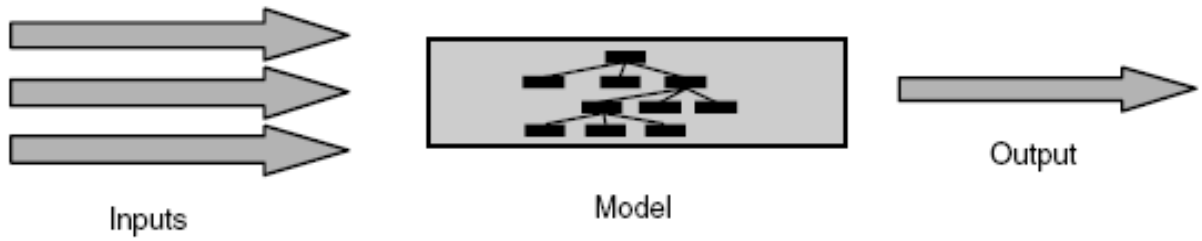


Figure 2.3: data mining model

2.5.1 Predictive model

Predictive modeling predicts the value of a particular attribute. A predictive modeling involves using a modeling database to discover a relation ship between two variables, a dependent and explanatory variable. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. By contrast, descriptive techniques, such as clustering, are sometimes referred to as unsupervised learning because there is no

already-known result to guide the algorithms (Two Crows Corporation, 1999). The following figure shows how the prediction model looks like (Thearling 2003).

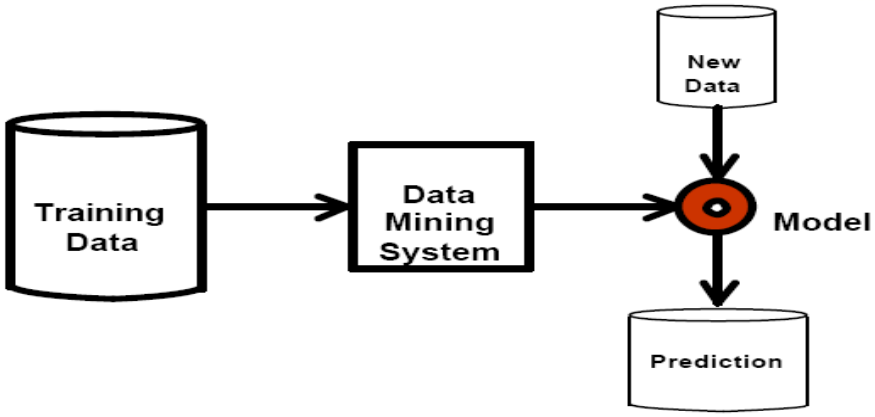


Figure 2.4 Predictive Modeling

The two most common predictive modeling tasks are called classification and regression. If the label is discrete (containing a fixed set of values), the task is called classification. If the label is a continuous value, the task is called regression.

Classification is the task of assigning a discrete label value to an unlabeled record. In doing so, records are divided into predefined groups. Han and Kamber (2001) describe classification as the process of finding a set of models (or functions) that describe and distinguishing data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. According to them, classification is a two-step process. In the first step, a model is built describing a predefined set of data classes or concepts. In the second step, the model is used for classification.

Regression is similar to classification, except that the label is not discrete. According to Two Crows Corporation (1999), regression uses existing values to forecast what other values will be. In the simplest case, regression uses standard statistical techniques, such as linear regression. For example, predicting salary or the price of stock is a regression, whereas predicting whether the salary is in a given range or whether a stock will go up or down is a classification task. Regression uses standard statistical techniques such as linear regression.

Hans and Kamber (2001) assert decision trees as one of the most commonly used algorithms used to perform classification. In line with Two Crows Corporation (1999), decision trees can also be used to perform regression. Based on this we do have two main types of decision trees: classification trees and regression trees. Decision trees, which are used to predict categorical variables, are called classification trees because they place instances in categories or classes. Decision trees which are used to predict continuous variables are called *regression trees.*

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification (Thearling, 2003).

In order to construct decision tree one has to pass several steps basically two of the major phases are tree-growing (splitting) phase, followed by a pruning phase. These phases will be discussed under decision tree development and pruning part of this chapter.

2.5.2 Descriptive modeling

While the goal of a predictive modeling is to predict the value of one column based on the value of other columns, the goal in descriptive modeling lies on discovering patterns and segments of the data. Descriptive modeling is unsupervised tasks. Unsupervised tasks provide insight to the data as a whole by showing patterns and segments that behave similarly. Two of the most common descriptive modeling tasks are association and clustering.

The task of association is to determine rules of implication between data attributes A and B, so that A implies B. Associations are used to find affinity groupings between attributes, such as: discover what items are usually purchased with others. The classic affinity grouping is market basket analysis, predicting the frequency with which certain items are purchased together.

Clustering algorithms segment the data into groups of records, or clusters that have similar characteristics. Clusters help to minimize data complexity. For example, it is probably easier to design a different marketing plan for each group of targeted customer clusters than to design a specific marketing plan for each million of individual customers (Gerritsen, 1999).

2.6 Modeling methods in data mining

In data mining environment there exist varieties of modeling methods, according to Balac (2003) these modeling methods are categorized as follows:

1. Decision Tree induction
2. Regression tree induction

3. Multivariate Regression Tree
4. Clustering
    - K-Means, EM, Cobweb
5. Neural Network
    - Back propagation
    - Recurrent
6. Others

2.6.1 Decision trees

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute which compares an attribute value with a constant, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions (Han and Kamber 2001).

A decision tree according to Berry & Linoff (2004) is a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decisions.

According to Witten and Frank (2005), Decision Tree is a “divide-and-conquer” approach to the problem of learning from a set of independent instances.

A decision tree contains variety of components; it composes of the root, branch and the leaf nodes Han and Kamber (2001). The first component is the top decision node, or root node, which specifies a test to be carried out. The results of this test cause the tree to split into two or more

branches, each representing one of the possible answers. The branches may be binary, if split in to two, or multi-way tree if more than two nodes exist, Two Crows (1999).

The nodes produced by the initial split (child nodes) are then split in the same manner as the root node. Once again, all input fields are considered as candidate splitters, even fields already used for splits (Berry & Linoff, 2004). Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. Leaf nodes give a classification that applies to all instances that reach the leaf or set of classifications, or a probability distribution over all possible classifications.

Decision trees are a way of representing a series of rules that lead to a class or value. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf. This navigation on the decision tree help to assign a value or class.

There are two main types of decision trees, as mentioned by Two Crows Corporation (1999). The first category of decision trees which are used to predict categorical variables are called *classification trees* because they place instances in categories or classes. The second categorization of decision trees is the one that is used to predict continuous variables and it is named regression trees (discussed on earlier topics).

2.6.1.2 Decision tree building

Decision trees are built through recursive partitioning, an iterative process of splitting the data up into partitions and then splitting it up some more (Berry & Linoff, 2000). After putting the entire training set in one box, the algorithm uses the data and makes split appropriately. According to Han and Kamber (2001), the iterative process is an induction algorithm that follows the following basic strategy in a top-down, recursive, divide-and-conquer manner.

- The tree starts as a single node representing the training samples.
- If the samples are all of the same class, then the node becomes a leaf and is labeled with that class.
- Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes the \test" or \decision" attribute at the node. In this version of the algorithm, all attributes are categorical, i.e., discrete-valued. Continuous-valued attributes must be discretized.
- A branch is created for each known value of the test attribute, and the samples are partitioned accordingly.
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node's descendents.
- The recursive partitioning stops only when any one of the following conditions is true:

- All samples for a given node belong to the same class, or
- There are no remaining attributes on which the samples may be further partitioned. In this case, majority voting is employed. This involves converting the given node into a leaf and labeling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored; or
- There are no samples for the branch. In this case, a leaf is created with the majority class in samples.

In order to get a decision tree model, one has to pass an exhaustive steps mentioned above. The decision tree process starts by taking each input variable in turn and measuring the increase in purity that result from every split suggested by that variable. After trying all the input variables, the one that yields the best split is used for the initial split, creating two or more children. If no split is possible (because there are too few records) or if no split makes an improvement, then the algorithm is finished with that node and the node become a leaf node. Otherwise, the algorithm performs the split and repeats itself on each of the children and is called a *recursive algorithm*.

2.5.1.3 Attribute selection

Which attribute must come first on the root node, second, etc…? This is the basic question on performing decision tree. The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split Han & Kamber (2003). The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute

minimizes the information needed to classify the samples in the resulting partitions and selects the least randomness or \impurity" in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

According to Han and Kamber(2003), the information gain of an attribute can be computed as:

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{V \in \text{Values}(A)} Sv / S \, (\text{Entropy}(Sv))$$

Where: S is the total sample

Sv/S is sample values of selected attributes over the entire number of samples.

Entropy (Sv) can be calculated as:

$$\text{Entropy } Sv = -[P\oplus \log_2 P\oplus] - [P\Theta \log_2 P\Theta]$$

Where: $P\oplus$ a positive probability that sample Sv belongs to the class

$P\Theta$ a negative probability that sample Sv belongs to the class

2.5.1.4 Pruning Decision Tree

In the absence of a stopping criterion, a tree model could grow until each node contains identical observations in terms of values or levels of the dependent variable. It is necessary to stop the growth of the tree at a reasonable dimension. A good decision tree, according to Giudici (2003),

composes of two properties; the final tree is expected to be both parsimonious and accurate. The first property implies that the tree has a small number of leaves, so that the predictive rule can be easily interpreted. The second property implies a large number of leaves that are maximally pure. The final choice is bound to be a compromise between the two opposing strategies.

According to Whitten & Frank (2001), Pruning methods are categorized in to two Pre-Pruning and Post-Pruning. By building the complete tree and pruning it afterward we are adopting a strategy of *post-pruning* (sometimes called *backward pruning*) rather than *pre-pruning* (or *forward pruning*). Pre-pruning would involve trying to decide during the tree-building process when to stop developing sub trees—quite an attractive prospect because that would avoid all the work of developing sub trees only to throw them away afterward.

Post-pruning remove branches from a "fully grown" tree. Here the node is pruned by removing its branches Han & Kambe (2003).

Post-pruning requires more computation, however, it seem to offer some advantages. For example, situations occur in which two attributes individually seem to have nothing to contribute but are powerful predictors when combined—a sort of combination-lock effect in which the correct combination of the two attribute values is very informative whereas the attributes taken individually are not. Most decision tree builders post-prune. The two techniques are however interleaved for a combined approach.

2.6 Application of data mining technologies

Data mining is increasingly popular because of the substantial contribution it can make. The relevancy of data mining has gained recognition by information intensive industries that maintain large databases. Many organizations use this technology to derive critical information from large and bulky databases (Seidman, 2001).

The two main reasons that data mining attracted organization are, make sense of their past and to make predictions about their future. Ultimately, they use the information taken from this computerized fortuneteller to make decisions about their futures. Data mining is an activity that offers business advantages, as well as solutions to some escalating problems associated with exploring the knowledge embedded within corporate databases.

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. A company can determine characteristics of good customers (profiling), similarity of customers, customers who have not bought a product, customers left a company, and customers who are at risk for leaving.

Insurance companies and stock exchanges are interested in applying this technology to reduce fraud. Medical applications are other fruitful areas of data mining applications such as predicting the effectiveness of surgical procedures, medical tests or medications. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances

that might be candidates for development as agents for the treatments of disease, Two Crows Corporation (1999).

Data mining is used in companies active in the financial markets to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons. (Two Crows Corporation, 1999).

So, while companies face bulky data that can not be processed easily, Data mining offers great promise in helping them uncover patterns hidden in their data. The following researches also witness the essential of data miming on their study.

Gashaw M. (2004), applied datamining inorder to support cutomer Insolvency Prediction at Ethiopian Telecommunication Corporation. On his study he used the neural network backpropagation algorithm, and MATLAB 6.5 neural network toolbox. His model can classify customers, well in advance, as potentially solvent or insolvent. Such a model can be used for decision making process on such areas. Then the organization can prdict what type of customer it is treating.

Denekew A. (2003), has conducted his research on assessing the application of data mining techniques to support Customer Relationship Management activities at Ethiopian Airlines. He derived new attributes from the existing original attributes, defining new attributes and then prepared new data tables. He used K-means clustering algorithm to segment individual customer

records into clusters with similar behaviors. He then classified those customers using J4.8 and J4.8 PART algorithms to develop a model that assigned new customer records into the corresponding segments. His prototype enabled to classify a new customer into one of the customer clusters.

Yoseph (2004) on his thesis also showed the application of data mining using decision trees on predictive model for Central Statistics Authority. He first identified variables as dependent and independent from experts' knowledge. He used neural network model for predictive assessment and KnowledgeStudio studio for decision tree modelling. Finally based on the result of the decision tree, he applied Visual Basic 6 to develop the predictive model. Yoseph also reported With 1,152 records he got 98.26% of accuracy.

In addition to the above areas, the data mining technology has many potential applications in other sectors such as transportation to determine the distribution schedules among outlets, and analyze loading patterns. As far as the transportation sector is concerned, the traffic movement and road safety has drawn a considerable attention of a good number of data miners. Since the research at hand is exploring the potential application of data mining in the traffic accident, the next sub-section reviews the possible applications of data mining in this area.

2.6.1 Application of Data mining in traffic accident

To implement data mining on certain specific area, there should be a record with considerable amount. The existence of bulky data on traffic accident encouraged many researchers to focus on the area. Varieties of studies are conducted using data mining on traffic accident. Generally the

importance of data mining can be categorized in to three different groups, Chong et al (2004). These applications of data mining in the area of road transport are:

- Traffic density analysis; measurement and investigation of traffic accident volumes.
- Traffic accident analysis; identifying determinant factors in Road Traffic Accidents and other related issues
- Injury severity analysis; modeling and predicting the severity of injuries resulting from traffic accidents.

Tibebe (2004), explored the possibility of data mining on traffic accident severity analysis. He considered 1165 records in order to asses the fatality of accident, such as death serious, light or property loss. He showed maximum accuracy of 87.47% while considering both 10 and 7 attributes of the AATO data on accident. His final data for experiment consists of attributes related to road, driver, vehicle, and environmental

The same study by Tibebe (nd) upgraded and published as, rule mining based on classifying Road Traffic Accidents using adaptive regression trees, also showed the same result. However he found that two types of accidents namely *denying pedestrian* and over speeding are contributing a lot to fatalities and serious injuries. Next to the accident type, *Accident_cause* was the important attribute in splitting the tree. Again, taking into consideration the two important accident types, denying pedestrian priority, driving with alcohol and over speeding are determinant cause of accident types.

Similarly Chong M. et al. from *Oklahoma State University, USA,* studied the National Automotive Sampling System, using automobile accident data from 1995 to 2000 of National Automotive Sampling System (NASS). They investigated the performance of neural networks and decision trees applied to predict drivers' injury severity in head-on front impact point collisions. They reported that the decision tree approach outperformed neural networks on injury (including fatality) classification. They categorized fatality based on possible injury, nonincapacitating injury, incapacitating injury, and fatal injury. Finally they reported their model showed good performance on fatal and non-fatal injury than other classes. They also gone through a further study of predicting fatal and non-fatal injury in relation to speed however the predication doesn't provide enough information on the actual speed, since speed for 67.68% of the data records' was unknown.

The assessment of severity of injury on traffic accident is also conducted by another person named Hossain M. in Thailand, the Asian institute of technology. His study was applied on 316, 868 records between the year 1999 to 2003. With an accuracy level of 76% he showed motorcyclists are the first cause of fatal accidents in Thailand.

This shows that traffic accident is diversified in several countries and data mining attracted researchers towards alleviating this problem.

CHAPTER THREE

Road Traffic Accident

3.1 Introduction

This chapter presents review of literatures made on trends in road transport, road and traffic system, road junctions, method of ranking accidents and accidents on Ethiopian roads, specifically on the city of Addis Ababa. This enables to easily understand the focus of the study and application of data mining for the problem domain.

3.2 Road Traffic Accident

The term Road Traffic Accident (RTA) is interchangeably used with various terms on different literature. Some of the phrases used to describe accidents as mentioned by (Wiki) include: auto accident, car crash, car smash, car wreck, fender bender, motor vehicle accident (MVA), personal injury collision (PIC), road accident, road traffic collision (RTC), road traffic incident (RTI), smash-up, and traffic collision.

According to SafeCarGuard(2004) a traffic accident is defined as any vehicle accident occurring on a public highway (i.e. originating on, terminating on, or involving a vehicle partially on the highway). These accidents therefore include collisions between vehicles and animals, vehicles and pedestrians, or vehicles and fixed obstacles. Single vehicle accidents, in which one vehicle alone (and no other road user) was involved, are included.

Traffic accidents occur in all roads but the degree of occurrence or severity may vary. At certain sites, the level of risk will be higher than the general level of risk in surrounding areas. Some locations are called dangerous crash locations or black spots than others. An accident blackspot is a term used in road safety management to denote a place where accidents are concentrated. Without a precise localization of road accidents, road administrators are not able to find, and effectively care for the accident blackspots on their road network. Inaccurate localizations mean misguided identifications and result in the loss of financial means and time. The effective evaluation of implemented countermeasures can also be influenced (Mikulík & Holló 2007).

Since the advent of vehicles, the number of Road Traffic Accident has risen proportionate to the number of vehicles manufactured. There are almost 885,000 deaths from RTA annually in the world (WHO, 1995). The term killed (in an RTA) is defined as any person who was killed outright or who died within 30 days as a result of accident (WHO, 1984).

In the last consecutive decades road safety has become a major concern for many governments. According to the National Highway Traffic Safety Administration (NHTSA): In 2004, there were an estimated 6,181,000 police-reported traffic crashes, in which 42,636 people were killed and 2,788,000 people were injured; 4,281,000 crashes involved property damage only. According to WHO (2004), the social cost of traffic safety in highly developed countries amounts to approximately 2% of their annual GDP. The identification of sites which are more dangerous than others (black spots) can help in better scheduling road safety policies (Tom Brijs, 2001).

The aggravation on traffic accident still exists today. As World Health Organization presented in (2008), More than 3000 people die on the world's roads every day. The process of rapid and unplanned urbanization has resulted in an unprecedented revolution in the growth of motor vehicles worldwide. The alarming increase in morbidity and mortality owing to Road Traffic Accidents over the past few decades is a matter of great concern globally. According to Ganveer & Tiwari, (2005), currently motor vehicle accidents rank ninth in order of disease burden and are projected to be ranked third in the year 2020. Worldwide, the number of people killed in road traffic crashes each year is estimated at almost 1.2 million, while the number injured could be as high as 50 million. The rate will rise to 2 million by 2020 unless new safety measures are taken, making road traffic injuries the third largest cause of death and disability.

The extent of death on traffic accident belongs to the top killers in Ethiopia. As the Ethiopian Traffic Enforcement Agency (nd), stated Traffic accidents are the leading cause of death; second only to AIDS. Severity of traffic accident is not limited to Ethiopia. Traffic deaths in Africa are mentioned as: The number “1” killer of African Children. The rate of death in Africa children is 100 times more compared to USA death rate for children. From 11% of the world's traffic deaths Africa constitutes with 4% of the world’s cars. The traffic accident all over the world is estimated to increase by 80% before 2020.

3.3 Contributing Factors to Road Traffic Accidents

A "factor" is a circumstance contributing to a result. Without this factor, the result would not exist but the factor alone is an element that, by itself, cannot produce the result. The term

"contributing factor" is meaningless if this definition is accepted. A factor must be contributing if it is present otherwise it is not a factor. The term "primary factor" is sometimes used by experts to indicate a factor that was strong in its contribution to the accident. This is misleading as there can be no one factor more important than any other if all factors must have been present to produce a result. No factor can be secondary, or less important than another if all are required for the result. Like the links of a chain, all must be present and none is more or less necessary than the other (Harris, nd).

Cause of road accident therefore constitutes a multitude factors. Many writers agree on the division of factors that cause traffic accidents. According to Sabel (2005) these factors are categorized in to three groups: the road environment, the condition of vehicles using the road system, and the skills, concentration and physical state of road users. This division is also supported by Guerts & wets, (2003).

Each of the above divisions is subject to a variety of modifiers. For instance, road factors include, but are not limited to lighting, view obstructions, recognizability, signs, signals, surface character, dimensions and protective devices. All factors are subject to modification by outside influences such as the road surface that becomes slick from rainfall, and road damage including pot holes. Modifying each of the listed road factors are weather, lighting, roadside devices, activities, visibility, road markings, surface deposits, damage, deterioration and age.

The demands of the road environment also vary due to factors such as traffic flow rates, geometric features of the road and type of road. Drivers normally adapt their performance level to the demands of the road system. A crash occurs when the driver's performance level is

insufficient to meet the performance demands of the road environment. Most of the time, driver capabilities exceed performance demands. Black spots are points of peak performance demand. Engineering improvements in the road network lower performance demands on the driver. This increases the safety margin between the driver's performance level and the performance demands of the road environment, and reduces the probability of a crash (Tom Bridge, 2001)

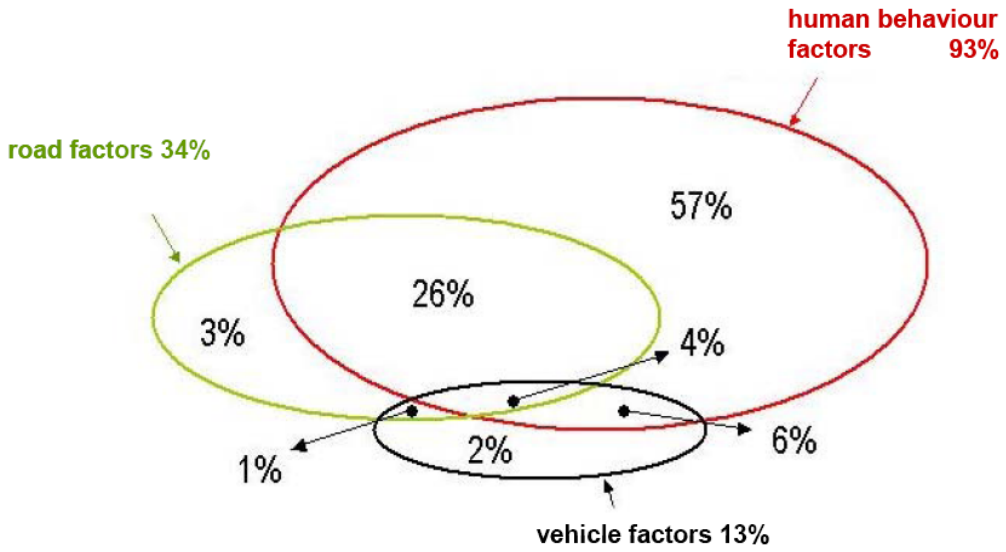
The second factor belongs to conditions related to vehicles. This factor includes equipment condition, view obstructions, distractions, instruments, signaling devices, control sensation, comfort, automatic controls and devices, weight, performance, dimensions and stability. Vehicle speed, as a factor, must exist. If neither vehicle had any speed, there could not have been a collision.

The Human factors, as a third, are without doubt the most complex and difficult to isolate because they are almost very temporary in nature. What existed at the time of the accident may not exist moments later. Consider sensory capabilities, knowledge, judgment, attitude and alertness, health, driving skill, age, customs, habits, weight, strength, fiddling with technical devices, talking on a cell phone, talking with passengers, eating or grooming in the car, dealing with children or pets in the back seat, or attempting to retrieve dropped items, and freedom of movement. Of these, the emotional factors are the greatest variable attributes and the most difficult to identify. Driver impairment by tiredness, illness, alcohol or other drugs, both legal and illegal are also included (Oshman & Mirisola, 2007).

Finally remote condition factors can also be considered when dealing with cause analysis although they are seldom of significance to the investigator on a single accident case. Remote

condition factors may involve a changing cultural climate in which the factors and their modifiers form. This includes moral influences, religion, beliefs, legal influence and values on vehicle designers, highway engineers, drivers and pedestrians.

Due to these reasons, Road Traffic Accidents (RTA) are complicated to analyze as they cross the boundaries of engineering, geography, and human behavior (*Sabel, 2005*). The human factor plays the lion's share of road accident causes, and then follows road condition, finally the car condition contribute its own impact. However one factor can not be treated isolated from the other. The following diagram taken from Mikulík (2007) shows the proportion of occurrences of these contributing factors.



(Source: PIARC Road Safety Manual, 2003)

Figure 3.1 Contributing Factors to TRA

Reducing the number of traffic accidents therefore requires an integrated approach, known as shared responsibility. For example, this can be carried out by improving the active and passive safety of cars, by raising awareness and forcing car drivers to be more careful and by reducing the hazardous condition of roads. The last option involves identifying sites presenting important accident risks so as to make the infrastructure changes needed to reduce the risks at the site. Furthermore, methods that can measure and produce comparable results regarding the risk of each site are of special interest for designing new roads or for enforcing rules (Tom Bridge, 2001).

3.4 Accident types and occurrences

The occurrence of accident or cause has different meaning from the cause of other incidences. For example a cause in the field of law is different from a cause in the rest of the world. As mentioned by the (experts group, nd) cause in the eyes of the courts is an issue of policy and not an instrument of factual analysis. The issue for the court is whether, as a policy decision, a defendant should be held liable for injuries or damages. In a traffic accident case is the resolution of that issue depends on whether the crash was a foreseeable result of a defendant's negligence or determined act.

Whether the accident is caused by one or several of the factors motioned above, it is possible to Classifying them according to their common features into several groups facilitates and defines the investigation process. Therefore, groups of accidents according to their occurrence and the types of collision are identified and used in accident analysis. The following list represents the

accident types used in the Czech Road Accident Typology (Mikulík, 2007), which is based on the Austrian version. These accidents are divided into the following 10 types:

- Single vehicle accidents
- Road accidents of vehicles driving in the same direction on the road section
- Road accidents of oncoming vehicles on the road section
- Road accidents of vehicles entering a junction from the same direction
- Road accidents of vehicles entering a junction from opposite directions
- Road accidents of vehicles entering a junction from neighboring lanes
- Road accidents of vehicles and pedestrians
- Road accidents with standing or parked vehicles
- Road accidents with animals and rail vehicles
- Other road accidents

Most other countries use similar typology of accident types with different number of accident types considered. For example in Germany, the typology contains somewhat less basic accident types – these accident types are discussed as follows:

Driving Accident: An accident in which the driver loses control of the vehicle because he or she was driving at a speed which was inappropriate for the layout, the cross-section, the incline or the conditions of the road, or because he or she did not realize how the road was laid out or that there was a change in the cross-section until it was too late. Driving accidents are not always

“one-party accidents” in which the vehicle leaves the road. They can also result in a collision with other road users.

Turning-off Accident: Turning-off accidents are those triggered by a conflict between a vehicle turning off a road and a road user traveling in the same or the opposite direction. This can happen at junctions and intersections with roads, at field tracks or cycle tracks, or at entrances to properties/car parks.

Turning-into/Crossing Accident: An accident triggered by a conflict between a vehicle which is obliged to give way, turning into a road or crossing the path of other traffic, and a vehicle which has right of way, is referred to as a “turning-into/crossing accident”. This can happen at junctions and intersections with roads, field/cycle tracks and railway crossings, or at entrances to properties/car parks.

Crossing-over Accident: An accident is triggered by a conflict between a pedestrian crossing the road, and a vehicle, provided the vehicle had not just turned off a road. This rule applies irrespective of whether the accident occurred at a site without any special pedestrian-crossing facilities or at a zebra crossing, a light-controlled crossing or similar installation.

Accident caused by Stopping/Parking: An “accident caused by stopping/parking” is an accident triggered by a conflict between a vehicle in moving traffic and a vehicle which is parked (parking) or has stopped (is stopping) on the road. Such accidents include accidents in which the moving traffic conflicted with a vehicle manoeuvring into/out of a parking position. It does not matter whether stopping/parking was permitted.

Accident in longitudinal traffic: An “accident in longitudinal traffic” is an accident triggered by a conflict between road users moving in the same or opposite directions, provided the conflict is not the result of a manoeuvre that corresponds to another accident type.

Other Accidents: These accidents are all those which cannot be assigned to any other accident type. The basic groups are subsequently divided according to the relevant conflict events into more detailed categories (Mikulík, 2007).

The accident types divisions of the AATO are almost similar to the division adopted by Czech Road Accident Typology mentioned earlier.

3.5 Types of Roads and Junctions

It is possible to categorize road types either by the road surface they are made from or by their junction type. According to Harford County Government, (1997), the road surface is divided in to three basic types, Asphalt, Tar and Chip, and earth.

Asphalt is the most common road surface. It is used on all new roads, roads throughout developments, mainline roads, and highways. The typical County residential roadway section is 4" of asphalt placed over 8" of stone. The typical County business roadway section is 5" of asphalt placed over 10" of stone. When the road needs to be resurfaced, 1-1/2 inches of asphalt is placed over the existing road. This surface usually lasts from 15 to 20 years.

Tar & Chip roads are constructed and are found mostly in rural areas of the County. Tar is placed on the road and stone is spread over the tar to provide a wear surface.

Earth roads consist of dirt and gravel. Very little traffic travels on these roads. When traffic increases or at the request of residents, these roads may be converted to tar and chip surfaces. Frequency of maintenance depends upon conditions such as traffic and weather.

The second alternative to categorize roads types is based on their junction. Road junctions on the context of this study are defined as the act or process of joining or the condition of being joined. Or a place where two things join or meet especially a place where two roads or railway routes come together and one terminates (Farlex, Inc Junction, 2008). A junction has a set of conflict points between vehicle paths. A good design should aim at minimizing the severity of potential accidents at these points. If the road has more conflict points on junction the accident rate will be maximized and if minimum the reverse will be true. Figure 3.2 shows conflict points on a road junction:

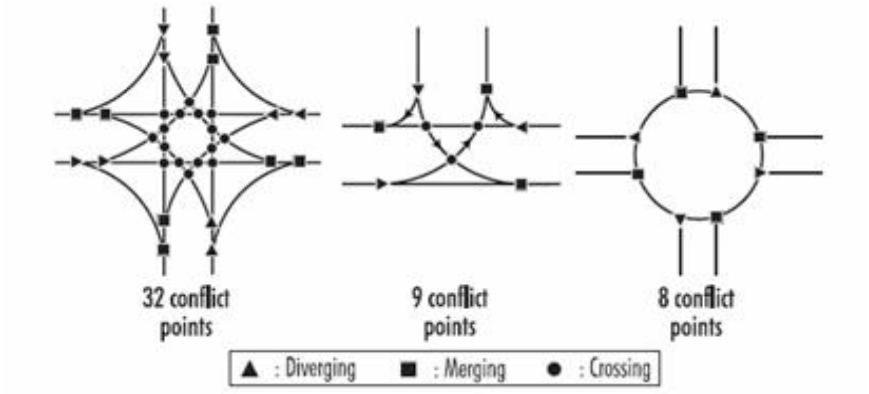


Figure 3.2 Number of conflict points at junctions and roundabouts

Road junction types may also be seen from their shapes and styles such as T-shape, 3-way junction, diamond interchange, traffic cycle, etc...(wiki).

From the user direction of sight junctions can be categorized as closed or open. A closed junction is one where your view is so limited that you have to be at the Give Way lines before you can see clearly into the road you intend to join. It may be parked vehicles, trees, hedges or even pedestrians. Essentially, it's anything that spoils your sight line on your approach to the end of the road.

An open junction is one where your view in to the main road is good on the approach and gradually improves as you get nearer the give way line. A good example of an open junction is a roundabout. A junction can only be considered open if your view is good in all directions on your approach.

3.6 ranking dangerous crash locations

Crashes do not occur equally on all sites. That is why selecting and ranking between locations is needed. Crashes will tend to be concentrated at relatively high-risk locations. Locations that have an abnormally high number of crashes are described as crash concentrated, high hazard, hazardous, hot spot or black spot sites. Sites with potentially hazardous features are sometimes described as grey spots (Guerts K. et al, 2003).

Accident locations may be prioritized based on different criteria; these criteria may vary according to the person who indulge the study. After identifying high crash zones, the next step is to prioritize or rank these high crash zones. High crash zones with higher ranks generally warrant greater attention in the safety enhancement programs. Three methods generally used to

rank high pedestrian crash zones are Crash Density Method, Crash Rate Method, and Crash Score Method (Pulugurtha and Nambisan, 2002, and Guerts & wets, 2003).

Ranking of high crash zones based on the amount of severity of the crashes and the area of high crash zone is called Crash Concentration Method. As described by Virtisen (2002), locations were ranked at first according to their reported number of accidents. Locations with a number exceeding a chosen threshold value were targeted as hot spots. That is selecting the site with the largest expected number of accidents (the 'worst' location) is above a chosen threshold value. However it is difficult to use total number of fatalities or casualties because of the vastly different population sizes and degrees of movement on different location.

The rules for targeting black spots described on the previous paragraph are based on the total number reported accidents at a site. However, road accidents are connected with a number of features, which may be considered to analyze the occurrence in variance between accident frequencies on different locations: the severity of the accident, e.g. fatal, injury or property damage only. Ranking of high crash zones based on the severity of crashes and population in the locality of the high crash zones by giving different weights to different age group of people is called Crash Rate Method. This also has its own side effect because the recording system in most Third World countries is not appropriate. This means that only fatalities are recorded to any reasonable degree of accuracy.

The Crash Score Method is based on normalizing the values to the same scale so as to obtain a score for each method (Pulugurtha, Nambisan and Uddaraju, 2005). Such a normalizing procedure is used to address the challenge of combining disparate components. The individual

scores for each component are normalized using a 0 to 100 scale and then summed to estimate the crash score for the zone.

Besides to the above mentioned methods it is also possible to prioritize sites by Saved accident costs. Locations are ranked depending on the way in which the estimated economical benefits of treating the location exceed a threshold value based on the allocated budget. Therefore, it is necessary to prioritize between sites and safety measures in order to utilize the limited funds as effectively as possible (Virtisen, 2002).

Ranking hazardous sites is an interesting means to get insight in dangerous locations, but with various approaches. Researchers have proposed several alternative methods for targeting and ranking black spots. However, there is no such thing as “the” correct ranking (Guerts & wets, 2003).

So, different researches tried several mechanisms to rank dangerous crash locations. Two of them are presented below.

Geurts K. et al (2003) has done identification and ranking of black spots and sensitivity analysis. They used historical record of Flanders in Belgium. They identified 1014 accident location as dangerous if a location has counted three or more accidents for the past three years. They used counting accidents at a location and multiplying by weight of (1-3-5) rule, developed by Flemish community for light, serious and death injuries respectively. If a site has a value more than 15 it will be considered as dangerous.

Brijs T. et al (nd.) ranked accident location using hierarchical Bayesian method. As they reported in their paper, they developed a hierarchical Bayes procedure for ranking sites. The procedure takes into account not only fatalities, but also injuries (serious and slight) and combines this information by means of a cost function in order to rank the sites. Their model not only ranks the sites but it also takes into account the variability of this ranking. The model suggested a 3-variate Poisson distribution with different covariance for each pair of variables. Their model is limited since it can not take in to consideration an accident produces more than one fatality.

3.7 Road Traffic Accidents at Addis Ababa

Ethiopia with a population of 73 million and 1.5 cars per 1,000 people has 109,000 cars which are involved in 1,800 fatal crashes per year, or one fatal crash for every 60 cars. A car driven by an Ethiopian is 134 times more likely to kill someone than a car driven by an Englishman. If all Ethiopians had cars, and if they had a life expectancy of 60 years, all Ethiopians would be killed in car accidents in less than 60 years." (Capua, 2006)

Ethiopia has one of the world highest road fatalities, with 114 deaths per 10,000 vehicles (AFP, 2008). Ethiopian News Agency (2008) presented the current situation of traffic accident based on news on the World Traffic Accident Victims' Day, which is celebrated at Addis Ababa, November 16, 2008. More than 2,000 people are killed in road accidents every year in Ethiopia while over 8,000 others suffer physical injuries. The report also added Road accidents also incur property damage amounting to over 500 million Birr per annum.

Addis Ababa, the capital city of Ethiopia, takes the lion's share of accidents occurred in the country. Only in 1999 E. C, 2224 accidents resulted in death, severe, and light injuries in the city. Out of 347 were dead, and others were severely and lightly injured. The traffic accident in the city is not limited to human beings but it also caused vast property damage. On the same year there was damage on property of car which was counted to 19, 573. This resulted in enormous amount of property loss. The estimated loss of property according to the estimation of the traffic office amounted to 23,094,667 Ethiopian birr (Addis Ababa Traffic Office, 1999).

CHAPTER FOUR

Experimentation

4.1 Introduction

This section of the thesis presents several activities done during the study including collecting and preprocessing the data for the experiment, running and evaluating the experiment, and finally selecting and testing the model.

4.2. Understanding the data

As mentioned in chapter one the data used for the study was obtained from the Addis Ababa Traffic office. Before the actual preprocessing started, the necessary data was extracted from the database of the institution. The data on MS Access database was in Amharic language.

As mentioned on chapter one the database consists of a single MS Access table named Table1, with 47 fields, 46 mentioned under annex one and a field “Autonumber” that serve as an ID for each row. The database at the time of data retrieval contained 12,441 rerecords of accidents occurred through the years 2005 to 2008.

To have a clear picture of the attributes’ they are categorized in to the following six groups:

- **Driver Related**: such as Driver_Sex, Driver_Age, Accused_Male, Accused_Female, Accused_Age, Educational_Background, License_Grade, etc.

- **Vehicle Related**: such as Types_Of_Vehicle, Plate_Code, and Ownership
- **Accident Place**: such as Sub_City, Kebele, Special_Place and Place
- **Road Related**: such as Road_Division, Road_Direction, Road_Joint, Types_Of_Road, and Road_Conditions
- **Pedestrian Related**: such as Injury_Age, Injury_Occupation, Health_Status, Movement_Of_Pedestrian etc.
- **Accident occurrence**: such as Month, Which_Week, Regno, Date, Date_Of_The_Week etc.

The attributes of the database with their data type and detail description are presented on annex one.

The data extracted from the accident database of the AATO had 76 records with missing values. Most of the missing values were reflected on more than one filed value. The other thing observed was availability of incorrect valued records; there were 12 records in which an age value was assigned for an accident type property damage/accident.

4.3 Data Preprocessing

Data preprocessing include various techniques which undoubtedly improve the overall data mining results. Attribute selection, data cleaning and data aggregation were performed on this step. The data extracted from MS Access was put into Ms Excel and then translated in to English language.

Data Selection

The first data reduction on the study was done by eliminating unnecessary attributes. The attributes were selected before cleaning the data, for there were many irrelevant attributes for the study. Reducing attributes before cleaning saves the time of cleaning irrelevant data too. At this phase attributes that best fit to the objective were selected.

The database of AATO didn't contain economical loss of each crash, due to this the researcher prioritize these crash locations based on their fatality. A consultation was done with the domain experts to select attributes related with road prioritization, three volunteer traffic polices available on the central office were consulted.

The experts on the area propose several fields as contributing factors for traffic accident rate and severity. These factors must be included on the environment because they have their own effect unless removed. Such attributes selected include weather and air conditions.

From the total fields, 11 attributes were selected as environmental factor in relation to road (see table 4.1 for detail). Where as 36 of them were removed, because they were unrelated and unnecessary for this study. The experts suggested the including of sub cities and their kebeles, however two reasons limited the researcher from including them: the first was, the absence of records from 3 sub cities, the second reason was absence of specific road name or road number on the databases.

Prioritizing road location can also have various variables such as economic loss of accidents, accident fatality, and number of accidents counted etc. Since no economical record of each

accident, fatality was used to prioritize locations. For fatality classification the necessary attributes were selected from the following four categories:

- Road related – all attributes under this category annex one.
- Pedestrian related – age category of injured person.
- Accident related – weather condition , Air condition, and year, and
- Place related – month, place and sub city.

The following table summarizes selected attributes and their description.

No	Attribute Category	Attribute Name	Data Type	Description
I	**Accident Place**			
1		Place	Text	Accident occurrence, known name such as institution, school, factory etc.
II	**Road Related**			
2		Road_Division	Text	Division of road such as Single or dual road divisions
3		Road_Direction	Text	Straight or curved directions
4		Road_Joint	Text	Shape of roads junction
5		Types_Of_Road	Text	Road surface moisture (external condition).
6		Road_Conditions	Text	Type of road surface, where it is made from.
IV	**Accident Related**			
8		Month	Numb	Month the accident occurred

			er	
9		Accident_Type	Text	Severity of Accident type
10		Weather_Conditions	Text	Weather condition at accident occurrence
11		Air_Conditions	Text	Sight of environment

Table 1: Attributes selected for ranking dangerous crash locations

Data cleaning

The data cleaning for the TRA was done after selecting the appropriate attribute for the study, because it was wastage of time and unnecessary to do cleaning on attributes not selected for the study.

Out of the total records, 76 records were found filled with values only on three of the fields (accident type, injured age and occupation of injured person), instead of filling values on the remaining eight attributes it was easier and logical to remove 0.63% records. Since the database had option of “unknown” value to be filled on eight of the fields, there were again 18 records that contained unknown vales on eight of the fields. These attributes constitute 0.12% of the total records and also removed from the dataset. Totally 0.75% records were completely removed.

The remaining 99.25% which amounted to 12,347 were kept for further preprocessing. For experimentation the attribute injured age was removed. The reason why age category was

removed from the beginning is it didn't characterize road location instead help to predict exposure group at the end.

There were some values missed on certain fields. One of the alternatives to fill missed values is filling the value manually with most probable value. 3 missed values from accident type were filled with "light injury", 5 missed values from place attribute were filled with "institution", and both are modal value of their attributes. The other values corrected on the dataset were eliminating values from the cells. There were 21 property accidents but had value on injury age column, that were un imaginary to have an accident type of property with age of a person these values were removed from the cell too.

Data transformation and aggregation

The method applied for data transformation is Data aggregation or summarization which is done on some instances to minimize the variation of attribute values on age groups; this is done to bring them in uniform age category.

Dataset Format

Converting the cleaned data to Comma Separated File (.CSV format) was the second to the last step on preprocessing. The .csv format needed either to use a string value of attribute or numeric value. Nominal values were re-phrased to single string. Most of this was done at the time of translation, however only a single value on the field road division was missed at the time of translation; this value was corrected from "one directional" road to "one_directional".

After all preprocessing is over and the file was converted to **.csv** format, Weka either process the .**csv** format itself or a file in the form of Attribute-Relation File Format (**.arff)**. for this study the data was given to the software in .arff format.

```
% header file contains:
        % @relation, @ attribute, and @data
        % attributes with their alternative values
@relation road
@attribute Place {factory, hospital, institution, market, recreational, religious, resident, school}
@attribute RoadDivision {Broken_line, mount_division, one_directional, two_directional, unBroken_line}
@attribute RoadDirection {curved, hill, direct_flat}
@attribute RoadJoint {Circular_joint, Cross_shaped, straight_one, T_shaped, Y_shaped}
@attribute TypesofRoad {fragmented_Asphalt, Smooth_Asphalt, stonish, Earth}
@attribute RoadConditions {dry, wet}
@attribute WeatherConditions {cloudy, cold, good_Air, rainy, semi_rainy}
@attribute AirConditions {day_light, night, sun_rise, sun_set}
@attribute TypesofAccidnet {death, serious, slight}
@data

resident,two_directional,curved,straight_one,fragmented_Asphalt,dry,good_Air,day_light,death
hospital,mount_division,direct_flat,Circular_joint,Smooth_Asphalt,dry,good_Air,day_light,death
institution,mount_division,direct_flat,Circular_joint,Smooth_Asphalt,dry,good_Air,day_light,death
institution,mount_division,direct_flat,Cross_shaped,Smooth_Asphalt,dry,good_Air,day_light,death
religious,unBroken_line,curved,straight_one,Smooth_Asphalt,dry,good_Air,day_light,death
school,Broken_line,direct_flat,straight_one,Smooth_Asphalt,dry,good_Air,day_light, seriousInjury
.
.
.
institution,mount_division,direct_flat,straight_one,Smooth_Asphalt,dry,good_Air,day_light,death
Institution,mount_division,direct_flat,straight_one,Smooth_Asphalt,dry,good_Air,day_light, lightInjury
```

Figure 4.1: Arff files for TRA data

Figure 4.1 shows a data preprocessed and finally given for the tool (Weka) on the experimentation of TRA. It is a sample to show how the data was encoded.

4.4 Running the experiment

Under this portion of the study decision tree building, attribute selection based on information gain, and model building on decision tree were done.

Based on the format mentioned on figure 4.1 the data cleaned and pruned was given to Weka software (version 3.5.8). The total number of instances was 12, 347 and the target class was designed to be "type of accident" on the experiment.

4.4.1 Input for decision tree and tree building

The data organized on .arff format was provided for Weka software. The decision tree model on all experiments were first trained, and then tested with percentage split, and finally with a 10-fold cross validation. No need of retaining separate data for test because Weka has inbuilt methods to train and test the data either by cross-validation (90% for training and 10% for testing) or percentage split of the data (66% for training and the reaming for testing).

Based on the result of the decision tree (such as accuracy obtained and number of leaves generated) and the information gain of attributes obtained on the software, several experiments were conducted. Figure 4.2 shows explorer window of Weka which contains different tabs and buttons for experimenting, list of attributes, and values graphically depicted from the data given.

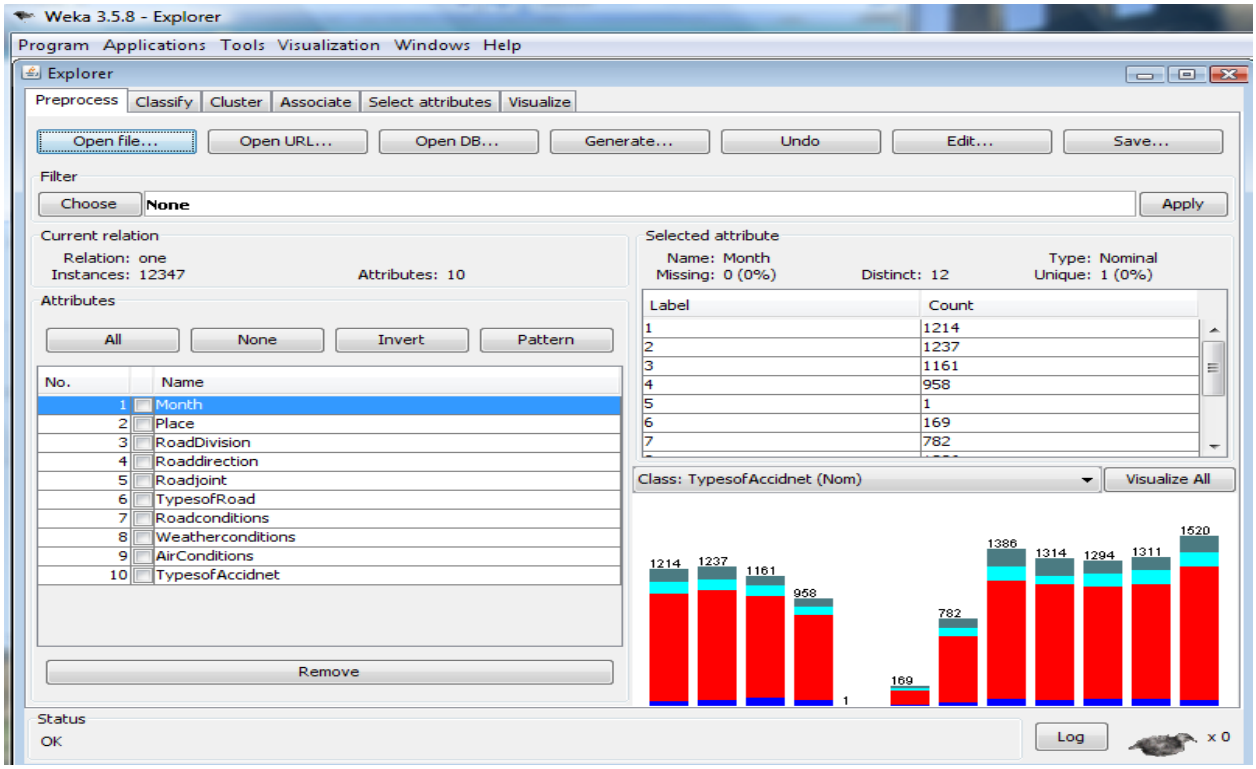


Figure 4.2: selected attributes on Weka explorer

4.4.2 Experimentation for Decision Tree Model Building

Based on the above input data, series of experiments were conducted until the study brought the best output. From several experiments conducted the detail of information of three of the experimentations are presented as follows:

Experiment one

On the first experiment, model building was tried on classification of trees – using J48 classifier. The whole data (i.e. 12347 records) with 10 attributes (on figure 4.2) were provided for the

software. The experiment brought 77.4034% accuracy, but single node of decision tree. The run information is presented on figure 4.3.

```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
: property (12347.0/2790.0)
Number of Leaves  : 1
Size of the tree :      1
Time taken to build model: 0.41 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        9557               77.4034 %
Incorrectly Classified Instances      2790               22.5966 %
Kappa statistic                          0
Mean absolute error                      0.1912
Root mean squared error                  0.3092
Relative absolute error                 99.9654 %
Root relative squared error            100    %
Total Number of Instances            12347
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0          0          0          0          0          0.498    death
                 1          1          0.774      1          0.873      0.5      property
                 0          0          0          0          0          0.499    serious
                 0          0          0          0          0          0.499    slight
=== Confusion Matrix ===

   a     b     c     d   <-- classified as
   0   555     0     0 |   a = death
   0  9557     0     0 |   b = property
   0  1034     0     0 |   c = serious
   0  1201     0     0 |   d = slight
```

Figure 4.3: run information on experiment one

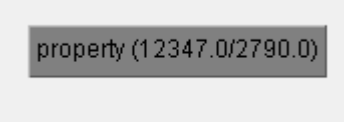


Figure 4.4: Decision Tree built on experiment one

As shown by the figure 4.4 above the result doesn't consider accident severity of death, serious, and light injuries. All records are classified under a single tree with a single leaf, "property"; however 2790 records are incorrectly classified under it as death, severe, and light injuries. This shows that it is impossible to get a pattern from the tree. The single leaf result may be because of the disproportional rate of data on the data set or attributes might not be selected well. From the total data set given, 9,557 were on property damage. Figure 4.5: shows this disproportional ratio of data.

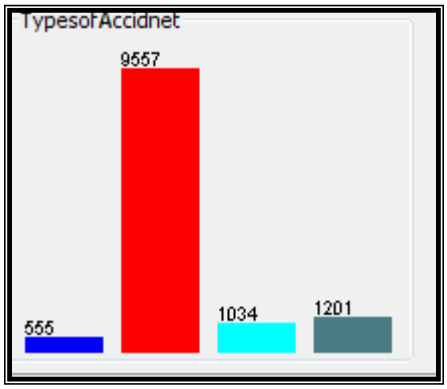


Figure 4.5: Accident disproportional graph

The first experiment was improved with selecting most important attributes from the 10 attributes experimented above. At this time of the study the "month" attribute was removed from the data set, because they can be replaced with the integration of weather and air conditions. The most important attributes were selected by considering the information gain ratio of the attributes for the class "accident type". The attributes with their relative selection value are presented in figure 4.6.

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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 AccidnetTypes):
        Information Gain Ranking Filter

Ranked attributes:
 0.221784    1 Place
 0.010741    8 AirConditions
 0.008856    7 Weatherconditions
 0.006501    4 Roadjoint
 0.003864    2 one_directional
 0.003297    5 TypesofRoad
 0.001855    6 Roadconditions
 0.00087     3 Roaddirection

Selected attributes: 1,8,7,4,2,5,6,3 : 8
```

Figure 4.6: Attribute selection on information gain

While the experiment was modified and repeated with all instances but with lesser attributes (the first five from figure 4.6); it brought the same accuracy and node of classification. Due to this conducting another experiment was essential.

Experiment two

This experiment consists two parts, the first is model building using all attributes on the above experiment and the second is by reducing number of attributes to the first five based on their information gain, but with less number of records.

Experiment two was done by applying data reduction that is minimizing the size of the data. By removing tuples on property damage, the ratio of data became proportional for the experiment.

- Accident records on humans were (2774) and
- Property damages were made (2774/3).
- The data remained with a total of 3699 (2774+925); taking property accidents data with the approximate average of the remaining three values

The attributes remained as they were on experiment one where as the exclusion of 8622 tuples from accident type property damage were done.

The tuples from property damage were selected by arranging the data based on property damage and picking 11 records on every 100 of the first 2500, and 10 on every 100 of the remaining. Table 2 shows proportion of data.

Class (Accident Type)	**Number of records**	**Percentile constitution**
Death injury	549	15%
Severe injury	1032	28%
Light injury	1193	32%
Proerty Damage	925	25%
Totlal	3699	100%

Table 2: distribution of records based on accident type

This experiment was conducted on a total of 3699 records and total attributes. It brought 55.5285%.accuracy and 222 total numbers of leaves. Again the accuracy level and the number of leaves obtained were worse to build a prediction model for crash locations. So another

experiment was done by further pruning the tree, Figure 4.7 show run information on this experiment.

```
Number of Leaves  :     222

Size of the tree :      291


Time taken to build model: 0.42 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2054               55.5285 %
Incorrectly Classified Instances      1645               44.4715 %
```

Figure 4.7: Result on experiment 2.1

From the above experiment the first five attributes were again selected based on their information gain ratio. At this time the number of leaves minimized to 159 on the other hand the accuracy diminishes to 55.3663%. Figure 4.8 depicts the attribute selection on the new data and results obtained on this experiment

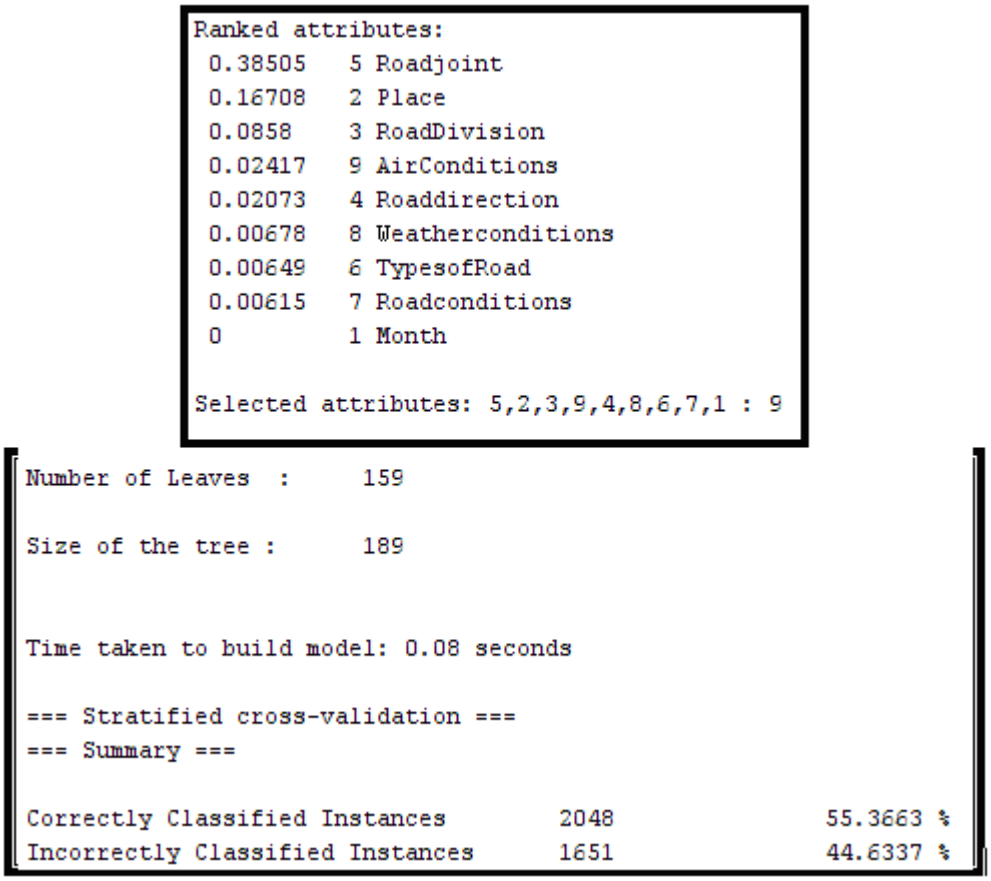


Figure 4.8 Attributes selected and Result on experiment 2.2

Further experiments were also done until the researcher got the best result on the study. Finally experiment four which is an improvement of experiment three is the best result observed out of all trials.

Experiment three

This experiment is another dimension on the research which constitutes full exclusion of property damages from the data set and concentrating on accidents happened on human beings

only. This was because if dangerous crash locations were identified based on severity only, it would be easy to predict age group exposure on that location.

This were experimented on the same ten attributes (figure 4.8); the experiment brought less number of leaves (100 leaves) where as lesser output accuracy (52.0375%).

```
Number of Leaves  :     100

Size of the tree :      132


Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1443               52.0375 %
Incorrectly Classified Instances      1330               47.9625 %
```

Figure 4.9 Result of experiment three

Experiment four

The final experiment is done using six attributes (shown in figure 4.10) with out considering property damage. The attributes were selected again based on their information gain ratio for the new data set. The result looks much better than the previous experiments.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     final selected attributes
Instances:    2774
Attributes:   6
              Place
              Road_division
              Roadjoint
              Weather
              AirConditions
              AccidnetTypes
Test mode:    10-fold cross-validation

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

```
Number of Leaves  :     110

Size of the tree :      143


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1707               61.5357 %
Incorrectly Classified Instances      1067               38.4643 %
Kappa statistic                          0.3791
Mean absolute error                      0.2915
Root mean squared error                  0.3878
Relative absolute error                 68.5867 %
Root relative squared error             84.1274 %
Total Number of Instances             2774

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.628     0.018      0.896     0.628     0.739      0.927    D
  0.451     0.208      0.562     0.451     0.5        0.727    S
  0.752     0.421      0.574     0.752     0.651      0.738    L

=== Confusion Matrix ===

   a   b   c   <-- classified as
 345  97 107 |   a = D
   9 465 558 |   b = S
  31 265 897 |   c = L
```

Figure 4.10 Run information for final experiment

This experiment resulted in 110 leaves with an accuracy level of 61.53%. From the experiment 1707 instances were correctly classified whereas 1067 instances moved to the wrong class. It is difficult to depict all 110 leaves of the decision tree on an A4 size paper due to this only the top four branching of the tree (figure 4.11.) and a segment/wing from the decision tree model is presented below the root node.

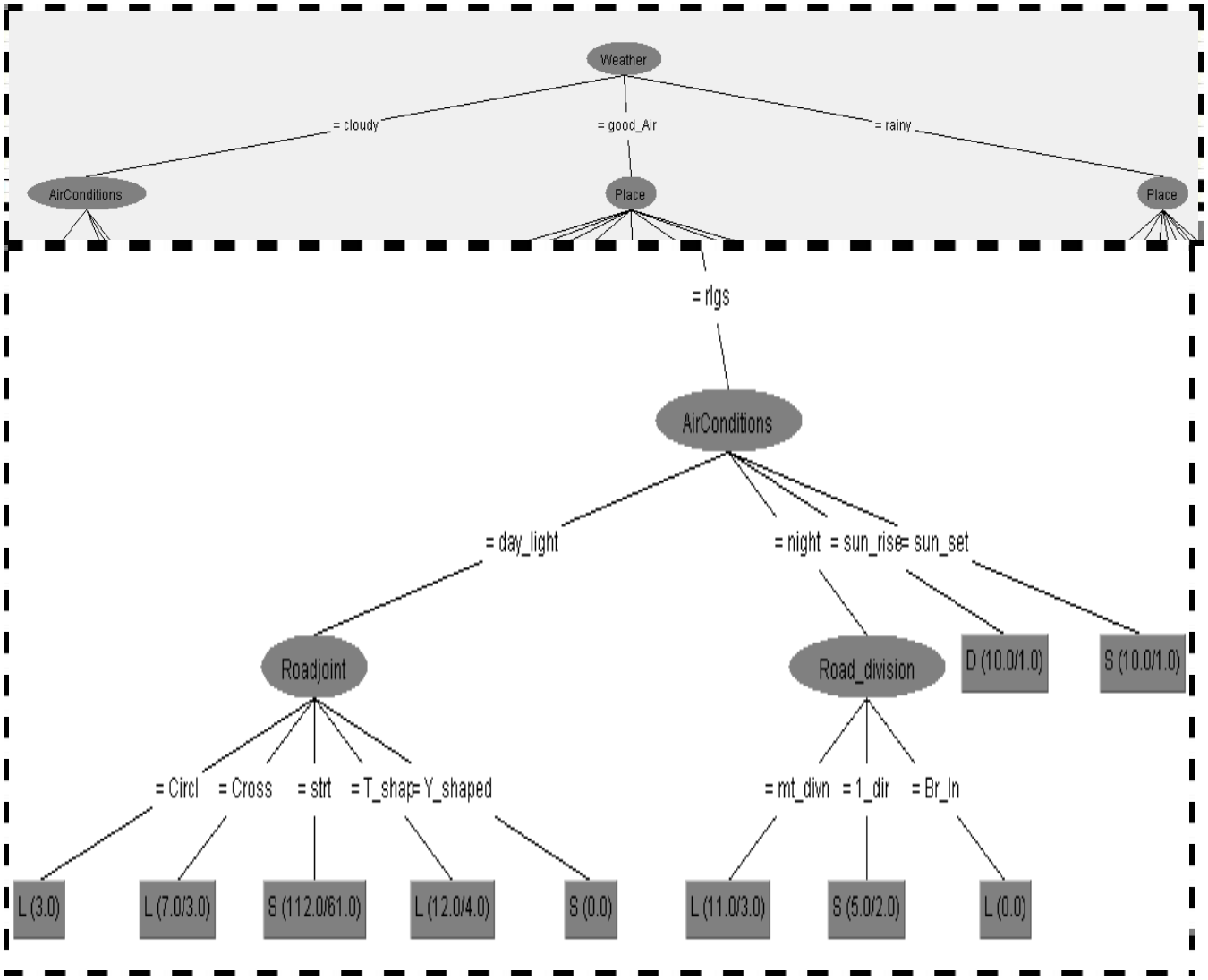


Figure 4.11: Decision tree model

The decision tree on figure 4.11 shows the segment of the model where the specific characterizations of the accidents on a location were:

1st – weather has to be in good Air condition

2nd the place has to be on religious environment, and

Finally the air condition and result of the accident based on severity is depicted as follows on figure 4.12.

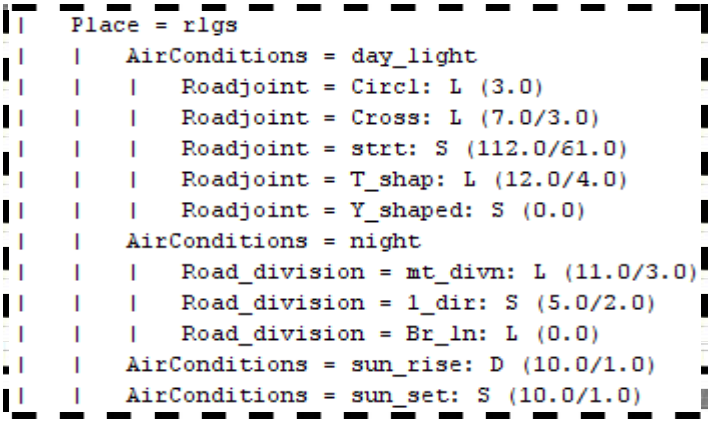


Figure 4.12 part of the J48 representation of the decision tree

For example figure 4.12 shows that, while the weather condition is good, the place is around religious location, and the air condition is at sun rise there will be 9 correctly classified death accidents and 1 wrongly classified instance.

4.4.3 Summery of experiments

Table 3 compares various experiments conducted so far. It includes information about the number of attributes given, the accuracy level observed, and the number of leaves obtained.

Experiment number	Attributes used	Accuracy on classification	Number of leaves observed
1	10	77.4%	1
2	8	54.06%	227
3	3	41.6%	110
4	4	54.06%	53
5	5	61.53%	110
6	5	52.03%	100

Table 3. Summery of experiments conducted

Based on the above table experiment 5, the one with average number of leaves and maximum accuracy level was selected as a good decision tree model.

4.5 Prioritizing crash locations

The basic purpose of building the decision tree model is to extract features for predictive model building. As mentioned earlier on chapter three, prioritization of high crash locations can have

several approaches. On this study severity of accidents was used to rank dangerous crash locations.

Based on this, locations with maximum number of death accidents ranked first, locations with maximum number of sever accidents ranked second, and locations with maximum number of light accidents were ranked third. The property damages were not taken in to account on the ranking steps. The two reasons why the study ignored property damage/loss from ranking were:

- Case one: it is inhuman to compare property damage with human beings for ranking
- Case two: even if the study tries, no recorded of cost estimation for the property damaged occurred on the TRA database.

4.5.1 J48 pruned Tree Based on Accident prioritization

The decision tree based on the selected algorithm, j-48, was used to see patterns of each accident type. The following prioritization shows rules generated based on their severity level. The j-48 algorithm result shows attributes prioritized first, on the external indention, where as the inner indention shows feature occurred as internal classification under the given attribute type.

1. **Death Locations**

Out of the 110 nodes identified on the experiment, the locations that resulted death of accidents are 15 nodes of the decision tree. The first three with 208, 17, and 14 numbers of most hazardous locations that resulted death are characterized as follows:

Weather = cloudy

| AirConditions = day_light
| | Roadjoint = Circl: D (7.0/2.0)
| AirConditions = night
| | *Place = instn: D (17.0)*
| *AirConditions = sun_rise: D (14.0)*
Weather = rainy
| Place = institution
| | Roadjoint = straight
| | | | Road_division = 1_dir: D (5.0)
| | Roadjoint = Y_shaped
| | | Road_division = mt_divn: D (4.0/2.0)
| Place = resident
| | Roadjoint = T_shap: D (2.0)
| *Place = schl: D (211.0/3.0)*

From the above J48 classifier result one can conclude that much of the death accidents occurred on:

- The First ranked death location was characterized as: (weather condition= rainy, place=school environment).
- The second exposed area of death is: (weather condition=cloudy, air condition=night, place=institution location)
- The Third ranked on death was: (weather condition=cloudy, air condition=sun rise).

Annex five shows the road locations where death occurs in detail. It is presented from highest death rate to lowest death rate.

2. Severe injury Accidents:

The second category of traffic accident is severe injury accidents. The following rules are derived from the last experiment decision tree of Weka.

Weather = good_Air
| Place = religious
| | AirConditions = day_light
| | | Roadjoint = strt: S (112.0/61.0)
| | AirConditions = night
| | | Road_division = 1_dir: S (5.0/2.0)
| | AirConditions = sun_set: S (10.0/1.0)
| Place = resident
| | Roadjoint = straight
| | | AirConditions = day_light: S (221.0/135.0)
| | | AirConditions = sun_set: S (5.0/2.0)
| | Roadjoint = T_shap
| | | AirConditions = sun_rise: S (3.0)
Weather = rainy
| Place = institution
| | Roadjoint = Circl: S (5.0)
| | Roadjoint = Cross: S (27.0)
| | Roadjoint = straight
| | | AirConditions = night
| | | | Road_division = mt_divn: S (36.0/1.0)
| | Roadjoint = T_shap: S (17.0/1.0)
| | Roadjoint = Y_shaped

From the above J48 classifier result one can see that, the first three severe accidents with 86, 61, and 35 injuries occurred on the following cases:

- The first severe accident location is characterized by: (weather condition=good air, location=around residents, road joint=straight line, air condition=during day light).
- The second severe accident occurrence is characterized by locations (weather condition=good, places=religious, air condition=day time, road joint=straight line).
- The third most severe location is characterized by: (Weather=rainy, place=around institution, road joint=straight, air condition=night, road division=mount division).

3. **Light injury locations:**

The third most hazardous location on traffic accident is the one which cause light injury. This location is characterized as follows:

```
Weather = good_Air
|   Place = institution
|   |   Roadjoint = Circle
|   |   |   Road_division = mt_divn: L (8.0)
|   |   |   Road_division = Br_ln: L (0.0)
|   |   Roadjoint = Cross
|   |   |   Road_division = mt_divn: L (17.0)
|   |   |   Road_division = 1_dir
|   |   |   |   AirConditions = sun_rise: L (4.0)
|   |   |   Road_division = Br_ln: L (0.0)
|   |   Roadjoint = straight
|   |   |   Road_division = mt_divn: L (485.0/188.0)
|   |   |   Road_division = 1_dir
```

| | | | AirConditions = day_light: L (366.0/180.0)
| | | Road_division = Br_ln: L (2.0)
| | Roadjoint = T_shap: L (128.0/36.0)
| Place = market
| | AirConditions = day_light: L (262.0/132.0)
| | AirConditions = night
| | | Roadjoint = strt: L (34.0/18.0)
| | AirConditions = sun_rise: L (15.0/7.0)
| | AirConditions = sun_set: L (1.0)

From the above J48 classifier result one can see that, the first light injury accidents with 297, 186, and 130 amounts occurred on the following cases respectively:

- The first location for light injury is: (weather condition=good air, place=around institutions, road joint=straight, road division=mount division
- The second ranked location on light severity is characterized by: (weather condition=good air, place=around institutions, road joint=straight, road division=one directional, air condition=day time).
- The third most severe location for light injury is the one characterized by: (weather=good air, place=around market, air condition=day time).

The above mentioned features on crash locations ranked most hazardous. Death, severe, and light also occurs on other places however these are the first three for each.

Best rules found for each accident type

For Death accident

Rule #1

If

Weather = rainy, Place = school:

Then # Death Injuries = (208)

Rule #2

If

Weather = good_Air, Place = factory:

Then # Death Injuries = (63)

Rule #3

If

Weather = cloudy, AirConditions = night, Place = institution:

Then # Death Injuries = (17)

Rule #4

If

Weather = cloudy, AirConditions = sun_rise:

Then # Death Injuries = (14)

Rule #5

If

Weather = good_Air, Place = religious, AirConditions = sun_rise:

Then # Death Injuries = (9)

Rule #6

If

Weather = good_Air, Place = resident, AirConditions = sun_rise:

Then # Death Injuries = (7)

Rule #7

If

Weather = cloudy, AirConditions = day_light, Roadjoint = Circl:

Then # Death Injuries = (5)

For Severe Accident

Rule #1

If

Weather = good_Air, Place = resident, Roadjoint = straight, AirConditions = day_light:

Then # Severe Injuries = (86)

Rule # 2

if

Weather = good_Air, Place = religious, AirConditions = day_light, Roadjoint = straight:

then # Severe Injuries = (51)

Rule #3

If

Weather = rainy, Place = institution, Roadjoint = straight, AirConditions = night, Road_division = mount division:

Then # Severe Injuries = (35)

Rule #4

If

Weather = rainy, Place = institution, Roadjoint = Cross Shaped:

Then # Severe Injuries = (27)

Rule # 5

If

Weather = good_Air, Place = institution, Roadjoint = straight, Road_division = 1_dir AirConditions = night:

Then # Severe Injuries = (19)

For Light accident

Rule #1

If

Weather = good_Air, Place = institution, Roadjoint = straight, Road_division = mount division:

Then # Light Injuries = (305)

Rule #2

If

Weather = good_Air, Place = institution, Roadjoint = straight, Road_division = 1_dir AirConditions = day_light:

Then # Light Injuries = (186)

Rule #3

If

Weather = good_Air, Place = market, AirConditions = day_light:

Then # Light Injuries = (130)

Rule #4

If

Weather = good_Air, Place = recreational, AirConditions = day_light:

Then # Light Injuries = (95)

Rule #5

If

Weather = good_Air, Place = institution, Roadjoint = T_shap,

Then # Light Injuries = (92)

4.5.2 Evaluation and Interpretation

The decision tree model was evaluated using the confusion matrix obtained from the j-48 algorithm. Confusion matrix shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong. It shows instances that were correctly classified, and wrongly classified under another, for each class. Based on the data input the confusion matrix of the model is represented as follows:

Actual	Predicted			Total	Score
	Death	Severe	Light		
Death	345	97	107	549	63.84%
Severe	9	465	558	1032	45.05%
Light	31	265	897	1193	75.19%
Total	385	827	1562	2774	61.53%

Table 4. Confusion matrix of accident severity

The table summarizes that out of 549 death records 345 are correctly classified, 97 are wrongly classified under severe injury, and 107 are wrongly classified under light injury. Similarly while 465 are correctly classified on severe injury, 567 records are wrongly classified, on light injury 897 records were correctly classified while 296 are wrongly classified. This shows that the classifier brought better results on death and light injuries.

The decision tree generated based on the above confusion matrix showed better result than the others with its accuracy level. As shown on table 4 it has 61.53% of accuracy when 6 attributes selected and it generated 110 leaves. These patterns of the decision tree were supported by experts on the area too. Experts also favored results on death, severe, and light injury accident locations.

CHAPTER FIVE

Conclusion and Recommendation

5.1 Conclusions

The explosion of information due to digital technology and availability of electronic documents has made the world full of information/data. The data in every sector is growing rapidly. Scientific, legal, business, medical records are increasing in enormous amount from time to time. This availability by itself created overload and difficulty on access and use of it. The consequence of overload of information is complex and poor decision making on areas that need proper analysis of data.

One scarce resource in this world is time. Organizations or individuals do not have time to go through all records and data collections. Organizational Mangers and individuals need filtered and simplified data from their long historical record. This retrieval system helps them to pass correct decisions on their daily activity or improve their future plan.

This problem can easily be alleviated by using data mining technology which filters out certain implicit feature from vast amount of data. The AATO contains much amount of data on traffic accidents records. These records help to get valid information by identifying features of

dangerous traffic locations, and predicting the possibility of pedestrians either to be free from accident or rate of severity if an accident occurs.

The aim of this study was to investigate application of data mining technology on traffic accident. The data needed for the study were collected from the database of AATO, which holds 12,441 records of accidents in Amharic language.

After all necessary preprocessing steps were done; the data was given for Weka software version 5.4.8. six experiments were done until the valid pattern was investigated. Finally with a record of 2,774 the final model was developed. This results in accuracy level of 61.54% (1707 records) which are predicted correctly. The characterizations of dangerous crash locations based on their severity were evaluated using decision tree.

As mentioned earlier quality of the road and its environment is not the only factor for the cause of Road Traffic Accidents. There are other causes such as quality of the car, skill of the driver, and road quality. The road quality includes several factors that affect specific location at a given time. Even With all these constraints, the result obtained is encouraging.

The result obtained was good especially for light injury accidents and death accidents (see the confusion matrix on table four) this shows that data mining technology can be applied on prioritization dangerous crash locations, however the complexity of factors for its cause brought decrease in its accuracy to some extent.

5.2 Recommendations

The accuracy level of the research can be improved and useful if the following mentioned recommendations are identified and thoroughly investigated by the organizations and other researchers too:

- On this study the researcher doesn't include property damage or loss to prioritize as dangerous crash locations this is because no such record for cost estimation on property. The AATO should keep record of the RTA in detail including the cost of damage on property, and injury level with estimated cost.

- Road and its quality is not the only factor for Traffic accident. It is better if further study should be done on integrating causes of traffic accidents both on driver quality and car quality with road and its environment.

- To be more accurate on traffic police assignment and detailed investigation on crash locations, the specific part of the city should be known; so further study which integrate the data with Geographical Information System are highly encouraged. This directly points out dangerous crash location from the map of Addis Ababa city.

- Further study on ranking dangerous crash locations should be conducted using cost function and specific location if the database on the institution is upgraded to include accident costs and road name/number.

REFERENCES

A Model For Identifying And Ranking Dangerous Accident Locations: A Case Study In Flanders. Available At URL: Http://Alpha.Uhasselt.Be/~Brijs/Pubs/Statistica%20Neerlandica%20-%20revised%20version.Pdf

Addis Ababa Transport Office (2009). Addis Traffic Office Report Manual. Addis Ababa

Agency France Press (2008), Availablle At URL: Http://Afp.Google.Com/Article/Aleqm5g4ocalopuovo0gkooorb2nhd9ohw

Berry, M. & Linoff, G (2004) Data Mining Techniques : For Marketing, Sales, And Customer Relationship Management 2nd Ed, Wiley Publishing, Inc., Indianapolis, Indiana.

Berson, A. & Smith, P. & Thearling, K. (Nd.)An Overview Of Data Mining Techniques: Available At URL: Http://Www.Thearling.Com/Text/Dmtechniques/Dmtechniques.Htm

Bigus, J. (1996), Data Mining With Neural Networks : Solving Business Problems – From Application Development To Decision Support I By

Brijs, T. & Karlis, D.(Nd.) A Bayesian Model For Ranking Hazardous Road Sites. Available At URL: Http://Alpha.Uhasselt.Be/~Brijs/Pubs/A882-R3-Final.Pdf

Capua, J. (02 March 2006), Putting The Brakes On Ethiopian Traffic Accidents, Available At URL: Http://Www.Voanews.Com/English/Archive/2006-03/2006-03-02-Voa20.Cfm?CFID=1851021&CFTOKEN=67737796

Chong, M. Et Al (Nd), Mtraffic Accident Analysis Using Decision Trees And Neural Networks Miao M. Chong, Ajith Abraham, Marcin Paprzycki

Chong, M. & Abraham, A. (Nd.)Traffic Accident Analysis Using Decision Trees And Neural Networks. Available At URL: Http://Arxiv.Org/Ftp/Cs/Papers/0405/0405050.Pdf

City Government Of Addis Ababa, Transport Authority. Facts About Addis Ababa City Transport. Available At URL: Http://Www.Telecom.Net.Et/~Aata/

Daniel, L. (2006) Data Mining Methods And Models, John Wiley & Sons, Inc

Denekew, A. (2003), The Application Of Data Mining To Support Custemer Relationship Management At Ethiopian Airlines. Denekew Abera Jembere.June, Unpublished Master's Thesis ,2003

Farlex, Inc Junction (2008), Available At URL: Http://Www.Thefreedictionary.Com/Junction

Fathers' Manifesto & Christian Party, (May 04, 1961). The Christian Party: Available At URL: Http://Christianparty.Net/Mvfr.Htm

Fayyad, U. &Smyth, P. (1996). From Data Mining To Knowledge Discovery In Databases. Available URL: Http://Citeseer.Nj.Nec.Com/Fayyad96from.Html

Ganveer, G & Tiwari, R. (2005), Injury Pattern Among Non-Fatal Road Traffic Accident Cases: A Cross-Sectional Study In Central India. Available At URL: Http://Www.Indianjmedsci.Org/Article.Asp

Gashaw, M. (2004), Application Of Data Mining Technology To Support Customer Insolvency Prediction At Ethiopian Telecommunication Corporation. Gashaw Mulatu Gessesse, Unpublished Master's Thesis, 2004

Geurts, K & Wets, G. (2003), Black Spot Analysis Methods. Available At URL: Http://Www.Ictct.Org/Workshops/04-Tartu/C4_Geurts.Pdf

Giudici, P. (2003), Applied Data Mining: Statistical Methods For Business And Industry / Paolo Giudici., Giudici, Paolo. By Biddles Ltd, Guildford, Surrey

Giustini, M., (2002).Traffic Fatalities In Italy, 1969-1998 Laboratory Of Epidemiology And Biostatistics Istituto Superiore Di Sanità, Rome. Available At URL:

Han, J. & Kamber, M. (2001). Data Mining: Concepts And Techniques. San Fransisco; Morgan Kufman Publishers.

Hand, D. & Mannila, H. & Smyth, P. (2001), Principles Of Data Mining, The MIT Press

Hossain, M. (Nd.), Application Of Data Mining In Road Safety, Asian Institute Of Technology. Available At URL: Http://Www.Library.Ait.Ac.Th/Thesissearch/Summary/Moinul%20Hossain.Pdf

Jacob, G. & Aeron-Thomas, A. (Nd.) A Review Of Global Road Traffic Accident Features. Available At Url: Http://Www.Transport-Links.Org/Transport_Links/ Filearea/

Publications/ 1_771_Pa3568.Pdf

James, O. Harris (Nd), Cause And Contributing Factors In Traffic Accidents. Available At URL: Http://Expertpages.Com/News/Cause_Contributing_Factors_Traffic_Accidents.Htm

Jha, N. & Shekhar, C. (Nd) Epidemiological Study Of Road Traffic Accident Cases: A Study From Eastern Nepal, Available At URL: Http://Www.Searo.Who.Int/EN/Section1243/Section1310/Section1343/Section1344/Section1836/Section1837_8158.Htm

Kumar, M. & Yadav, B.(2005), Involvement Of Children In Road Traffic Accidents In Eastern Nepal Available At Url: Http://Www.Icfmt.Org/Vol3no2/Easternepal.Htm

Leul, W. (3003), The Application Of Data Mining On Crime Prevention: The Case Of Oromya Police, AAU.

Mikulík, J.& Holló, P. (2007), Piarc Road Accident Investigation Guidelines For Road Engineers, World Road Association Piarc Technical Committee -3.1 "Road Safety"

New York Personal Injury Lawyers (2007) Traffic Accident Causes, Available At URL: Http://Oshmanlaw.Com/Personal_Injury/Traffic_Accident_Causes.Html

Oshman & Mirisola 2007 , LLP | 42 Broadway 10th Floor New York, NY 10004 New York Personal Injury Lawyers SEO By Consultwebs.Com: Law Firm Website Designers - Design By Ejustice

Pal, N. & Jain, L. (2005), Advanced Techniques In Knowledge Discovery And Data Mining, Springer-Verlag London Limited

Palous, J. (Nd). Machine Learning And Data Mining. Prague: Gerstner Laboratory For Intelligent Decision Making And Control Czech Technical University. Available At URL: Http://Citeseer.Nj.Nec.Com/506615.Html

Pande, A. & Abdel-Aty, M. (2005), Identification Of Rear-End Crash Patterns On Instrumented Free Ways: A Data Mining Approach. Available At URL: Http://Ieeexplore.Ieee.Org/Iel5/10189/32528/01520155.Pdf

Queen's University Of Belfast, Maintained By Alan Rea Data Mining Techniques: Available At URL: Http://Www.Pcc.Qub.Ac.Uk/Tec/Courses/Datamining/Stu_Notes/Dm_Book_4.Html

Sabel, C. & Kingham, S. & Nicholson A. & Bartie P. (2005),Road Traffic Accident Simulation Modelling - A Kernel Estimation Approach

Safetynet, Junctions (2007) Available At URL: Http://Www.Erso.Eu/Knowledge/Content/15_Road/Junctions.Htm

SINGAPORE POLICE FORCE (2006), Road Traffic Situation. Available At URL: Http://Www.Spf.Gov.Sg/Stats/Traf2004_Concern.Htm

Singhal, A. (2007), Data Warehousing And Data Mining Techniques For Cyber Security, Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA

Srinivas, P. & Shashi, N. & Vanjeeswaran K (2007). New Methods To Identify And Rank High Pedestrian Crash Zones. Available At URL: Http://Www.Sciencedirect.Com/Science?_Ob=Articleurl&_Udi=B6V5S-4MV1H56-1&_User=5270405&_Rdoc=1&_Fmt=&_Orig=Search&_Sort=D&View=C&_Acct=C000066709&_Version=1&_Urlversion=0&_Userid=5270405&Md5=30636076c8572f40dc602e273981a605

Statsoft Inc (1984), Data Mining Technique: Available At URL: Http://Www.Statsoft.Com/Textbook/Stdatmin.Html

Tesfaye, H. (2002). Predictive Modeling Using Data Mining Techniques In Support Of Insurance Risk Assessment. Unpublished Master's Thesis. Addis Ababa University.

The Ethiopian Traffic Enforcement Agency (No Date), The Sad Facts Available At URL: Http://Ethiopiantraffic.Com/Facts.Htm

Thearling, K. (2003) An Introduction To Data Mining

Tibebe, B. (2005), Application Of Data Mining Technology To Support Road Traffic

Accident Severity Analysis At Addis Ababa Traffic Office, Unpublished Master's Thesis, AAU.

Tibebe, B. & Abrham, A. (Nd.) Rule Mining And Classification Of Road Traffic Accidents Using Adaptive Regression Model. Available At URL: Http://Www.Softcomputing.Net/Ijsst1.Pdf

TWO CROWS Corporation (1999), Data Mining Glossary: Available At URL: Http://Www.Twocrows.Com/Glossary.Htm

Whitten, I. & Frank, E. (2000). Data Mining: Practical Machine Learning Tools And Techniques With Java Implementations

WHO (2008), Road Traffic Injuries. Available At URL: Http://Www.Who.Int/Violence_Injury_Prevention/Road_Traffic/En/

Witten, I &Frank, E. (2005), Data Mining: Practical Machine Learning Tools And Techniques /. – 2nd Ed.

World Bank Group, Nd. Road Safety: Accident Counter Measures At Hazardous Locations. Available At URL: Http://Www.Worldbank.Org/Transport/Roads/Saf_Docs/Haz_Locs.Htm

Yoseph, A.(2004) Application Of Predictive Data Mining Model For Central Statistics Authority. Unpublished Master's Thesis, Addis Ababa University

APPENDICES

Annex 1: traffic accident database fields, values and description

No	Attribute Category	Attribute Name	Data Type	Description
I	**Driver Related**			
1		Driver_Sex	Text	Gender of driver
2		Driver_Age	Numeric	Age of the driver
3		Accused_Male	Numeric	Gender of criminal
4		Accused_Female	Numeric	Gender of criminal
5		Accused_Age	Numeric	Age of the driver
6		Educational_Background	Text	Qualification of the driver
7		Relationship_With	Text	Driver relation with car owner
8		Experience	Numeric	Years of driving
9		Accused_No	Text	Record number
10		License_Grade	Text	Drivers license number
II	**Vehicle Related**			
11		Types_Of_Vehicle	Text	Classification of vehicle
12		Plate_Code	Text	Plate code of the vehicle
13		Ownership	Text	To whom the car belongs
14		Depreciation_Year	Numeric	Past Service year of the car
15		Automotive_Status	Text	Deficiency of vehicle
III	**Accident Place**			
16		Sub_City	Text	Accident occurrence, Sub city of Addis Ababa
17		Kebele	Numeric	Accident occurrence, Kebele of the sub city
18		Special_Place	Text	Accident occurrence, village name
19		Place	Text	Accident occurrence, known name
IV	**Road Related**			
20		Road_Division	Text	Single or dual road divisions

21		Road_Direction	Text	Straight or curved directions
22		Road_Joint	Text	Shape of roads junction
23		Types_Of_Road	Text	Road surface moisture
24		Road_Conditions	Text	Type of road surface
V	**Pedestrian Related**			
25		Injury_Age	Numeric	Age of the injured
26		Injury_Occupation	Text	Occupation of injured person
27		Health_Status	Text	injured person earlier health condition
28		Movement_Of_Pedestrian	Text	Pedestrian movement at the time of accident
19		Injury_Persons	Numeric	Total Number of injured at an accident
30		Injury_Persons_Male	Numeric	Total muscular gender injured
31		Injury_Persons_Females	Numeric	Total feminine gender Injured
32		Injury	Text	Injured person category, driver pedestrians others
VI	**Accident Related**			
33		Month	Numeric	Month the accident occurred
34		Which_Week	Numeric	Week of the month
35		Regno	Text	Accident record number
36		Date	Text	Accident occurrence date
37		Date_Of_The_Week	Text	Day of the week Mon. –Sun.
38		Time	Date	Accident time in 24 hours
39		Acupuncture_With	Text	Other object crashed with
40		Types_Of_Accident	Text	
41		No_Of_Car_Crushed	Numeric	Participants on crashing
42		Accused_Vehicles_Movement	Text	Movement of vehicle at accident
43		Accident_Type	Text	Severity of Accident type
44		Cause_Of_Accident	Text	Driver situation and act
45		Weather_Conditions	Text	Weather condition at accident occurrence
46		Air_Conditions	Text	Sight of environment

Annex 2: j48 pruned tree result

Weather = cloudy
| AirConditions = day_light
| | Roadjoint = Circl: D (7.0/2.0)
| | Roadjoint = Cross: L (1.0)
| | Roadjoint = strt: L (25.0/9.0)
| | Roadjoint = T_shap: L (2.0)
| AirConditions = night
| | Place = instn: D (17.0)
| | Place = mrkt
| | | Roadjoint = strt: S (10.0)
| | | Roadjoint = T_shap: L (4.0)
| AirConditions = sun_rise: D (14.0)
| AirConditions = sun_set: S (20.0)
Weather = good_Air
| Place = instn
| | Roadjoint = Circl
| | | Road_division = mt_divn: L (8.0)
| | | Road_division = 1_dir: D (3.0/1.0)
| | Roadjoint = Cross
| | | Road_division = mt_divn: L (17.0)
| | | Road_division = 1_dir
| | | | AirConditions = day_light: D (2.0)
| | | | AirConditions = night: D (4.0)
| | | | AirConditions = sun_rise: L (4.0)
| | Roadjoint = strt
| | | Road_division = mt_divn: L (485.0/188.0)
| | | Road_division = 1_dir
| | | | AirConditions = day_light: L (366.0/180.0)

| | | | AirConditions = night: S (26.0/7.0)
| | | | AirConditions = sun_rise: S (19.0/6.0)
| | | | AirConditions = sun_set: S (13.0)
| | | Road_division = Br_ln: L (2.0)
| | Roadjoint = T_shap: L (128.0/36.0)
| Place = fct: D (63.0)
| Place = mrkt
| | AirConditions = day_light: L (262.0/132.0)
| | AirConditions = night
| | | Roadjoint = strt: L (34.0/18.0)
| | | Roadjoint = T_shap: S (13.0/2.0)
| | AirConditions = sun_rise: L (15.0/7.0)
| | AirConditions = sun_set: L (1.0)
| Place = rcrtl
| | AirConditions = day_light: L (169.0/74.0)
| | AirConditions = night
| | | Road_division = mt_divn: L (17.0/9.0)
| | | Road_division = 1_dir: S (21.0/7.0)
| | AirConditions = sun_rise: L (1.0)
| | AirConditions = sun_set: L (1.0)
| Place = rlgs
| | AirConditions = day_light
| | | Roadjoint = Circl: L (3.0)
| | | Roadjoint = Cross: L (7.0/3.0)
| | | Roadjoint = strt: S (112.0/61.0)
| | | Roadjoint = T_shap: L (12.0/4.0)
| | AirConditions = night
| | | Road_division = mt_divn: L (11.0/3.0)
| | | Road_division = 1_dir: S (5.0/2.0)

| | AirConditions = sun_rise: D (10.0/1.0)

| | AirConditions = sun_set: S (10.0/1.0)

| Place = rsdnt

| | Roadjoint = Circl: L (4.0)

| | Roadjoint = Cross: L (5.0/1.0)

| | Roadjoint = strt

| | | AirConditions = day_light: S (221.0/135.0)

| | | AirConditions = night: L (43.0/25.0)

| | | AirConditions = sun_rise: D (8.0/1.0)

| | | AirConditions = sun_set: S (5.0/2.0)

| | Roadjoint = T_shap

| | | AirConditions = day_light: L (20.0/12.0)

| | | AirConditions = sun_rise: S (3.0)

| Place = schl

| | Road_division = mt_divn: D (3.0)

| | Road_division = 1_dir

| | | AirConditions = day_light: L (95.0/35.0)

| | | AirConditions = night: S (18.0/5.0)

| | | AirConditions = sun_set: S (4.0)

| Place = hspt

| | Road_division = mt_divn: L (7.0)

| | Road_division = 1_dir: L (6.0)

| | Road_division = Br_ln: D (6.0/1.0)

Weather = rainy

| Place = instn

| | Roadjoint = Circl: S (5.0)

| | Roadjoint = Cross: S (27.0)

| | Roadjoint = strt

| | | AirConditions = day_light: L (12.0/3.0)

| | | AirConditions = night
| | | | Road_division = mt_divn: S (36.0/1.0)
| | | | Road_division = 1_dir: D (5.0)
| | | AirConditions = sun_rise: S (6.0)
| | Roadjoint = T_shap: S (17.0/1.0)
| | Roadjoint = Y_shaped
| | | Road_division = mt_divn: D (4.0/2.0)
| | | Road_division = 1_dir: L (6.0/3.0)
| Place = fct: S (3.0)
| Place = mrkt
| | AirConditions = day_light: L (8.0/2.0)
| | AirConditions = night: S (6.0)
| | AirConditions = sun_set: S (1.0)
| Place = rcrtl: L (5.0/1.0)
| Place = rlgs: L (3.0)
| Place = rsdnt
| | Roadjoint = strt
| | | AirConditions = day_light: S (4.0/2.0)
| | | AirConditions = night: L (2.0)
| | | AirConditions = sun_set: S (2.0)
| | Roadjoint = T_shap: D (2.0)
| | Roadjoint = Y_shaped: L (1.0)
| Place = schl: D (211.0/3.0)
| Place = hspt: S (16.0)

Annex 4: Data set Sample

Place	Road division	Road junction	Weather	Air condition	Accidnent Type
rcrtl	mt_divn	Circl	good_Air	day_light	Light
rlgs	mt_divn	Circl	good_Air	day_light	Light
rlgs	mt_divn	Circl	good_Air	day_light	Light
rlgs	mt_divn	Circl	good_Air	day_light	Light
rsdnt	mt_divn	Circl	good_Air	day_light	Light
rsdnt	mt_divn	Circl	good_Air	day_light	Light
rsdnt	1_dir	Circl	good_Air	day_light	Light
rsdnt	1_dir	Circl	good_Air	day_light	Light
instn	mt_divn	Circl	good_Air	night	Light
instn	mt_divn	Circl	good_Air	night	Light
rlgs	mt_divn	Circl	good_Air	night	Light
rlgs	mt_divn	Circl	good_Air	night	Light
instn	1_dir	Circl	good_Air	night	Light
instn	mt_divn	Circl	Cloudy	day_light	Severe
instn	mt_divn	Circl	Cloudy	day_light	Severe
instn	mt_divn	Circl	Rainy	day_light	Severe
instn	mt_divn	Circl	Rainy	day_light	Severe
instn	mt_divn	Circl	Rainy	day_light	Severe
instn	mt_divn	Circl	Rainy	night	Severe
instn	mt_divn	Circl	Rainy	night	Severe
mrkt	mt_divn	Circl	Rainy	night	Severe
mrkt	1_dir	Circl	Rainy	night	Severe
instn	1_dir	Cross	good_Air	day_light	death
instn	1_dir	Cross	good_Air	day_light	death
instn	1_dir	Cross	good_Air	night	death

instn	1_dir	Cross	good_Air	night	death
instn	1_dir	Cross	good_Air	night	death
instn	1_dir	Cross	good_Air	night	death
instn	mt_divn	Cross	Cloudy	day_light	Light
instn	mt_divn	Cross	good_Air	day_light	Light
instn	mt_divn	Cross	good_Air	day_light	Light
instn	mt_divn	Cross	good_Air	day_light	Light
instn	mt_divn	Cross	good_Air	day_light	Light
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
rlgs	mt_divn	Cross	good_Air	day_light	Light
rlgs	mt_divn	Cross	good_Air	day_light	Light
rlgs	mt_divn	Cross	good_Air	day_light	Light
rlgs	mt_divn	Cross	good_Air	day_light	Light
rsdnt	mt_divn	Cross	good_Air	day_light	Light
instn	mt_divn	Cross	good_Air	night	Light
instn	mt_divn	Cross	good_Air	night	Light
rlgs	mt_divn	Cross	good_Air	night	Light
rlgs	mt_divn	Cross	good_Air	night	Light
rsdnt	1_dir	Cross	good_Air	night	Light
instn	1_dir	Cross	good_Air	sun_rise	Light
instn	1_dir	Cross	good_Air	sun_rise	Light

Annex 5: Accident location (for Accident Type = Death)

Rank No	**Road division**	**Road joint**	**Place**	**Weather**	**Air condition**	**Accident Type**	**Correctly classified**
1			School	Rainy		Death	208
2			Factory	Good air		Death	63
3			Institution	Cloudy	Night	Death	17
4				Cloudy	Sun rise	Death	14
5				*Good air*	Sun rise	Death	8
6		Straight	Resident	Good air	Sun rise	Death	7
7		Circle		Cloudy	Day light	Death	5
7		Straight	Institution	Rainy	Night	Death	5
9	*Broken line*		*Hospital*	*Good air*		*Death*	*4*
9			*Institution*	*Good air*	*night*	*Death*	*4*
11	*M_division*		*School*	*Good air*		*Death*	*2*
11	*1_direction*	*Circle*	*Institution*	*Good air*		*Death*	*2*
11	*1_direction*	*Cross*	*Institution*	*Good air*	*Day light*	*Death*	*2*
11	M_division	Y shape	Institution	Rainy		Death	2
11		T- shaped	Resident	Rainy		Death	2
Total							345

Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university

Haleluya Kiflu

January, 2009

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

Rahel Bekele (PhD)