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

APPLYING DATA MINING TO IDENTIFY DETERMINANT FACTORS OF DRIVERS AND VEHICLES IN SUPPORT OF REDUCING AND CONTROLLING ROAD TRAFFIC ACCIDENT: IN THE CASE OF ADDIS ABABA CITY

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER'S OF SCIENCE IN INFORMATION SCIENCE.

BY

GETNET MOSSIE ZELEKE

APRIL, 2009

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

APPLYING DATA MINING TO IDENTIFY DETERMINANT FACTORS OF DRIVERS AND VEHICLES IN SUPPORT OF REDUCING AND CONTROLLING ROAD TRAFFIC ACCIDENT: IN THE CASE OF ADDIS ABABA CITY

BY

GETNET MOSSIE ZELEKE

Approved by the board member of the examiners	Signature
Ato Lemma Lessa Chairperson of the Examination	
Dr. Manoj V.N.V Advisor	
Dr.Eng. worku Alemu External Examiner	

DEDICATION

This research work is dedicated to my beloved and kindhearted mother,

Abeba Ejigu

TABLE OF CONTENT

ABSTRACT

Road transport plays vital roles in the effort of enriching the economic growth of the society, especially in developing countries. An efficient transport system is decisive factor to promote socio-economic development of Ethiopia. Although the transport sector is important in facilitating economic growth and development, a very negative phenomenon, namely road traffic accident, has increased thereby highly threatening the safety of every traveler in Ethiopia, in particular at Addis Ababa city.

Traditionally, simple manual and statistical techniques are used for traffic accident analysis at Addis Ababa traffic control and investigation office. These methods are inefficient and impractical as the volume of road traffic accident data increases. Thus this research work will discuss how to investigate the potential application of data mining tool and techniques to develop models that can support to reduce and control road traffic accident by identifying and predicting the major drivers and vehicles determinant risk factors (attributes) that causes road traffic accident.

The methodology used for this research work had three basic steps namely, data collection, data preprocessing and model building and evaluating. The dataset used for this research work was collected from Addis Ababa traffic control and investigation office, 6107 road traffic accident records. Since the collected dataset was not suitable as it is for experiment, data preprocessing activities were done. In data preprocessing steps data cleaning and data reduction were undertaken. To build models decision tree and rule induction techniques were employed using Weka, version 3-5-8, data mining tool.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

The incidence of road traffic accidents is rising in the world-wide. Every day thousands of people are killed and injured on the roads. Men, women, or children walking, biking or riding to school or work, playing in the streets or setting out on long trips, will never return home, leaving behind shattered families and communities. Millions of people each Year will spend long weeks in hospital after severe crashes and many will never be able to live, work or play as they used to do.

Traffic accidents are major public health problems all over the world. It is estimated that 1.2 million road traffic deaths occur and about 50 million people are injured in the world every year; most of them are in developing countries. In addition to this loss, the economic cost of road traffic crash and injuries is estimated to be 1% of the gross national product (GNP) of low-income countries. The global cost is estimated to be US $ 518 Billion per year. Low-income countries account for US $ 65 Billion, which is more than they receive in development assistance [38].

According to World Health Organization (WHO) and world Bank Report "The Global Burden of Disease", deaths from non communicable diseases are expected to climb from 28.1 million a year in 1990 to 49.7 million by 2020 - an increase in absolute number of 77%, traffic accidents are the main causes of this rise. Road traffic accidents are expected to take third place in the rank order of disease burden by the year 2020.In traffic accident, according to WHO, Ethiopia has the highest rate of fatalities per vehicles in the world [28].

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).This accident therefore include collusions between vehicles and animals, vehicles and pedestrian, vehicles and fixed obstacles or vehicles and vehicles [28].

With regard to road safety, the accident being happened in least developing countries is high by the time it is decreasing in developed countries. In Ethiopia for instance different reports revealed that on average, per 10,000 vehicles a death for 180 people takes place every year. This makes the country to be one of the few countries with high accident fatality in spite of less population of vehicles in the country. The main factor for such significant problems of road safety in Ethiopia is mainly the problem with the human factors, mainly the drivers' ethics and capability rather than mechanical problems [32].

Peter Termeulen, head of the International Road safety Academy in the Netherlands says, also strengthen the above idea, for every 10,000 vehicles in Ethiopia 180 people die in traffic related accident. That compares to the United States where 21 people die in traffic related accidents for every 100,000 vehicles. Peter Termeulen says it is not simply a matter of Ethiopians being bad drivers. It is more a matter of their "not being educated" about road safety [36].

The traffic accident rate in Ethiopia is growing at alarming rate especially since 1996/97. In 1996/97 the registered traffic accident was 11524, but after 8 and 9 year, the figure reached 17

457 in year 2003/04. Out of these registered accident figures, the traffic accident with fatality has increased by 12% per annum on average [32].

Due to the occurrences of such large number of accident in Ethiopia, enormous amount of traffic accident data were collected and stored in large databases. Thus it is beyond human ability for understanding the cause and solution of the accident without powerful tools from this large volume of accident database. This explosive growth of stored data has generated an urgent need for new techniques and automated tools that can intelligently assist in transforming the vast amount of accident data into useful information and knowledge.

As data volumes grow dramatically, data analysis based on manual and statistical methods with small portion of data is becoming completely inefficient and impractical in many domains. To evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data and to discover new patterns and relationships hidden in the data Prather et.al (2001) [29]. One of the potential tools is data mining.

Data mining is usually defined as searching, analyzing and sifting through large amounts of data to find relationships, patterns, or any significant statistical correlations. It is a new generation of computerized methods for extracting previously unknown, valid, and actionable information from large volume of database and then using this information to make critical decision [4]. The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge [15].

Applying data mining technologies to model traffic accident data records can help to understand the characteristics of drivers, vehicles, road type and weather conditions that were causation with different injury severity. This can help the decision makers to formulate better traffic safety control policies.

Data mining technology has enabled traffic control and investigation office to identify and search previously unknown, actionable information from large volume of traffic accident database and to apply it to improve the quality and efficiency of reducing and controlling road traffic accident activities. And various researchers from different parts of the world have been trying to study different issues on road safety and road traffic accident, and used different approaches to identify determinant factors of road traffic accidents using data mining tools. Since the rate of traffic accident is increasing at alarming rate in Ethiopia, in particular in Addis Ababa, related research in different approach should be done to address and solve the problem of traffic accident. In connection with this [1], [11], [20], [21] used data mining tools and techniques to solve real problem in Ethiopia in their Graduate study programme at Addis Ababa University department of Information science. Reviewing these works gave an in-depth idea, techniques and approaches to my research work.

In Ethiopia Traffic control and investigation office is responsible to control any traffic activities. Addis Ababa has its own traffic control and investigation office, which was established in 1900 E. C. with the introduction of Motor vehicles. Now the office is located at Hayahulet Mazoria. The Transport Authority, Road Authority and Police commission give support to Addis Ababa Traffic

control and Investigation Office to reduce and control road traffic accident, enforce all traffic laws and regulations, and develop and implement strategies that will improve the flow of traffic.

The Addis Ababa city Transport Authority, It got this name in 1995 E.C, gives support for the Addis Ababa Traffic control and Investigation Office in reducing and controlling traffic accident and these two organizations have the following main and common functions related to road traffic safety [2]:

1. To ensure the safety of the citizen by promoting safe and orderly flow of traffic on the street and highways.
2. To enforce all laws and regulation as they relate to each of different forms of road and highway.
3. To reduce the number of road traffic accident in the city.
4. To control road traffic accident by technical investigation of the vehicle and their power with the weight they carry.
5. To develop and implement strategies that will improve the flow of traffic, and others activities.

At the present time the traffic office with the support of Transport Authority started a radio program on FM 96.3 to teach the society about the effect of road traffic accident and to discuss how to control such severe accident in Addis Ababa city and in the neighborhood of Addis Ababa.

contribute to create accident. It was also studied that most of the accidents are created and controlled by person/drivers. Abebe (2004) [34] stated that 81% of the vehicle accident caused by the driver fault. Any driver can't drive any type of vehicles; it should be allowed based on their performance. Unless the capability of the driver on a specific vehicle is assured an unexpected accident may occur. Currently the Transport Authority implemented a driver qualification certification license proclamation, Proclamation No.600/2008.This proclamation was constructed based on drivers and vehicles characteristics. Thus, this research will answer the following questions to address the identified problem and to provide an input that used to emphasize, support and construct such type of proclamation.

- What are the main determinant risk factors (attributes) of drivers and vehicles that cause to traffic accident?
- What are the most interesting patterns or rules generated using determinant risk factors of drivers and vehicles that can used as a traffic rules and policies?
- Which Data mining techniques perform well in developing a model that can identify and predict drivers and vehicles determinant risk factors?
- How can we reduce and control road traffic accident that has various impacts on the societies in general?

Currently the Addis Ababa Traffic control and Investigation Office used small portion of data (daily, weekly, monthly, and yearly statistical data report) for its activities like decision making, revising the existing rules, policies and to induce new policies. But this has limited capacity to discover new and unforeseen pattern and relationships that are hidden in conventional database Plate et.al (1997) [29]. The effort of reducing and controlling such unmanageable road traffic

accident by identifying major determinant risk factors that causes accidents, take preventive measures on such factors and improve the quality of life manually and statistical method (daily, weekly, monthly, and yearly statistical data report) with small portion of data is time consuming, doesn't give motivating result, it is error prone and a difficult task.

Although, the existence of a severe and unmanageable number of road traffic accident were shown by different studies and road traffic accident data were gathered periodically by the Addis Ababa Traffic control and Investigation Office, due to lack of appropriate data analysis tools this historical and accumulated data, major source of solution, was not used to assess and analyze the determinant risk factors of the problem, more specifically drivers and vehicles determinant risk factors that causes a great loss of life and economy, and to find possible solution. That is, it is difficult for the traffic control office and Transport Authority to assess the determinant risk factors of drivers and vehicles and to find the hidden pattern or relationship of these factors using statistical and manual method.

1.3 JUSTIFICATION OF THE STUDY

All the above discussion both in the background and problem statement shows the wide-spread and severity of road traffic accident in the world, in Ethiopia and particularly at Addis Ababa city. One of the major tasks of Addis Ababa Traffic control and Investigation Office is to reduce and control road traffic accidents. So far the office collects road traffic accident data periodically and used only statistical method manually in small portion of data for different activities. But since the severity and the magnitude of the accidents increases time to time, the Addis Ababa Traffic

control and Investigation Office together with Transport Authority should used advanced techniques for assessing and identifying major drivers and vehicles determinant risk factors to reach at measurable and actionable recommendations in strategic planning, decision making and inducing new policies. That is the office can easily revise the existing rules, polices and induces new polices based on the assessment of determinant attributes.

Using weather conditions, road type, driver and vehicles attributes [34] developed a model using decision tree that can predict the severity level of the accident at Addis Ababa city. But I, the researcher, don't believe that such unmanageable problem will be solved by this research only. So another research and approach should be done to assess and predict the determinant risk factors of drivers and vehicles so that measures can be undertaken in advance to prevent accidents. As traffic rules, like Drivers Qualification Certification License Proclamation No.600/2008, are constructed based on drivers and vehicles factors, this study focused on drivers and vehicles details alone.

Assessing and identifying the major drivers and vehicles determinant risk factors that causes accident from huge traffic accident data automatically will save life of the society and economy of the country, and gives support for the traffic control office and Transport Authority in reducing and controlling traffic accident .So, doing a research to propose a solution for such a crucial problem is the most justifiable work and it gives sound.

1.4 OBJECTIVES

The general and specific objectives of this research are as follows:

1.4.1 GENERAL OBJECTIVE

The general objective of this research is to develop a model that can support to reduce and control road traffic accident by identifying major determinant factors (attributes) of drivers and vehicles that causes to road traffic accident using data mining tool and techniques.

1.4.2 SPECIFIC OBJECTIVES

In order to achieve the general objective the following specific objectives were planned during the research work.

1) To review different literatures that can support for the study.
2) To Preprocess and prepare the raw data into a suitable dataset for experiment.
3) To explore and select appropriate data mining tool, techniques and functionalities.
4) To build models using the preprocessed dataset, selected tool and techniques.
5) To evaluate the models built and selects the best one.

1.5 METHODOLOGIES ADOPTED

In order to achieve the proposed research objectives I used the following methodologies:

1.5.1 LITERATURE REVIEW

An in-depth literature review was made to get more insight to the concept of data mining and its application, especially in road traffic accident preventive activities and applications.

1.5.2 DATA SOURCE AND COLLECTION

For this research I used secondary data from Addis Ababa Traffic control and Investigation Office. The data is traffic accident records, which is the collection of daily accident report filled and reported by the traffic police officers. It consists of the full details about a given accident.

1.5.3 DATA PREPROCESSING

The data to be mined was collected and arranged into a new database to make it suitable for the experiment and for the selected data mining tool. That means a new database was prepared by analyzing the collected data using data preprocessing tasks like data cleaning and data reduction.

1.5.4 EXPERIMENTATION

As it is indicated in the research objective the researcher applied data mining techniques to develop a model that predict risky factors of drivers and vehicles that cause road traffic accident. So, in the experimentation part predictive models was developed by using the selected data mining tool and techniques.

In connection with understanding the data mining goal which is predictive modeling, the key task is the selection of right data mining tool and technique(s) for conducting the experiment of this research. The data mining tools support different methods for different steps of the experiment to be carried out. Even though there are several data mining tools that may fulfill the research objectives ,techniques and tasks used Weka data mining tool of version 3-5-8 was selected based on the researcher's familiarity to this tool and freely available in the internet both the software and related documentation. Decision tree and Rule induction techniques were used to build the predictive models and to generate rules.

By using such tool and techniques models were built using the prepared data set. After building the models the next step to be taken was evaluating, assessing and interpreting the models. This step will help in selecting the best model that finds an interesting pattern or rules.

1.6 SCOPE AND LIMITATION OF THE STUDY

The scope of the research is limited to assess the potential application of data mining tool and techniques, specifically predictive modeling, in supporting traffic accident reducing and controlling activities in Addis Ababa city. More specifically the research bordered to find the major drivers and vehicles determinant risk factors that are cause of road traffic accident and assessing the accident risk level using these factors.

The limitation faced during this research work was to get the data on time. I suffered more to get permission for accessing the data. Another limitation in this study was the values of the attributes recorded in interval form have an unequal width; such type of data representation has some influence in the result. The time gap given to undertake this research was also the limitations of this research, which is not sufficient.

1.7 EXPECTED BENEFITS OF THE RESEARCH RESULT

This research will give benefit in reducing and controlling of road traffic accident and improve the life of the society in general. It will support the Addis Ababa Traffic control and Investigation Office and Transport Authority in revising the existing rules, policies, and inducing new rules and policies. More specifically this research work can be used to revise and strengthen the newly implemented Drivers Qualification Certification License Proclamation No.600/2008 because it focuses on drivers and vehicles details. They also used the research result, the predictive model, as a risk assessment tool in their role of reducing and controlling traffic accident like taking preventive measure in advance on determinant risk factors of drivers and vehicles based on the assessment of traffic accident data. In general the society, the government, the researchers, domain experts, policy makers, Addis Ababa Traffic control and Investigation Office and Transport Authority will get benefit from the research result.

1.8 THESIS ORGANIZATION

This thesis report is organized in six chapter .The first chapter discuses about the background of the research, the problem of the statement, objectives and methodology used in this research work. It also shows the expected benefits of this research.

The second chapter deals about data mining technology, available techniques and function. It also discuses the general applications of data mining technology in various fields, more specifically its application in road traffic accident controlling activities.

The third chapter discuses about road transport, its role, road traffic safety and road traffic accident. It also deals the current condition of Addis Ababa city in road traffic accident and its current traffic accident controlling activities.

The fourth chapter deals about data preparation using data preprocessing techniques like data cleaning and data reduction.

The fifth chapter explains the experiment done in this research and the last chapter presents conclusions and recommendations of the research.

CHAPTER TWO

KNOWLEDGE DISCOVERY IN DATABASE (KDD) AND DATA MINING

2.1 INTRODUCTION

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. The rapid growth and integration of databases provides scientists, engineers, and business people with a vast new resource that can be analyzed to make scientific discoveries, optimize industrial systems, and uncover financially valuable patterns [6].

The progresses in digital data acquire and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human attempt, from the ordinary (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more exotic(such as images of astronomic bodies, molecular databases, and medical records) [33]. The large size and complexity of such data in many scientific domains leads impractical to manually analyze, explore, and understand the data.

Having concentrated so much attention on the accumulation of data, the problem was what to do with this valuable resource/data? It was recognized that information is at the heart of business operations and that decision makers could make use of the data stored to gain valuable insight into the business [3]. Thus, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing

5. Data mining- an essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation- to identify the interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation- where visualization and knowledge representation techniques are used to present the mined knowledge to user.

The need to scales up human analysis capabilities to handing the large number of bytes that we can collect is both economic and scientific. Businesses use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in. Because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to discover meaningful patterns and structures from the massive volumes of data. Hence, KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: data overload [10].

2.3 DATA MINING

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The fast growing, tremendous databases, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large databases become "data tombs". Consequently, important decisions are often made based not on the information rich data stored in databases but rather on a decision makers intuition,

The process of data mining consists of three stages [30].

1. Exploration – This stage usually starts with data preparation which may involve cleaning data, data transformation, selecting subsets of records and performing some preliminary features selection operations to bring the number of variables to a manageable range.
2. Model building and validation-This stage involves considering various models and choosing the best one based on their predictive performance.
3. Deployment –The final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimate of the expected out come.

Data mining is an interactive and iterative process involving data preprocessing, search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps of data mining. In general, using a variety of techniques data mining can be used to identify nuggets of information or decision making knowledge from bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, predication, forecasting and estimation[3].

2.4 DATA MINING FUNCTIONALITIES

Data mining functionalities are used to specify the kind of pattern to be found in data mining tasks. In general data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the

database, it visualize the information. Predictive mining tasks perform inference on the current data in order to make predictions, it tell us the future information [15].

Data mining methods or techniques may be classified by the function they perform or according to the class of application they can be used in. The most popular used data mining tasks are described below:

2.4.1 CLASSIFICATION

Classification is learning a function that maps (classifies) a data item into one of several predefined classes. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from a historical database. The main objective of classification is to identify the characteristic that indicate the group to which each case belong. This pattern can be used both to understand the existing data and to predict how new instance will behave. That is the system take a case or records with certain known attribute values and able to predict what class this case belongs to.

2.4.2 PREDICTION

Prediction can be viewed as the construction and use of a model to asses the class of unlabeled sample is likely to have [15].That means, it predict unknown or missing class value.

The concept of predictive data mining refers to the application of a model for prediction or classification to new data. After a satisfactory model or set of models has been identified (trained) for a particular application, one usually wants to deploy those models so that predictions or predicted classifications can quickly be obtained for new data.

The difference between classification and prediction was explained by Gregory Piatetsky-Shapiro (Ph.D., Data Mining and Analytics Expert) for KDnuggets News based on the following question [42]

If one does a decision tree analysis, what is the result? A classification? A prediction?

Dr. Gregory Piatetsky-Shapiro said the decision tree is a classification model, applied to existing data. If you apply it to new data, for which the class is unknown, you also get a prediction of the class. He also said the assumption is that the new data comes from the similar distribution as the data you used to build your decision tree. In many cases this is a correct assumption and that is why you can use the decision tree for building a predictive model.

2.4.3 ASSOCIATION

Given a collection of items and a set of records, each of which contain some number of items from the given collection, an association function is an operation against this set of records which return affinities or patterns that exist among the collection of items . These patterns can be expressed by rules such as '72% of all the records that contains item A, B and C also contain items D and E [3].

Association rule mining finds interesting association or correlation relationship among a large set of data items. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision making processes, such as catalog design and cross marketing. A typical example of association rule mining is market basket

analysis. This process analyses of customer buying habits by finding associations between the different items that customers place in there shopping baskets [15].

2.4.4 CLUSTERING

Clustering is defined as the processes of creating a partition so that all the members of the partition are similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Clustering divides a database into different from each other, and whose members are very similar to each other [3]. Unlike classification, one doesn't know what the clusters will be when clustering start, or by which attributes of the data will be clustered, it is unsupervised learning.

2.4.5 TIME SERIES DATA ANALYSIS

Time series data often arise when monitoring industrial processes or tracking corporate business metrics. Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for analysis [13], [26]. One of the main goals of time series data analysis is to forecast future values of the series. In general, an effective approach to time-critical dynamic decision modeling should provide explicit support for the modeling of temporal processes and for dealing with time-critical situations.

2.4.6 SUMMARIZATION AND VISUALIZATION

Before building good predictive models one must understand the data. Summarization and visualization involves methods for finding a compact description for a subset of data by gathering

a variety of numerical summaries (including descriptive statistics such as averages, standard deviations and graphs) and looking at the distribution of the data [35].

In this research classification data mining task was applied to develop predictive model and to identify determinant factors of drivers and vehicles by sharing Dr. Shapiro [42] idea.

2.5 DATA MINING TECHNIQUES

A database is a store of information but more important is the information which can be inferred from it. In order to do this a wide variety of data-mining methods or techniques should be used. There is no particular rule that would tell you when to choose a particular technique over another one. Sometimes those decisions are made relatively arbitrary based on the availability of data mining analysts who are most experienced in one technique over another. These techniques can be used for either discovering new information within large databases or for building predictive models [23]. Decision trees, Rule Induction (Rule Learner), Neural Network, Clustering, and Association Rule Mining are some of data mining techniques that are used in most cases. Among these different techniques the researcher used Decision Trees and Rule Induction techniques in this research because their result is simple to explain for end user and these techniques support the selected data mining tasks for this research.

2.5.1 DECISION TREES

Decision trees defined as simple knowledge representation and they classify examples/records to a finite number of classes, the nodes are labeled with attribute names, the edges are labeled with

1. Classification trees- Takes categorical values and label records and assign them to the proper class. The classification tree reports the class probability, which is the confidence that a records is in a given class.
2. Regression tree- Estimates the value of a target variable that takes on numeric values.

The process of building a tree starts with a training set consisting of pre classified records. Pre classified means that the target field, or dependent variable, has a known class. The goal is to build a tree that distinguishes among the classes. That is, the tree can be used to assign a class to the target field of a new record based on the values of the other fields or independent variables [23].

2.5.1.1 DECISION TREE MODELING PROCESSES

All decision tree construction methods are based on the principle of recursively partitioning the dataset until homogeneity is achieved. The construction of decision tree involves the following three phases [16]:

1. Construction phase: - The initial decision tree is constructed in this phase, based on the entire training dataset. It requires recursively partitioning the training set into two or more sub-partitions using a splitting criterion, until a stopping criterion is met. The basic strategy of construction phase described as follows[15]:
    a. The tree starts as a single node representing the training samples.
    b. If the samples are all of the same class, then the node becomes a leaf and is labeled with that class.
    c. Otherwise, the algorithm uses entropy-based measures known as information gains as a heuristic for selecting the attribute that will best separate the samples into

individual classes. The information gain measure is used to select the test attribute at each node in the tree. The attribute with the highest information gain is chosen as the test attribute for the current node.

2. Pruning phase: the pruning phase involves removing some of the lower branches and nodes to improve performance.
3. Processing the pruned tree: in this step decision tree is processed to improve understandability.

Various listed decision tree algorithms such as CHAID , C4.5/C5.0, CART , ID3 and many others with less familiar algorithms produce trees that differ from one another in the number of splits allowed at each level of tree, how those splits are chosen when the tree is built [33].

2.5.2 RULE INDUCTION (RULE LEARNER)

Rule induction (sometimes called rule learner) is one of the major forms of data mining techniques and is perhaps the most common form of knowledge discovery learning systems. It is also perhaps the form of data mining that most closely resembles the process that most people think about when they think about data mining, namely "mining" for gold through a vast database. The gold in this case would be a rule that is interesting - that tells you something about your database that you didn't already know and probably weren't able to explicitly articulate [33].

Rule induction on a database is a process undertaking using intelligent software where all possible patterns are systematically pulled out of the data. In this process the accuracy is added to

them that tell the user how strong the pattern is and how likely it is to occur again [33]. Rule induction has been widely used to represent knowledge in expert systems and they have the advantage of being easily interpreted by human experts because of their modularity [4], [12].

Rule induction systems are highly automated and are probably the best of data mining techniques for exposing all possible predictive patterns in a database. They can be used in prediction problems but the algorithms for combining evidence from a variety of rules come from practical experience [33].

In rule induction systems the rule itself is of a simple form of "if this and this and this then this". In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule. These are accuracy and coverage. Just because the pattern in the database is expressed as rule does not mean that it is true all the time. Thus just like in other data mining algorithms it is important to recognize and make explicit the uncertainty in the rule, i.e. how often is the rule correct. This is what the accuracy of the rule means. The coverage of the rule has to do with how much of the database the rule "covers" or how often the rule applies [33].

2.5.2.1 RULE INDUCTION FOR PREDICTION

Data mining techniques are based on data retention and data distillation. Rule induction belong to the logical, pattern distillation based approaches of data mining. These technologies extract patterns from dataset and use them for various purposes, such as prediction of the value of a dependent field. After the rules are created and their interestingness is measured there is also a call for performing prediction with the rules. When the rules are mined out of the database the

rules can be used either for understanding better the problems domains that the data reflects or for performing actual predictions against some predefined prediction target [25].

2.5.2.2 RULE INDUCTION ALGORITHMS

Rule induction seeks to go from the bottom up and collect all possible patterns that are interesting and then later use those patterns for some prediction target. Rule induction systems retain all possible patterns even if they are redundant or do not aid in predictive accuracy. For instance, consider that in a rule induction system if there were two columns of data that were highly correlated (or in fact just simple transformations of each other) they would result in two rules [33].

Rule induction technique applies an iterative process to generate a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process. Some algorithms of these kinds of systems which are supported by weka software are described below [22]:

PART: PART is a separate-and-conquer rule learner proposed by Eibe and Witten. The algorithm produce sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning.

Decision Table: Decision Table algorithm builds a simple decision table classifier as proposed by Kohavi. It summarizes the dataset with a 'decision table' which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table.

ConjunctiveRule: ConjuctiveRule algorithm implements a single conjunctive rule learner that can predict for numeric and nominal class labels. A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification/regression. In this case, the consequent is the distribution of the available classes in the dataset. If the test instance is not covered by this rule, then it's predicted using the default class distributions/value of the data not covered by the rule in the training data. This learner selects an antecedent by computing the Information Gain of each antecedent and prunes the generated rule using Reduced Error Pruning (REP) or simple pre-pruning based on the number of antecedents. OneR, Ridor and JRip (RIPPER) are also algorithms of rule induction technique.

2.6 POTENTIAL APPLICATIONS OF DATA MINING

Various fields of data mining application were described below [3]:

1. Retail/marketing – Data mining used in marketing/retail to identify buying patterns from customers, find associations among customer demographic characteristics, predict response to mailing campaigns and market basket analysis.
2. Banking – In this field data mining has the functions of detecting patterns of fraudulent credit card use, identify loyal customers, predict customers likely to change their credit card affiliation, determine credit card spending by customer groups, find hidden correlation between different financial indicators and identify stock trading rules from historical market data.

3. Insurance and Health care–The potential applications of data mining in this area are claims analysis i.e. which medical procedure are claimed together, predict which customers will buy new policies ,identify behavior patterns of risky customers and identify fraudulent behavior.

4. Transportation –In this field data mining can be used to determine the distribution schedule among outlets, analyze loading patterns, and analyze the performance of the engines of airplane and to analyze road traffic accident.

5. Medicine-Data mining has also good applications in this field to characterize patient behavior to predict office visits and identify successful medical therapies for different illnesses.

2.7 APPLICATION OF DATA MINING IN ROAD TRAFFIC ACCIDENT CONTROLLING ACTIVITIES

Patterns involved in dangerous crashes could be detected if we develop accurate prediction models capable of automatic classification of type of injury severity of various traffic accidents. These accident patterns can be useful to develop traffic safety control policies. I believe that to obtain the greatest possible accident reduction effects with limited budgetary resources, it is important that measures be based on scientific and objective surveys of the causes of the accidents and severity of injuries.

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. Thus applying data mining techniques to model traffic accident data records can help to understand the problem and to find possible solutions. This can help the decision makers to formulate better traffic safety control policies. According to Chong et.al [34] the application of data mining in the area of road transport can be classified as:

1. Traffic density analysis; measurement and investigation of traffic accident volumes.
2. Traffic accident analysis, identifying determinant factors in road traffic accidents and other related issues.
3. Injury severity analysis; modeling and predicting the severity of injuries resulting from traffic accident.

Identifying in advance the area in which traffic accident volume is high, identifying determinant risk factors of vehicles, drivers, weather conditions, road type and pedestrian, and analyzing the injury severity of the accident using traffic accident data periodically and automatically needs data mining tools and techniques. Thus, this will help to facilitate and simplify the activities of reducing and controlling road traffic accident.

2.8 RELATED RESEARCH WORKS

Reviewing related works that were done using data mining tools and techniques at different place and time in the same problem domain gave to the researcher an in-depth insight for this research. The following paragraphs discuss about related research works done on the same problem domain.

Neural network used to detect safe driving patterns that have less chance of causing death and injury when a car crash occurs. The research indicates that by controlling a single variable such as the driving speed, or the light conditions they potentially could reduce fatalities and injuries by up to 40% [41].

A study was done using neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy. The researchers feed-forward MLP using BP learning. The model had 10 input nodes for eight variables (day or night, traffic flows circulating in the intersection, number of virtual conflict points, and number of real conflict points, type of intersection, accident type, road surface condition, and weather conditions). The output node was called an accident index and was calculated as the ratio between the number of accidents for a given intersection and the number of

accidents at the most dangerous intersection. Results showed that the highest accident index for running over of pedestrian occurs at non-signalized intersections at nighttime [24].

Poisson regression was used to analyze the association between the fatal crash rate (fatal crashes per vehicle mile traveled) and the speed limit increase and found that the speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in Washington State [27]. A log-linear model was also developed to clarify the role of driver characteristics and behaviors in the causal sequence leading to more severe injuries. They found that driver behaviors of alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries [17].

Abebe (2004) [34] tried to assess points that need immediate attention with respect to road safety, He stated that 81% of the accident all over the country, in Ethiopia, is due to drivers fault and the other is due to vehicle, pedestrians and road types. He also mentioned the main road safety problem. These are drivers not respecting pedestrians' priority, over speeding, poor skill and undisciplined behavior of drivers, weak traffic law enforcement and less engineering effort in road design to consider safety.

An automated traffic information system was proposed for Addis Ababa region traffic office with the aim of helping the office in information handling, which facilitates better traffic accident controlling mechanism [19]. Decision tree was used in developing a model that can support road traffic accident severity analysis. The researcher used road types, weather conditions, and vehicle type and driver determinant factors to build a model. In his study he found that attributes such as accident causes, accident type, driver age, road surface type, road condition, vehicle type and light condition are important variables for the classification of accident severity level [34].

All the above related research works used as a guideline to do this research. Besides this input having a brief idea about road transport and road traffic accident, and reviewing and discussing the existing traffic control mechanisms of Addis Ababa city will simplify the task of the research. So, the next chapter will discuss about road transport, road traffic accident and in general how Addis Ababa Traffic control and Investigation Office works.

CHAPTER THREE

ROAD TRANSPORT AND ROAD TRAFFIC ACCIDENT

3.1 ROAD TRANSPORT AND ITS ROLE

It is appreciated that transportation is an integral part of the functioning of society, like arteries and veins of blood system in human being. It exhibits a very close relationship to the style of life, the range and location of productive and leisure activities and the goods and services which will be available for consumption. In early days, horses, mules, and carts were used as a means of transport. But with technological development, different kinds of modern transportation, road (vehicle), air, rail and marine came in to existence.

Majority of the transport operation services in almost all countries of the world is provided by one mode of transport, namely road transport. Road transport is (alternatively road transportation) is transport on roads, i.e. transport over land which is not rail transport in the wide sense [18]. The road transport mode is named from the infrastructure that the vehicles use, which is road. The road is specifically designed and surfaced highways for the passages of wheeled vehicles, each vehicle being controlled and guided independently by a driver.

Road transport plays vital roles in the effort of uplifting the economy of the society. It facilitates more the conveying of passengers, by products of agricultural industry, and freight from origin to the destination. Dynamic and efficient transport sector is a decisive factor, without which it is

impossible to reach the goal of national socio-economic development. For instance, in the case of developing and agrarian countries, like Ethiopia, road transport facilitates the communication between people and it is an important sector to facilitate different socio-economic activities.

In Ethiopia road transport is growing and becoming the dominant sub sector over the others sub modes for the last couples of decades. Presently the share of road transport sub sector accounts for about 90% revealing that the massive transport is becoming carried out through it in the country [32]. Ethiopia is one of the land locked country in Africa, this implies that the country's socio-economic development is expected to relay on the road transport in conveying good and services through out the country. Despite of all these facts the sub sector is trapped by so many complex factors. The traffic safety problem is one of these factors and is ringing in every citizen's mind. So, efficient and safety transportation service of adequate quality and capacity should be provided to avoid some of the most disturbing effects like pain, damage to property, injuries and loss of life brought by unsafe acts and unsafe events in the transportation system.

3.2 ROAD TRAFFIC SAFETY AND POLICIES

There is a great need for a systematic arrangement of the major elements of street and highway safety which will encourage user-friendly channels of communication and cooperation. The ultimate goal of road traffic safety is to enhance the movement of people and cargo on a street and highway system, and reduce the incidence of traffic accidents through a basic understanding of the factors involved. Road traffic safety is promoted through education, regulation and law enforcement, as a result it attain a most efficient and safe road transport system.

Road traffic safety aims to reduce the harm (deaths, injuries, and property damage) resulting from crashes of road vehicles. It deals exclusively with road traffic crashes-how to reduce their number and consequences. Road traffic safety can be improved by reducing the chances of driver making an error, or by designing vehicles to reduce the severity of crashes that do occur. Safety interventions focusing on motorized vehicles and their drivers. For instance by compulsory training and licensing (periodic retests of drivers); compulsory safety testing of vehicles over a certain age i.e. a well-designed and well-maintained vehicle, with good brakes, tires and well-adjusted suspension will be more controllable in an emerging and thus be better equipped to avoid collusions [40].

According to WHO (2004) [34] some of the important points that need attentions in understanding and dealing with road safety are:

- Road safety policy must be based on a sound analysis and interpretation of data, rather than on anecdote.
- Road safety is a public health issue that intimately involves a range of sectors including that of health. All have responsibilities and all need to be fully engaged in injury prevention.
- Since human error in complex traffic systems can't be eliminated entirely, environmental solutions (including the design of roads and vehicles) must help in making road traffic safety system.

Appropriate road safety policy is one of the essential elements of well-balanced overall transport and public health policy. Cost-efficient traffic and transport regulation depends on a sound road safety policy and management. All traffic measures should first be evaluated, not only in terms of road safety but also from the economic and environmental point of view. Not respecting these demands brings unwanted consequences.

Road safety policies obeyed by the traffic law includes all police activities relating to the observation of traffic violations and the police actions to be taken, such as warning, reporting, summonsing, and arresting. Whenever possible the form of enforcement used should be designed to educate those who have violated the law and others who may be influenced by their examples so that such unlawful and unsafe driving will not be repeated [7].

3.3 ROAD TRAFFIC ACCIDENT

Traffic accidents are a major problem in both developed and developing countries. The increase in the number and gravity of traffic accidents is directly related to the increase in the use of motorized transport. This problem originated in developed countries in the first decades of this century. Early on, accidents were considered either an act of God, or an unavoidable consequence of modern life. The first action to be taken to control traffic accident was seeing the accident as a man-made problem and no longer as a question of fate. Therefore, the accident could be prevented. The second major change that followed was seeing the accident as a public health problem [7].

Traffic accidents cause several physical damages (life and property) to those who survive. Some of these effects are temporary, some are permanent, some of people will be totally impaired for most activities, and others will be partially.

Traffic accident is increasing at alarming rate due to the invasion of motor vehicles. Motorized vehicles are much more damaging in face of the much higher kinetic energy involved and its harming potential. When motorized vehicles are large (trucks, buses, and automobiles) or travel at high speeds (automobiles and motorcycles) danger of severe harm is much higher, especially against pedestrians and cyclists. Human beings face various types' accidents, among these road traffic accident also called car accident cause considerable damage to both human life and property.

As discussed in the background section it was estimated that 1.2 million people were killed world wide as a result of road traffic collision in 2002. About 20 to 50 million people were injured and about 5 million people were disabled for life. In 1990 road traffic injury was the 9th leading contributor to the global burden of disease. If current trends continue it will be the 3rd leading contributor to the global burden of disease by the year 2020 [38].

In general, road traffic accidents can be prevented and their consequences can be alleviated if the appropriate policies, strategies safety regulation and guidelines are in place. There are strategies and experiences that have been tested and proven to be effective in developed countries, which developing countries can successfully adapt to their own settings.

3.4 ROAD TRAFFIC ACCIDENT IN ADDIS ABABA

Ethiopia is one of those developing countries with low level of income accompanied by high rate of population growth. As part of the developing world, Ethiopia is predominantly an agrarian country with low level of urbanization. The economic performance of different sectors of the national economy is very low. This low performance is due to a number of constraints such as low level of investment in different sectors of the national economy. Among these the existing transport could be mentioned as one [2].

Addis Ababa is the seat of government of the Federal Democratic Republic of Ethiopia (FDRE). It is also home to the African Union, the Economic Commission for Africa and other International Organizations. Urban transport in Addis Ababa is carried out by the mixture of ownership structures, of which public and private operators are predominantly contenders for business. The mode of transport systems in Addis Ababa are categorized into motorized and non-motorized traffic. As such the modes of transport include public bus, minibus, taxis, and non-motorized transport, walking and animal carts dominant the periphery [2].

The scarcity of infrastructures, the growth of population and vehicles together with the low standing of drivers and vehicles, and problems in executing traffic law has made Addis Ababa to cover the largest share of traffic accident in the country [2].

As it can be seen from various reports of WHO and statistical reports of Addis Ababa traffic control office, at Addis Ababa the total number of road traffic accident (deaths and injuries) at all levels are growing at alarming rate. Addis Ababa is experiencing more than its share of road accidents. It is always important to remember that an accident could happen as a result of any of

the following elements or a combination of these elements; the road environment, the vehicle, the driver, weather condition, road user or pedestrians. Road traffic accident in Addis Ababa is a very crucial problem due to these several and complex factors. Thus, there should be an effort and mechanism to control it.

3.5 ROAD TRAFFIC ACCIDENT CONTROLLING ACTIVITIES IN ADDIS ABABA

Traffic police means member of the police force assigned to enforce the observance of traffic rules [8]. The primary duty of the police is to protect the life and property of the society. Traffic police are specialized unit under traffic control office to facilitate traffic enforcement, which takes care of road traffic safety. The main objective of traffic control office is to facilitate traffic controlling activity, which is an attempt to solve the problem of human survival on the roads in our mechanical and electronically era, i.e. to protect road users. The control includes the protection of all road users against each other and even the individual road user against him. This kind of protection demands from the special organization traffic police, particular knowledge of different sciences, special analysis methods, planning and assessing police road safety process, special education and training programs, special police tactics, police control devices and foreign police experiences. The office gives such training to build the capacity of the traffic polices.

The second partner that involve in the traffic accident controlling activity in Addis Ababa is the Addis Ababa city road Authority; which is responsible for selecting, installing and maintaining traffic control devices. These traffic control devices are used in regulating, warning, and guiding

or channeling traffic. To achieve these purposes, the traffic signs and road marking are regularly maintained, renewed and replaced whenever necessary by Addis Ababa city road authority.

The third partner that is responsible for traffic accident controlling system is Addis Ababa city Transport Authority. The following are the main objectives of Transport Authority [2]:

1. Facilitate transportation system.
2. Register vehicles, annual vehicle investigation and give vehicle ownership.
3. Implement efficient traffic management and it gives capacity building for the traffic control office.
4. It gives driving license based on the rules and regulations.
5. It works to prevent traffic accident by investigating vehicles performance and give parking services.

Based on the responsibility given by proclamation No.41/1985 to the Transport Authority, it works to enforce traffic rules and regulations' using proclamation No.5/1990.This traffic proclamation has six levels of punishment for laws offender. The reason for implementing these laws is the traffic accident is rising at alarming rate in the city, the existing traffic rules should be revised based on the current conditions and there is a need of modern traffic laws that supports and facilitates the socio-economic development of the city, and the country in general. Currently the Transport Authority also implemented new proclamation, proclamation No. 600/2008, which is used to provide for driver's qualification certification license. The purpose of this proclamation, which has 27 articles, is to [8]

1. Ensure that drivers operate vehicles in appropriate condition by acquiring adequate driving skill to achieve safe transport service,
2. Set nation wide driving qualification standard and establish a system for the issuance of driving license qualification certification free from forgery, corruption and bureaucratic red tape.
3. Ensure bilateral and multilateral agreements relating to qualification of driving and movements of traffic on any Ethiopia roads are observed by drivers.

On Article 7 the proclamation deals about the drivers' qualification certification license in different category of drivers' qualification certificate permit and type of vehicle operated. For instance where a holder of drivers' qualification certification license wants to obtain a driver's qualification certification license of different category the driver shall be required to take the theoretical and practical training and the test specified for such category of license. The proclamation also deals about age and educational background of drivers requirements to a specific category of vehicles. For instance, in Article 12(2) in the case of tanker, bus or taxi driver's qualification certification license, the driver should have completed at least eight grade educations and attained the age of not less than 24 year.

The proclamation also include Article 19 that deals about validity and renewal of driver's qualification license, for instance, Article 19(1) a drivers qualification certification license of any category shall be valid for a period of four years from the data of its issuance. According to Article 20(1) suspension and revocation of driver's qualification certification licensing rule is explained i.e. where based on the traffic offence records of a license holder or on other sufficient

grounds, it is proved that either his physical fitness or driving skills deficient, the licensing body may suspend the drivers qualification certification license or require him to undergo medical examination or order him to take driving qualification test or both. Article 25 was also included for the purpose of punishment for offences committed by drivers; the record of petty and aggravated offences shall be valid for one and three years respectively. Thus, in this study emphasis was given how to construct such types of rules or proclamation using data mining technology at the same time Proclamation No.600/2008 will be strengthen and revised, since it was constructed based on drivers and vehicles characteristics.

In general, the authority has given a responsibility to control road traffic safety and facilitate better traffic management activities in the city; as a result the city will have modern traffic safety system. The authority also has a responsibility to provide sufficient transportation system in the city.

So, all these partners are working together tightly to control traffic accident in Addis Ababa. In addition to these partners, currently WHO supports to strengthening the road traffic injury data management capacity of traffic police of Addis Ababa city. The project focuses on; the development of an easy to use data collection form; a computer-based database system; training the traffic police officers on data management; and promoting collaboration among key stakeholders in road traffic safety.

Currently the traffic control office uses computer based database management system (access database in Amharic language) to store traffic accident data. The accident details described using

47 attributes/fields .Each traffic accident that results in personal injury and property damage requires the traffic officer to identify the detail characteristics about the crash and completing a narrative description of the accident events in a form prepared for this purpose. During the accident, characteristics of the crash, the vehicles, people involved, weather conditions and road type information are recorded on the form and reported to the traffic control officer's. The report also includes the result of the officer's investigation about the accident itself. The accident report form asks for information such as the time and location of the accident, the number of vehicles involved, the vehicle type, drivers' information, injury severity etc.

According to IHT (1990) [19] accident reporting system activity constitutes the root of the whole accident investigation process. Accident data collection forms should be carefully designed so as to capture all the necessary information required to perform an overall as well as an in-depth accident data analysis.

The accident record is basically used for various purposes in the office and for other stakeholder. National and regional transport offices use the data in directing their focuses of attention in decision and policy makings with regard to road safety. Different health offices and non-governmental organizations working in this area use the data in determining and managing health problem in society [34].

The purpose of any traffic accident analysis system is to find the possible causes of accidents related to vehicle, roadways, drivers, pedestrians and weather conditions, and to plan measures to protect the road traffic accident by reducing the frequency and severity of the accidents. Accurate

accident reports and records are the foundation for analysis and prevention of traffic accident. These records serve in shaping traffic law enforcement and traffic education policies and procedures .So analyzing the collected traffic accident data of Addis Ababa city will improve the accident prevention mechanism and to revise the existing traffic laws. Thus, in the next chapter data preparation using data preprocessing techniques on the collected road traffic accident data will be done to make it suitable for experiment.

CHAPTER FOUR

DATA PREPARATION FOR EXPERIMENT

4.1 PREREQUISITE TASKS OF DATA MINING

An important requirement to any data mining activity is the data itself. Real-world data tend to be dirty, incomplete and inconsistent. Data preprocessing techniques can improve the quality of data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step and requirement in the data mining process, since quality decisions must be based on quality data [15]. Data understanding, data preprocessing and preparing suitable database for the selected data mining tool and techniques are the most important prerequisite tasks for data mining process.

4.2 GOAL OF THE RESEARCH

Data analysis is the basis for investigations in many fields of knowledge, from science to engineering and from management to process control. Data on a particular topic are acquired in the form of symbolic and numeric attributes. The source of these data varies from human beings to sensors with different degrees of complexity and reliability. Analysis of these data gives a better understanding of the phenomenon of interest [9].It also gives a clear understanding to the goal of the research.

The goal of data mining is to produce new knowledge that the user can act upon. It does this by building a model of the real world based on the data collected from a variety of sources which may include corporate transaction, process control data, credit information, weather data [35].The goal of this research work is to develop a model that can help the Addis Ababa Traffic control and Investigation Office in the effort of reducing and controlling road traffic accident and generating rules, rules related to drivers and vehicles. To solve such crucial issues understanding the problem is the first important step, which was briefly discussed in the statement of the problem section, besides understanding the goal of the research.

4.3 DATA SOURCE AND DATA COLLECTION

There are two keys to success in data mining. First is coming up with a precise formulation of the problem you are trying to solve. The second key is using the right data [35].That is analyzing and understanding the content and structure of the collected data is one of the most important tasks that need attention in data mining process.

The data for this research work was taken from Addis Ababa Traffic control and Investigation Office. The data was road traffic accident data, which was filled and reported by the traffic officers. It contains full details for each accident. That is, the data contains details information about drivers, vehicles, weather condition, road types and victim pedestrians. The Addis Ababa traffic control and investigation office stores this data using access database in Amharic language. The data set used for this research is the data which was collected and reported by the traffic officers from 28/9/99 to 10/7/2000 E.C.

4.4 DATA DESCRIPTION AND DATA UNDERSTANDING

The Addis Ababa Traffic control and Investigation Office keeps detail information about an accident using different attributes. The attributes describes details of drivers and vehicles characteristics, road and weather conditions, date and time of the accident, accident type and investigated accident causes. The attributes and their description are shown in Table 4.1.

S.No	Attribute	Description	Type
1	Reg- No	Registration Number of the accident	Number
2	Month	Month in which the accident occurred.	Number
3	week	Week in which the accident occurred.	Number
4	Date	Date in which the accident occurred.	Date/Time
5	Time	Time in which the accident occurred.	Date/Time
6	Date of the week	Date of the week	Text
7	Driversex	Sex of the driver	Text
8	Driverage	Age of the driver	Number
9	Accused No.	No. of accused persons	Number
10	Accused male	No. of accused male	Number
11	Accused female	No. of accused female	Number
12	Accused age	Age of the accused person	Number
13	Educationalbackground	Educational background of the driver	Text
14	Relationshipwith	Driver relationship with vehicle	Text
15	Experience	Driving experience of the driver	Number
16	Typeofvehicle	Types of vehicle	Text
17	Platecode	Plate code of the vehicle	Number
18	Ownership	Owner of the vehicle	Text
19	Vehicleserviceyear	Vehicle service year	Number
20	Automotivestatus	Mechanical status of the vehicle	Text

21	Sub city	Sub city of the accident area	Text
22	Kebele	Kebele of the accident area	Text
23	Place	Place of the accident area	Text
24	Special place	Special place of the accident area	Text
25	Road _segmentation	Types of road segmentation	Text
26	Road_ direction	Types of road direction	Text
27	Road_joint	Types of road joint	Text
28	Type of road	Types of road	Text
29	Road _condition	Types of road condition	Text
30	Weather condition	Weather condition during the accident	Text
31	Air_condition	Air condition during the accident	Text
32	Accused vehiclemovement	Vehicle motion during the accident	Text
33	Typeofaccident	Type of accident	Text
34	Accident upon with	Accident upon with	Text
35	No of crushed	No of crushed vehicles	Number
36	Year	Year	Date
37	Cause of accident	Cause of accident	Text
38	Licensegrade	License grade of the driver	Number
39	Injury	Type of crush	Text
40	Injury age	Age of the injury person	Number
41	Injury occupation	Injury occupation	Text
42	Health status	Health status of the injury	Text
43	Movement of pedestrian	Movement of pedestrian	Text
44	No of injury person	No of injury person	Number
45	Injury person female	No. of injury females	Number
46	Injury person male	No. of injury males	Number
47	Investigator	Investigator name	Text

Table 4.1 Attributes used to store road traffic accident information

The success of the objective of a given research is a result of proper data and problem understanding. Understanding the collected data clearly will help to do an efficient data preprocessing, as a consequence the efficiency and accuracy of the data mining result will be improved. That is a good understanding of the data at hand lead to a better success in achieving the data mining goal, more specifically the objective of the research. The collected data for this research was not clean and suitable for experiment as it is. So, the compulsory step to exercise data mining activity on data, data preprocessing took most of the research time. To do so different data preprocessing techniques were used to prepare suitable database for the experiment.

4.5 DATA PRE PROCESSING

Data pre processing step consists of the core of the data mining and take anywhere from 50% to 90% of the time and effort of the entire knowledge discovery process [34]. Data preprocessing techniques used to improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent data mining process [15].

In order to solve data problems such as noisy data, irrelevant or missing attributes and values in the dataset, one may be interested, in learning more about the nature of the data, or changing the structure of the data to prepare the data for a more efficient data analysis [9]. Data preprocessing consists of all the action taken before the actual data analysis process starts. Data preprocessing is defined as a transformation that transforms the raw real world data to a set of new data [9].

The basic limitation in data collection and data analysis are due to the quality and completeness of the data. Inaccuracies in the measurement of inputs or incorrect feeding of the data to data

analysis tool, e.g. classifiers could cause various problems. It is therefore the primary task in data analysis to identify these insufficiencies. The preprocessing of the data is a time consuming, but critical, the most compulsory activity in the data mining process [9].

The first action that was done after collecting the data, in this research, was transferring the data from access database to excel format to take advantage of easier data manipulation and also compatible interaction with the selected tool, weka data mining tool, which was selected to do the experiment. The data that was collected and stored in excel format passed through important steps of data preprocessing such as data cleaning and data reduction. Detail activities that were done during data preprocessing are described below.

4.5.1 DATA CLEANING

Data cleaning is a routine work to "clean" the data by filling in missing values, smoothing, and noisy data, identifying or removing outliers, and resolving inconsistencies. Dirty data can cause confusion for the mining procedure, resulting in unreliable output [15]. The reliability on output of data mining procedure highly depends on cleanliness of the data.

The total number of records taken for this task was 6107.In the collected dataset all records that had unknown values for most of the attributes due to the drivers' goby after creating an accident were removed and record that contains missing values in most of the attributes also removed.

Missing or insufficient attributes, values are examples of the data problems that may complicate data analysis tasks such as learning and hinder accurate performance of most data analysis

system. For example, many data analysis applications involve splitting the data into training and testing sets. Although the splitting process may be iterated several items, missing attributes values may cause inaccurate evaluation of the results [9].

4.5.2 DATA REDUCTION

Data reduction defined as a data preprocessing technique that used to obtain a reduced representation of the dataset that is much smaller in volume, yet produces the same (or almost the same) analytical results [15]. There are a number of strategies for data reduction, these include data aggregation, dimension reduction (removing irrelevant attributes, data compression using encoding schema such as minimum length encoding and sampling).

The dataset of this research contains 47 attributes; some of them are irrelevant to the mining task according to the research objective specified. The research focused on drivers and vehicles determinant risk factors of road traffic accident, which are used to generate traffic accident controlling rules. So, some attributes that doesn't have relation to the specified objective were removed.

Reducing the dimensionality of (through eliminating irrelevant data) may improve the performance of a data analysis tool, since the number of training examples needed to achieve a desired error rate increases with the number of measured variables or features increases [9].

In data translation, the data are transferred or consolidated into forms appropriate for mining process. The collected dataset for this research was in Amharic language and it was translated into English language. All data translation (It is not a data transformation, it is simple data encoding task) process result is given in Table 4.2.

Attribute name =DriverAge	
Old value	**Transformed value**
<18 years	Groupone
18-30 years	Grouptwo
31-50 years	Groupthree
≥51 years	Groupfour
Attribute name = Experience	
<1 years	Lev_ one
1-2 years	Lev_ Two
2-5 years	Lev_ Three
5-10 years	Lev_ Four
≥ 10 Years	Lev_ Five
Attribute name =Vehicleserviceyear	
<1 years	Ser_ level one
1-2 years	Ser _level two
2-5 years	Ser_ level three
5-10 years	Ser_ level four
≥ 10 Years	Ser_ level five
Attribute name=Educationalbackground	
<6 grade level	level one
7-8 grade level	level Two
9-12 grade level	level Three
>12 grade level	level Four

Attribute name =LicenseGrade	
1st	First level
2nd	Second level
3rd	Third level
4th	Fourth level
5th	Fifth level

Table 4.2 Table that shows data translation values

Some drivers doesn't have license and some have special license .These two values were represented by NoLicense and special during transformation of Licensegrade attribute values.

Although data preprocessing is useful and in many application it is a compulsory step in order to perform a meaningful data analysis, if proper techniques are not selected, it may result in loss or change of useful information to be discovered during the analysis. To perform meaningful data preprocess, either the domain expert should be a member of the data analysis team or the domain should be extensively studied before the data is preprocessed [9]. Involvement of the domain expert, traffic officers, in this research work would result in some useful feedback to verify and validate the use of particular data preprocessing techniques. The data preprocessing activities in this research was done by considering all these points and facts.

Fig 4.1 shows output and tasks performed in data preparation.

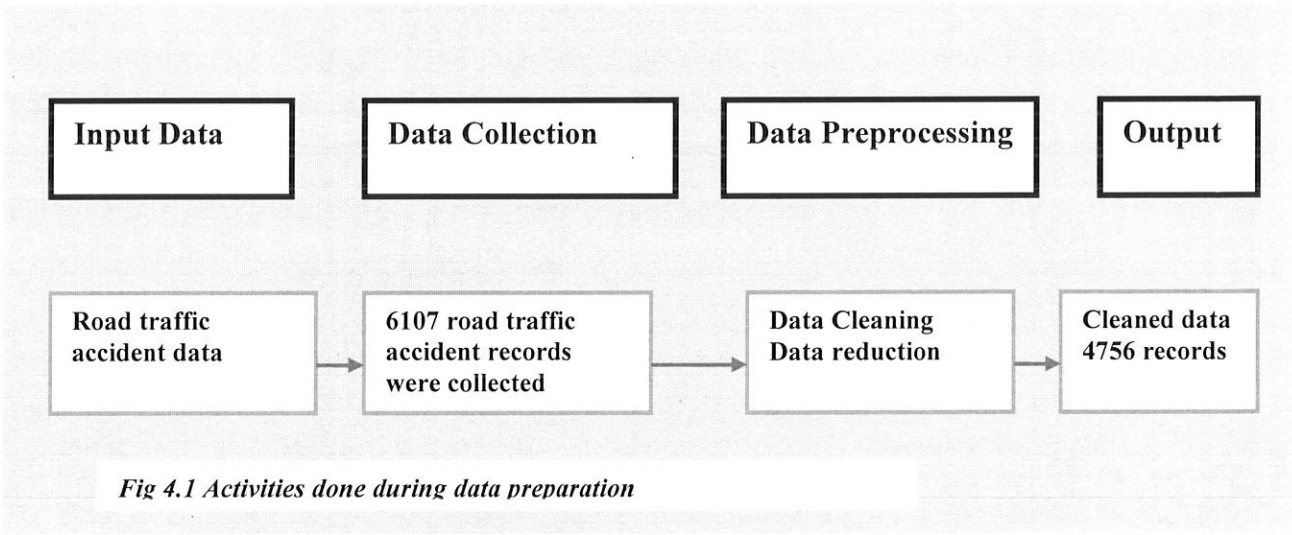

Fig 4.1 Activities done during data preparation

By understanding the problem statement clearly, considering the research objective and through an in-depth discussion with the traffic officers the following attributes, Table 4.3, were taken as a candidate for the experiment to reach the main objective of the research.

S.No	Attribute name	Description	Type
1	DriverSex	Sex of the driver	Text
2	DriverAge	Age of the driver	Text
3	Educationalbackground	Educational background of the driver	Text
4	Relationshipwith	Driver relationship with vehicle	Text
5	Experience	Driving experience of the driver	Text
6	Typeofvehicle	Types of vehicle	Text
7	Ownership	Owner of the vehicle	Text
8	Automativestatus	Mechanical status of the vehicle	Text
9	Vehicleserviceyear	Vehicle service year	Text
10	LicenseGrade	License grade of the diver	Text
11	TypeofAccident	Type of accident	Text

Table 4.3 Table that shows selected attributes for experiment

After the data preprocessing was completed the final dataset used for experiment had 4756 records described by 11 attributes. The dataset has four class label values namely, Fatal_injury, Severe_injury, Slight_injury and Property_damage. In the dataset there are 531(11.2%) Fatal_injury records, 861(18.1%) Slight_injury records, 658(13.8%) Severe_injury records and 2706(56.9%) Property_damage records. This cleaned, reduced and manageable dataset was used for experiment to build predictive models, which will be discussed in the next chapter.

CHAPTER FIVE

EXPERIMENTATION

5.1 PREDICTIVE MODELING

Predictive modeling is a process used in a predictive analytics to create a model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. A predictive model is made up of a number of predictors, variable factors that are likely to influence future behavior or results. In marketing, for example a customer's gender, age and purchase history might predict the likelihood of a future sale [39].

To build a predictive model, data is collected, and preprocessed, model is built or formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The first challenge in building predictive models is gathering together enough preclassified data. In preclassified data the outcomes are already know and because these known examples will be used to teach the model about the data.

The basic steps in building a predictive model were described as follows [23]

1. The model is trained using preclassified data in a subset of the model set called the training set. In this step, the data mining algorithm find pattern of predictive value.
2. The model is refined using another subset called the test set in order to prevent the model from memorizing the training set, thereby ensuring that the model is more general and will work better on unseen data.

3. Estimate the performance of the model, or compare the performance of several models, evaluation.
4. The model is applied to the score set. The score set is not pre classified and is not part of the model set. We don't know the outcomes for this data.

The basic process for building predictive models is the same, regardless of the data mining technique being used. Success depends more on the process than on the techniques. And this process depends critically on the data being used to generate the model.

5.2 OVERVIEW OF WEKA SOFTWARE

As described in the methodology section weka data mining tool of version 3-5-8 was used for experiment in this research. Thus, an overview of weka software is given below.

Weka stands for Waikato environment for knowledge analysis. It is a collection of machine learning algorithms written in Java. Weka contains abundant algorithms for data preprocessing, classification, clustering, association rule, and visualization i.e. it contains data analysis and predictive modeling, together with graphical user interface for easy access to this functionality [5].

The explore interface has several panels that give access to the main components of the workbench. The preprocess panel has facilitates for importing data from a database and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data and make it possible to delete instance and attributes according to the specific criteria.

The classify panel enables the user to apply classification and regression to the resulting dataset, to estimate the accuracy to the resulting predictive model; and to visualize it, like decision tree. The associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data. The cluster panel gives access to the clustering techniques in weka. The next panel is select attributes provide algorithms for identifying the most predictive attributes in a dataset. The last panel is visualize, which shows a scatter plot matrix [5], [37].

Weka's main user interface is the explorer, but essentially the same functionality can be accessed through the component-knowledge flow menu and from command line. There is also experimenter, which allows the systematic comparisons of the predictive performance of weka's machine learning algorithms on a collection of datasets.

Weka supports input documents, Arff and CSV format. The Arff format is standard format for weka input and it will be mainly used in the experiment [5]. Thus in this research the cleaned and prepared excel format data were converted to CSV and supplied to weka and then changed to Arff format so that the data can be easily manipulated.

5.3 MODEL BUILDING AND RULE GENERATING

In order to accomplish this research work, the researcher used two data mining techniques. These are decision trees (using J48 algorithm) and Rule induction (using PART algorithm). Different models will build using these algorithms by changing the composition of the variables utilized and parameters so as to discover the most important model and generating the most interesting

rules. Weka data mining tool has facility to generate rule sets using decision trees and rule induction techniques.

The experiment was conducted using 11 attributes that were selected during the data preparation phase. These are DriverSex, DriverAge, Educationalbackground, Relationshipwith, Experience, Typesofvehicle, Ownership, Vehicleserviceyear, Automotivestatus, LicenseGrade and TypesofAccident. The TypeofAccident is the dependant variable and the rest are independent variables or predictors. The experiments that were done using the selected techniques and algorithms are explained below.

5.3.1 MODEL BUILDING USING J48 ALGORITHM

Weka has implementation of numerous classification and prediction algorithms to develop decision tree. J48 algorithm of decision tree technique is one of these algorithms which support both numeric and nominal predicators and nominal class attribute values [5].

J48 algorithm is an implementation of the C4.5 decision tree learner. The algorithm for induction of decision trees uses the greedy search technique to induce decision trees for classification. There are many parameters which can be adjusted in order to obtain better models with respect to the accuracy (or other parameters which can be used as measure for the quality of the model). These parameters allow greater control of the user in the process of learning the models [14].

The basic ideas behind all of the decision trees algorithms have general approach to develop decision tree models and the steps how decision tree algorithms work were explained under section 2.5.1.

An important feature of J48 is its facility of generating outputs both in tree form and rule sets. Rule sets are generally easier to understand since each rule describes a specific context associated with a class. It also shows the hierarchy of the determinant factors or attributes.

To build a model, the first task performed was importing the cleaned and prepared dataset of arff format into weka software. All the selected attributes and other dataset statistics are shown in figure 5.1.

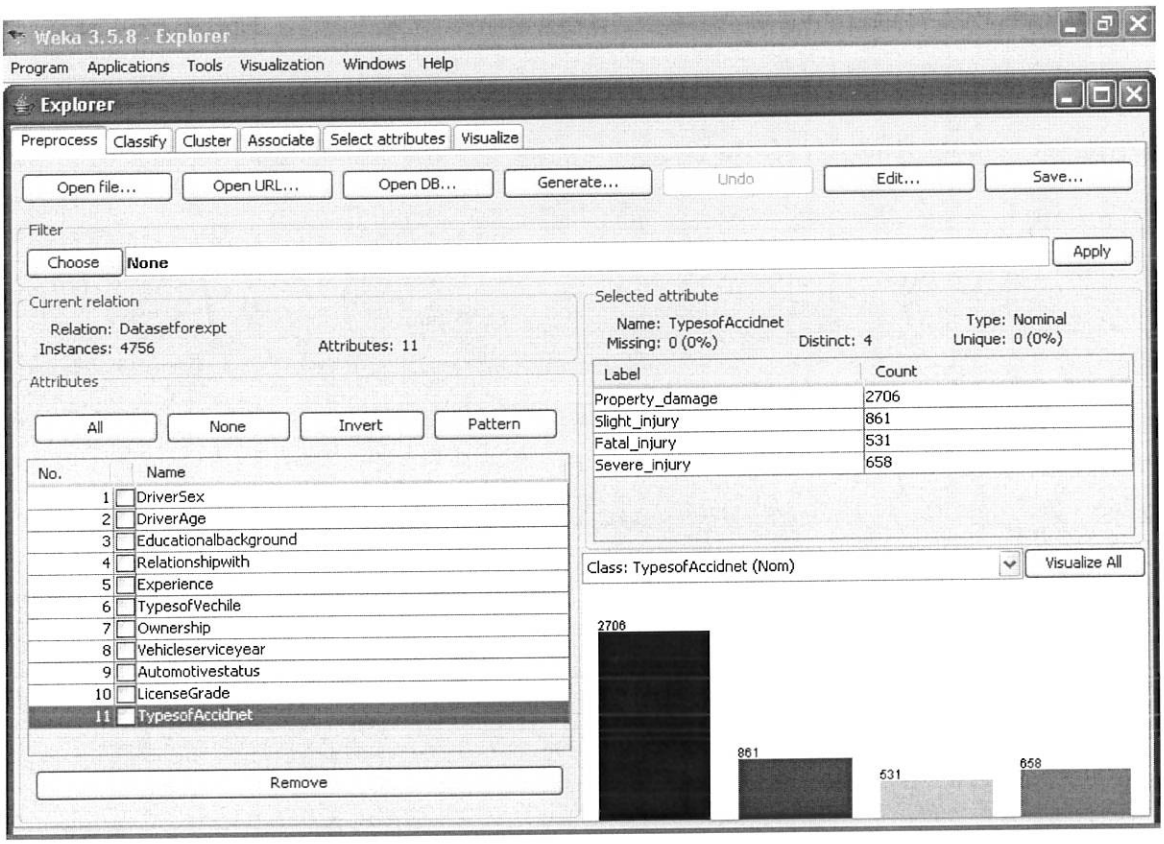

Fig 5.1 *Screenshot that shows attributes prepared for experiment*

In the first experiment all the 11 attributes were used to build model which are seen in the above screenshot.

To start building the model J48 algorithm was selected, by activating the classify panel, from the classifiers package. This algorithm has different default parameters. I used these default parameters in the first experiment. Throughout this research experiments I used the k-fold (k=10) cross validation test options because the dataset has unbalanced number of dependent class values. By doing so the partition and the experiment could be more reliable. In this test option the accuracy estimate is the overall number of correct classifications from the k iteration divided by the total number of samples, which is k. After deciding the values of the parameters the algorithm was run to start building the model.

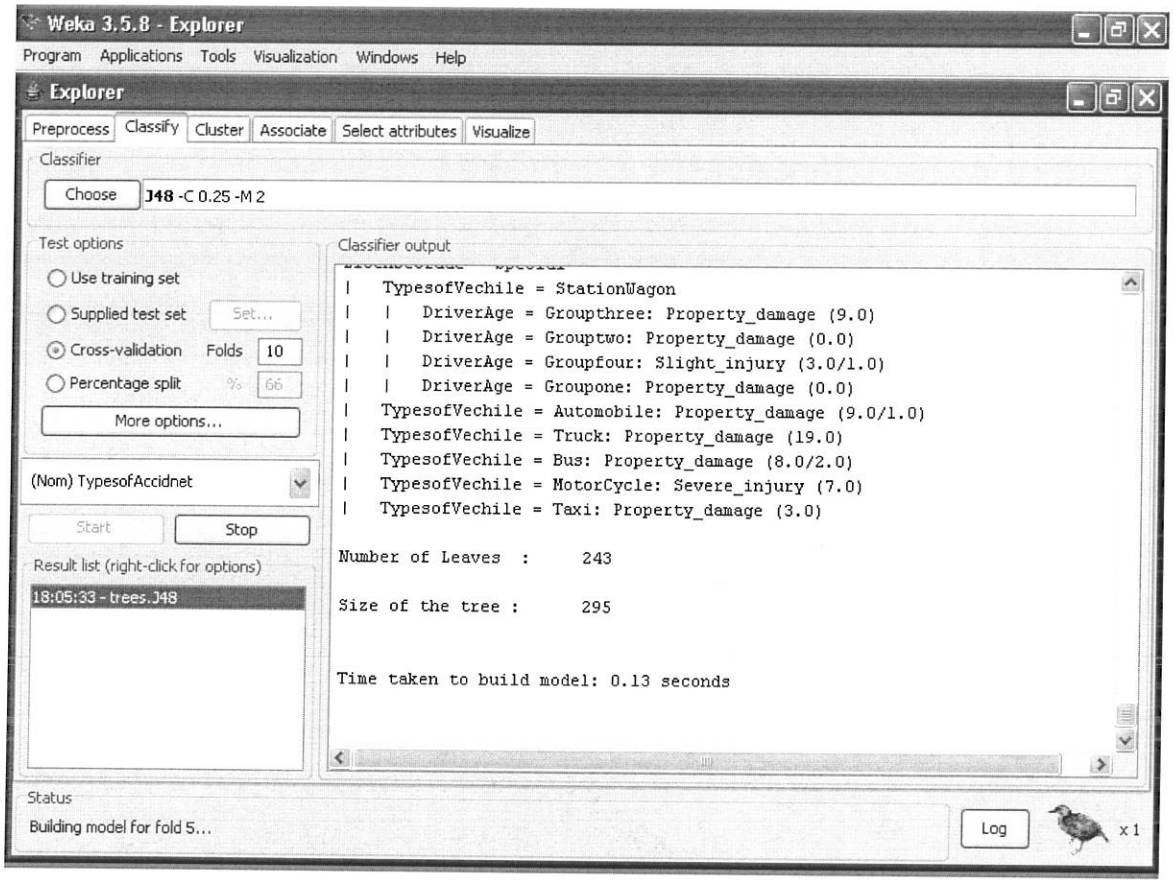

Fig 5.2 During building a model using J48 algorithm

Figure 5.3 shows the statistical summary of the first experiment that was done using all the selected attributes.

```
=== Summary ===

Correctly Classified Instances        3798               79.857  %
Incorrectly Classified Instances       958               20.143  %
Total Number of Instances             4756


=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2593   68   36    9 |    a = Property_damage
  301  458   23   79 |    b = Slight_injury
  135   21  328   47 |    c = Fatal_injury
  147   45   47  419 |    d = Severe_injury
```

***Fig 5.3** Statistical summary of the first experiment of J48 algorithm*

As indicated in the summary above, out of 4756 records of the dataset 3798 records were classified correctly and the model has an accuracy of 79.85%. The confusion matrix also shows that 328 out of 531 Fatal_injury, 2593 out of 2706 Property_damage, 419 out of 658 Severe_injury and 458 out of 861 Slight_injury records were classified correctly.

In the first experiment DriverSex and Automotivestatus variables were included only in a few line of the J48 pruned tree. That means these attributes were considered as insignificant to discriminate records. Thus in the next experiment the number of variables used was reduced to 9 by pruning DriverSex and Automotivestatus attributes to get interesting rules.

The second experiment was done using 9 attributes including the dependant variables by excluding DriverSex and Automotivestatus. This experiment was also carried out using the default parameter values. The output of this experiment shows that 3778 records were classified correctly from 4756 records, which is 79.4% accuracy. This approximates that the importance of DriverSex and Automotivestatus is insignificance in model building.

The statistical summary of the second experiment is given below.

```
=== Summary ===

Correctly Classified Instances        3778               79.4365 %
Incorrectly Classified Instances       978               20.5635 %
Total Number of Instances             4756
```

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2593   70   34    9 |    a = Property_damage
  303  457   22   79 |    b = Slight_injury
  141   20  309   61 |    c = Fatal_injury
  148   45   46  419 |    d = Severe_injury
```

Fig 5.4 *Statistical summary of the second experiment of J48 algorithm*

Part of the decision tree constructed by J48 algorithm in the second experiment is shown in Figure 5.5.

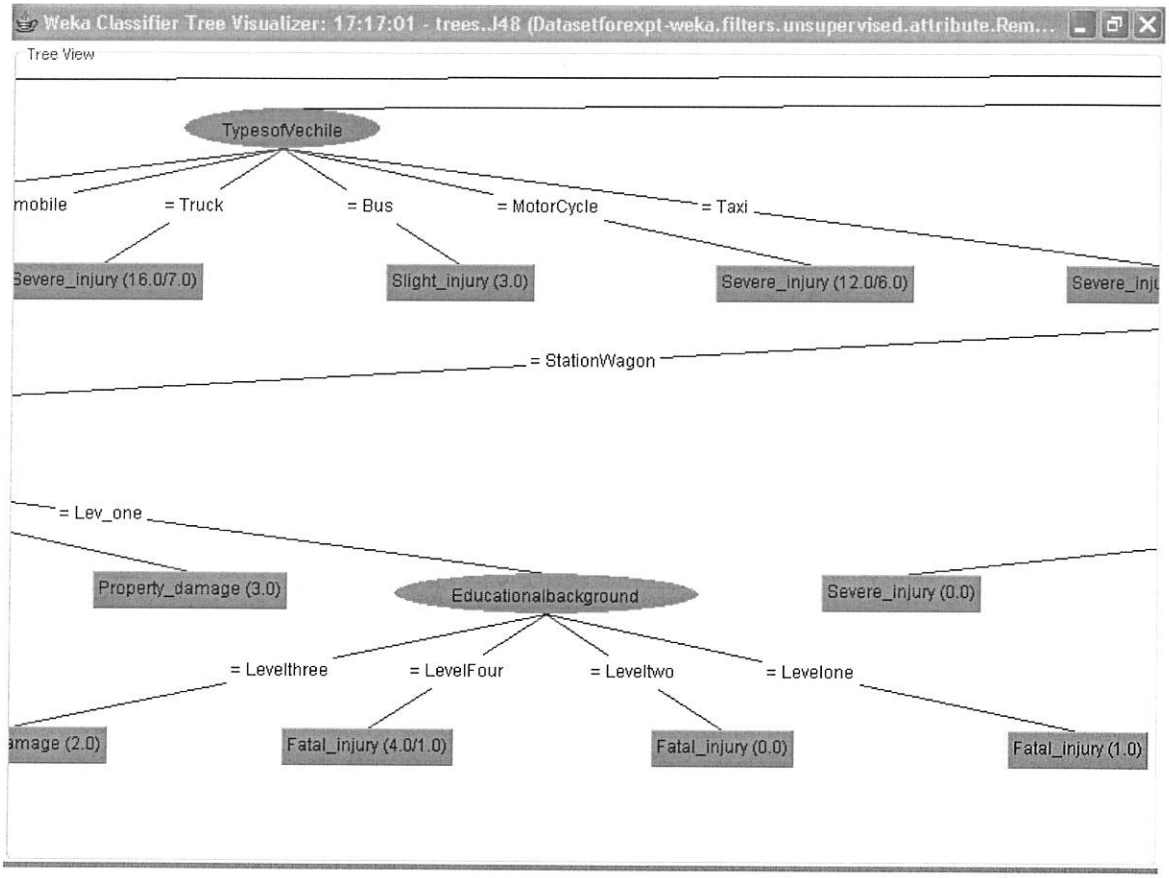

Fig 5.5 Part of the decision tree constructed

In the process of building model and finding the best model measures like adjusting the values of the parameters were also taken. That is experiment was also carried out by varying the parameters

of the J48 algorithm, but changing the default parameters was not resulting in a significant change to the performance of the model. Therefore the researcher decided to proceeds the experiment by using the default parameters as the default parameters values work reasonably well in most case and which are strongly recommended by weka software. Besides this activity, changing the composition of inputs variables was done. So, an attempt has been made to conduct the experiment by using different composition of variables by excluding and including some of the attributes to see if the accuracy of the learning scheme could be improved. Thus, the third experiment was done by removing Ownership and Relationshipwith attributes. Attributes used in the third experiment are DriverAge, Educationalbackground, Experience, Typesofvehicle, LicenseGrade, Vehicleserviceyear and the dependant variable TypesofAccident. The model obtained using these 7 attributes measures the accuracy of 76.8%.These shows that variables Ownership and Relationshipwith have some significance in the classification process.

```
=== Summary ===

Correctly Classified Instances        3654               76.8293 %
Incorrectly Classified Instances      1102               23.1707 %
Total Number of Instances             4756

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2532   89   53   32 |    a = Property_damage
  329  428   26   78 |    b = Slight_injury
  157   23  291   60 |    c = Fatal_injury
  172   43   40  403 |    d = Severe_injury
```

Fig 5.6 *Statistical summary of the third experiment of J48 algorithm*

By removing Experience, Typesofvehicle, Licensegrade and Vehicleserviceyear attributes one at a time the accuracy of the learning scheme showed 75.9%, 71.7%, 71.5%, and 63.9% respectively. This shows that the exclusion of these variables resulted in less performing model, which indicates the importance of the excluded variables in building an efficient model.

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

In decision tree each rule is taken by reading the J48 pruned tree following the path from the root node to each leaves that contains the dependent class values. From the above experiments which were done using J48 algorithm of decision tree technique variables Licensegrade, Vehicleserviceyear, Typesofvehicles and Experience are the most determinant factors to predict the type of accident. In addition to these variables DriverAge, Relationshipwith and ownership are also important variables to be taken into consideration to prevent and control road traffic accident.

5.3.2 MODEL BUILDING USING PART ALGORITHM

The second data mining technique used in this research work was Rule induction. As discussed in the literature review rule indication is one of the next generation data mining techniques that apply an iterative process to generate rules. The basic approach is the same for all Rule Induction algorithms to generate rules and was discussed under section 2.5.2.One of the rule induction algorithm used in this study is PART algorithm. The algorithm produce set of rules called "decision lists". In weka software PART algorithm is categorized under Weka. Classifiers. rules. PART package.

The first experiment was done using 9 attributes, since DriverSex and Automotivestatus are insignificant variables. Running the PART algorithm on the supplied dataset generates rules in

plain text form, which is simple to understand. The learning scheme of the built model in this experiment has an accuracy of 78.02%. Figure 5.7 shows the summary of this experiment.

```
=== Summary ===

Correctly Classified Instances        3711               78.0278 %
Incorrectly Classified Instances      1045               21.9722 %
Total Number of Instances             4756

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2534  103   49   20 |    a = Property_damage
  301  450   25   85 |    b = Slight_injury
  133   25  305   68 |    c = Fatal_injury
  135   44   57  422 |    d = Severe_injury
```

Fig 5.7 *Statistical summary of the first experiment using PART algorithm*

Following the previous experiment, an attempt was made to test by varying the composition of the input variables. Thus, an experiment was conducted using the above 7 attributes used in J48 algorithm. The accuracy of the learning scheme of the built model is 75.35%.That is 3584 instances were classified correctly out of 4756 records.

```
=== Summary ===

Correctly Classified Instances        3584               75.3574 %
Incorrectly Classified Instances      1172               24.6426 %
Total Number of Instances             4756

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 2481  109   63   53 |    a = Property_damage
  335  413   27   86 |    b = Slight_injury
  150   29  287   65 |    c = Fatal_injury
  167   43   45  403 |    d = Severe_injury
```

Fig 5.8 *Summary of the second experiment using PART algorithm*

Rules or decision lists which are generated using PART algorithm are more clear and understandable. Portion of the rules generated in the first experiment using PART algorithms are presented in appendix II.

Using PART algorithm, experiment also showed that Vehicleserviceyear, LicenseGrade Typesofvehicle and Experience are the most important variables to classify records to their predefined class.

5.4 MODELS EVALUATION AND DISCUSSION OF THE RESEARCH

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

In this experiment evaluation of models was taken place based on performance/accuracy of models and confusion matrix, discussion with the domain expert and based on the soundness of the rules generated. Thus the models built using 9 attributes in both algorithms were taken by considering these criteria.

Even though J48 algorithm is slightly better than PART algorithm, some interesting rules from PART decision list and data visualization method were also considered to discuss the research result. Some of the rules extracted from J48 pruned tree and PART decision list, outputs of the experiments done using 9 attributes that have accuracy of 79.4% and 78.02% respectively, are given below.

Rules from J48 algorithm

1. LicenseGrade=second level
   Vehicleserviceyear=Ser_levelfive
   Ownership=private
   Experience=Lev_three: Fatal_injury (82.0/6.0)

2. LicenseGrade=second level
   Vehicleserviceyear=Ser_levelfour
   Typesofvehicle=Truck: Fatal_injury (63.0/5.0)

3. LicenseGrade=Third level
   Vehicleserviceyear=Ser_levelfive
   Typesofvehicle=Taxi
   Experience=Lev_Two: Severe_injury (109.0/29.0)

4. LicenseGrade=Fourth level
   Experience=Lev_Five
   Educationalbackground=Level four: Slight_injury (6.0/2.0)

5. LicenseGrade=Fourth level
   Experience=Lev_three
   Typesofvehicle=Truck
   Ownership=private: Fatal_injury (47.0/21.0)
   Ownership=Government: Slight_injury (4.0)

6. LicenseGrade=NoLicense
   Typesofvehicle=Taxi: Severe_injury (12.0/6.0)
   Typesofvehicle=Truck: Severe_injury (16.0/7.0)
   Typesofvehicle=Motorcycle: Severe_injury (16.0/7.0)

7. LicenseGrade = Second level
   Vehicleserviceyear = Ser_levelone
   TypesofVehicle = StationWagon
   Experience = Lev_Four: Slight_injury (12.0/2.0)

8. LicenseGrade = Third level

Vehicleserviceyear = Ser_levelthree

TypesofVehicle = Truck

Ownership = Private

Experience = Lev_Two: Severe_injury (10.0/2.0)

Experience = Lev_Four: Slight_injury (12.0/2.0)

Rules from PART algorithm

9. LicenseGrade = Third level AND
Experience = Lev_One AND
Vehicleserviceyear = Ser_levelfive AND
TypesofVehicle = Taxi AND
Educationalbackground = Levelthree: Severe_injury (14.0/8.0)

10. LicenseGrade = Third level AND
Experience = Lev_Four AND
Educationalbackground = Leveltwo
DriverAge=Grouptwo: Slight_injury (7.0/4.0)

11. LicenseGrade = Third level AND
Vehicleserviceyear = Ser_levelfour AND
Ownership = Private AND
TypesofVehicle = Taxi AND
Relationshipwith = Employed AND
Experience = Lev_Four AND
Educationalbackground = Levelthree: Property_damage (87.0/19.0)

12. LicenseGrade = Third level AND
Vehicleserviceyear = Ser_levelthree AND
TypesofVehicle = Taxi AND
Experience = Lev_Five: Slight_injury (28.0/9.0)

13. LicenseGrade = Second level AND
    TypesofVechile = StationWagon AND
    Vehicleserviceyear = Ser_levelfour AND
    DriverAge = Groupthree: Fatal_injury (11.0/4.0)

14. LicenseGrade = Third level AND
    Vehicleserviceyear = Ser_levelfour AND
    TypesofVehicle = Taxi AND
    Relationshipwith=Employed AND
    Experience = Lev_Three: Property_damage (182.0/44.0)

15. LicenseGrade = Third level AND
    Experience = Lev_Three AND
    Ownership = Private AND
    DriverAge = Grouptwo: Fatal_injury (9.0/4.0)

16. Vehicleserviceyear = Ser_levelthree AND
    TypesofVehicle = Automobile AND
    Ownership = Public_Agency AND
    DriverAge = Grouptwo: Property_damage (9.0/2.0)

As it is observed from the above rules, the classifier has used some attributes to construct rules and provided the class predicted by the model. The numeric values which appeared in bracket next to the class label indicates the number of correctly and incorrectly classified records respectively.

For example Rule 1 is interpreted as driver who has second level driving license and Lev_three experience, and drove a private vehicle that gave Ser_levelfive was grouped as Fatal_injury

(82.0/6.0). Meaning in the dataset there are 82 records that exactly satisfy this rule and 6 records are misclassified to this rule.

As the research indicated, for instance rule 11, 14 and 15, Typesofvehicle, Ownership and Relationshipwith are important and unforeseen factors in accident analysis. In these rules private taxies and trucks which are droved by employee driver cause most of the accidents. But employee drivers of government and international agency vehicles create less accident. In this research the data visualization showed that vehicle owners are almost safe drivers.

In this research it is observed that those who haven't driving license create severe and fatal accident, as it is seen in rule 6. So rules should be highly strict on those who drive without having license. Actions should be also taken on vehicle like suspending for a certain period of time if it is droved by non licensed persons. The data visualization method also showed that drivers who have second and third driving license create most of the accident.

In this research data visualization showed that vehicles that gave services ser_levelfour and Ser-levelfive and drivers who have educational background level two and level three cause most of the accidents. Thus, vehicles should get technical investigation, follow up and vehicles which are supposed to cause accident should be suspended or blocked not to give service. It was also proved that in Insurance risk assessment study at Nyal Insurance Company [31], as vehicle service year increases the risk exposure also become higher. The Transport authority, in Article 12 of the newly implemented proclamation, Proclamation NO.600/2008, should also consider experience of the driver besides educational background because experience is the most important factor. So,

during vehicle ownership registration and license upgrading activities, experiences of drivers should be considered to a specific vehicle, especially on Truck and Taxi, which cause most of severe and fatal accident.

Using data visualization the researcher observed that most of the accident caused by age group of Grouptwo (18-30) and Groupthree (31-50). But since the interval used to encode the age value is very wide and not equal, it is difficult to get the extreme importance of this variable in accident analysis and to differentiate the type of accident created by each age group. If the office uses for instance a 5 widths interval, the significance of this attribute will be extremely high to improve the performance of the models studied in this research.

In general interesting rules were generated both by J48 and PART algorithm that indicates the possible condition in which an accident will result in different accident types. The rule also shows that attributes such as LicenseGrade, Vehicleserviceyear, Typesofvehicle, Experience, Relationshipwith, DriverAge and Ownership, are found to be important variables to classify the accident type. This will help the traffic control and investigation office and transport authority to focus their attention on these factors during revision and construction of rules and policies.

This study was concerned with drivers and vehicles factors and how they eventually determine the occurrence of road traffic accidents. An important relationship between dependent and independent variables was observed and a general function can therefore be developed as:

TOA=P ({LG, VSY, TV, E, DA, DE, O, R}), is a function that maps a combination of attributes values to a unique accident type.

Where, TOA=Type of accident/Accident severity

LG=License grade level

VSY=Vehicle service year

TV=Types of vehicle

E=Experience

DA=Driver age

DE=Driver educational background

O=Ownership

R=Relationship with vehicle

P= function

The function P evaluates or measures the magnitude of the accident severity based on the values of the independent variables selected. For instance based on the observation of the rules generated, TOA is at high risk, fatal or severe injury if LG is second level or Nolicense, VSY is Ser_levelfive or Ser-levelfour, TV is Truck or Taxi, E is Lev_one or Lev_Two etc. On the other hand TOA is at low risk if LG is Fourth level or Fifth level, VSY is Ser_levone or Ser_levtwo, E is Lev_four or Lev_five etc. This shows that Predictive models are functions that map a given case or record to a specific class.

Using this study and the above generalization I defined a general function called predictive function as follows:

A function P defined by $C=P(\{ a_1(b_1,b_2,...,b_m), a_2(c_1,c_2,...,c_n),...,a_k(x_1,x_2,...,x_r)\})$ *is called predictive function that maps a combination of attributes values to a specific predefined class type based on learning scheme from the historical data used.*

Where $a_1, a_2 ,..., a_k$ *are attributes,* $b_1,b_2,...,b_m,c_1,c_2,...,c_n,...,x_1,x_2,..,x_r$ *are values of the attributes ,k,m,n,r are natural numbers and C is the predicted class type.*

This predictive function maps a single record, which is one possible combination of attributes values, to a unique predefined class type. This is a simple definition that used to clarify the meaning of classification and prediction. Thus, predictive models work their prediction based on the definitions of this predictive function.

5.5 COMPARISON OF J48 AND PART ALGORITHMS

Weka software has a facility to compare the performance of two or more algorithms at the same time. Thus, the next experiment shows the comparison of the performance of J48 and PART algorithms using Experimenter interface of weka software.

Experimenter interface of weka software was used to compare the performance of J48 and PART algorithms to conform the above experiment results. The two algorithms run at the same time, on the same dataset, the same and equal number of attributes (9 attributes). Figure 5.9 shows the setup of the experiment that contains the dataset, the two algorithms and other parameters.

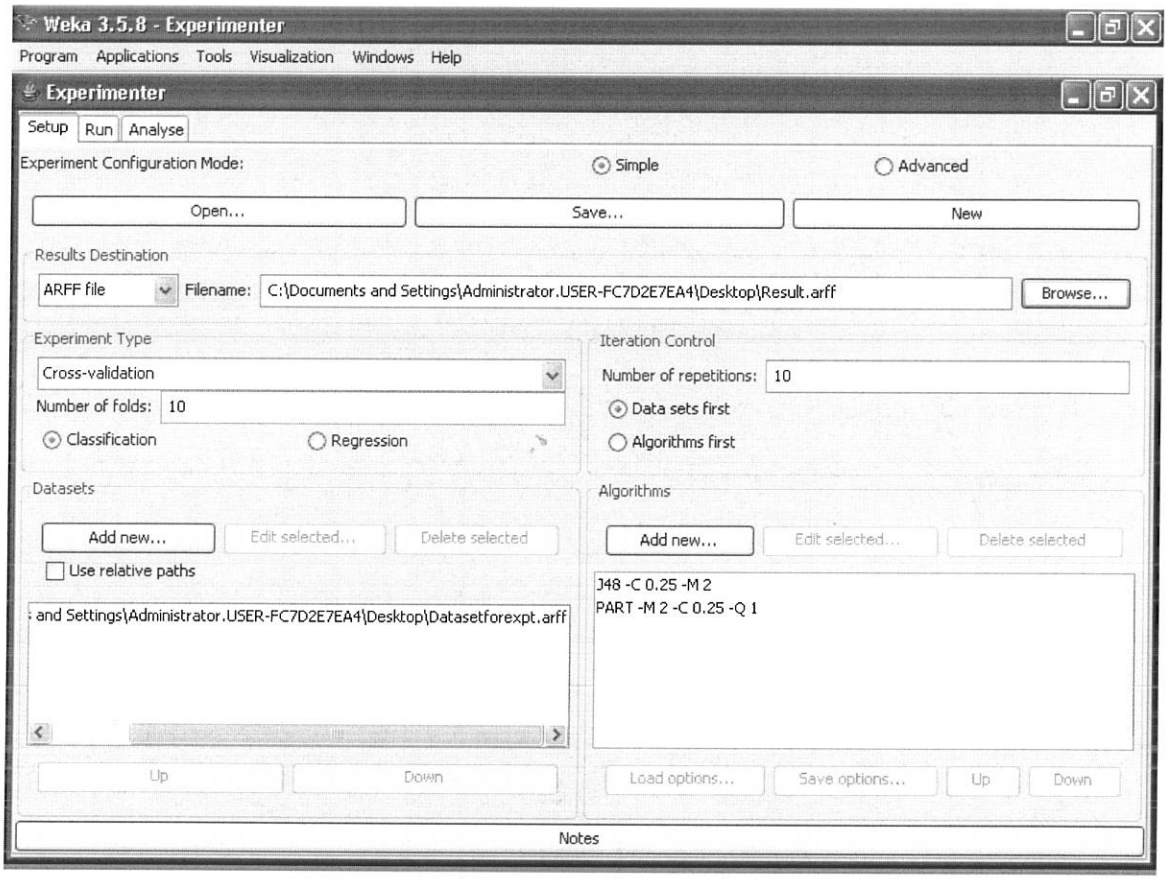

Fig 5.9 *Screenshot that shows selected algorithms, dataset and other parameters*

After finishing the necessary setup the algorithms were run to do the experiment. To analyze the experiment the percent_correct comparison measurement factor was selected from comparison field combo box as shown in figure 5.10. Then the perform test button was clicked to compare the performance of the two algorithms based on the criteria selected.

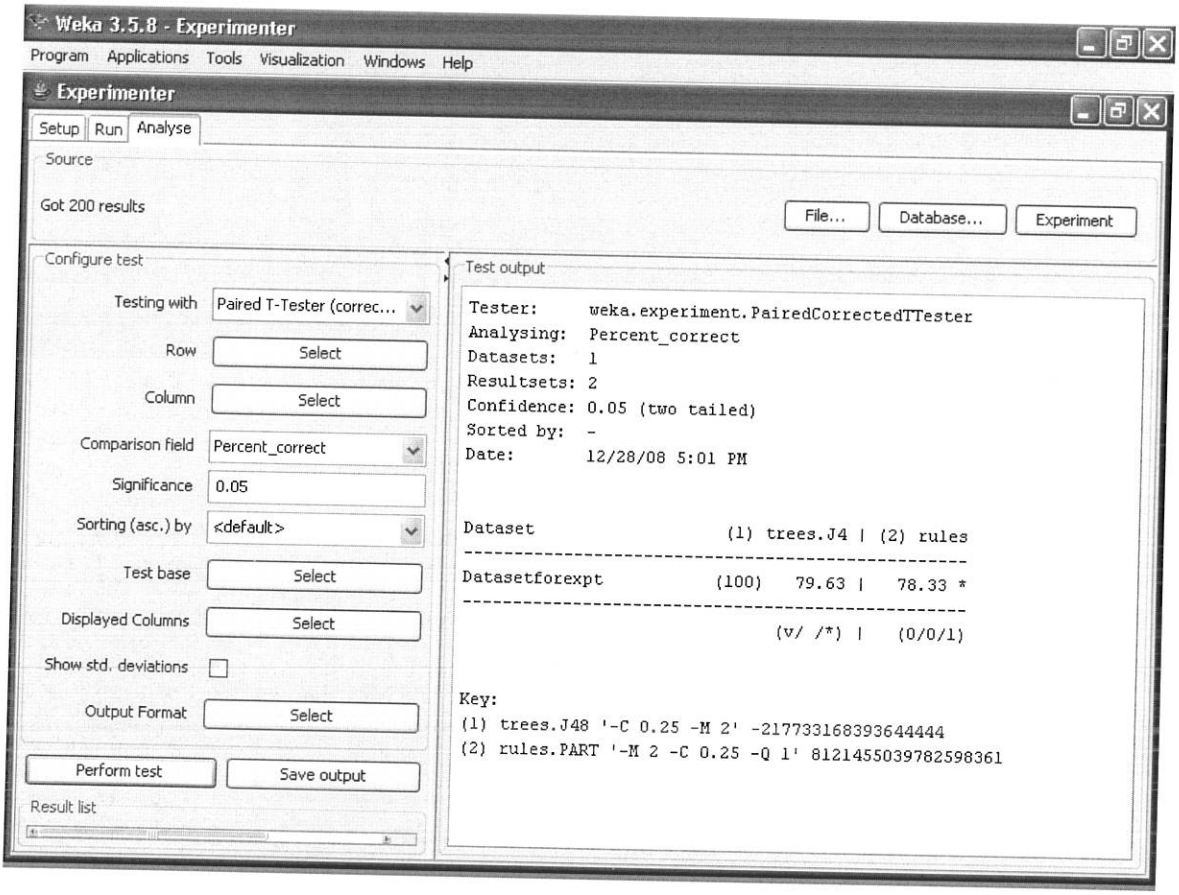

Fig 5.10 *Screenshot that shows the comparison result*

The comparison result in the above figure shows that J48 algorithm has an accuracy of 79.63% while the accuracy of PART algorithm is 78.33%. Thus performance of the model built using J48 algorithm is slightly better than PART algorithm to classify records to their predefined classes.

All the above activities and results done during experimentation are summarized by fig 5.11

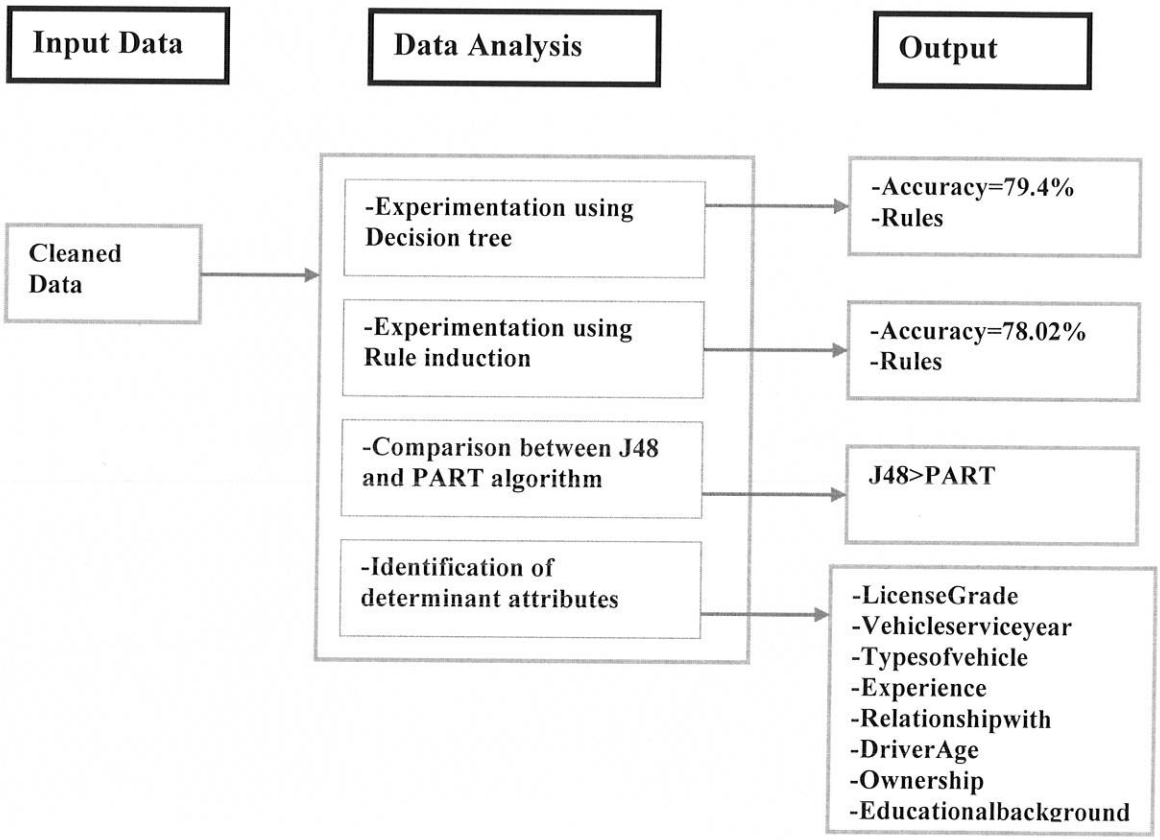

Fig 5.11 Activities done during experimentation

In this research the above overall model building process done with decision tree and rule induction techniques demonstrated how to support rood traffic accident prevention and controlling activities, at Addis Ababa city. From the results obtained it can be generalized that the road traffic accident predictive model can be developed using drivers and vehicles determinant factors from collected road traffic accident data. In addition to this, although both decision tree and rule induction showed comparable performance in predicting the risk of road traffic accident

and the accuracy of the decision tree is slightly better than rule induction, both techniques can be used in parallel to solve this crucial problem.

Fig 5.12 shows the overall activities and outputs of this research work.

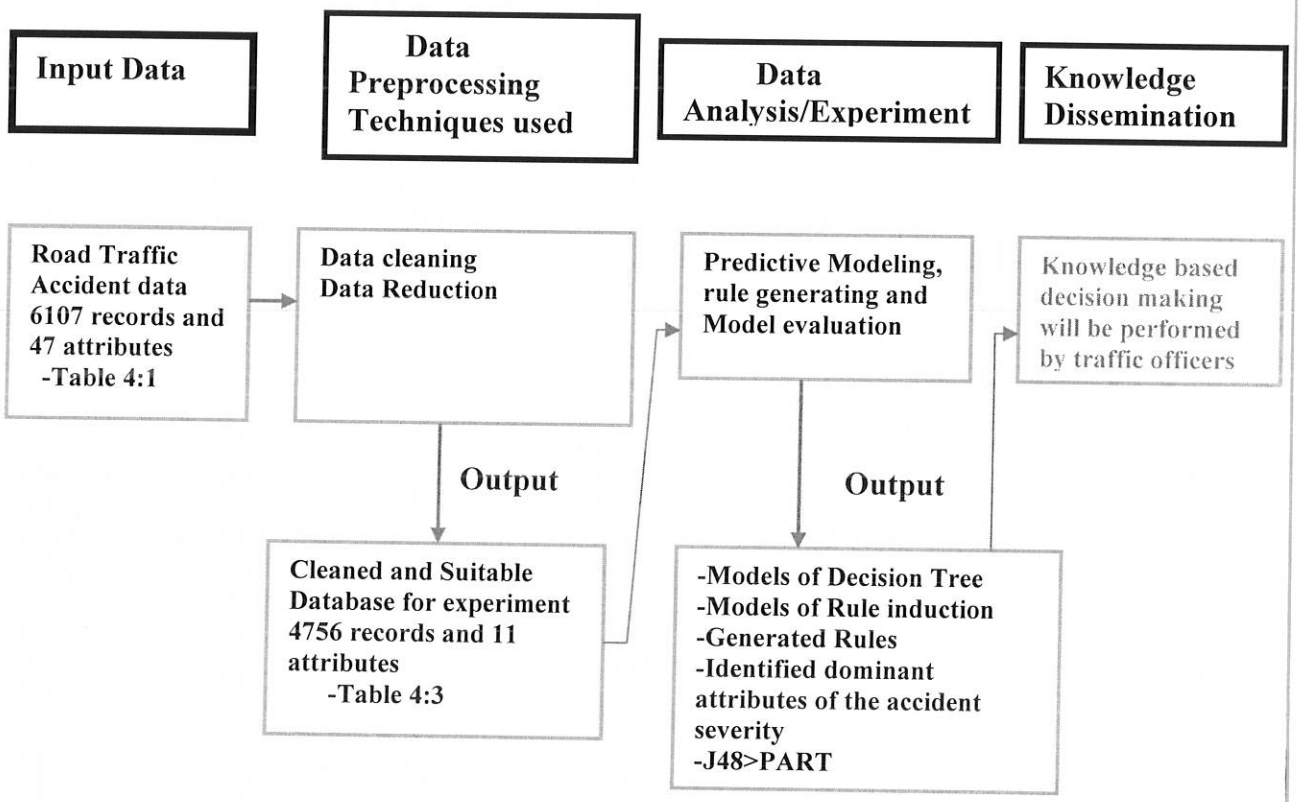

Fig 5.12 Process Diagram of the research work

All the above experiment result shows that data mining technology in road traffic accident analysis can help to understand drivers and vehicles factors that were causally connected with the different injury severity. This can in turn help decision makers to induce and revise traffic control rules. Conclusions and recommendations of the research findings are discussed in the next chapter.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 CONCLUSIONS

The amount of data being collected in databases today far exceeds our ability to analyze these data as they might contain some precious information for decision support. As data volumes grow dramatically, for instance explosive growth of road traffic accident data, data analysis based on manual and statistic methods is becoming completely impractical. Due to these challenges searching for knowledge in large database and our inability to interpret these accumulated data needs new generation of tool, data mining, for automated and intelligent database analysis, which is useful in decision making, problem solving and rule construction.

In this research an attempt was made to investigate the possible application of data mining technology in road traffic accident prevention and controlling activity at Addis Ababa city, by developing a predictive model that could help traffic control and investigation office to identify the determinant risk factors of drivers and vehicles so that they can take measures before expensive loss of life and property occurred. Attempt was also undertaken to generate rules that can be used as traffic accident controlling rules and policies.

Various experiments were made iteratively by making adjustments of the parameters and composition of variable in both techniques to come up with meaningful output. In this research major risk factors of drivers and vehicles were identified and using these variables models were built and rules also generated with decision tree (using J48 algorithm) and rule induction (using

PART algorithm) techniques. The experiment of this research also showed that the performance of J48 algorithm is slightly better than PART algorithm. In addition to this comparison, the results of the decision tree and rule induction showed an interesting pattern which might be an indication that the combination of the models could result in a better classification performance.

The determinant drivers and vehicles risk factors that cause accident were identified in this research. These are Licensegrade, Vehicleserviceyear, Typesofvehicles, Experience, Ownership, DriverAge, Relationshipwith and Educationalbackground. In this research some variables were identified which were not initially suspected to be very important in road traffic accident prevention, for instance Ownership and Relationshipwith.

As trend showed most of the time, in this department decision tree and neural networks techniques were widely used to do researches using data mining tool .In this research attempt was made to show how to apply Rule induction techniques to do research and to solve real problem of the society.

In general, encouraging results were obtained by employing both decision tree and rule induction techniques as rules generated by J48 and PART algorithms are much understandable to explain the prediction outcome easily. Thus, the result obtained in this research has proved the applicability of data mining in road traffic accident preventing and controlling activities. More specifically it is highly supportive to construct, evaluate and update traffic rules and policies especially using decision tree and rule induction techniques as rules are provided in plain text.

6.2 RECOMMENDATIONS

In this research work the researcher apply data mining tool and techniques to predict the accident type or risk based on drivers and vehicles factors alone. In the process of this research, it was learnt that more research efforts need to be conducted to enable the full exploitation of data mining technology in road traffic accident prevention and controlling activities. Thus in the course of doing this study and based on the research findings of this research work, the researcher would like to make the following recommendations that needs more consideration in the future work

1. Data handling system of Addis Ababa traffic control and investigation office is computerized database management system. In this system the Addis Ababa traffic control and investigation office encodes values of DriverAge attribute in four interval, <18 years, 18-30 years, 31-50 years and ≥ 51 years. Since the interval used is wide and unequal, a given age group of drivers grouped under fatal or severe or slight injury or property damage i.e. a given age group maps to both accident types. So, to get it's extremely significance a maximum of 5 width interval is recommended. The same is true for others variables that are encoded in interval form. An interesting accuracy and result will be investigated if all numeric values are encoded in continuous or non-interval form.
2. Since road traffic accident is a very complex problem domain to reduce and control this problem, the Addis Ababa traffic control and investigation office should construct rules and polices based on the research results, especially research results of data mining techniques of decision tree and rule induction. For instance the newly implemented drivers' qualification certificate license should be modified to include the risky attribute

values identified in this research and discussed in the research findings section to construct the most efficient traffic rules.

3. In this research work variables like Vehicleserviceyear, Licensegrade, Experience and Typesofvehicle are the most determinant factors for causes of road traffic accident. So responsible partner, the Addis Ababa Transport Authority should work tightly with Addis Ababa traffic control and investigation office by doing technical investigation of vehicles which gave service for long years and revise their licensing system.
4. The road traffic accident is a very complex problem domain. Thus, effects or contributions of different component in the accident should be seen and studied separately to observe and understand their effects.
5. In this research work data mining was used to assess and predict the accident type (severity) by using drivers, vehicles and relation between drivers and vehicles detail variables. Other researchers however can consider a number of other variables to investigate further effect of those variables in road traffic accident.
6. The models in this research work was developed using small percentage of dataset, 4756 records. So, efforts to build more meaningful and comprehensive models by using large number of dataset records are highly recommended.
7. Both decision tree and rule induction approach resulted in an encouraging output, especially simplicity of the output of both techniques to explain for the end user. The application of other data mining techniques like neural networks, Bayes techniques and others can also be applied and tested if they could be more applicable to this problem domain.

8. Finally, identifying problems can be considered as solving the problem half way. The government has to be inflexible with drivers who cause road traffic accidents, such drivers should be suspended for a certain period of time. The problem of road traffic accidents can't be left to the government alone. There is a need for attitude change among all people towards road safety and all stakeholders are expected to give attention to reduce the existing road traffic accident in general.

REFERENCES

[1]. Abinet Tale (2005). Predictive Modeling using data mining technology in support of cement quality assessment. A Master's thesis submitted to Information science department, A.A.U

[2]. Addis Ababa city Transport Authority. Facts about Addis Ababa City Transport. Accessed date: September 4, 2008 http://www.telecom.net.et/~aata/

[3]. Alan Rea (1995). Data Mining an Introduction, Version 2.0, Queens University Belfast http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html

[4]. Cabena p, A. Corma, M. Moliner, J.M. Serra, *P.* Serna . (1998).Discovering data mining from concept to implementation Prentice Hall, New Jersey:

[5]. Christie Ezeife (2006). The Exploration and Application of WEKA Data Miner, University of Windsor .Accessed date: September 12, 2008 http://cs.uwindsor.ca/~zhu19/60539-project-weka.doc.

[6]. Hand D.J., Mannila H., and Smyth P. (2001).Principles of data mining. The MIT press, London

[7]. Eduardo A Vasconcellos (2005), Traffic accident risks in developing countries: Superseding biased approaches. Campo Grande

[8]. FDRE Negarit Gazeta (2008). Drivers Qualification Certification License Proclamation No. 600/2008, Addis Ababa

[9]. Famili, and Turneyl (1997). Data Preprocessing and Intelligent Data Analysis, Institute for Information Technology, National Research Council Canada. http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-40166.pdf

[10]. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth (1996). From Data Mining to Knowledge Discovery

Accessed date: September 12, 2008

http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf

[11].Haileleoul Gudeta (2004). A Gas Turbine engine performance classifier: Neural Network Approach. A Master's thesis submitted to Information science department, A.A.U

[12]. Helen M. Moshkovicha, Alexander I. Mechitova and David L. Olson (2002). Rule induction in data mining: effect of ordinal scales, 30 January, Elsevier Science Ltd.

[13]. Hussein Arshan (2005). Time-Critical Decision Making for Business Administration.

Accessed date: October 15, 2008

http://home.ubalt.edu/ntsbarsh/stat-data/Forecast.htm

[14]. I. Witten and E. Frank (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2nd Edition

[15]. J. Han and M. Kamber (2001): Data Mining Concepts and Techniques. New York: Morgan Kaufman.

[16]. Khalid Sheikh (2003). Using Decision Trees to predict customer behavior

Accessed date: October 12, 2008

http://www.expresscomputeronline.com/20040412/technology01.shtml

[17]. Kim, K., Nitz, L., Richardson, J., & Li, L (1995). Personal and Behavioral Predictors of Automobile Crash and Injury Severity. Accident Analysis and Prevention, Vol 27, No.4.

[18]. Knowledgerush.com (2003). Road transport

Accessed date: November 10, 2008

http://www.knowledgerush.com/kr/encyclopedia/Road_transport/

[19].Mekitew Mola (2000). Traffic Information System: The case of Addis Ababa Traffic police department. A Master's thesis submitted to school of information studies for Africa., A.A.U.

[20].Mesfin Fikre (2005).Predictive Data Mining technique in Insurance: The case of Ethiopia Insurance Corporation. A Master's thesis submitted to Information science department, A.A.U

[21]. Leul Woldu (2003).The application of Data Mining in crime prevention: The case of Oromia police commission. A Master's thesis submitted to Information science department, A.A.U

[22] MyDataMining's Weblog (2008). Rule learner (or Rule Induction) , April 14.

Accessed date: December 2, 2008

http://mydatamining.wordpress.com/2008/04/14/rule-learner-or-rule-induction/

[23]. Michael J.A. Berry and Gordon Linoff (2000) .Mastering Data Mining: The Art and Science of customer Relationship Management. New York: John Wiley & Sons, Inc.

[24]. Mussone, L., Ferrari, A., & Oneta, M. (1999). An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention, Vol 31

[25]. Nikolaos F. Matsatsinis, E. Ioannidou, E. Grigoroudis (1998). Customer satisfaction using data mining techniques

Accessed date: October 12, 2008

http://www.erudit.de/erudit/events/esit99/12753_p.pdf

[26]. Nist Sematech e-book (2006). Time series data analysis.

Accessed date: October 12, 2008

http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.html

[27]. Ossiander, E. M., & Cummings, P. (2002). Freeway speed limits and Traffic Fatalities in Washington State. Accident Analysis and Prevention, Vol 34

[28]. Safecarguide.com (2004). Injury and fatality statistics

Accessed date: September 4, 2008

http://www.safecarguide.com/exp/statistics/statistics.htm

[29]. Shegaw Anagaw (2002).Application of data mining technology to predict child morality Patterns: The case of Butajira Rural Health Project (BRHP).

[30]. StatSoft, Inc. Data Mining Techniques

Accessed date: September 4, 2008

http://www.statsoft.com/textbook/stdatmin.html

[31]. Tesfaye Hintsay (2002). Predictive Modeling using Data Mining Techniques in Support of Insurance Risk Assessement: The case of Nyal Insurance Corporation. A Master's thesis submitted to Information science department, A.A.U

[32]. Temesgen Aklilu (2007). Ethiopian road transport development: problems, challenges and constraints. Paper presented at the 5th international conference on the Ethiopian Economy

[33]. Thearling, Alex Berson and Stephen Smith, (2003). An Overview of Data Mining Techniques.

Accessed date: September 4, 2008

http://www.Thearling.com/text/dmtechniques/dmtechniques.htm

[34]. Tibebe Beshah (2005). Application of Data mining technology to support Road Traffic Accident Severity Analysis at Addis Ababa Traffic office. A Master's thesis submitted to Information science department, A.A.U

[35]. Two Crows Corporation (2005). Introduction to Data Mining and Knowledge Discovery, Third Edition

[36].VOA.com (2006). Putting the breaks on Ethiopian Traffic accident, March 6

[37].Wekadocs.com. Weka Docs. Analyzing the output

Accessed date: November 15, 2008

http://wekadocs.com/node/13

[38]. WHO-World health organization (2004). Reports on the celebration of world health day on the theme "Road safety Is no accident"

Accessed date: September 12, 2008

http://www.Who.int/countries/eth/news/speeches/070404/en/index.html

[39]. Whatis.com (2005).Data Management Definitions-Predictive Modeling, Last updated: Nov 10.

Accessed date: November 20, 2008

http://searchdatamanagement.techtarget.com/sDefinition/0,,sid91_gci809473,00.html

[40].Wikipedia.org. Car accident

Accessed date: September 12, 2008

http://en.wikipedia.org/wiki/Car_accident

[41]. Yang, W.T., Chen, H. C., & Brown, D. B. (1999). Detecting Safer Driving Patterns by a Neural Network Approach, Vol 9

[42]. KDnuggets News. Classification vs. Prediction

http://www.kdnuggets.com/faq/classification-vs-prediction.html

APPENDICES

APPENDIX I

Some part of Rules generated by J48 algorithm.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Datasetforexpt-weka.filters.unsupervised.attribute.Remove-R1,9
Instances:    4756
Attributes:   9
              DriverAge
              Educationalbackground
              Relationshipwith
              Experience
              TypesofVechile
              Ownership
              Vehicleserviceyear
              LicenseGrade
              TypesofAccidnet
Test mode:    10-fold cross-validation

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

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

LicenseGrade = Second level
|   Vehicleserviceyear = Ser_leveltwo
|   |   TypesofVechile = Automobile: Property_damage (353.0)
|   |   TypesofVechile = StationWagon: Property_damage (31.0)
|   |   TypesofVechile = Truck
|   |   |   Educationalbackground = Levelthree: Slight_injury (10.0)
|   |   |   Educationalbackground = LevelFour: Slight_injury (10.0/2.0)
|   |   |   Educationalbackground = Levelone: Property_damage (4.0)
|   |   |   Educationalbackground = Leveltwo: Slight_injury (0.0)
|   |   TypesofVechile = Bus: Property_damage (0.0)
|   |   TypesofVechile = Taxi: Property_damage (66.0/11.0)
|   |   TypesofVechile = MotorCycle: Property_damage (0.0)
|   Vehicleserviceyear = Ser_levelfive
|   |   Ownership = Core_Deplomatic: Property_damage (2.0)
|   |   Ownership = Government: Property_damage (17.0/2.0)
|   |   Ownership = InternationalAgency: Property_damage (1.0)
|   |   Ownership = Military: Severe_injury (1.0)
|   |   Ownership = Police: Fatal_injury (0.0)
|   |   Ownership = Private
|   |   |   Experience = Lev_Four
|   |   |   |   Educationalbackground = Levelthree: Fatal_injury (46.0/11.0)
```

```
|   |   TypesofVechile = StationWagon
|   |   |   Experience = Lev_Four: Slight_injury (12.0/2.0)
|   |   |   Experience = Lev_Three
|   |   |   |   Relationshipwith = Employed: Slight_injury (7.0/1.0)
|   |   |   |   Relationshipwith = Veh_owner: Property_damage (4.0)
|   |   |   |   Relationshipwith = Others
|   |   |   |   |   DriverAge = Groupfour: Property_damage (0.0)
|   |   |   |   |   DriverAge = Groupthree: Property_damage (2.0)
|   |   |   |   |   DriverAge = Grouptwo: Slight_injury (2.0)
|   |   |   |   |   DriverAge = Groupone: Property_damage (0.0)
|   |   |   Experience = Lev_Five: Slight_injury (20.0)
|   |   |   Experience = Lev_Two: Property_damage (17.0/4.0)
|   |   |   Experience = NoLicense: Slight_injury (0.0)
|   |   |   Experience = Lev_one: Slight_injury (5.0/1.0)
|   |   TypesofVechile = Truck: Property_damage (4.0)
|   |   TypesofVechile = Bus: Property_damage (0.0)
|   |   TypesofVechile = Taxi: Property_damage (0.0)
|   |   TypesofVechile = MotorCycle: Property_damage (0.0)
|   Vehicleserviceyear = Ser_levelthree
|   |   TypesofVechile = Automobile: Property_damage (174.0/54.0)
|   |   TypesofVechile = StationWagon
|   |   |   Ownership = Core_Deplomatic: Property_damage (2.0)
|   |   |   Ownership = Government: Fatal_injury (7.0)
|   |   |   Ownership = InternationalAgency: Property_damage (4.0/1.0)
|   |   |   Ownership = Military: Fatal_injury (2.0)
|   |   |   Ownership = Police: Property_damage (0.0)
|   |   |   Ownership = Private: Property_damage (30.0/1.0)
|   |   |   Ownership = Puplic_Agency: Fatal_injury (2.0)
|   |   TypesofVechile = Truck: Property_damage (43.0/2.0)
|   |   TypesofVechile = Bus: Property_damage (0.0)
|   |   TypesofVechile = Taxi: Property_damage (0.0)
|   |   TypesofVechile = MotorCycle: Property_damage (0.0)
LicenseGrade = Third level
|   Vehicleserviceyear = Ser_leveltwo: Property_damage (141.0/13.0)
|   Vehicleserviceyear = Ser_levelfive
|   |   TypesofVechile = Automobile: Severe_injury (131.0/6.0)
|   |   TypesofVechile = StationWagon
|   |   |   Experience = Lev_Four
|   |   |   |   Ownership = Core_Deplomatic: Severe_injury (1.0)
|   |   |   |   Ownership = Government: Property_damage (2.0/1.0)
|   |   |   |   Ownership = InternationalAgency: Property_damage (5.0)
|   |   |   |   Ownership = Military: Severe_injury (0.0)
|   |   |   |   Ownership = Police: Severe_injury (0.0)
|   |   |   |   Ownership = Private: Severe_injury (13.0/1.0)
|   |   |   |   Ownership = Puplic_Agency: Severe_injury (0.0)
|   |   |   Experience = Lev_Three: Property_damage (27.0/1.0)
|   |   |   Experience = Lev_Five: Property_damage (15.0/2.0)
|   |   |   Experience = Lev_Two: Property_damage (4.0)
|   |   |   Experience = NoLicense: Property_damage (1.0)
|   |   |   Experience = Lev_one: Property_damage (2.0)
|   |   TypesofVechile = Truck
|   |   |   Experience = Lev_Four: Fatal_injury (27.0/12.0)
|   |   |   Experience = Lev_Three: Property_damage (49.0/12.0)
|   |   |   Experience = Lev_Five: Fatal_injury (3.0/1.0)
|   |   |   Experience = Lev_Two: Property_damage (20.0/7.0)
|   |   |   Experience = NoLicense: Property_damage (0.0)
|   |   |   Experience = Lev_one: Property_damage (4.0)
```

APPENDIX II

Some part of Rules generated by PART algorithm.

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


Experience = Lev_Three AND
Vehicleserviceyear = Ser_levelfive AND
Ownership = Private AND
LicenseGrade = Third level: Severe_injury (189.0/46.0)

LicenseGrade = Third level AND
TypesofVechile = Taxi AND
Experience = Lev_Four: Property_damage (93.0/19.0)

LicenseGrade = Third level AND
Experience = Lev_Three AND
Ownership = Private AND
DriverAge = Grouptwo: Fatal_injury (9.0/4.0)

LicenseGrade = Third level AND
Experience = Lev_one AND
Vehicleserviceyear = Ser_levelfive AND
TypesofVechile = Taxi AND
Educationalbackground = Levelthree: Severe_injury (14.0/8.0)


Experience = NoLicense AND
DriverAge = Grouptwo AND
Vehicleserviceyear = Ser_levelfour: Severe_injury (10.0/3.0)

Experience = NoLicense AND
Vehicleserviceyear = Ser_levelfive AND
Educationalbackground = Levelone: Severe_injury (9.0/3.0)

Experience = NoLicense AND
Vehicleserviceyear = Ser_levelfive AND
Educationalbackground = Levelthree AND
Relationshipwith = Employed AND
DriverAge = Grouptwo: Fatal_injury (4.0/1.0)

Experience = NoLicense AND
DriverAge = Grouptwo AND
TypesofVechile = Taxi: Severe_injury (3.0/1.0)

TypesofVechile = Taxi AND
DriverAge = Groupthree AND
Educationalbackground = Levelone: Property_damage (9.0/3.0)

Vehicleserviceyear = Ser_leveltwo AND
TypesofVechile = Truck: Severe_injury (4.0/2.0)
```

```
Vehicleserviceyear = Ser_levelthree AND
Experience = Lev_Two AND
Educationalbackground = Levelone: Slight_injury (4.0/1.0)

Vehicleserviceyear = Ser_levelthree AND
Experience = Lev_Five: Fatal_injury (13.0/6.0)

Vehicleserviceyear = Ser_levelthree AND
Ownership = Puplic_Agency AND
LicenseGrade = Third level: Property_damage (6.0/1.0)

Vehicleserviceyear = Ser_levelthree AND
Ownership = Private AND
Experience = Lev_Four: Slight_injury (17.0/5.0)

Vehicleserviceyear = Ser_levelthree: Property_damage (12.0/5.0)

Experience = Lev_Two AND
LicenseGrade = Third level AND
Relationshipwith = Employed: Severe_injury (5.0)

Experience = Lev_Two AND
DriverAge = Groupthree AND
Relationshipwith = Veh_owner: Severe_injury (4.0/1.0)

Experience = Lev_Two AND
LicenseGrade = Third level: Fatal_injury (4.0/1.0)

Experience = NoLicense AND
TypesofVechile = Truck: Severe_injury (9.0/3.0)

Experience = Lev_Two: Slight_injury (7.0/3.0)

Experience = NoLicense AND
Vehicleserviceyear = Ser_levelfour: Property_damage (7.0/3.0)

Experience = Lev_Five AND
TypesofVechile = Bus AND
DriverAge = Groupthree: Property_damage (5.0/2.0)

TypesofVechile = Bus: Slight_injury (9.0/4.0)

TypesofVechile = Automobile AND
LicenseGrade = Third level: Severe_injury (5.0/1.0)

TypesofVechile = Automobile AND
LicenseGrade = Fourth level: Property_damage (3.0)

TypesofVechile = Taxi AND
DriverAge = Groupfour: Property_damage (7.0)

Vehicleserviceyear = Ser_levelfive AND
TypesofVechile = Truck AND
LicenseGrade = Third level AND
Ownership = Private AND
Educationalbackground = Levelthree AND
```

```
DriverAge = Groupthree: Fatal_injury (7.0/3.0)

TypesofVechile = Taxi AND
Educationalbackground = Leveltwo AND
Vehicleserviceyear = Ser_levelfive: Severe_injury (6.0/3.0)

TypesofVechile = Taxi AND
Educationalbackground = Levelthree AND
Relationshipwith = Employed AND
DriverAge = Grouptwo AND
Vehicleserviceyear = Ser_levelfive: Property_damage (4.0/2.0)

TypesofVechile = Taxi AND
Educationalbackground = Levelthree AND
Relationshipwith = Employed: Slight_injury (5.0/1.0)

LicenseGrade = Fourth level AND
Ownership = Government AND
Educationalbackground = Levelthree: Property_damage (5.0/1.0)

LicenseGrade = Fourth level AND
Educationalbackground = Levelthree AND
DriverAge = Groupthree AND
Vehicleserviceyear = Ser_levelfive AND
Experience = Lev_Five: Slight_injury (6.0/3.0)

LicenseGrade = Fourth level AND
Experience = Lev_Five: Slight_injury (6.0/1.0)

LicenseGrade = Fourth level AND
DriverAge = Grouptwo AND
Educationalbackground = Levelthree: Severe_injury (4.0)

LicenseGrade = Fourth level AND
DriverAge = Groupthree AND
Vehicleserviceyear = Ser_levelfour: Slight_injury (6.0/3.0)

Ownership = Puplic_Agency AND
Vehicleserviceyear = Ser_levelfour: Fatal_injury (3.0/1.0)

Ownership = Government AND
Experience = Lev_Four AND
Vehicleserviceyear = Ser_levelfive: Fatal_injury (8.0/3.0)

Ownership = Government AND
Experience = Lev_Four: Slight_injury (6.0/2.0)

Ownership = Private AND
LicenseGrade = Fourth level AND
DriverAge = Groupthree: Fatal_injury (3.0/1.0)

Ownership = Private AND
LicenseGrade = Fourth level AND
DriverAge = Grouptwo: Slight_injury (3.0)
```

DECLARATION

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

Getnet Mossie Zeleke

April 2009

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

Dr. Manoj V.N.V (Assistant Professor)

April 2009