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

BAYESIAN APPROACH FOR ANALYSIS OF ROAD TRAFFIC ACCIDENT (THE CASE OF ADDIS ABABA)

A THESIS SUBMITTED TO

THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR

THE DEGREE OF MASTER OF SCIENCE

IN INFORMATION SCIENCE

BY

ALEMAYEHU TABOR FEYISSA

January 2009

ADDIS ABABA

Dedication

To my darling wife, Belen Mekonnen for her unquestioning support and to my loving daughter, Salem Alemayehu and to my and Belen's family for setting a wonderful example in all that they do.

Acknowledgements

I sincerely wish to thank my advisor Dr. Rahel Bekele for her invaluable support and for the significant contribution she has made to the path of my research. Three key contributions I'd like to mention are: her perseverance in providing study material, her skill as technical peer and her meticulous refereeing of my written work.

I have undoubtedly benefited considerably from the experience and expertise of my academic teachers along with the guidance of colleagues at the Addis Ababa University.

Finally, I'd like to express my deep gratitude towards my colleagues for their keen participation in my study work.

Alemayehu Tabor

TABLE OF CONTENTS

List of Figures

List of Tables

List of appendices

Abstract

Road traffic accidents are among the top leading causes of deaths and various levels of injuries in the world. Ethiopian is one of the developing countries where the situation is becoming worse and worse form time to time. The country is experiencing highest rate of such accident result in fatalities and a high economic loss. Addis Ababa, the capital city of Ethiopia, accounts for approximately more than 21% of the fatal accidents, 42% of the injury accident and 65% of the total accidents reported in the whole country.

This thesis reports the study carried out to develop accident predictive models based on the data collected on road accident in Addis Ababa. The BN power predictor and constructor were used for prediction and model construction purposes respectively. As a result relating or finding interrelatedness between the road and traffic flow explanatory variables and building a significant accident predictive models was possible. In doing so the potential applicability of Bayesian network to help traffic accident data analysis in decision-making process was explored. In the thesis, the process of building a model using Bayesian network tools and techniques from historical road accident record data is explained. Different tools and techniques are also used for the purpose of data analysis. The methodology adopted consisted of basic steps of data collection in which all the records are selected and extracted from Addis Ababa traffic office; data preparation which includes tasks such as data transformation, deriving of new attributes, and handling of missing values, and finally model building and validation using the selected tools and techniques.

In the first experiment, a best learned model that can classify accidents well with a better accuracy as serious, crash or property damage was selected and evaluated. The second experiment was also conducted after the necessary input of the domain experts is added. Experiment results reveal that the model built with mentioned techniques and tools are very

much helpful in identifying the potential contributors or causes of this ever-growing challenge of the road transport in Addis Ababa and their interrelatedness. The whole research process can be a good input for further research works.

ACRONYMS

BN	Bayesian Network
WHO	world Health Organization
AATA	Addis Ababa Transport Authority
UK	United Kingdom
GNP	Gross National Product
DAG	Direct Acyclic Graph
CPT	Conditional Probability Table
CIDR	Conditional Independence and Dependence Relation
TPDA	Three Phase Dependency Analysis
IID	Independent and Identically Distributed
RATA	Road And Transport Authority

CHAPTER ONE

INTRODUCTION

1.1 Background

As we are all aware transport spread throughout the whole of civilized world, like the arteries and veins in the human body. Transport services take people to places where they want to go and deliver goods to places where people require them efficiently, economically and safely. Among the various types of transportation systems road transport or road transportation is the major one which transport on roads passengers or goods. According to Journal of Transportation Engineering (2002) transport on roads can be roughly grouped into two categories: transportation of goods and transportation of people. In many countries licensing requirements and safety regulations ensure a separation of the two industries. The nature of road transportation of goods depends, apart from the degree of development of the local infrastructure, on the distance the goods are transported by road, the weight and volume of the individual shipment and the type of goods transported. For light and small shipments in small distances, a van or pickup truck may be used. For large shipments, even if it is less than a full truckload, a truck is more appropriate. People (Passengers) are transported on roads either in individual cars or automobiles or in mass transit/public transport using bus. Despite this broad goal, there are many factors that commonly disturb or damage these systems as a result cause road traffic accidents.

James O. Harris (2002) asserted in his studies of “Cause and Contributing Factors in Traffic Accidents”, that there are various factors that cause traffic accidents. Among them are:-:

- Driver distraction, including fiddling with technical devices, talking with passengers, eating or grooming in the car, dealing with children or pets in the back seat, or attempting to retrieve dropped items (Harris ,2002).

- Driver impairment by tiredness, illness, alcohol or other drugs, both legal and illegal (Harris, 2002).
- Mechanical failure, including flat tires or tires blowing out, brake failure, axle failure, steering mechanism failure (Harris, 2002).
- Road conditions, including foreign obstacles or substances on the road surface; rain, ice, or snow making the roads slick; road damage including pot holes (Harris, 2002).
- Speed exceeding safe conditions, such as the speed for which the road was designed, the road condition, the weather, the speed of surrounding motorists, and so on (Harris, 2002).
- Road design and layout. Some roads are notorious for being accident "black spots" for a whole variety of reasons, many subtle and not necessarily immediately obvious. These include alignment, visibility, camber and surface conditions, road markings, etc (Harris, 2002).

According to World Health Organization (WHO) road traffic accidents result in the deaths of an estimated 1.2 million people worldwide each year, and injure about forty times this number (WHO, 2004).

Each year, an estimated 1.2 million people are killed in road crashes and up to 50 million injured worldwide (WHO , 2004) Road traffic injuries are currently ranked 9th globally among the leading causes of disease burden, in terms of disability adjusted life years lost. In the year 2020, road traffic injuries are projected to become the 3rd largest cause of disabilities in the world (Murray, 1996).

Developing countries bear the brunt of the fatalities and disabilities from road traffic crashes, according to WHO accounting for more than 85% of the world’s road fatalities, and about 90% of the total disabilities are due to road traffic injuries. The problem is increasing in these countries at a fast rate, while it is declining in all industrialized nations (Western Europe, North America, Japan, Australia and New Zealand).

The tragedy is more or less similar in Ethiopia especially in Addis Ababa. The rate of traffic accidents in Addis Ababa goes up together with the increase of motor vehicles and population size. The rise in automobile ownership together with the poor condition of the roads has resulted in high level of traffic safety and congestion problems.

At a country level, above 1,800 people died while above 7,000 were crippled or injured annually. Moreover the death rate is 136 per 10,000 vehicles and Ethiopia is loosing over 400 million birr yearly as a result of road traffic accidents (Abebe, 2004). The share of Addis Ababa city in the total number of accidents was 60 percent in 1989 with annual average traffic accident growth 31.4 percent (Addis Ababa Transport Authority, AATA, 2004). Now a day Addis Ababa is experiencing around 700 accidents per month resulting in various level of injury severity.

1.2 Statement of the Problem

Ethiopia with a population of 73 million and 1.5 cars per 1,000 people has 109,000 cars which are involved in 1,800 fatal crashes per year, or one fatal crash for every 60 cars. The UK with a population of 60.4 million and 434 cars per 1,000 people has 26.2 million cars which are involved in 3,262 fatal crashes per year, or one fatal crash for every 8,031 cars. This fact shows us that a car driven by an Ethiopian is 134 times more likely to kill someone than a car driven by an Englishman that means Ethiopian drivers 134 times more deadly than English drivers (Fathers' Manifesto & Christian Party 2008). In addition to this as some reports from World health Organization indicates Ethiopia is one of the most dangerous places to drive in Africa. Specifically according to world health organization, Ethiopia has the highest rate of fatalities per vehicle in the world. Uganda ranks second in road fatal accident rates in the world next to Ethiopia (WHO ,2008) and of all the accidents that occurred in Ethiopia, Addis Ababa takes the largest share (Tibebe, 2005). In this regard statistical data from the Addis Ababa traffic police office shows that the city is experiencing from 700 to 800 accidents per month and this number has been increasing from time to time. As some reports from the Addis Ababa traffic office shows road traffic crashes are the major cause of death among young people between 10 and 24 years and thousands more are injured or disabled. This shows that the country is losing a lot of productive young people almost in every tragic road accident situations which has a great impact on socio economic aspect of the country. According to some reports of the world health organization for many low- and middle-income countries, the cost of road crashes represents between 1-1.5% of GNP and in some cases exceeds the total amount the countries receive in international development aid. This shows that unless more comprehensive action is taken, the number of deaths and injuries is likely to rise significantly.

The Addis Ababa traffic office and Addis Ababa transport authority tries their best on taking different measures to reduce and to manage these traffic accidents. Some of the activities that are taken so far are to adjust the traffic rules and laws with different time intervals and penalizing those who broke the laws. In addition to these permanent programs about the traffic accidents, causes and rules and laws are transmitted using mass Medias. Having all these done still the problem remains unsolved. That means the accident rate is increasing from time to time.

The attempt made by Addis Ababa traffic office and Addis Ababa transport authority to collect and document traffic accident data is somehow encouraging but a lot needs to be done to draw some valuable knowledge and information from these historical records and so that the research will contribute it’s share in alleviating the problem.

Analyzing these data manually is very difficult because it is very time-consuming and messy (Gregory, 1993). Consequently, because of the increasing traffic flow and vehicles, it has become necessary to use computer based data analysis techniques and tools. This allows the Addis Ababa traffic office authority not only to monitor the traffic state at each site but also to get real-time section related traffic accident information.

A research made by Tibebe (2005) is one of the attempts that focused on developing adaptive regression trees to build a decision support system to handle road traffic accident analysis for Addis Ababa city traffic office using data mining tools, but the crude point that we understand from the traffic data is that the number of accidents on a given road section during a certain period of time is probabilistic in nature and is a non-negative integer. Despite the fact that accidents are random and unpredictable at micro level, statistical models can predict reliable estimates of expected accidents by relating aggregates of accidents to various explanatory measures of flow, site characteristics, and road geometry at macro level. Numerous empirical relationships between vehicle accidents and these explanatory variables have been established in several previous studies. Such accident predicting models are useful in identifying the most critical variables to safety, assessing design and management alternatives, and improving the safety standards for new roads. Because of the complex and uncertain nature of accident events and the requirement of unattainable quality of accident, traffic, and road data together with lack of skilled manpower and resources such types of studies are very limited in developing countries. The application of research results obtained from developed countries to these countries is questionable due to differences in the natures of the problem.

Some of the points that justify the complexity and uncertainly nature of the problem are its partial observability (road state, other drivers' plans, etc.), noisy sensors (traffic reports), uncertainty in action outcomes (flat tire, etc.), immense complexity of modeling and predicting traffic. Therefore due to these facts uncertainty management techniques need to be used in order to deal with such uncertain knowledge.

One of the uncertainty management techniques is the Bayesian Network model (also called belief networks, Bayesian belief networks or casual probabilistic networks) that has proven to be a valuable tool of encoding, learning and reasoning about probabilistic relationships. In addition to this the factors that cause traffic accidents are highly interrelated. For instance, road conditions are influenced by the weather. Traffic flow depends on the time of the day, whether it is a weekday or weekend, and weather conditions. The characteristics of people involved (e.g., age, sex, experience) can often be related to the speed of the vehicles in an accident and the use or non-use of seat belts. The outcome of an accident is, by and large, dependent on the speed of the vehicles involved. So in order to model the interdependence among the variables related to accidents and to combine domain knowledge with the statistical data Bayesian network seemed particularly useful sine they are based on a combination of probability theory, which deals with uncertainty with complexity (interrelatedness). These networks are an important tool in the design and analysis of machine learning algorithms and are based on the idea of modularity whereby a complex system is built by combining simpler parts (Neapolitan, 2002). Probability theory connects parts and ensures the consistency of the system as a whole while providing the possibility of interfacing the models with the data (Andrew Y. 2000).

So the researcher attempted to draw rational conclusions about a road traffic accident which can be viewed as a problem in uncertain reasoning about a particular event, to which developments in the modeling of uncertain reasoning for artificial intelligence can be applied to get the predicted model so that a better understanding between the relationship of injury severity and risk factors as well as identification of the contributing factors affecting crash severity with broad considerations of driver characteristics, roadway features, vehicle types, pedestrian characteristics and crash characteristics and their interrelatedness is made using Bayesian belief network model for better decision making. In addition to this the model paves the way to develop better parameters in all aspect of traffic control system which will improve the traffic control system. The research also be a ground for other researchers to extend the research in other direction on the basis of the result found. The research can be also adapted to other domains that have dynamic interdependency between different factors. Examples can be crime prevention and identification system, city traffic Monitoring systems, and fire fighting systems.

1.3 Research Questions

From the discussion of the problem statement, the research questions that guided the research work are mentioned as follows.

1. What are the methods or techniques used to identify and measure the factors (variables)?
2. What are the variables related to the type of traffic accidents?
3. How can one find the dependency between these variables?
4. Which factors have significant contributions for most traffic accidents?
5. How can one build the Bayesian network based on these variables?

1.4 Objectives

The general objective of this research work is to come up with a Bayesian model that can support road traffic accident analysis in the effort of preventing and controlling vehicle accident at the city of Addis Ababa.

1.4.1 Specific Objectives

The specific objectives of this research work are stated as follows:

- To study existing techniques and traffic control system.
- To develop a Bayesian network model using the selected tools, techniques and technologies.
- To assess the potential application of Bayesian Networks in identifying the major causes (influencing factors) for traffic accidents and to identify their dependencies, which help the police officers what kind of measures that they should take and adjusting the law accordingly.
- To identify the variables (factors that cause accidents) and their relationship with the type of traffic accidents.
- To assess which factors have significant contributions for most traffic accidents?

1.5 Methodology /Approaches

In order to achieve the general and specific objectives mentioned before, the researcher employed different methods in three consecutive phases of this work. Detailed explanations of this part are also presented in Chapter 4 of this thesis work.

1.5.1 Data collection/data preparation

Collecting data, which are relevant to the problem domain, was one of the important parts of the research work. Studying the traffic scene situations or environments to point out what the actual scenario needs from the analysis using Bayesian Networks facility, collecting data and arranging into a form that will be suitable for the particular analysis of the BN power soft tool was the major activities carried out in this first phase.

So the relevant data was collected from Addis Ababa traffic office which was stored using MS Excel file format. Accessing all attributes was not easy and actually was not important since some of the fields contain some values which are not relevant to this study. So formal and informal discussions were made with the traffic officers in order to get a clear understanding of some of the values of the fields and to resolve some privacy issues. Detail discussions on data collection phase are made in Chapter 4 of this thesis.

1.5.2 Data analysis

After collecting the data, cleaning of data was done using MS Excel. Here data cleaning activities which includes detecting the missing values of the fields and filtering irrelevant data values was made. In addition data transformation and data reduction was also carried out. Finally we have remained with 1956 records for further analysis. All these activities are discussed in detain in chapter four.

1.5.3 Build and train model

In this phase before building the Bayesian network model based on the inputs from the first and the second phases, training set was done with 10-fold cross validation technique which is highly recommended for large datasets (Ricardo, 2001).

The number of records used in this phase is 1956 and testing and building the Bayesian model was implemented using BN Power Constructor before and after eliciting the domain experts. The training and testing was conducted repeatedly ten times. Its confusion matrix was generated correspondingly. Then the average prediction accuracy was calculated for the ten folds before and after eliciting the domain experts, here the result found in the experiment showed that there is no such big difference before and after reinforce the BN model. This implies that BN model can perform in a promising way even with the absence of elicitation process. The detail of the experimentation is discussed in chapter four.

1.6 Scope and limitation of the study

In general the scope of the research is limited to assessing the possible application of Bayesian network technology at Addis Ababa traffic offices; specifically it is limited to identifying different factors (variables) that cause and highly associated with road accidents and to identify the interdependency between these factors or causes using the potentials of Bayesian network techniques by developing appropriate model.

However getting access to resources such as literature related to the study especially on Addis Ababa case is one limitation. Moreover, even if considering each and every law set for any traffic accidents are important topics and have some contribution in this research; this work does not go into the detail of these issues, as the topic is very wide and demands to undertake a separate exploration.

In addition to this, due to time limitation, only prediction and BN network model were made and inference on the model was not considered.

1.7 Thesis organization

This thesis report is organized into five chapters. Chapter two presents introductory background on Bayesian network which forms the foundation for the subsequent chapters. In this chapter, definition of Bayesian network, modeling (representation) and reasoning issues in Bayesian network are discussed. Chapter three covers overview on traffic accidents and traffic accident severity, road accidents at Addis and the issues related to traffic accident analysis in Addis Ababa. The fourth chapter presents general overview on data collection, data prepossessing steps, variables selections and the detail of developing the model by incorporating the description of tools used in developing the model, selection of the modeling technique and the result obtained when evaluating the model. Finally, the fifth chapter wraps-up the whole thesis and points out possible rooms of improvement in the future by providing conclusion and recommendation.

CHAPTER TWO

BAYESIAN NETWORK

2.1 Introduction

Bayesian networks are an increasingly popular method of modeling uncertain and complex domains such as ecosystems and environmental management. At best, they provide a robust and mathematically coherent framework for the analysis of this kind of problems (Laura, 2001). The main goal in using Bayesian network is to capture dependencies that exist in real decision–making problems (Rahel, 2005). In this chapter basic concepts on Bayesian network and their potential application in different areas are discussed.

2.2 Fundamentals of Bayesian Probability

2.2.1 Events

The language of probabilities consists of statements (propositions) about probabilities of events. The probability of an event "a" denoted by $p(a)$. An event can be considered as an outcome of an experiment (e.g., a coin toss), a particular observation of a value of a variable (or set of variables), an assignment of a value to a variable (or set of variables), etc. As a probabilistic network defines a probability distribution over a set of variables, V , in our context an event is a configuration, $x \in dom(X)$, (i.e., a vector of values) of a subset of variables $X \subseteq V$

2.2.2 Axioms

The following three axioms provide the basis for Bayesian probability calculus (Bayes, 1763; Grinstead, 1993).

Axiom 1 For any event, a, $0 \le p(a) \le 1$, It simply means that a probability is a non-negative real number less than or equal to 1, and that it equals 1 if and only if the associated event has happened for sure.

Axiom 2 For any two mutually exclusive events a and b the probability that either a or b occur is

P(a or b) $\equiv p(a \cup b) = p(a) + p(b)$. In general, if events $a_1, \ldots\ldots, a_n$ are pair wise incompatible, then

$p\left(\bigcup_{i}^{n} a_i\right) = p(a_1) + \ldots.. + p(a_n) = \sum_{i}^{n} p(a_i)$. It means that if two events cannot co-occur, then the probability that either one of them occurs equals the sum of the probabilities of their individual occurrences.

Axiom 3 For any two events a and b the probability that both a and b occur is

P(a and b) $\equiv p(a,b) = p(b \mid a)\, p(a) = p(a \mid b)\, p(b)$ is called the **joint** probability where it is the probability of both events a and b together. This sometimes referred to as the fundamental rule of probability calculus.

2.2.3 Prior and Conditional probability

The basic concept in the Bayesian treatment of uncertainty is that of conditional probability (Krzysztofowicz, 2001, Bayes, 1763). Given event B, the conditional probability of event A can be written as $p(A|B)$. This means that if B is true and everything else known is irrelevant for A, then the probability of A is determined.

The conditional probability $p(A|B)$ can be calculated using the formula:

$$p(A|B) \equiv \frac{p(A \cap B)}{p(B)}$$, where $p(B) > 0$;

On the other hand the prior probability of event A which can be written as $p(A)$ is the probability of A regardless of any other information.

2.2.4 Conditional Independencies

If two events A and B are independent then one can clearly define the probability of one of them regardless of any information about the other. That means their conditional probability is the same as their prior probability. This can be written mathematically as:

$p(A \mid B) = p(A)$ and $p(B \mid A) = p(B)$.

Where as if two events A and B conditionally independent given event C then the joint probability of A and B can be calculated using the following mathematical formula. $p(A = a, B = b|C = c) = p(A = a|C = c) * p(B = b|C = c)$.

Similarly the joint probability of ABC, $p(A, B, C)$ where A and B conditionally independent, given C, can be calculated as:

$$p(A, B, C) = p(A|C) * p(B|C) * p(C)$$

2.2.5 Bayes' Theorem

Basically the concept Bayes' Theorem is derived from the concept of conditional probability. In order to derive the theorem let us see first the probability of event A given B and the probability of B given A in the following equations.

$$p(A|B) \equiv \frac{p(A \cap B)}{p(B)}$$, where $p(B) > 0$; is the probability of event A given B

Similarly the probability of event B given A can be calculated as:

$$p(B|A) = \frac{p(A \cap B)}{p(A)}, \text{ where } p(A) > 0$$

From the above two equations, we can derive the following equation.

$$p(A|B) * p(B) = p(B|A) * p(A).$$

From this we can find the probability of A given B as:

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)}, \text{ which is called Bayes' Theorem}$$

2.3 Basics of Bayesian networks

The need for a method for knowledge representation and reasoning under uncertainty, led to the definition of Bayesian networks (Pearl 1988).A Bayesian network can be described briefly as directed acyclic graphs where nodes represent random variables or uncertain quantity which take two or more possible values and arcs represent causal connections or direct influences between nodes (pearl, 1998). Associated with each node there is a probability table that provides conditional probabilities of the node's possible states given each possible state of its parents (Huang 1993). Given the observed values of any subset of the random variables, the posterior probability distribution can be calculated for the unknown variables for example a Bayesian network could represent the probabilistic relationship between type of road traffic accident and its cause. Given the type of road traffic accidents, the network can be used to compute the probabilities of the presence of various causes of accident. (Neil 2005). The following figure illustrates how the direct acyclic graph (DAG) between nodes seems like.

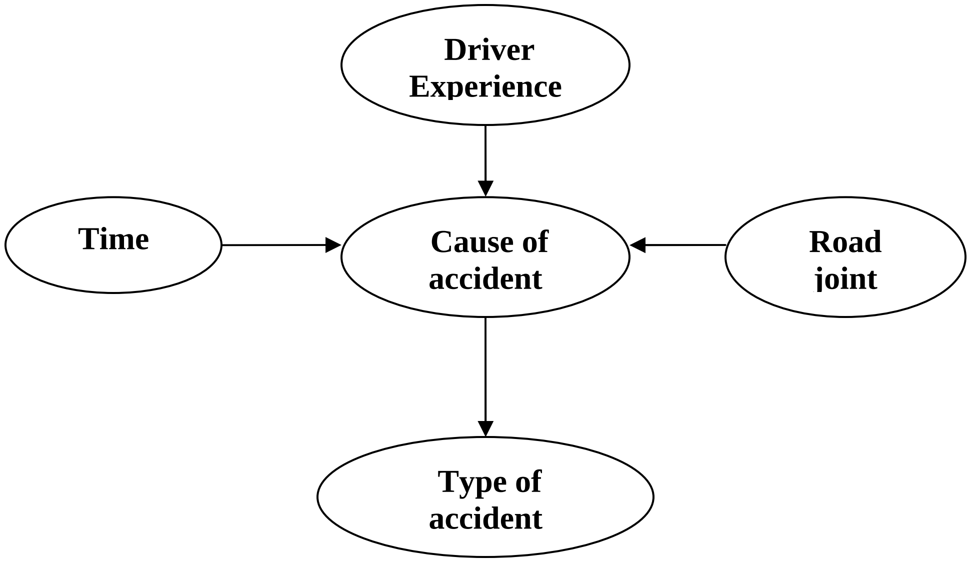

Figure 1.1 A Bayesian network diagram

Mathematically the relation ship between the variables and their values in the arcs can be expressed as follows. The acyclic directed graph (DAG) which defines a factorization of a joint probability distribution over the variables that are represented by the nodes of the DAG, where the factorization is given by the directed links of the DAG. More precisely, for a DAG, G = (V, E), where V denotes a set of nodes (or vertices) and E a set of directed links (or edges) between pairs of the nodes, a joint probability distribution, $p(x_v)$ over the set of (typically discrete) variables x_v indexed by V can be factorized as

$$p(x_V) = \prod_{v \in V} p(X_v \mid X_{pa(v)})$$

Where $X_{pa(v)}$ denotes the (preferably small) set of parent variables of variable X_v for each node $v \in V$ .The factorization in the above Equation expresses a set of independence assumptions, which are represented by the DAG in terms of pairs of nodes that are not directly connected to one another by a directed link. It is the existence of such independence assumptions and the small set of parents for each node that makes it possible to specify the conditional probabilities and to perform inference efficiently in a Bayesian network.

Each conditional probability distribution $p\left(X_v \mid X_{pa(v)}\right)$ represents a set of "rules", where each "rule" (conditional probability) takes the form

If $X_{pa(v)} = x_{pa(v)}$ then $X_v = x_v$ with probability z,

Where x_v and $x_{pa(v)}$ denote, respectively, a value assigned to X_v and a vector of values assigned to the parent variables of X_v. For example, if one of five possible values can be assigned to X_v and it has four parents each of which can be assigned one of three possible values, and then $p\left(X_v \mid X_{pa(v)}\right)$ represents a collection of $5 \times 3^4 = 405$ rules (Norvig, 2003). For example if we assume "n" variables then the amount of information needed to specify each conditional probability table will be at most

2.4 Causality

Assume that the occurrence of some event c is known to cause the effect e and that the relationship between c and e is known to be deterministic (logical). Then, obviously, observing c we can conclude e. observing e, on the other hand, does not make us able to conclude c, unless c is known to be the only cause of e. Thus, in formulating the causal relationship between c and e as a rule we would obviously want to formulate it as "if c then e" (i.e. c → e) rather than "if e then c". Form this point of view we can see that causality is key in the process of constructing probabilistic network models.

From this insight we conclude that rules like R1 and R2 express causal relationships, where s1, say, plays the role of the cause and s2 the role of the effect, the only possible exception being if s1, as an effect, only can be caused by s2. Violating the "causal direction" in formulating rules is, however, not advisable.

2.4.1 Basic connections in causality

The Connections between nodes in a DAG is not only helpful in knowing the causal relationship between the nodes but also to know the conditional independence.

The basic three connections in Bayesian network are the following.

I. Serial Connections

According to Lugg (Lugg et al, 1995) in Serial connection the causal effect between nodes is expressed as $X \rightarrow Y \rightarrow Z$ and it means X causes Y and Y causes Z, X and Z are not probabilistically independent, and X and Z are conditionally independent given Y.

For example, the serial connection Driver Sex → Driver Experience → Type of Accident shows that driver sex causes driver experience in turn driver experience causes type of accident occurred. So here type of accident is conditionally independent of driver sex given driver experience. As we can see there is no direct connection between driver sex and type of accident occurred so we said driver sex and type of accident occurred are direction separated (D-separated).

II. Diverging Connections

In Diverging connection the causal relationship between variable may be expressed as

X ←Y →Z, where variable Y causes variable X and variable Z. Here Z is conditionally independent of X given Y. That means once the state of Y is known further knowledge about Z is irrelevant to X.

III. Converging Connections

In converging connection the causal relationship between variable can be expressed as

X → Y ←Z, where X and Z are independent causes of Y. Here if evidence on Y or one of its descendants is available then X and Z are relevant to each other.

2.5 Learning in Bayesian Networks

According to Kelly (Kelly , 2001) learning a Bayesian network as the task of identifying a DAG structure G and a set of conditional probability distributions P with parameters Θ on the basis of database of cases D = $\{c^1 . . . c^N\}$ and possibly some domain expert background knowledge. When learning Bayesian networks, the correctness of the learned network of course depends on the amount of training data available (Mitchell, 2006).When training data is scarce, it is useful to employ various forms of prior knowledge about the domain to improve the accuracy of learned models. For example, a domain expert might provide prior knowledge specifying conditional independencies among variables, constraining or even fully specifying the network structure of the Bayesian network.

In addition to helping specify the network structure, the domain expert might also provide prior knowledge about the values of certain parameters in the conditional probability tables (CPTs) of the network, or knowledge in the form of prior distributions over these parameters.

2.5.1 Structure Learning From Data

Structure learning from data is the task of inducing the structure, i.e., the graph, of a Bayesian network from a source of data. There are two main approaches in structural learning of Bayesian network namely scoring-based and constraint-based .The Scoring-Based approach goal is to find the graph that best matches the data by first introducing a scoring functions that evaluates each network with respect to the data and then searching for the best network according to this score whereas in constraint-based approach one tries to match the conditional independence relation observed between variables in the data with those entailed by a graph, based on statistical independence tests , that means in the constraint-based approach, the DAG G of a Bayesian network N = (X, G, P) is considered as an encoding of a set of (conditional) dependence and independence relations (CIDRs) M_G, which can be read off G using the d-separation criterion (Lauritzen et al. 1990b, Geiger, Verma & Pearl 1990). Structure learning is then the task of identifying a DAG structure that (best) encodes a set of CIDRs. The set of CIDRs may, for instance, be derived from the data source by statistical tests. Based on D alone, we can at most hope to identify an equivalence class of graphs encoding the CIDRs of the generating distribution p_o.

A constraint-based structure learning algorithm proceeds by determining the validity of independence relations of the form $I\left(X,Y \mid S_{XY}\right)$ (i.e., X is independent of Y given subset S_{XY} where $X,Y \in x$ and $S_{XY} \subseteq x$). The structure learning algorithm will work with any information source able to provide such information.

Learning the structure of a sparse graph is computationally less involved than learning the structure of a dense graph where the number of edges is used as a measure of the density of the graph. Inducing a graph from a sample of cases that require the induced graph to be dense is computationally more expensive than inducing a graph from a sample of cases that require the induced graph to be sparse. In addition, domains that require the induced graph to be dense may be difficult to represent as a Bayesian network as inducing the graph is computationally expensive, representing a dense graph requires a lot of storage, and inference in dense graphs may be intractable. The size of the space of possible DAGs grows super-exponentially with the number of vertices in the graph. Robinson (1977) gives the following recursive formula for calculating the number $f(n)$ of DAGs on n vertices:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{(n-i)!i!} 2^{i(n-1)} f(n-i)$$

For example, $f(10) \approx 4.2 * 10^{18}$

2.5.2 Structure Constraints

Prior to the testing phase, background knowledge of domain experts in the form of constraints on the structure of the DAG can be specified. It is possible to specify the presence and absence of edges, the orientation of edges, and a combination.

If the background knowledge is assumed to be consistent with the underlying DAG G_0 of the generating distribution P_0, then it is not necessary to test the validity of the background knowledge. Hence, specifying background knowledge may reduce the number of statistical tests. Unfortunately, this may, in practice, produce unwanted behavior of the edge-orientation algorithm. This implies that background knowledge should often be used with caution. (Mitchell, 2006)

2.5.3 Algorithm for Learning Bayesian Network

The Bayesian Network can be developed using different algorithms. The Three-Phase-Dependency-Analysis (TPDA) learning algorithm is one of them developed by Cheng et al (Cheng, J., 2001) using the idea of information theory is the one incorporated in the BN PowerConstructor software and the researcher used since it is recommended as a good learning algorithm. In using the TPDA algorithm some assumptions are required about the input data. These are:

- The data set is "Independent and Identically Distributed (IID)"
- The cases in the data are drawn IID from a DAG faithful distribution
- The attributes of a table have discrete values
- There are no missing values in any of records
- The quantity of data is large enough for the CI test being reliable

TPDA consists of three phases of TPDA algorithm, i.e.: drafting, thickening and thinning, as described below, and complete TPDA algorithm can be found collectively in the three phase (Cheng, J., 2001):

1. **Drafting Phase:** produces an initial set of edges based on sufficient mutual information test. The draft is a singly-connected graph (a graph without loop) which is found using the Chow-Liu algorithm
2. **Thickening Phase:** adds edges to the current graph when the pairs of nodes cannot be separated using a set of relevant CI tests. The graph produced will contain all the edges of the underlying dependency model
3. **Thinning Phase:** each edge is examined and it will be removed if the two nodes of the edge are found to be conditionally independent. Finally, TPDA runs procedure for orienting edges.

2.6 Inference in Bayesian Networks

Contrary to rule-based systems with certainty factors, inference in Bayesian networks is always consistent and the ability to handle the explaining-away[A] problem is embedded naturally in the way in which inference is performed in Bayesian networks. However, in general, it is an hard task to solve the inference problem in Bayesian networks (Cooper 1990); even approximate inference is NP-hard[B] (Luby 1993). Fortunately, efficient inference algorithms have been developed such that inference in Bayesian networks can be done in fractions of a second even for large networks containing hundreds of variables (Lauritzen & Spiegelhalter 1988, Jensen, Lauritzen & Olesen 1990). Efficiency of inference, however, is highly dependent on the structure of the DAG, so networks with a relatively small number of variables sometimes resist exact inference, in which case approximate methods must be applied.

As Bayesian networks most often represent causal statements of the kind X → Y, where X is a cause of Y and where Y often takes the role of an observable effect of X, which typically cannot be observed itself, we need to derive the posterior probability distribution P(X|Y = y) given the observation Y = y using the prior distribution P(X) and the conditional probability distribution P(Y |X) specified in the model. Reverend Thomas Bayes (1702–1761) provided the famous Bayes' rule for performing this calculation:

$$p(X \mid Y = y) = \frac{p(Y = y \mid X)\, p(X)}{p(Y = y)}$$

[A] 'Explaining away' is a common pattern of reasoning in which the confirmation of one cause of an observed or believed event reduces the need to invoke alternative causes. The opposite of explaining away also an occur, where the confirmation of one cause increases belief in another.

B A problem is 'NP-hard' if an algorithm for solving it can be translated into one for solving any NP-problem (nondeterministic polynomial time) problem. NP-hard therefore means "at least as hard as any NP-problem," although it might, in fact, be harder.

Where $p(Y=y)=\sum_x p(Y=y \mid X=x)p(X=x)$. This rule (or theorem) plays a central role in statistical inference because the probability of a cause can be inferred when its effect has been observed. Olmsted (1983) and Shachter (1986) developed a method for inference in Bayesian networks, which involved multiple applications of Bayes' rule. Lauritzen & Spiegelhalter (1988) and Jensen et al. (1990) developed inference methods for Bayesian networks based on message passing in a tree structure (junction tree) derived from the structure of the Bayesian network. The latter approach is the prevailing inference method used in modern software packages for inference in probabilistic networks.

2.7 Applications of Bayesian networks

A number of researches have been conducted to proof the immense application of Bayesian network on the area of business and finance, causal learning, computer games, computer vision, Natural Language Processing, biology and medicine. "Computer-Assisted Learner Group Formation Based on Personality Traits" (Rahel, 2005) "Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management" (Ezawaet al, 1995) "A conceptual approach to assessing risk to habitat" (Taylor, 2003) and "developing a technique for learning causal relationships among genes by analyzing gene expression data" (Friedman, 2000) can be mentioned as an example among the various research works made on Bayesian belief network.

In general a number of researches have been conducted to show the possible applications of Bayesian network in many areas and organizations in different parts of the world, this is because, Bayesian networks have been found very useful for real life applications since it has the following advantages (Pourret, 2006):

- Bayesian networks can readily handle incomplete data sets.
- Bayesian networks allow one to learn about causal relationships

- Bayesian networks readily facilitate use of prior knowledge
- Bayesian methods provide an efficient method for preventing the over fitting of data (there is no need for pre-processing) so that Contradictions do not need to be removed from the data and data can be "smoothed" such that all available data can be used
- Bayesian Networks visually represent all the relationships between the variables in the system with connecting arcs.
- Bayesian Networks is easy to recognize the dependence and independence between various nodes. Humans can easily understand the network structures and experts can modify them to obtain a better predictive model.

From the foregoing discussion it is apparent that the application areas of Bayesian network as a field of study is growing from time to time. And it is getting more and more acceptance in areas where uncertainty exists.

CHAPTER THREE

ROAD TRAFFIC ACCIDENT AND INJURY ANALYSIS

Under this chapter concepts like trends in road transport, road traffic accidents, traffic system, road safety and accident injury severity are discussed. This will help a great deal in understanding the domain area where Bayesian network modeling technology is going to be applied.

3.1 Introduction

The road transport system is one of the transportation sectors and named from the infrastructure that the vehicles use which is road, today roadways are mainly asphalt or concrete and specifically designed for the passage of wheeled vehicles. Road transport plays an important role in facilitating the socio-economic activities in the national economy. The term "Traffic" consist of pedestrians, ridden or herded animals, vehicles, streetcars and other conveyances, either singly or together, while using the public way for purposes of travel. In general, the movement of people along with transportation means is referred to as traffic.

3.2 Road Traffic Accidents in Addis Ababa

Road Traffic accidents have been recognized as one of the major causes for human and economic losses both in developed and developing countries. This problem is of great concern in developing countries, because of its seriousness and the limited resources to develop feasible countermeasures for reducing this ever-growing challenge. Developed countries have designed and implemented different strategies to reduce the scale and severity of this problem through education, enforcement, and engineering.

Each year, an estimated 1.2 million people are killed in road crashes and up to 50 million injured worldwide (WHO , 2004) Road traffic injuries are currently ranked 9th globally among the leading causes of disease burden, in terms of disability adjusted life years lost. In the year 2020, road traffic injuries are projected to become the 3rd largest cause of disabilities in the world (Murray, 1996). According to WHO (WHO, 2004) some 50% of road traffic fatalities worldwide involve young adults aged 15-44 years corresponding to the most economically productive segment of the population.

Developing countries bear the brunt of the fatalities and disabilities from road traffic crashes, according to WHO accounting for more than 85% of the world's road fatalities, and about 90% of the total disabilities are due to road traffic injuries. The problem is increasing in these countries at a fast rate, while it is declining in all industrialized nations (Western Europe, North America, Japan, Australia and New Zealand).

The tragedy is more or less similar in Ethiopia, above 1,800 people died while above 7,000 were crippled or injured annually. Moreover the death rate is 136 per 10,000 vehicles and the country is loosing over 400 million birr yearly as a result of road traffic accidents (Abebe, 2004). The situation is worse especially in Addis Ababa, since the city is the political and commercial centre of Ethiopia and represents the most complex motor and pedestrian traffic system in the country. Unlike many modern cities, it is growing without adequate plan and control, which led to the existing mixed-up land uses. The city's road network is not adequately planned to meet the current traffic demand. Road hierarchies are not well established and arterial roads are not access controlled. The traffic management system in the city is inadequate to cope with the rapidly increasing motor and pedestrian traffic. Traffic controls, signs, and markings as well as pedestrian facilities are not in place. The rapidly increasing traffic on the undeveloped road infrastructure coupled with the unsafe behavior of road users results in a highly conflicting traffic that consequently leads to the occurrence of many road accidents in the city. Addis Ababa accounts for approximately more than 21% of the fatal accidents, 42% of the injury accidents, and 65% of the total accidents reported in the whole country (Berhanu, 2000, Addis Ababa Transport Authority, AATA, 2004). Arterial roads of the city are amongst transport facilities, which experience greater number of traffic accidents.

According to the sources of the road and transport authority of the city (RATA ,2004) , the leading causes of road traffic accidents in Ethiopia are, not respecting speed limit, driver characteristics, not giving priority for pedestrian, and vehicle defects.

Even if a traffic rules and laws are set for annual technical investigation of vehicles and speed limit of vehicles in different road conditions and types, due to the problems in implementing the rules and low level of awareness of the drivers , the factors are remain as major causes of road traffic accidents.

3.3 Road Safety and Accident Injuries Severity

Reports of WHO and Traffic Office at Addis Ababa showed that the total number of road traffic deaths and injuries at all levels are growing at an alarming rate. It should get due attention to avert the problem with its devastating human impact and large economic cost to the country.

Road traffic safety deals exclusively with road traffic crashes and how to reduce their number and their consequences. Generally it aims to reduce the harm (deaths, injuries, and property damage) resulting from crashes of road vehicles. Harm from road traffic crashes is greater than that from all other transportation modes (air, sea, space, off-terrain, etc.) combined (UNECE, 2006).

According to WHO (WHO, 2004), speeding, alcohol, non-use of helmets, seat belts and other restraints, poor road design, poor enforcement of road safety regulations, unsafe vehicle design, and poor emergency health services are put as the major risk factors. The organization also tried to advocate a systematic approach as strategy to road safety, which takes in to consideration the key aspects of the road user, the vehicle and the infrastructure.

Some of the important points mentioned as the new strategy in understanding and dealing with road safety are summarized as follows:

- Crash injury is largely predictable preventable. It is a problem amenable to rational analysis and remedy
- Road safety policy must be based on a sound analysis and interpretation of historical and current data
- Road safety is a public health issue that intimately involves a range of sectors including that of health. All have their responsibilities and all need to be fully engaged in injury prevention.
- Since human error in complex traffic systems cannot be eliminated entirely, environmental solution (including the design of roads and of vehicles) must help in making road traffic systems safer.

In general a better understanding of the interdependency and relationship between the potential causes of road traffic accident will help a lot in reducing road traffic accidents.

3.4 Accident Data Analysis at Addis Ababa Traffic Office

It is crystal clear that collecting the raw facts like road type, the road way and its related environmental features, the vehicle information, the driver information, causality details and traffic characteristics related to time and location will provide good information about the traffic accident analysis study.

After accidents are reported, the Addis Ababa traffic office will assign an investigator in order to collect the necessary detail information about the accident. The investigation and recording of the raw facts will assist to find detailed and accurate information about cause of the accident and determine whether there was a violation of the law or not, but due to the time gap between the accident and the arrival of the investigator traffic officer, some details of the information, like the severity level and the true cause of the accident, may not be identified effectively.

The Addis Ababa traffic office and other governmental and non-governmental organizations use the data collected for various purposes. The Addis Ababa traffic office actually tried to classify accident records by accident time and date, age and sex of drivers, education level and experience of the driver, road and light conditions and by possible cause of accident summary of report using the collected information where as the national and the regional transport authority offices basically used the data for setting different traffic rules, laws and policies with regard to road safety. However the increase of the accident number showed that to use simply the raw facts will not bring a dramatic change in preventing and minimizing the accidents.

As it is compared to the complexity and magnitude of the road accident problem in the country we will found only very few research works with respect to the problem, but in the developed countries, because they are able to conduct a lot of researches by putting all the necessary resources and efforts they are in a position to control the problem. Therefore the country needs to give a serious attention and conduct a compressive and targeted researches need to be conducted using different techniques and tools until a satisfactory result is found.

A research work by Abebe (2004) is the one among the few research work which tried to assess points that need close attention with respect to road safety. He stated that 81% of the accident all over the country is due to drivers fault and the other 29% is due to vehicle, pedestrian and road faults. The researcher listed the following as the main road safety problems.

- Divers ignoring pedestrian priority
- Over speeding
- Utilization of freight vehicles as passenger transportation
- Poor skill and undisciplined behavior of drivers

- Poor road design
- Poor vehicle conditions
- Pedestrian ignorance on traffic rules and negligence
- Weak traffic law enforcement
- Lack of proper emergency medical service

Finally he concludes that, road safety publicity, targeted traffic law enforcement, hazardous location identification, increasing the pedestrian awareness, upgrading drivers skill and behavior both technically and with respect to keeping rules should get due consideration. Girma (2003) had also tried to develop accident predictive mathematical model based on the data collected on arterial roads in Addis Ababa. He use Poisson and Negative Binomial Regression methods to relate the discrete accidents data with the road and traffic flow explanatory variables. The result of the research showed that the existing inadequate road infrastructure and poor road traffic operations are the potential contributors of this ever-growing challenge of the road transport in Addis Ababa. The results also indicate that improvements in roadway width, pedestrian facilities, and access management are effective in reducing road traffic accidents. Wondwossen (1999) had also tried to conduct a study on car accidents in Addis Ababa. He used Chi-square Test and Logistic Regression Analysis in identifying correlates of car accidents. The research indicates that in addition to other variables like light condition of the road, 'not giving priority to pedestrians' has the highest association with physical damage. Taddele and Larson (1991) on their study found out that drivers involved in motor vehicle injuries are more likely to be male, young and less experienced. With respect to vehicle characteristics associated with involvement in motor vehicle injuries, elevated odds ratios were found for newer government owned vehicles , for taxis and buses. Tibebe (2005) on his part studied rule mining and classification of road traffic accidents using Adaptive Regression Trees. He used data mining techniques and tools to generate some rules using decision trees.

As it is visible from the above discussion the unavailability of comprehensive research, and from the magnitude of the problem makes the need for some advanced technology in constructing model that shows the interdependency of the variables which cause the traffic accidents is unquestionable.

After clear understanding of the domain area and the Bayesian network technology the researcher tried to build the model that will be helpful in making sound and informed decision in the effort of saving life., the next chapter focuses on the actual work of the research in attaining the objectives stated earlier.

CHAPTER FOUR

EXPERIMENTATION

This chapter explains the detail steps that are carried out in preparing the data for analysis and the experimentation.

4.1 Overview

In order to get a good result from Bayesian networking tools and techniques, the data needed to be cleaned and organized. Data collections, data cleaning, attribute selection, data formatting and transformation, and dimensionality reduction are the most important activities under data preparation process.

4.2 Data collection

Collecting, analyzing and understanding the content and structure of the data available is one of the most crucial activities that need to be carried out in order to properly utilize the Bayesian tools.

The raw data was collected from Addis Ababa traffic office from the office's flat database and from daily accident record file. Careful analysis of the data was done together with the domain experts by evaluating the relationship of the data with the problem at hand and the particular research task to be performed. This activity was very much helpful in understanding the data and viewing it from different angle to explore its potential.

The attributes and their description in the initial data source are listed in the following table.

No	**Attribute Name**	**Data Type**	**Description**
1	Accident Date	Date/Time	Date of an accident
2	Accident Day	Text	The day of an accident in week days
3	Accident Time	Date/Time	The time of the accident

4	Accident Id	Number	A number to identify a given accident uniquely
5	Driver Name	Text	Name of the driver who cause an accident
6	Driver FName	Text	Driver's Father name
7	Driver Age	Number	Age of the driver
8	Driver Sex	Text	Gender of the driver
9	Driver Edu_Level	Text	Educational level of driver
10	Driver License Grade	Text	Status of the driver license
11	Driver-Vehic-Rel	Text	The relation of the driver and the vehicle
12	Driver Experience	Number	Driving experience of the driver
13	Vehicle Type	Number	The type of the vehicle
14	Vehicle Age	Text	Service year of the vehicle
15	Vehicle possession	Text	Possession of the vehicle
16	Vehicle Defect	Text	Problem or defect of the vehicle
17	Accident Area	Text	The area where an accident occurred
18	Accident Road-Name	Text	The name of the road where an accident occurred
19	Road Joint	Text	The road segment separation
20	Road Condition	Text	Road surface condition at the time of accident
21	Light condition	Text	Light condition at the time of accident
22	Weather condition	Text	The weather condition at the time of accident
23	Category	Text	The owner category of the vehicle
24	Profession	Text	The profession of the driver
25	Health condition	Text	The health condition of the driver
26	Number of victims	Number	The number of victims in the accident
27	Plate number	Number	The plate number of the vehicle
28	Estimated cost	Number	The estimated damage cost due to the accident
29	Cause of accident	Text	The first seen cause of accident
30	Type of accident	Text	The type of accident happened

Table 1 attributes with their description and data type

4.3 Data pre-processing

After the data was collected the pre-processing step was conducted. The important activities carried out in this step are discussed as follows.

4.3.1 Data Cleaning

In order to get the required output, the necessary relevant input data need to be feed to the Bayesian tool. In line with this depth exploration of the data and frequent consultation with the domain experts carried out. Accordingly, correcting errors like filling missing values removing records with different vague values are carried out. In this regard missing values of nominal variables, like ‘sex’, are filled based on the idea of observing neighboring record values. On the other hand, missing values for the numerical attributes like, driver age and driver experience are filled by using the average value of the corresponding attribute values. Removing records with incomplete and /or invalid data or missing values under each column is also another data cleaning process. Consequently the size of the dataset was reduced to 1956 records.

4.3.2 Attribute/Variable selection

The other activity that was conducted under data pre-processing was variable selection.

Here unwanted and irrelevant variables were removed because some of these data have no significant importance for prediction and model construction of accident severity and some of them did not have complete and reliable values. Those attributes which are not selected and which are selected are shown in the following tables consecutively.

No	Attribute	Data type	Description
1	Driver name	Text	The name of the driver
2	Driver Fname	Text	Driver’s Father name
3	Category	Text	Owner category of the vehicle
4	Road Condition	Text	Road surface condition at the time of accident

5	Accident day	Text	The day at which the accident happened
6	Profession	Text	The profession of the driver
7	Health condition	Text	The health condition of the driver
8	Number of victims	Number	The number of victims in the accident
9	Plate number	Number	The plate number of the vehicle
10	Estimated cost	Number	The estimated damage cost due to the accident
11	Light condition	Text	Light condition at the time of accident
12	Accident Road-Name	Text	The name of the road where an accident occurred
13	Accident Date	Date/Time	Date of an accident
14	Accident Id	Number	A number to identify a given accident uniquely
15	Vehicle possession	Text	Possession the vehicle
16	Vehicle Age	Text	Service year of the vehicle
17	Accident Road-Name	Text	The name of the road where an accident occurred
18	Accident Area	Text	The area where an accident occurred
19	Vehicle Defect	Text	Problem or defect of the vehicle

Table 2 attributes which are not selected to build the model

No	Attribute Name	Type	Description
1	Time	Number	The time of the accident
2	Driver Sex	Text	Gender of the driver
3	Driver Age	Number	Age of the driver
4	License Grade	Number	Status of the driver license
5	Driver Educational Status	Text	Educational level of driver
6	Driver Experience	Number	Driving experience of the driver
7	Type Of Vehicle	Text	The type of the vehicle

8	Cause Of Accident	Text	The first seen cause of accident
9	Road Joint	Text	The type of the road where the accident happen
10	Weather Condition	Text	The weather condition at the time of accident
11	Type Of Accident	Text	The type of accident happen

Table 3 attributes selected to build the model

4.3.3 Data transformation

Some of the attributes were with values which contradict with the input criteria of the Bayesian network tool. Therefore from the row data of the Addis Ababa traffic some attributes were derived and their values are reduced to manageable size in order to get results that can be easily interpretable. The following table summarizes the transformed attributes.

No	Derived Attributes	Values
1	Time	M A E N
2	Driver Experience	A-F
3	Driver Age	18-30 31-50 50 and above Unknown
4	Type of Accident	C P S

Table 4 derived attributes with their values

4.4 Building the Bayesian network

In order to build the network structure and to find the prediction accuracy the Bayesian network Power soft tool and the Three Phase Dependency Analysis (TPDA) algorithms was used. The package in Belief Network PowerSoft includes three applications: BN PowerConstructor, BN PowerPredictor, and Data PreProcessor.

Data PreProcessor is the tool to be used with PowerConstructor and PowerPredictor for pre-processing the training data. It is used for discretize data, convert other data formats to *.MDB (MS-Access) format, and partition the training data into internal training data and internal test data.

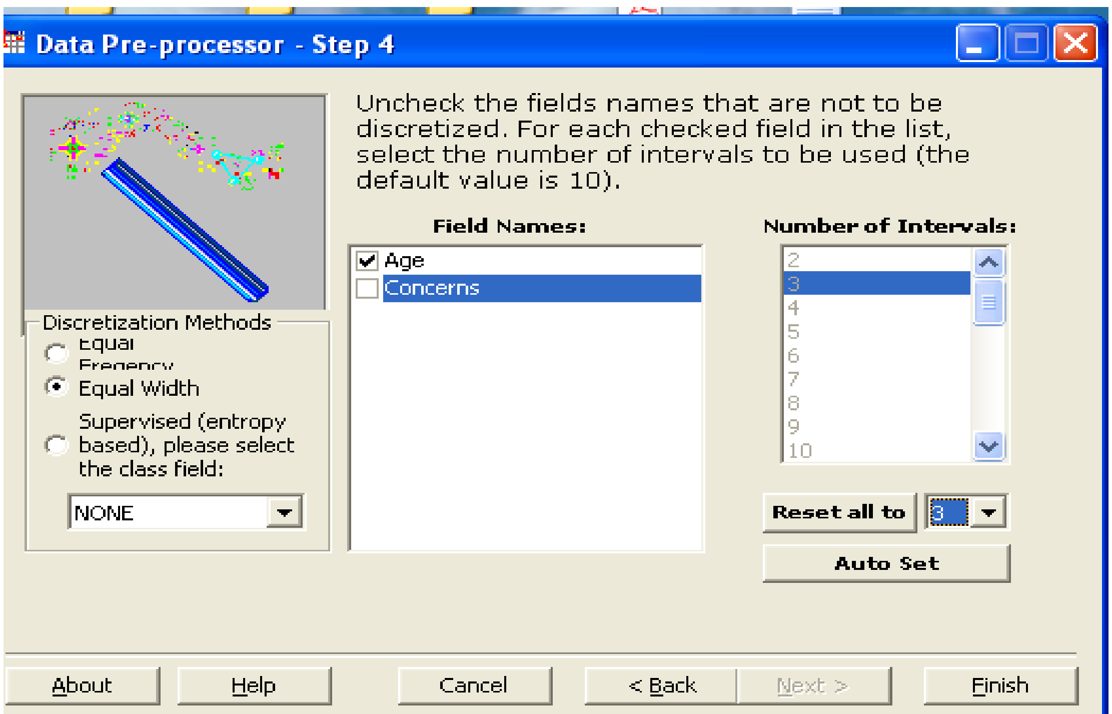

Figure 2: The data preprocessor of the BN powersoft

PowerConstructor is an efficient system that learns Bayesian belief network structures & parameters from data. Which is designed based on three-phase belief network (BN) construction algorithms.

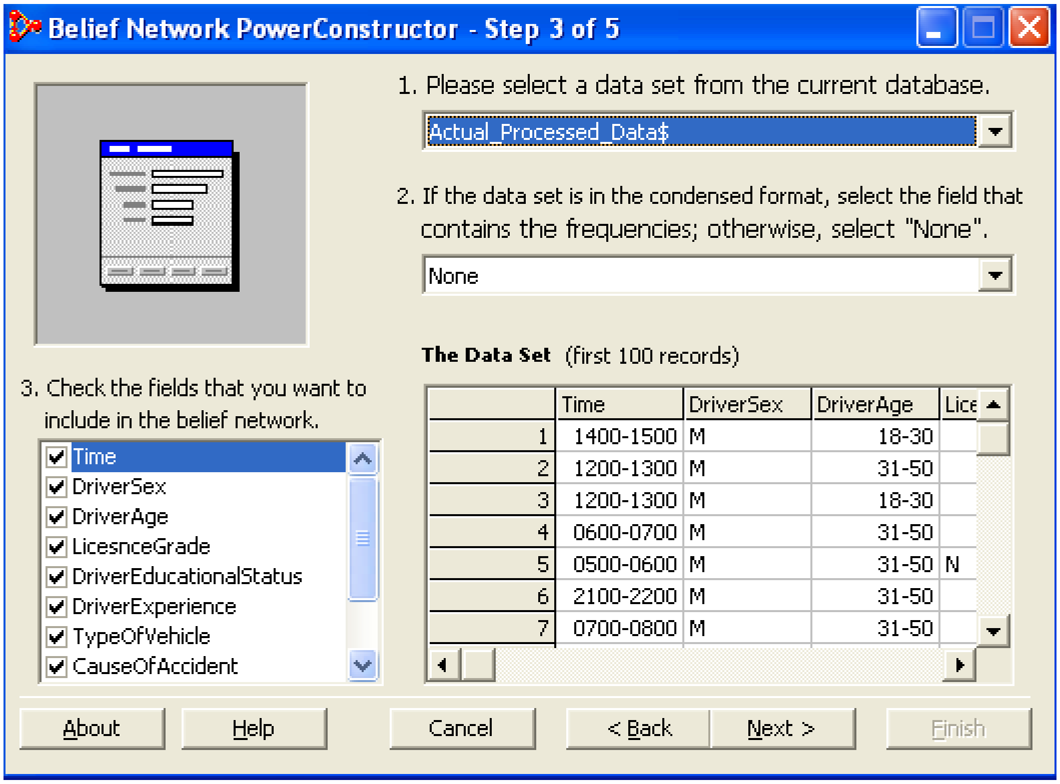

Figure 3: The BN power constructor of the BN powersoft

PowerPredictor is a system and which is an extension of our BN learning system (BN PowerConstructor) to BN based classifier learning and using. It can learn general Bayesian network classifiers and Bayes multi-net classifiers from training data and use these classifiers to classify new data. The system can also perform feature subset selection automatically.

After the training data sets are processed by the data preprocessor, BN Power Constructor and BN Power Predictor were employed for model building and accuracy prediction purpose respectively.

Experiments were carried out ten times by splitting the data into ten partitions (10-fold cross validation). The following sections explain in detail about the experimentation and the results found.

4.4.1 Training and testing in experiment one

The experiment was carried after the dataset is divided using ten-fold cross validation technique. 1760 dataset were used for training and 195 data set were used for testing each experiment. The training and testing experiment was conducted ten times. The best prediction accuracy result is shown in the following sub section and the rest prediction accuracy results are shown in the appendix part.

4.4.1.1 Prediction Accuracy of experiment one

After the training data set is preprocessed and "type of accident" is considered as target class the prediction accuracy test is carried out. Among the ten experiments conducted the best result was found on the fourth training and its confusion matrix is shown below.

		Predicted		
Actual		Crash	Property	Severe
	Crash	**2**	51	87
	Property	0	**1399**	8
	Severe	0	66	**148**

Table 5 Confusion Matrix for the fourth Partition

The confusion matrix shown above can be read horizontally and vertically in the following way.

<u>Reading the table horizontally</u>

Among 140 actual records in class crash:-

- 2 of these records are correctly classified into class crash.
- 51 of these records are wrongly classified into class property and
- 87 of these records are wrongly classified into class severe.

Among 1407 actual records in class property:-

- 0 of these records are wrongly classified into class crash.
- 1399 of these records are correctly classified into class property.
- 8 of these records are wrongly classified into class severe.

Among 214 actual records in class severe:-

- 0 of these records are wrongly classified in to class crash.

- 66 of these records are wrongly classified into class property.
- 148 of these records are correctly classified into class severe.

Reading the table vertically

Among 2 records classified into class crash:

- 2 of these records are correctly classified into class crash.
- 0 of these records are wrongly classified into class crash.

Among 1516 records classified into class property:

- 117 of these records are wrongly classified into class property.
- 1399 of these records are correctly classified into class property.

Among 243 records classified into class severe:

- 95 of these records are wrongly classified into class severe
- 148 of these records are correctly classified into class severe.

Generally there are about 1548 of the records were classified correctly and only 212 records are misclassified. The estimate prediction accuracy was 87.96% at 95% confidence level.

The summary of the prediction accuracy for each partition is show in the following table.

Partition #	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Prediction accuracy (%)	87.73	87.68	87.22	87.96	87.27	87.56	87.39	87.84	87.33	87.74

Table 6 summary of prediction accuracy results for the first experiment

The average estimation prediction accuracy of the first experiment is **87.57%.**

4.4.1.2 Constructing BN model for Best learned model

After the dataset was stored and ready in the database, constructing the BN Model of the causes of traffic accidents was done using the poweConstructor for the best learned model without the eliciting process with the domain experts.

The best learned model is shown below.

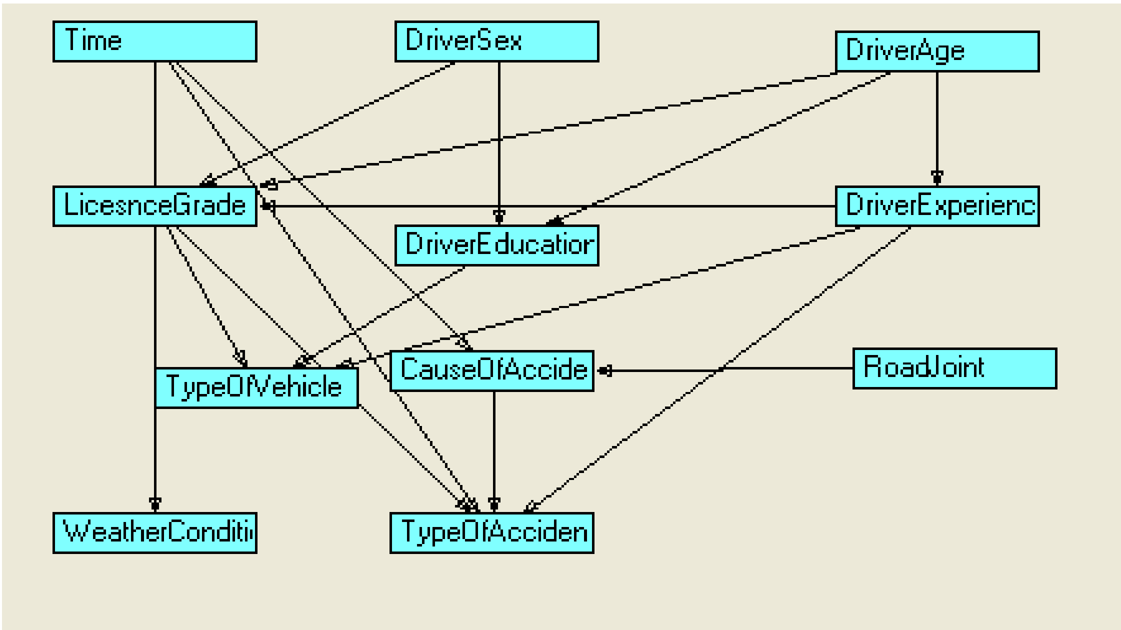

Figure 4: The network structure for best learned model

We can see the following dependency between variables in the first model.

- Type of accident is directly influenced by four factors, which are license grade, time, and cause of accident and driver experience.
- Driver age directly affecting license grade, driver education and driver experience but its direct influence on cause of accident is not seen.
- Road joint affects only cause of accident whereas it is believed by the domain expert that it can also affect type of accident.
- Driver sex was affecting driver education and license grade but it was highly believed by the domain experts that it can also affect cause of accident (directly influence cause of accident)

- Weather condition was not affecting any variables whereas the experts do believe that it has a direct effect on cause of accidents and type of accidents.
- Type of vehicle was not affecting any node but the domain experts believe that it has direct influence on type of accident and cause of accident, and so on.

4.4.2 Training and testing in experiment two

Experiment two was carried out after elicitation of the domain experts with the same dataset in which the ten-fold cross validation technique was applied. Here 1760 dataset were used for training , 195 data set were used for each experiment and the training and testing experiment was conducted ten times. The following sections explain details of the experimentation results.

4.4.2.1 Constructing BN model after eliciting the domain expert

To avoid some of the limitations in secondary data, the second experiment was conducted after eliciting the domain experts. Here the domain experts identify the contributing factors for the traffic accidents and suggest the interdependency between these major influencing factors. Rearranging the nodes based on the domain expert idea is conducted accordingly.

The network diagram after elicitation is shown below.

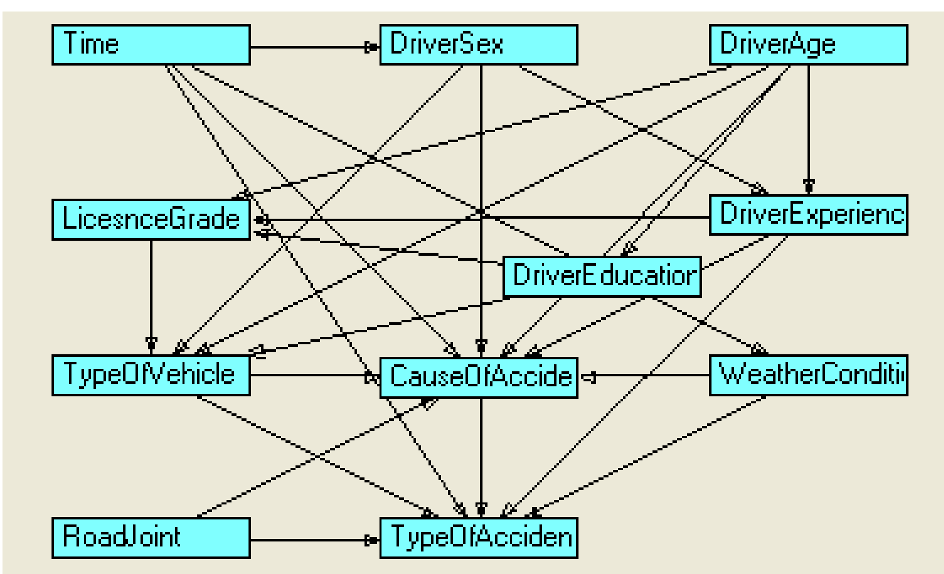

Figure 5: The network structure after elicitation of domain expert

The major changes made in the above network diagram as it is compared to the best learned networks are the following:

- Type of accident is highly influenced by weather condition, road joints and type of vehicle. Such dependency relation is not shown in the best train model of the first experiment
- In the best learnt network, we have driver age directly affecting license grade, driver education and driver experience, whereas in the second network diagram it also affects type of vehicle, cause of accident.
- Road joint in the previous model was seen to affect only cause of accident whereas in the second model, it was seen to affect the type of accident.
- Driver sex was affecting driver education and license grade in the first model, but it was seen as it can also affect cause of accident.
- Weather condition was not affecting any variables in the first model whereas in the second model weather condition affects cause of accidents and type of accidents.
- Driver education affects only type of vehicle in the first model but in the second model it also affects the license grade.
- Type of vehicle was not affecting any node in the first model, but in the second model it directly influences the type of accident and cause of accident.
- In addition to this driver age, weather condition, type of vehicle, driver experience and driver sex are also seen as a direct influence factors for causes for traffic accidents in the second model.

4.4.2.2 Prediction Accuracy of experiment two

The prediction of accuracy in experiment two was conducted after the elicitation of the domain expert is done. The experiment was done with 10-fold cross validation.

The following table shows the estimated prediction accuracy of the best result that was found in the first training dataset. The remaining prediction accuracy results of experiment two are shown in the appendix part.

	Predicted			
		Crash	Property	Severe
Actual	Crash	0	148	0
	Property	0	1413	0
	Severe	0	199	0

Table 7 summary of prediction accuracy results for the second experiment

The confusion matrix shown above is read horizontally and vertically in the following way.

Reading the table horizontally

Among 148 actual records in class crash:-

- 0 of these records are correctly classified into class crash.
- 148 of these records are wrongly classified into class property and
- 0 of these records are wrongly classified into class severe.

Among 1413 actual records in class property:-

- 0 of these records are wrongly classified into class crash.
- 1413 of these records are correctly classified into class property.
- 0 of these records are wrongly classified into class severe.

Among 199 actual records in class severe:-

- 0 of these records are wrongly classified in to class crash.

- 199 of these records are wrongly classified into class property.
- 0 of these records are correctly classified into class severe.

<u>Reading the table vertically</u>

Among 0 records classified into class crash:

- 0 of these records are correctly classified into class crash.
- 0 of these records are wrongly classified into class crash.

Among 1760 records classified into class property:

- 347 of these records are wrongly classified into class property.
- 1413 of these records are correctly classified into class property.

Among 0 records classified into class severe:

- 0 of these records are wrongly classified into class severe
- 0 of these records are correctly classified into class severe.

Generally among 1760 cases of the test dataset, 1413 cases were classified correctly and only 347 records are misclassified. The estimate prediction accuracy is 80.28 at 95% confidence level. The summary of the prediction accuracy of experiment two for each partition is show in the following table.

Partition #	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Prediction accuracy (%)	80.28	79.50	79.09	79.90	79.03	79.43	79.77	80.11	79.32	79.98

Table 8 summary of prediction accuracy results for the second experiment

The average estimation prediction accuracy of the second experiment is 79.64%. There is 7.93% prediction accuracy difference between the first and the second average prediction accuracy results. This shows that Bayesian network can perform in a promising way and it can also be used for real life decision making purposes. In addition, if it is applied in situation where the data has complete attributes and well documented, we can find more reliable and accurate results.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

A significant amount of information was reviewed and summarized for the creation of the model that has been built in the previous chapter. This chapter includes more general conclusions about the experiment that has been conducted. Recommendations are also provided as a response to each of the conclusions.

5.1 CONCLUSIONS

Bayesian networks (belief networks) are gaining an increasing popularity as modeling and efficient predictive tools for complex problems nowadays. In this research work , since road traffic accidents are probabilistic , unpredictable , complex and uncertain , it is realized that Bayesian networks can be fruitfully applied in this problem as it is compared with other well-known statistical methods. In this research work, the researcher finds out that many traffic laws exist solely to govern the safe movement of traffic on the roads; stop signs, speed limits, etc…, and may concur with an accident factor that existed. For the police officer, it is convenient to match a law violation with a "cause" but this practice often results in errors as it tends to ignore other existing factors that may have been present and were equally contributing to an accident. Therefore Identifying these factors that are traditional thought to be causes of traffic accident through research based work and to do scientific analysis on identifying the interdependency between these factors, and analyzing whether these factors are common law violation or legally relevant proximate cause of an accident was the major task and contribution of this research work. Based on this fact a valid model of knowledge for road accident prediction and variable relationship had been extracted.

The experimentation part of this research was conducted using the belief network powersoft software in which the BN power constructor was employed in developing the model and the BN power predictor was also employed for the purpose of evaluating the predication accuracy of the model. The three phase dependency analysis algorithm was the one used in the process. The experiment was conducted by adopting the ten-fold cross validation technique after and before the elicitation of the domain experts.

The best prediction accuracy found in the first experiment (before the elicitation of the domain expert) was 87.96% and the average of all the test datasets in the first experiment was87.57%.

According to the first experiment type of accident is directly influenced by four factors, which are license grade, time, and cause of accident and driver experience.

In the second experiment (after elicitation of the domain expert) the best prediction accuracy was 80.28% and the average predictive accuracy of all the ten test dataset results was 79.64%.

According to the second experiment the type of accident is highly influenced by weather condition, road joints and type of vehicle. Such dependency relation is not shown in the best train model of the first experiment. In addition to this driver age, weather condition, type of vehicle, driver experience and driver sex are also seen as a direct influence factors for causes for traffic accidents.

In general, the estimated Bayesian network can be regarded as a compact and structured representation of the given database of car accidents. The results shown here are encouraging and point to possible directions for improvement, such as including more variables and larger datasets that cover more years. Extending the Bayesian network (with good performance results) into a decision network is another possibility.

More importantly, the significant road traffic accident predictive models developed in this study are applicable in road safety improvements and could serve as a basis for further research works in Ethiopia.

5.2 RECOMMENDATIONS

Through this research work, an attempt has been made to find out the potential applicability of Bayesian network to support the traffic control activity at Addis Ababa city abased on the data accumulated on accidents. Based on the findings, the researcher would like to make the following recommendations.

- The research attempts to assess the applicability of Bayesian belief network to support the identification of interrelatedness between the determinant factors for the causes of traffic accidents using some set of variables that were considered important by experts. However it has to be open for investigation in order to consider the interdependency of other necessary variables to build a model with better performance and accuracy than the model built in this research work. So the accident record's data base need to be revised in order to incorporate as many attributes as possible to create a more comprehensive database.
- For efficient data analysis process, availability of electronic data is very crucial. Even though there is a promising start, Addis Ababa traffic office and the Addis Ababa road authority should exert their effort to store their records in an electronic format and make all decisions based on the collected records.
- Bayesian belief network could contribute a lot in identifying potential interdependencies between the attributes that are the possible causes of traffic accidents which will result in fatal and serious injuries as well as property damages. Thus, it could be more important

to use Bayesian belief network techniques and tools for the decision making process especially in revising the road traffic rules and regulations and creating awareness for drivers and pedestrians.

- Further research works by extending this research should be made on the same dataset by including the specific inference results (including the inference process) and one or more techniques and tools in order to get a better prediction performance.

Reference:

1. Abebe Asrat (2004). *Basic points requiring close attention on road safety in Ethiopia. Unpublished document. Addis Ababa*
2. Addis Ababa transport authority (2004). *Addis transport. Biannual magazine.1 (1). Addis Ababa*
3. Bayes, T. (1763). *An Essay towards Solving a Problem in the Doctrine of Chances.*
4. Berhanu, G., (2000) *Effects of Road and Traffic Factors on Road Safety in Ethiopia.*
5. Belief Network PowerConstructor: *documentation* available at http://www.*http://www.cs.ualberta.ca/~jcheng/bnpc.htm*
6. Charles M. Grinstead, (1993) *introduction to probability*
7. cheol kim (2001), *using Bayesian Networks to Analyze Medical Data*
8. David Heckerman ,(1997), *Bayesian network for data mining*
9. Finn V. Jensen (1996), *Introduction to Bayesian Networks*
10. Friedman, N., Geiger, D. & Goldszmidt, M. (1997). *Bayesian Network classifiers,*
11. Geiger, D., Verma, T. & Pearl, J. (1990). *Identifying independence in Bayesian Networks*
12. Giacomo Gamberoni1,(2001*) Bayesian Networks Learning for Gene Expression Datasets*
13. Girma Berhanu (2003*) Models relating traffic safety with road environment and traffic flows on arterial roads in Addis Ababa*
14. Han, J., Kamber, M (2001). *Data Mining*: *Concepts and Techniques*
15. Honglak Lee, *Andrew Y.* (2000) *A Dynamic Bayesian Network Mode*
16. Houkuan Huang (2005), *An Improved Bayesian Network Learning Algorithm Based on Dependency Analysis*
17. H. Wong a , P. Harris (2003), *Artificial Intelligence In Medicine*

18. Heckerman, D. (1986*). Probabilistic interpretations for MYCIN's certainty factors, Uncertainty in Artificial Intelligence*

19. Heckerman, D. (1993). *Causal Independence for Knowledge Acquisition and Inference*

20. Heckerman, D. & Breese, J. S. (1994). *A New Look at Causal Independence*

21. Heckerman, D. (1995). *A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06. Microsoft Research.*

22. Heckerman, et al., (2000). *Dependency Networks for Inference,Collaborative Filtering, and Data Visualization. Journal of Machine Learning Research, 1: p. 49-75.*

23. J. Polpinij, C. Jaruskulchai & P. Haddawy (2002) *A Bayesian Classification Approach to Personal Automotive Marketing Analysis.*

24. James Orr *(2002), Bayesian Networks and Data Mining.*

25. Jensen, F. V., Lauritzen, S. L. & Olesen, K. G. (1990*). Bayesian updating in causal probabilistic networks by local computations*

26. *Journal* of *Transportation* Engineering (2002) LR 1057, *Transport* and *Road* Research Laboratory, Crowthorne, U.K.; Paterson, W. (1987).

27. Jonathan Kelly (2001) *Learning Bayesian Networks from Data: An Information-Theory Based Approach*

28. Judea Pearl (2000) *Causalility, Models Reasoning and Inference*

29. Judea Pearl (2000) *Bayesian networks*

30. Kohavi, Ron (1995). *"A study of cross-validation and bootstrap for accuracy estimation and model selection".*

31. Laura E. Brown (2001) *The max-min hill-climbing Bayesian network structure learning algorithm*

32. Lauritzen, S. L. (1995). *The EM algorithm for graphical association models with missing data,*

33. Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H.-G. (1990*). Independence properties of directed Markov fields, Networks*

34. Lugg, J.A., (1995) *Introduction to Causal Modeling, Bayesian Theory and Major Bayesian Modeling Tools*

35. Martin Neil (2005) *A new Bayesian Network approach to Reliability modeling*

36. Norvig W.J (2003), *Real time assessment of drinking water systems using a dynamic Bayesian network*

37. Olivier Pourret (2006) *Bayesian Networks: A Practical Guide to Applications*

38. Pearl (1988), *Probabilistic Reasoning in Intelligent Systems*

39. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* available at http://citeseer.ist.psu.edu/kohavi95study.html. (Morgan Kaufmann, San Mateo)

40. Rahel Bekele, (2005). *Computer-Assisted Learner Group Formation Based on Personality Traits*

41. Ricardo Gutierrez-Osuna (2001). *learning for classification*

42. Richard E. Neapolitan (2002) *learning Bayesian networks*

43. Russell and Gregory (1993*) Symbolic and Quantitative Approaches to Reasoning and Uncertainty*

44. Spiegelhalter, D. & Lauritzen, S. L. (1990). *Sequential updating of conditional probabilities on directed graphical structures*

45. Taddele , d. and larson , c.p (1991) *the occurrence and driver characteristics associated with motor vehicle injuries in Addis Ababa*

46. Tim Huang, Gary Ogasawara, and Stuart Russell (1993) *Symbolic Traffic Scene Analysis Using Belief Networks..*

47. Tom M. Mitchell (2006) *Bayesian Network Learning with Parameter Constraints*

48. Uffe B. Kjærulff • Anders L. Madsen (2008) *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*

49. UNECE, (2006).*Improving global road safety*

50. Wermuth & Lauritzen 1983, Verma & Pearl 1992, Spirtes et al. 2000 *Graphical and recursive models for contingency table*

51. WHO (2004), *World report on road traffic injury prevention. Switzerland. Geneva*

52. Wondwossen Mulugeta (1999) *correlates of car traffic accident: the case of Addis Ababa*

Appendix

Confusion matrix for the first experiment

Observed		Predicted		
		Crash	Property	Severe
	Crash	1	57	90
	Property	0	1405	8
	Severe	0	61	138

First partition with 87.73 prediction accuracy

observed		Predicted		
		Crash	Property	Severe
	Crash	1	55	89
	Property	0	1395	5
	Severe	0	68	148

Second partition with 87.68 prediction accuracy

observed		Predicted		
		Crash	Property	Severe
	Crash	2	54	95
	Property	0	1385	7
	Severe	0	69	148

Third partition with 87.22 prediction accuracy

observed		Predicted		
		Crash	Property	Severe
	Crash	2	56	92
	Property	0	1384	7
	Severe	0	69	150

Fifth partition with 87.27 prediction accuracy

observed		Predicted		
		Crash	Property	Severe
	Crash	2	53	90
	Property	0	1390	8
	Severe	0	68	149

Sixth partition with 87.56 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	2	59	89
Property	0	1396	8
Severe	0	66	140

Seventh partition with 87.39 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	2	54	83
Property	0	1402	8
Severe	0	69	142

Eighth partition with 87.84 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	2	55	92
Property	0	1389	7
Severe	0	146	146

Ninth partition with 87.33 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	2	55	93
Property	0	1396	6
Severe	0	61	140

Tenth partition with 87.74 prediction accuracy

Confusion matrix for the second experiment

observed	Predicted		
	Crash	Property	Severe
Crash	0	143	0
Property	0	1400	0
Severe	0	216	0

Second partition with 79.50 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	0	151	0
Property	0	1392	0
Severe	0	271	0

Third partition with 79.09 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	0	140	0
Property	0	1407	0
Severe	0	214	0

Fourth partition with 79.90 prediction accuracy

observed	Predicted		
	Crash	Property	Severe
Crash	0	150	0
Property	0	1391	0
Severe	0	219	0

Fifth partition with 79.03 prediction accuracy

Observed		Predicted		
		Crash	Property	Severe
	Crash	0	150	0
	Property	0	1404	0
	Severe	0	206	0

Sixth partition with 79.43 prediction accuracy

observed		Predicted		
		Crash	Property	Severe
	Crash	0	139	0
	Property	0	1410	0
	Severe	0	211	0

Seventh partition with 79.77 prediction accuracy

Observed		Predicted		
		Crash	Property	Severe
	Crash	0	149	0
	Property	0	1396	0
	Severe	0	215	0

Eighth partition with 80.11 prediction accuracy

Observed		Predicted		
		Crash	Property	Severe
	Crash	0	150	0
	Property	0	1402	0
	Severe	0	201	0

Ninth partition with 79.32 prediction accuracy

observed		Predicted		
		Crash	Property	Severe
	Crash	0	145	0
	Property	0	1398	0
	Severe	0	217	0

Tenth partition with 79.98 prediction accuracy

Author's Declaration

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

Alemayehu Tabor Feyissa
January 2009

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

Dr. Rahel Bekele
January 2009