Using Data Mining Techniques to Analyze Crash Patterns in Sri Lanka Road Accident Data

U L A S Perera
158768U

Faculty of Information Technology
University of Moratuwa
February 2019

Using Data Mining Techniques to Analyze Crash Patterns in Sri Lanka Road Accident Data

U L A S Perera
158768U

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology

February 2019

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information delivered form the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of the Student	Signature of the student
U. L. A. S. Perera	....
	Date:

Supervised by

Name of Supervisor	Signature of Supervisor
S. C. Premaratne	....
	Date:

Acknowledgements

I would like to express sincere gratitude to my project supervisor, Mr. S. C. Premaratne, Senior Lecturer in University of Moratuwa, who spent his valuable time for guiding this research to make it a success. I would also like to thank Prof. Asoka Karunananda and Dr. M. F. M. Firdhous who are lecturers in charge of literature review, research methodology and thesis writing subjects which were the basis for this research. Not only that my thanks should go to all the lecturers in MSc in Information Technology, all the facilitators and batch mates of MSc IT 2015/2016 batch.

Moreover a very special thanks should go to DIG Traffic Administration, Mr. Hemantha and his staff members of Sri Lanka Police for assisting me in collecting dataset and giving me their valuable comments on my research.

Finally I wish to thank my wife Kumudu for her understanding and guidance throughout this project and my two sons, Venura and Sandaru for their love and support.

Abstract

The road safety has been identified as a major factor that influences the sustainable development worldwide. This growing interest in road safety, is reflected by including it in Sustainable Development Goals of United Nations as “Halve the number of global deaths and injuries from road traffic accidents by 2020”. According to road accident statistics published by Sri Lanka traffic police in 2015, every three and half hours a person is killed due to a road accident and two are seriously injured. This shows that travelling on local roads becoming more and more unsafe and risky. When improving the road safety conditions, it is necessary to identify the major factors contributing to road crash injuries and deaths, in order to take appropriate safety measures.

The Sri Lanka Police department uses MAAP (Microcomputer Accident Analysis Package) system for the storage and analysis of Road Traffic Accidents (RTA) data. However MAAP has its own limitations of analysis of accident data.

In the area of road traffic accident analysis, data mining technique has been recognize as a reliable technique which can be used beyond the conventional techniques. When analyzing road traffic accidents, different models were developed to identify factors affecting the severity of a traffic accident.

The objectives of this study are to explore the underlying factors influencing on injury severity, to identify the human, environment and vehicle factors influencing the road traffic accident severity and to identify crash proneness of road segments using available road and crash factors. In this study, data mining classification model is used to detect factors which influence on road accidents. We conducted an experiment with road accident data in 2015, provided by Sri Lanka Police.

In this research we proposed an accident severity model based on selected data mining techniques to identify influential factors for the severity of road traffic accidents. The solution model is developed using Weka software tool.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The road safety has been identified as a major factor that influences sustainable development worldwide. The World Health Organization (WHO) reveals that More than 1.2 million people are affected by road traffic accidents globally each year and it is a major public health concern, also it has a huge impact on the sustainable development of the country. Due to growing recognition of road safety to health, development of a country and broader environmental objectives [1], the United Nations (UN) has included, the road safety concerns “to halves the road traffic deaths and injuries by 2020” and “to provide access to safe, affordable, accessible and sustainable transport systems for all by 2030”, in their sustainable development goals(SDGs).

According to road accident statistics published by Sri Lanka traffic police in 2015, every three and half hours a person is killed due to a road accident and two are seriously injured. This shows that travelling on local roads becoming more and more unsafe and risky [2].

1.2 Background and Motivation

The Sri Lanka Police department uses MAAP (Microcomputer Accident Analysis Package) Software to store the Road Accident Data (RTA), which is also capable of analyzing road accidents. However MAAP has its own limitations of analysis of accident data. It uses cross-tabulation analysis and conventional statistical methods to analyze road accidents. Police department also uses Excel Pivot tables to analyze RTA data

Due to the rapid growth in collecting and storing huge amounts of data in recent years, data mining has emerged as a multidisciplinary technique to analyze data. The prediction and pattern recognition can be identified as major applications of data mining in the field of data analysis [3].

Knowledge discovery form data or Data Mining can be defined as: “A nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [4] or “A novel technique to extract hidden and previously unknown information from the large amount of data” [5].

In the area of road traffic accident analysis, data mining technique has been recognized as reliable technique which can be used beyond the conventional techniques. When analyzing road traffic accidents, different models were developed based on different data mining techniques. Among them, most of the models were developed to identify factors affecting the severity of a traffic accident.

The application of Data Mining Techniques to RTA will develop a new insight to road accident analysis in Sri Lanka and will help to improve road safety.

1.3 Problem Definition

Based on the literature, the research problem of this project has been defined as identification of risk factors influence on road traffic accidents which contributes to road safety decision making in Sri Lanka.

1.4 Aim and Objectives

The aim of this project is to discover hidden crash patterns in road traffic accident data by applying data mining techniques. Based on the problem statement analysis, the following individual objectives have been identified for this research initiative:

- To explore the underlying factors influencing on injury severity
- To identify the human , environment and vehicle factors influencing the road traffic accident severity
- To identify crash severity of road segments based on available accident related factors

1.5 Proposed Solution

Data mining is a proven technique to discover the underlying factors of a road accident which contributes to the severity of the incident. In this research we proposed an accident severity model based on data mining techniques to identify the contributing factors to the severity of road traffic accidents.

The dataset obtain from the traffic headquarters of Sri Lanka Police is cleansed within the pre-processing stage by filling in missing values, smoothing the noisy data, and resolving the inconsistencies in the data. A relevant data sets with independent class variables are identified to explore the underlying factors influencing on injury and accident severity. Since the original dataset has large set of attributes, but only few of them are relevant next step is to reduce the dataset by attribute subset selection/feature selection. For the attribute/feature subset selection, the wrapper method of Weka is used with J48 classification algorithm. After that Random forest was applied to measure the variable importance. The intention was to identify the high impact variables individually from each of the datasets. Finally to further analyze the contributory factors that have an impact on traffic accident severity, FURIA a fuzzy rule based algorithm is used to generate rules.

1.6 Structure of Dissertation

The overall dissertation structure as follows. Chapter 1 gives an introduction to full project with objectives, background, problem and solution. Chapter 2 critically reviews the literature on road accident analysis and application of data mining techniques on road accident data. Chapter 3 is about details of data mining technology by showing its relevance to mining road traffic accident data. Chapter 4 presents the approach with users, input, output, process and features. Chapter 5 is the design of the data mining solution, while Chapter 6 is of implementation of the solution. Chapter 7 reports on the evaluation of the solution. Chapter 8 concludes the solution with a note on further works.

1.7 Summary

This chapter presented a brief introduction to the research study with specific reference to identification of research problem, aims and objectives of research initiative and proposed solution. The next chapter critically reviews the road traffic accident data analysis and usage of data mining techniques.

Chapter 2

Discovering Crash Patterns in Road Accident Data

2.1 Introduction

This chapter critically reviews the application of data mining techniques on road accident data to discover crash patterns. In this sense first we discuss the road safety and traffic accidents. Subsequently road accident analysis will be discussed. Then we identify unsolved issues and concerns in data mining in Road Traffic Accident (RTA). Finally we define our research problem to be addressed in this dissertation. This chapter also identifies the possible technologies which can be used to solve the problem. This chapter is organized under the headings of road safety and traffic accidents, road accident analysis, and data mining techniques in in RTA.

2.2 Road Safety and Traffic Accidents

The road safety is broader concept. *United Nations* (UN) recognizes it as "a five pillars of action plan which consist of safe roads, safe vehicles, people, post-crash care and efficient crash management" [6]. The road safety has been identified as a major factor that influences sustainable development worldwide. The World Health Organization reveals that More than 1.2 million people are affected by road traffic accidents globally each year and it is a major public health concern, also it has a huge impact on the sustainable development of the country. Due to growing recognition of road safety to health, development of a country and broader environmental objectives [1], UN has included, the road safety concerns "to reduce the road traffic deaths and injuries 50% by 2020" and "to provide access to safe, affordable, accessible and sustainable transport systems for all by 2030" in their sustainable development goals(SDGs) [1].

WHO defines the road accidents as "complex events occurring due to the interaction of human beings, vehicles, and the road environment", further it stresses that these interactions can occur at any time and any place, therefore it is essential to develop a strong insight into the instruments to find appropriate solutions [6] .

As reported by WHO the growing number of motor vehicles is one of major factor contributing to the increase in global road crash injuries. Ensuring appropriate road safety measures are also accompany this growth [7].

2.3 Road Accident Analysis

Road accident analysis requires the knowledge of the factors affecting them, as they are uncertain and unpredictable incidents. These data are defined by set of variables and most of them are of discrete nature. The heterogeneous nature of the accident data is another major problem arise when analyzing them [8] . As reported by literature, the major factors that influence road crashes can be broadly categorized into “Human”, “Road and Environment and “Vehicle” related factors.

According to the literature, when analyzing road accidents, basically two types of models have been developed. The frequency based models and severity based models. Frequency based model takes a particular segment of roadway and determines what factors are associated with a specific number of crashes occurring during a time period. The Poisson distribution is used which much more properly describes frequency data. Injury or crash severity models were developed to find out which factors affect the injury severity in crashes. Much of the analysis of crashes are based on severity [9].

Lee and et al [10] state that analysis of freeway crashes and relationship between crash involvement with traffic, geometric, and environmental factors, were widely based on statistical models. Among them, Linear and Poisson regression models were developed for the purpose of crash prediction models.

Chen and Jovanis [11] argue that certain problems can arise when analyzing large dimensional datasets using traditional statistical techniques due to sparse data in large contingency tables. Further, violation of model specific assumptions in statistical models can lead to some erroneous results.

The literature reveals that due to the limitations of statistical approach to road accident analysis, which tends to use of data mining approach to road accident analysis. Data mining approach can be applied to expose novel, implicit and hidden information from large volumes of data.

Therefore data mining approach can be identified as prominent approach with multiple techniques to analyze road accidents when dealing with large and heterogeneous accident datasets.

2.4 Data Mining Techniques in Road Traffic Accident

When considering the increasing ability of collecting and storing data in recent years, data mining emerged as a novel technique in data analysis. The prediction and pattern recognition are the widely used data mining techniques among others.

The application of data mining techniques to RTA analysis will develop new insight to road accident analysis in Sri Lanka and will help to improve road safety.

As stated earlier, it is necessary to study various factors affecting road accidents in the area of traffic accident analysis. The previous studies on road traffic accidents have adopted different data mining models with different scenario. It is a common practice to study the human, vehicle, road and environmental related factors connected with different accident incidents in any road traffic accident.

Data mining can be recognized as reliable technique to answer the problem of discovering the factors behind traffic accident and analyzing the severity level of it. To resolve the traffic accident severity problem, there are well established data mining techniques which could be used as optimal techniques, based on road traffic accident scenario [12].

Castro and Kim [13] describe, a factor assessment model based on classification technique. This study focused on behavior of different factors on injury risk. The study mainly concerned on technical factors instead of human factor. The factors with the greatest influence on car accidents are identified by using three data mining classification models. The severity of the road accident is predicted by applying the Bayesian network, J48 decision tree and Multi-Layer Perception. According to the results revealed, the frequent factors are light conditions, road type and vehicle maneuver, also the study exposed that the age of the vehicle and weather conditions were not significantly influence the degree of injury. Further it is difficult to predict the severity level more accurately [13].

Beshah and Hill [14] investigated the accident severity and contribution of road and environmental related factors, when predicting the severity. This research was conducted based on road accident data in Ethiopia and different predictive models. They examined the applicability data mining techniques to expose the relation between road related characteristics and accident severity in Ethiopia. Classification models were built using Decision tree, Naive Bayes, and K-nearest neighbor and finally a set of classification rules were extracted based on PART algorithm [14].

In contrast, Pakgohar *et el* [15] conducted a research concerning the human contribution to the accident severity of road crashes in Iran. The Logistic Regression (LR), Classification and Regression Trees (CART) are used to predict the severity of road accident. Moreover the CART and LR approaches were used to discover human contribution in accident severity. The degree of injury severity had a noticeable relationship with the factors “Driving License” and “Safety Belt”, also revealed that the factor “gender of drivers” had no significant relationship with it [15].

The key factors can be identified rapidly and efficiently and instructional methods can be provided to the road traffic accident prevention and reduction by studying on road traffic accident causes, also resultant personal casualty could be reduced greatly. Road safety management level can be improved effectively with the traffic data analysis methods. Finally, the accuracy level of defining the traffic accident severity using Decision tree, Random forest, ID3, Functional Tree, J48 and Naïve Bayes has concerned over different datasets. Among other techniques , the J48 recoded the higher accuracy level [16].

2.5 Summary of Challenges

The discussion in the previous sections has identified number of unsolved in RTA data mining in general and some selected RTA data mining challenges. The table 2.1 summarizes the challenges pertaining to RTA data mining. Among others, the challenges stated in the Table were selected by considering the technique of data mining.

Project	**Technology used**	**Algorithms**	**Findings**	**Influential factors**
[13] Y. Castro and Y. J. Kim,	Classification models	Bayesian J48 MLP	"Most accurate model is Bayes Network against J48 decision tree and MLP model"	"Road type, Light conditions, Vehicle maneuver, Propulsion code, Weather conditions, Age of vehicle, Road surface conditions"
[14] T. Beshah and S. Hill	Classification models	Decision tree Naive Bayes K-nearest neighbor		"Subcity, ParticularArea, RoadSeparation, RoadOrientation, RoadJunction, RoadSurfaceType, RoadSurfCondition,WeatherCondition ,LightCondition, AccidentSeverity"
[15] A. Pakgohar, R. S. Tabrizi, M. Khalili, and A. Esmaeili,	Classification and Regression	CART LR	"CART has higher accuracy than LR method"	"Driving License and Safety Belt"
[16] M. Singh and A. Kaur	Classification and Regression	Naive Bayes J48	"Reveals that value of Naive Bayes and J48 techniques are approximately same accuracy"	
[17] A. Tavakoli Kashani, A. Shariat-Mohaymany, and A. Ranjbari	Classification and Regression	CART		"Injury Severity, Gender, Age, Seat Belt, Cause Of crash, Collision type, Vehicle Type, Location type, Lighting conditions, Weather conditions, Road surface condition, Occurrence, Shoulder type, Shoulder Width"

Table 2-1 Summary of literature

2.6 Problem Definition

Sri Lanka Police uses MAAP system for recording and storing of road accident data. MAAP has its own limitations of analysis of accident data. It uses cross-tabulation and conventional statistical methods to analyze road accidents.

The existing RTA data can be fully utilized for policy level decision making on road safety. As the existing system is unable to perform extensive and detailed analysis on RTA data, it is further to uncover valuable information. Therefore the risk of road traffic injuries and deaths can be reduced based on these information.

Identification of risk factors of road traffic accident is a major step of road safety decision making. However it is yet to discover the significant factors affecting injury and accident severity.

Based on the literature, the research problem of this project has been defined as identification of risk factors influence on road traffic accidents which contributes to road safety decision making in Sri Lanka.

2.7 Aim and Objectives

The aim of this project is to discover hidden crash patterns in road traffic accident data by applying data mining techniques. Based on the problem statement analysis, the following individual objectives have been identified for this research initiative:

- To explore the underlying factors influencing on injury severity
- To identify the human , environment and vehicle factors influencing the road traffic accident severity
- To identify crash severity of road segments based on available accident related factors

2.8 Summary

This chapter presented a comprehensive critical review of using data mining techniques on road traffic accident data analysis with specific reference to identification of crash severity and influential factors. We reported, development in road accident analysis in a broad spectrum of disciplines and concerns in data mining in Road Traffic Accident (RTA) analysis. The next chapter describes the data mining technology in more detail which is selected to analyze the data.

Chapter 3

Technology Adapted

3.1 Introduction

Chapter 02 presented a comprehensive critical review of road accident data analysis and using datamining techniques over conventional techniques on RTA Data with specific reference to identification of crash and injury severity. This chapter presents the data mining technology in more detail which is selected to analyze the data.

3.2 Data Mining

Han *et al* [18] defines the data mining as "Data mining is a *process* of discovering interesting patterns and knowledge from *large* amounts of data" . For the purpose of decision making, data mining techniques can be used to discover hidden patterns and relationships on large volumes of data.

The Knowledge Discovery form Data (KDD) is another term used in the field of data mining. Although data mining and knowledge discovery in database are frequently treated as synonyms, KDD is much broader concept. Data mining can be considered as a single step of knowledge discovery process,

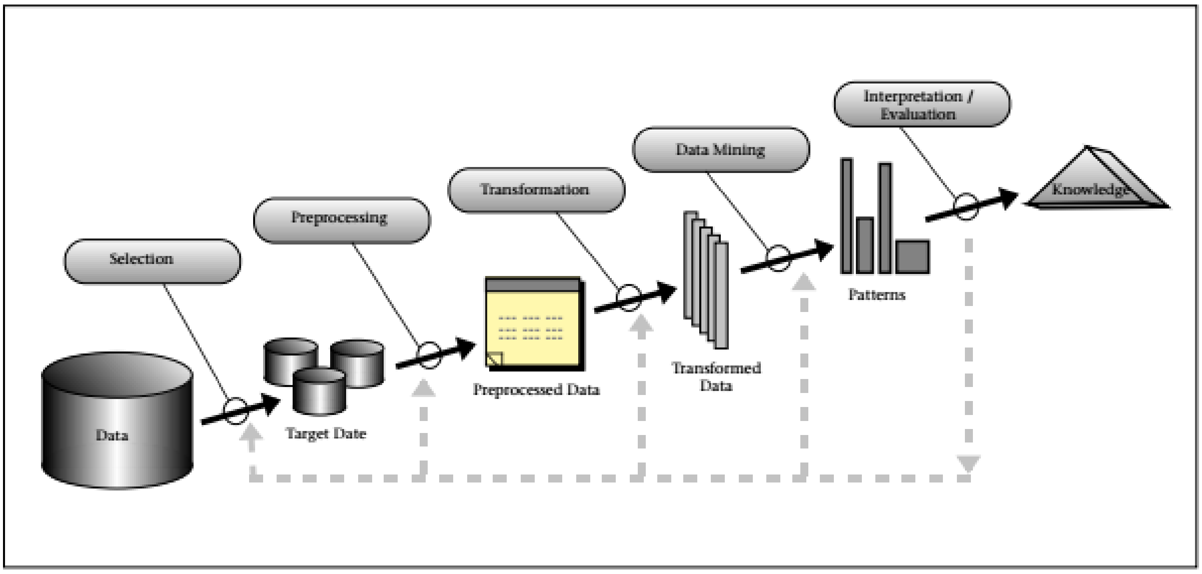


Figure 3-1 Steps of Knowledge Discovery Process

(Source: From Data Mining to Knowledge Discovery in Databases)

3.3 Steps of Knowledge Discovery Process

The Fayyad [19] Knowledge Discovery Process (KDP) model consists of nine steps, which are outlined as follows:

- **Developing and understanding the application domain.** Acquire the appropriate preceding knowledge.
- **Creating a target data set**. A target dataset is created by selecting a subset of variables and data points. The discovery task is performed based on the selected subset.
- **Data cleaning and preprocessing.** The data is cleansed to remove outliers and noise and missing data values are addressed.
- **Data reduction and projection.** The application of dimension reduction and transformation methods are performed to find out useful attributes and invariant representation of the data is examined.
- **Choosing the data mining task.** The specific data mining method for example classification, regression, clustering, etc.is matched with the goals defined in step 01
- **Choosing the data mining algorithm.** Methods are selected in order to search for patterns in the data and decisions are made on appropriate models and parameters of the methods.
- **Data mining.** A specific representational form such as such as classification rules, decision trees, regression models, trends, etc., is used to generate patterns.
- **Interpreting mined patterns**. Visualization is performed by using extracted patterns and models.
- **Consolidating discovered knowledge.** The extracted knowledge is incorporated into performance system and it is documented and reported to the interested parties.

3.4 Data Mining Models

The intension of data mining is either to formulate a descriptive or predictive model. Different data mining techniques can be used to achieve the goal of predictive and descriptive model.

The purpose of descriptive model is to detect the patterns or relationships in data and to find out the properties of the observed data. Summarization, Clustering, Sequence discovery, Association rule, are some of models, which belong to this category.

In contrast, predictive model is to forecast about unidentified data values based on the identified values. Classification, Regression, Prediction, Time series analysis models belong to this category. [20].

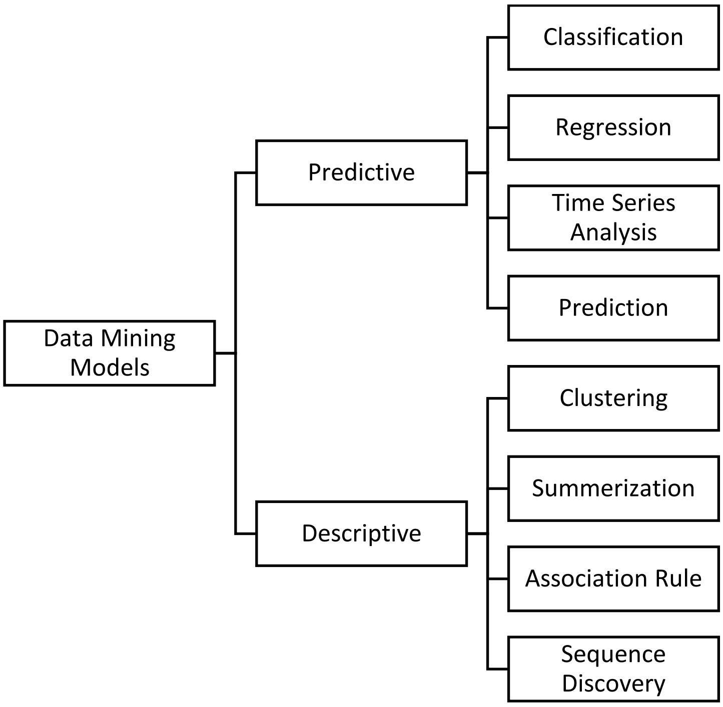


Figure 3-2 Data Mining Models

Classification: The data is mapped into predefined groups or classes by using classification model. As the classes are determined in advance it is referred to as supervised leaning. It is required to define the classes based on attribute values of data, when using the classification algorithms. Classes are described by looking the characteristics of data already known which belong to these classes, for an instance pattern recognition is a one of classification technique, which classify an input pattern into one of predefine classes based on similarity.

Regression: Regression maps a data item to a prediction variable with real value. It engaged in, the leaning of the function that does this mapping. With the assumption of target data fit into some known type of function like liner, logistic, the regression determines the best function that models the given data of this type. Moreover, the best function is determined, based on some type of error analysis. For an instance standard linear regression is a one of regression technique.

Time Series Analysis: The purpose of time series analysis is to inspect the value of an attribute which deviates over time. Typically the values are captured uniformly spaced time points. For instance daily, weekly, hourly, and so on. The time series visualization is achieved by using a time series plot. Three basic functions of time series analysis can be identified. Firstly, measure the distance in order to determine the similarity between different time series. Further, to determine its behavior, the structure of the line is examined. Thirdly, predict future values based on historical time series plot.

Prediction: Prediction is a data mining task or technique which can be considered as a type of classification. Although the prediction task is a type of prediction model, it differ from prediction model which predicting a future state instead of current state. The application of prediction task includes machine learning, patter recognition flooding, and speech recognition.

Clustering: Clustering is also known as unsupervised learning or segmentation which is similar to classification apart from the groups are not predefined, preferably data is defined by itself. The data is mapped into groups are not predefined, which can be considered as partitioning or segmenting data into groups. These groups might or might not be disjointed. Typically by determining the similarity among the data, clustering is achieved on predefined attributes.

Summarization: Summarization also known as characterization or generalization, maps data into subsets with associated simple descriptions. By retrieving portions of the data, it derives representative information about the database. Moreover, it can derive summary type information like the mean of some numeric attribute from the data.

Association Rules: Association also known as affinity analysis or link analysis is used to uncover relationship or association among data. Association rule mining is of this kind of application to explore association rules which identifies specific type relationships with in data. Associations rule are used to identify the items that are frequently purchased together in the retail sales community.

Sequence Discovery: In sequence discovery, sequential patterns are discovered in data, based on a time sequence of actions. Although discovered patterns are found to be related like associations in that data, these relationships are based on time. Moreover, in market basket analysis the items need to be purchased at the same time. In contrast, the items are purchased over time in some order with sequence discovery [21].

3.5 Major Applications of Data Mining

In the field of science and business data mining plays a major role. The data mining techniques are adapted in numerous fields, due to the fast access to the data and the capability of exploring valuable information form large mount data. Marketing, telecommunication, fraud detection, finance, and education sector, and medical are some of fields which use data mining applications. The table 3.1 list the some of the data mining application areas [22].

Area	Application of Data Mining Techniques
Education	Student learning behavior, performance and dropouts and predict the results
Banking and Finance	Predict credit card fraud, estimate risk, trend analysis and profitability.
Marketing	explore the frequently purchase items of customers
Telecommunication	Improve marketing efforts, fraud detection and better management of telecommunication networks
Agriculture	Analysis of crop harvest a with respect to parameters like year, rainfall, production and area of sowing
Cloud Computing	Useful information is extracted from virtually integrated data warehouse in order to reduce the infrastructure and storage cost

Table 3-1 Application of Data Mining Techniques

3.6 Data Mining in Road Accident Analysis

In the area of road traffic accident analysis, data mining technique has been recognize as reliable technique which can be used beyond the conventional techniques. When analyzing road traffic accidents, different models were developed based on different data mining techniques. Among them most of the models were developed to identify factors affecting the severity of a traffic accident.

3.7 Tools used for Data Mining

We have used, mainly the tools provided with Weka software to pre-process data, and to create, train and test the data models which were proposed in this research initiative.

WEKA

Weka includes a set machine learning algorithms for data mining tasks. Using Weka, data mining algorithms can be directly applied to a dataset. Also it can be called from a Java code. The tools provided with Weka can be used for data mining tasks such as pre-processing of data, classification, regression, clustering, association rules, and visualization. Moreover, it is possible to develop new machine leaning schemes with Weka [23].

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities [24].

3.8 Data Mining Algorithms used

J48 Algorithm

J48 is implementation of C4.5 algorithm. C4.5 builds decision trees with the help of information entropy. At every node of the tree, attribute is selected which is most effectively splitting itself into multiple subset. Splitting is done based on Information Gain (IG) value. For decision making, the attribute with highest normalized IG is used. This algorithm has the limitation of handling numeric data only.[25]

Random Forest Algorithm

Random Forest algorithm is a supervised classification algorithm.

Random Forest corresponds to a collection of combined Decision Tree {**hk**(**x**,**Tk**)}, where **k** = 1,2,...,L where **L** is number the tree and **Tk** is the training set built at random and identically distributed, **hk** represents the tree created from the vector **Tk** and is responsible for producing an output **x** [26]**.**

FURIA Algorithm

FURIA is a novel fuzzy rule-based classification method, which is short for Fuzzy Unordered Rule Induction Algorithm. FURIA built upon RIPPER algorithm and extends the futures of it. RIPPER is a rule learning algorithm which generate simple and comprehensible rule sets without global optimization. Moreover a number of modifications and extensions are also included in FURIA. Instead of conventional rules it learns fuzzy rules also instead of rule lists it learns unordered rule sets. In addition, it makes use of an efficient rule stretching method to deal with uncovered examples. As per the research study, FURIA surpass the performance based on classification accuracy of original RIPPER significantly, and also other classifiers like C4.5 [27].

3.8 Summary

This chapter presented the technology used to implement the solution with specific reference to data mining models and tasks .We also reported the specific tools and data mining algorithms which were used to build the solution. The next chapter describes the novel approach to analyze the road accident data.

Chapter 4

A Novel Approach to Analyze Road Traffic Accident Data

4.1 Introduction

Chapter 3 discussed the technology for analyzing road traffic accident data to identify the injury and accident severity based on underlying factors. This chapter presents our approach to analyze road traffic accident data using data mining techniques in detail. This chapter consist of following sections, hypothesis, input, output, process, users and features.

4.2 Hypothesis

Application of data mining techniques on road traffic accident analysis will discover hidden factors on accident severity. A predictive model is built, which will allow, the value of a dependent variable is predicted based on known values of independent variables. This model can be used to predict the factors affecting injury, accident severity and crash proneness.

4.3 Input

The initial input for the road traffic accident analysis is obtain form Sri Lank Police Traffic Division. The input data covers all the road accident details reported by MAAP database of Sri Lanka Police. These data can be divided into three tables: attendant circumstance, vehicle details and casualty details. The total dataset for this study contains traffic accident records of year 2015, stored in .mdb Microsoft database file format.

4.4 Output

As the output of the process predict the factors affecting accident severity and injury severity

4.5 Process

In the process of road traffic accident data analysis, standard steps of knowledge discovery process form data selection to evaluation are accomplished.

4.5.1 The Proposed Model

The Proposed model has 4 steps, namely Data Pre-processing, Attribute/Feature selection, Measure the variable importance, Classification rule extraction

1. Data Pre-processing

Pre-processing is a significant phase in data mining, which involves the transformation of variables in the given data set into an understandable format. Dataset is cleansed within the pre-processing stage. For instance, missing values are filled, data inconsistences are resolved and noisy data is smoothed.

2. Attribute/Feature Selection

Since the original dataset has large set of attributes, but only few of them are relevant next step is to reducing the dataset by attribute subset selection. Two attributes have been selected as class labels to full fill the first and second objective. The feature selection was conducted through application of Wrapper method. The J48 algorithm has been used as the classifier. A classification model is built by using training data to assign a class label to testing data set

3. Measure the variable importance

Random forest was applied to measure the variable importance. The intention was to identify the high impact variables individually from each of the datasets.

4. Classification rule extraction

Finally, classification and regression trees have been applied for rule extraction. The intension was to extract the knowledge based on classification rules which is more understandable to end users.

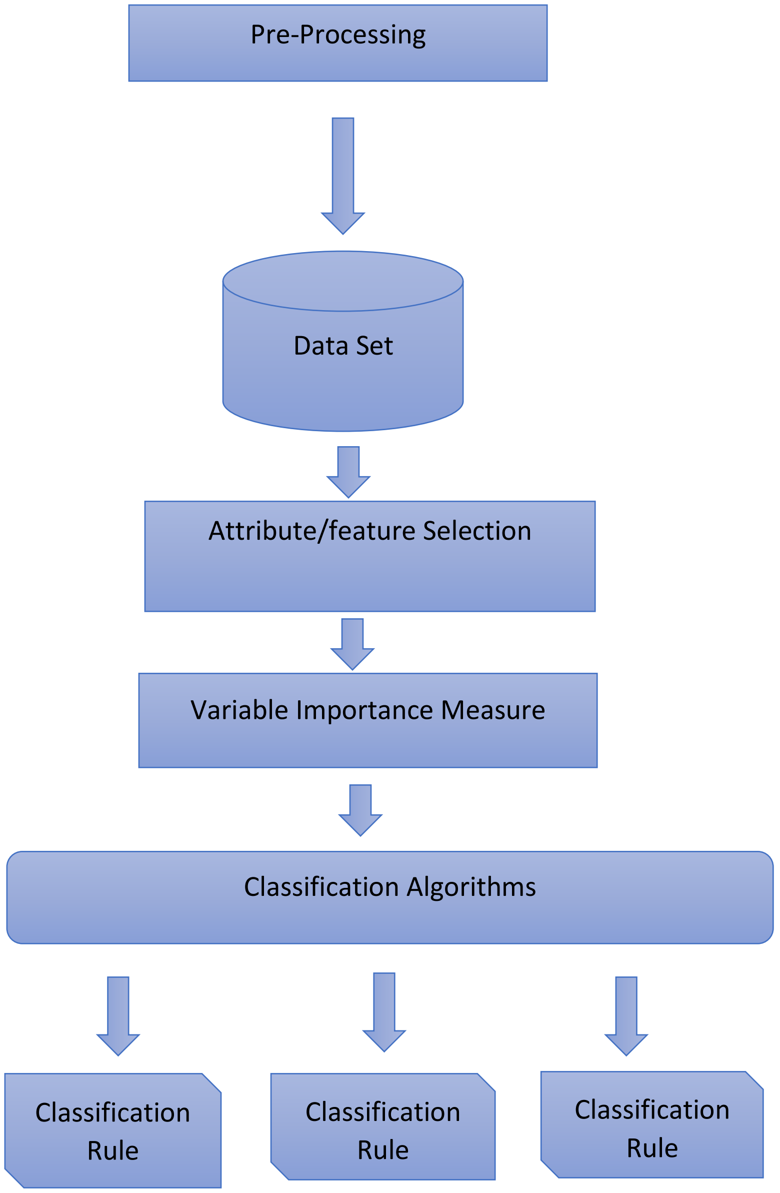

Pre-Processing
Data Set
Attribute/feature Selection
Variable Importance Measure
Classification Algorithms
Classification Rule
Classification Rule
Classification Rule

Figure 4-1 Proposed Model

4.6 Users

The Government of Sri Lanka, policy makers of road safety, Sri Lanka Police are the users who should pay their attention to the results of this data analysis.

4.7 Summary

This chapter presented our novel approach to analyze road traffic accident data, to identify the underlying factors influence on road traffic accidents which contribute to road safety decision making. The next chapter presents the system design of the novel approach to analyses road accident data.

Chapter 5

Research Design for Analyzing RTA Data

5.1 Introduction

In the previous chapter we briefly discussed our approach to analyze the road traffic accident data, based on data mining process and also the proposed model. This chapter describes the system design and expends the process in the approach towards the solution.

5.2 Research Design

This research of analyzing road traffic accidents using data mining techniques to explore the underlying factors influencing on injury and accident severity, requires a scientific method to follow systematic research.

In scientific method the knowledge is acquired based on observations, which includes hypotheses formulation, by means of induction. Testing of deductions based on observations and measurement are drawn from the hypotheses. One of the scientific research method is experimentation which is used to decide the characteristics of the relationship in between independent and dependent variable. A researcher manages one or more variables and controls and measures any change in other variables in experimental method [28] [29].

The Figure 4-2 describes the steps we have followed form beaning to end of research. This study start from data collection and end with prediction of the road accident severity. Once the dataset is collected from Sri Lanka Police, initial set of attributes based on literature and availability of data is fed into feature selection process to select the desired attributes which includes one dependent variable for each research objective. After the feature section, selected attributes of the dataset are checked for missing values, duplication, and outliers. Once the preprocessing has done, the dataset is separated into training and testing data sets.

Next step is applying the J48 classification algorithm on the training data set to train the model and test classifier which is trained and generate results. After that Random Forest algorithm is applied to measure the variable importance. Then accident severity prediction is done.

Finally to find out the relationship between the underlying factors that frequently affect the severity of a traffic accident we have applied FURIA algorithm.

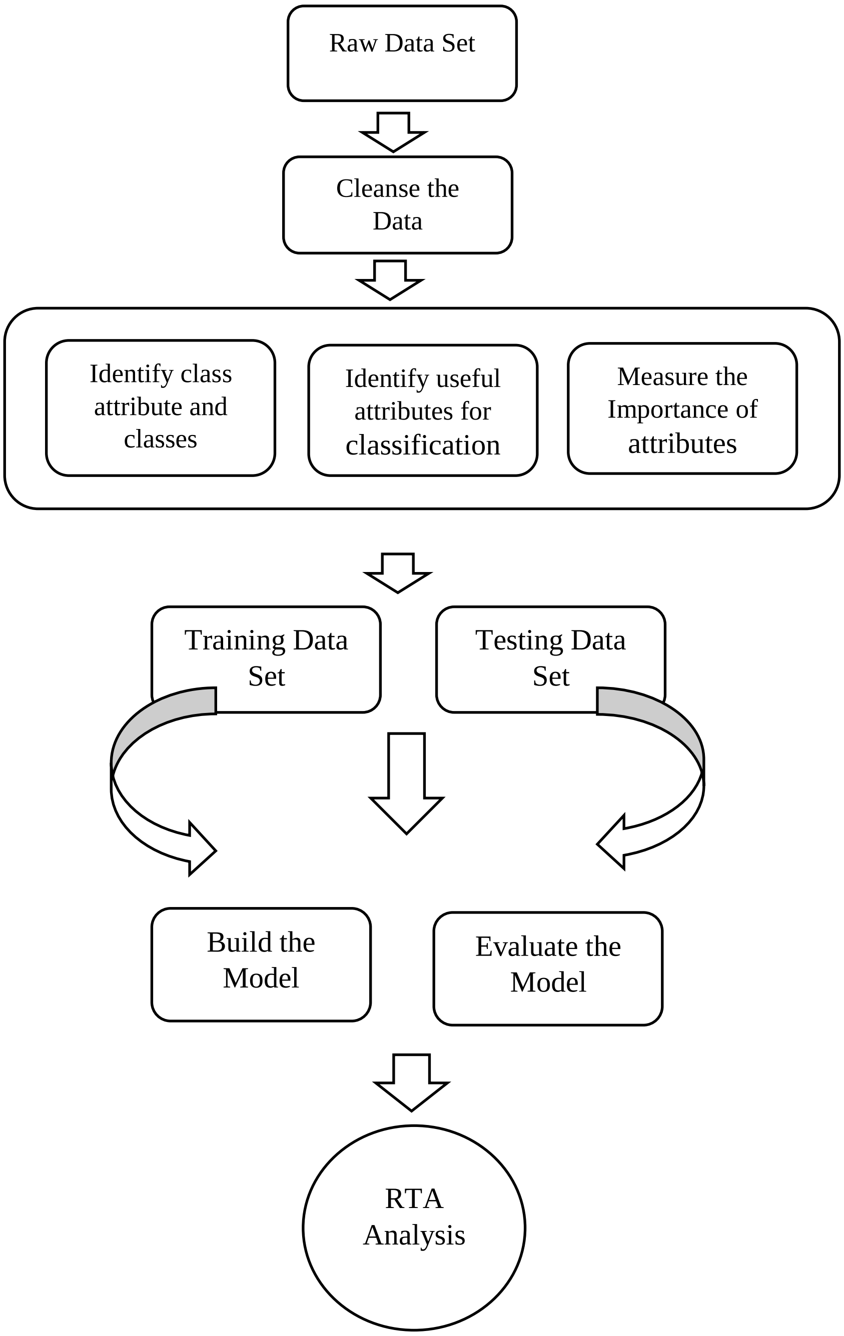

Raw Data Set
Cleanse the Data
Identify class attribute and classes
Identify useful attributes for classification
Measure the Importance of attributes
Training Data Set
Testing Data Set
Build the Model
Evaluate the Model
RTA Analysis

Figure 4-2 Overall System design of the Proposed Solution

5.3 Summary

This chapter provided the details of research design and applicability of research method for this research initiative. In this research some of the most significant capabilities of data mining techniques were leveraged trough a model for crash pattern analysis. Next chapter provides the implementation details of the design.

Chapter 6

Implementation

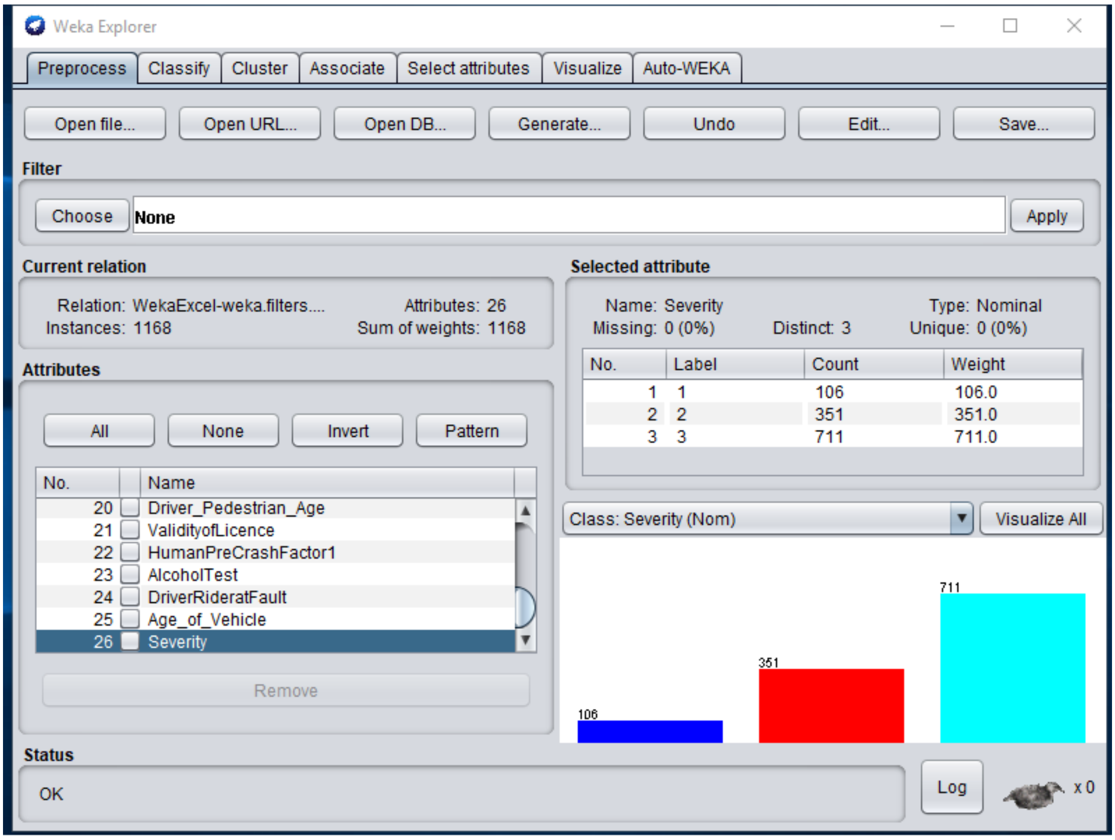
6.1 Introduction

In chapter 5 the top level design of the solution has been described in the terms of what attributes are used to represent the data mining model, model formulation, training and testing of models for analyzing road accident data. This chapter describes the implementation of each research objective concerning software, algorithms, methods, etc.

6.2 Solution for the Research Objective

The solutions for the research objectives which are stated in chapter two were implemented with the help of Weka software using steps mentioned below.

The Weka software shown in figures 4-3 is a "collection of machine learning algorithms for data mining tasks".



Filter

None

Current relation

Relation: WekaExcel-weka.filters.... Attributes: 26
Instances: 1168 Sum of weights: 1168

Selected attribute

Name: Severity Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	1	106	106.0
2	2	351	351.0
3	3	711	711.0

Attributes

No.		Name
20	□	Driver_Pedestrian_Age
21	□	ValidityofLicence
22	□	HumanPreCrashFactor1
23	□	AlcoholTest
24	□	DriverRideratFault
25	□	Age_of_Vehicle
26	□	Severity

Class: Severity (Nom)

Status

OK

Figure 4-3 Weka GUI

It includes tools basically for data pre-process, classification, regression, clustering, association rules mining, and visualization. Practically Weka can be run on any platform which is developed using Java. Using Weka, data mining algorithms can be applied directly to a dataset. Moreover it can called from code written in Java.[23] The Weka provides useful tools to for data preprocessing.

6.2.1 Data Pre-processing

Pre-processing of data is a significant phase in data mining. It is essential to avoid incomplete, noisy and inconsistent data before formulating a data model. RTA Data for the analysis is collected form Sri Lanka Police. The RTA data at Sri Lanka Police is maintained in Microsoft Database (Mdb) file format. As the Weka prefers to load data in ARFF (Attribute-Relation File Format) format, Mdb data file has been converted into arff file format with the help of MS Access and WekaExcel package of Weka software. In road traffic accident data set, there are missing data values due to not recording it at the incident point for several attributes.

6.2.1.1 Data Pre-processing for the First Research Objective

As the solution for the first research objective, to explore the underlying factors influencing on injury severity, the proposed model limits to predict injury severity of pedestrians using available data and considering single vehicle accidents with pedestrians only. The initial sample dataset consist of RTA data of Colombo Division.

6.2.1.2 Data Pre-processing for the Second Research Objective

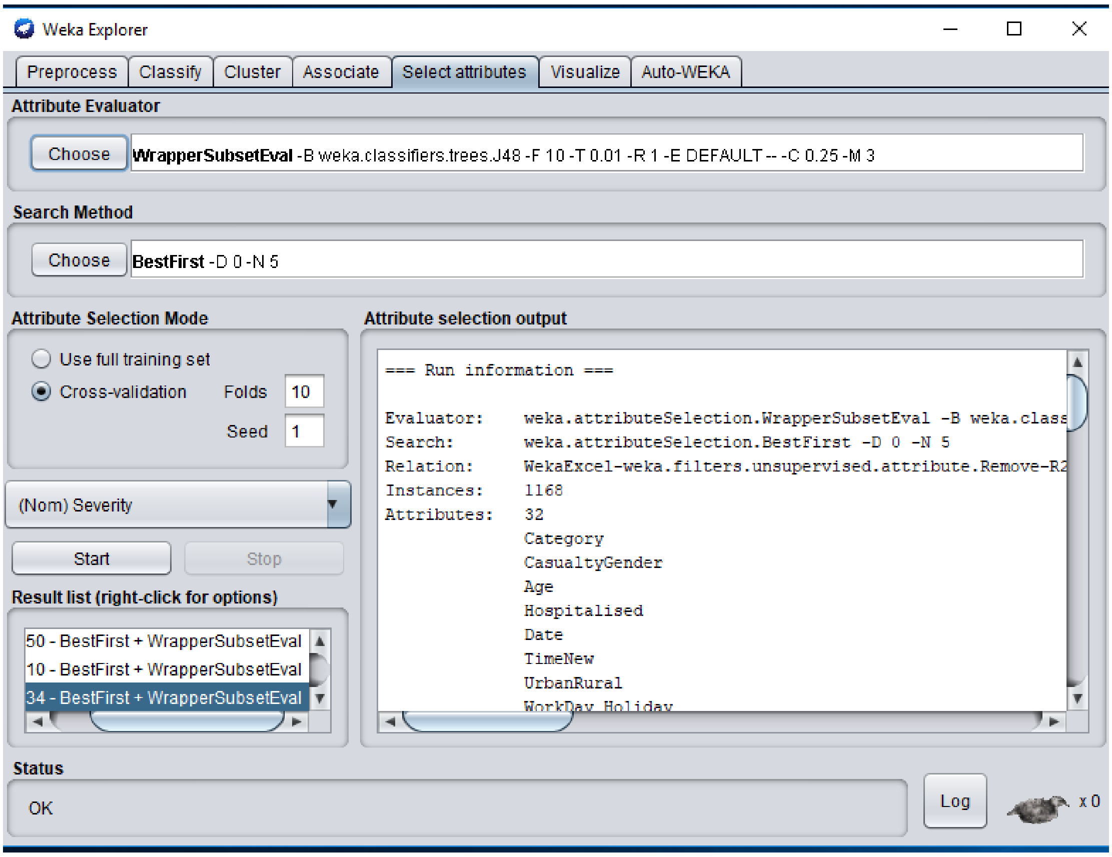
As the solution for the second research objective, to identify the human, environment and vehicle factors influencing the road traffic accident severity, the proposed model limits to predict overall accident severity of single vehicle accidents. The initial sample dataset consist of RTA data of Colombo Division.

6.2.2. Attribute/Feature Selection

Since the original dataset has large set of attributes, but only few of them are relevant next step is to reducing the dataset by attribute subset selection/feature selection. Feature selection is a process of selecting a subset of feature form original dataset for data mining

We have used the wrapper approach for the initial attribute selection process. Wrapper approach uses a classification algorithm to measure the importance of features set, therefore the feature selection depends on the classifier model selected. Although wrapper approach is very expensive for large dimensional database, when considering the computational complexity and time consumption as each feature set considered must be evaluated with the classification algorithm used, it generally outperforms the filter approach [30].

For this research, Weka attribute selector tool is used to for the initial attribute selection process. This process consist of into two steps. Selecting the **Attribute Evaluator** and **Search Method.** A subset of attributes are assessed based on Attribute Evaluator method. **Wrapper Subset Evaluator** is used to assess, subsets using a J48 classifier with 10-fold cross validation. **Best First** Search Method with backward search direction is used to navigate attribute subsets. Search Method is a well-organized way of navigating the search space of possible attribute subsets based on attribute evaluator.



Attribute Evaluator

WrapperSubsetEval -B weka.classifiers.trees.J48 -F 10 -T 0.01 -R 1 -E DEFAULT -- -C 0.25 -M 3

Search Method

BestFirst -D 0 -N 5

Attribute Selection Mode

○ Use full training set

● Cross-validation Folds 10

Seed 1

(Nom) Severity

Result list (right-click for options)

50 - BestFirst + WrapperSubsetEval
10 - BestFirst + WrapperSubsetEval
34 - BestFirst + WrapperSubsetEval

Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.WrapperSubsetEval -B weka.class
Search:       weka.attributeSelection.BestFirst -D 0 -N 5
Relation:     WekaExcel-weka.filters.unsupervised.attribute.Remove-R2
Instances:    1168
Attributes:   32
              Category
              CasualtyGender
              Age
              Hospitalised
              Date
              TimeNew
              UrbanRural
              WorkDay Holiday
```

Status

OK

Figure 4-4 Weka Attribute Selector Tool

6.2.2.1 Attribute Selection for the First Research Objective

For the attribute/feature subset selection, the wrapper method of Weka is used with J48 algorithm as the classifier. Initially 32 attributes were selected, which includes 31 independent variables and one dependent variable (‘Severity’) based on literature review and availability of data which relevant to build the model. The class label ‘Severity’ has three nominal values which are: ‘Fatal,’ ‘Grievous,’ and ‘Non Grievous’.

After the initial attribute selection process, 11 attributes, which includes 10 independent variables and one dependent variable (‘Severity’) were selected to build the model. Following table lists the selected attribute set.

#	**Variable Name**	**Values**
1	Hospitalized	1. Injured and admitted to hospital
		2. Injured but not admitted to hospital
2	Weather	1. Clear
		2. Cloudy
		3. Rain
		4. Fog/Mist
3	Light Condition	1. Daylight
		2. Night ,No street lighting
		3. Dusk, Dawn
		4. Night Improper Street Lighting
		5. Night , Goof Street lighting
4	Pedestrian_Location	1. on Pedestrian crossing
		2. Pedestrian crossing within 50 meters
		3. Pedestrian crossing beyond 50 meters
		4. Pedestrian over-pass bridge or under pass tunnel within 50 meters
		5. Hit outside side walk
		6. Hit on side walk
		7. Hit on road without Side walk
5	Location_Type	1. Stretch of road, no junction within 10 meters
		2. 4-leg junction
		3. T-Junction

		4. Y-Junction
		5. Roundabout
		6. Multiple Road junction
		7. Entrance by private road
		8. Rail Road Crossing
6	SpeedLimitPosted	1. Yes
		2. No
7	Element _Type	00. Unknown
		01. Car
		02. Dual purpose vehicle
		03. Lorry
		04. Cycle
		05. Motor cycle, Moped
		06. Three wheeler
		07. Articulated vehicle, prim
		08. SLTB bus
		09. Private bus
		10. Intercity bus
		11. Land vehicle/ Tractor
		19. Others
8	Driver_Pedestrian_Gender	1. Male
		2. Female
9	HumanPreCrashFactor1	01. Speeding
		02. Aggressive / negligent driving
		03. Error of judgment
		04. Influenced by alcohol / drugs
		05. Fatigue / fall asleep
		06. Distracted / inattentiveness (handling radio, mobile phone, mental stress etc.)
		07. Poor eye sight
		08. Sudden illness
		09. blinded by another vehicle / sun
10	AlcoholTest	1. No alcohol or below legal limit
		2. Over legal limit

		3. Not tested
	Severity **(Class Variable)**	1. Fatal
		2. Grievous
		3. Non Grievous

Table 4-1 Attributes selected after Wrapper Method

6.2.2.2 Attribute Selection for Second Research Objective

For the attribute/feature subset selection, the wrapper method of Weka is used with J48 algorithm as the classifier. Initially 25 attributes were selected, which includes 24 independent variables and one dependent variable (class label 'Highest_Severity') based on literature review and availability of data which relevant to build the model. According to severity of accidents there are four categories. (Fatal, Grievous, Non Grievous, Damage Only). But 'Damage only' accidents are not reported to the police due to the interferences of insurance agents and settling without reporting to police. Therefore that accident category was excluded as these accidents may create incorrect results.

After the initial attribute selection process, 12 attributes, which includes 11 independent variables and one dependent variable ('Highest_Severity') were selected to build the model. Following table lists the selected attribute set.

#	**Variable Name**	**Values**
1	Time	1-> 00:00 – 03:00
		2-> 03:00 – 06:00
		3-> 06:00 - 09:00
		4-> 09:00 - 12:00
		5-> 12:00 – 15:00
		6-> 15:00 – 18:00
		7-> 18:00 – 21:00
		8-> 21:00 – 24:00
2	Day_of_Week	1 - Sunday

		2 - Monday
		3 - Tuesday
		4 - Wednesday
		5 - Thursday
		6 - Friday
		7 - Saturday
3	Road_Surface	1 - Dry
		2 - Wet
		3 - Flooded with water
		4 - Slippery surface
4	Weather	1 - Clear
		2 - Cloudy
		3 - Rain
		4 - Fog/Mist
		9 - Other
5	LightCondition	1 - Daylight
		2 - Night ,No street lighting
		3 - Dusk, Dawn
		4 - Night Improper Street Lighting
		5 - Night , Goof Street lighting
6	Pedestrian_Location	1 - on Pedestrian crossing
		2 - Pedestrian crossing within 50 meters
		3 - Pedestrian crossing beyond 50 meters
		4 - Pedestrian over-pass bridge or under pass tunnel within 50 meters
		5 - Hit outside side walk
		6 - Hit on side walk
		7 - Hit on road without Side walk
7	SpeedLimitPosted	1 – Yes 2 - No
8	Hospitalised	1 - Injured and admitted to hospital
		2 - Injured but not admitted to hospital
9	Element _Type	00 Unknown
		01. Car
		02. Dual purpose vehicle
		03. Lorry

		04. Cycle
		05. Motor cycle, Moped
		06. Three wheeler
		07. Articulated vehicle, prim
		08. SLTB bus
		09. Private bus
		10. Intercity bus
		11. Land vehicle/ Tractor
		19. Others
10	AlcoholTest	1. No alcohol or below legal limit
		2. Over legal limit
		3. Not tested
11	HumanPreCrashFactor1	01. Speeding
		02. Aggressive / negligent driving
		03. Error of judgment
		04. Influenced by alcohol / drugs
		05. Fatigue / fall asleep
		06. Distracted / inattentiveness (handling radio, mobile phone, mental stress etc.)
		07. Poor eye sight
		08. Sudden illness
		09. blinded by another vehicle / sun
	Highest_Severity **(Class Variable)**	1. Fatal
		2. Grievous
		3. Non Grievous

Table 4-2 Attributes selected after wrapper method

6.2.3. Measure the Variable Importance

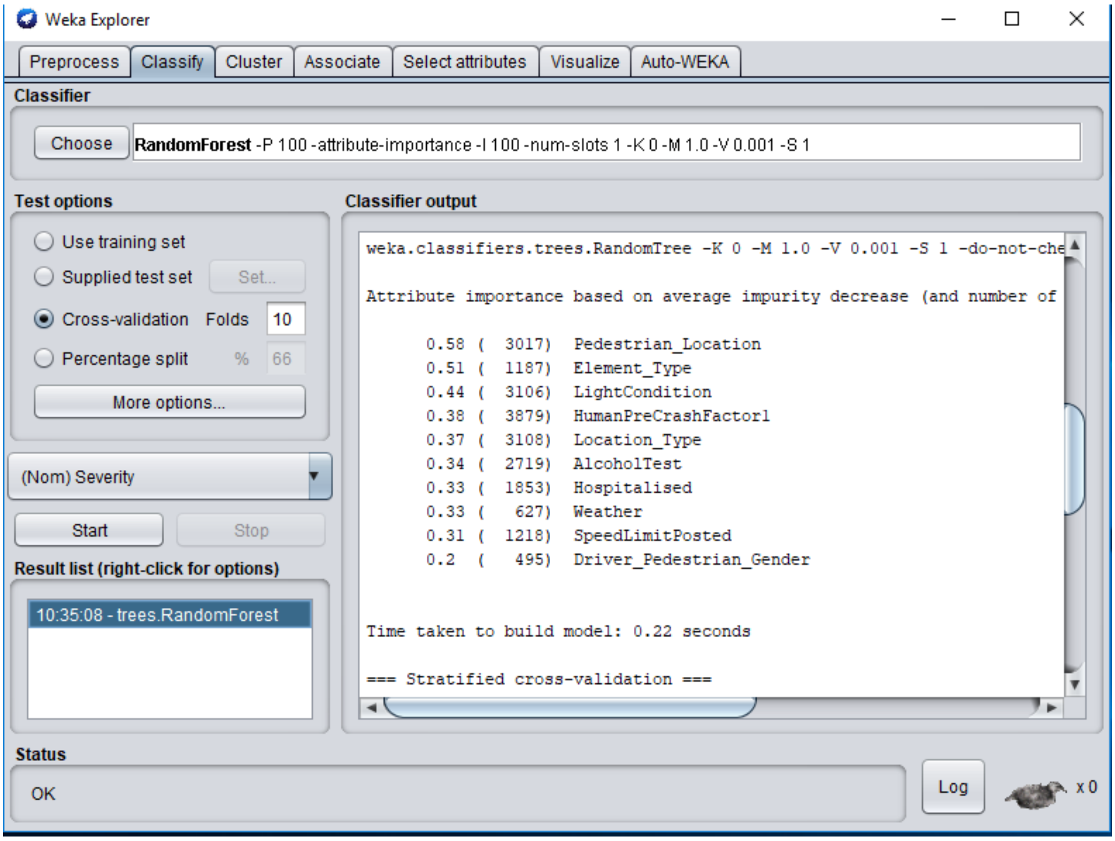
Random forest was applied to measure the variable importance. The intention was to identify the high impact variables individually from each of the datasets.

In the context of ensembles of randomized trees, Breiman [31] proposed to evaluate the importance of a variable X_m for predicting Y by adding up the weighted impurity decreases $p(t)$ Δi (s_t, t) for all nodes t where X_m is used, averaged over all N_T trees in the forest:

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = X_m} p(t) \Delta i(s_t, t)$$

and where $p(t)$ is the proportion N_t=N of samples reaching t and $v(s_t)$ is the variable used in split s_t.[32]

Predicting a value of dependent variable based on a set of independent variables is an important task in many scientific fields. These predictions are made not only to make the most accurate predictions of the dependent but also to identify which predictor variables are the most important to make predictions. Random Forests and its variants are used in different scientific areas as they can be applied to a wide range of problems and the capability of building accurate models and providing variable importance measures simultaneously, become a major data analysis tool with success [32].



RandomForest -P 100 -attribute-importance -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

● Cross-validation Folds 10

(Nom) Severity

10:35:08 - trees.RandomForest

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-che

Attribute importance based on average impurity decrease (and number of

      0.58 (  3017)  Pedestrian_Location
      0.51 (  1187)  Element_Type
      0.44 (  3106)  LightCondition
      0.38 (  3879)  HumanPreCrashFactor1
      0.37 (  3108)  Location_Type
      0.34 (  2719)  AlcoholTest
      0.33 (  1853)  Hospitalised
      0.33 (   627)  Weather
      0.31 (  1218)  SpeedLimitPosted
      0.2  (   495)  Driver_Pedestrian_Gender


Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
```

OK

Figure 4-5 Measure the attribute importance with RF

In Weka the Random Forest uses Random Tree, which employs Entropy rather than Gini index. The tables 4-3 and 4-4 illustrate the attribute importance based on average impurity decrease using Random Forest.

Attribute Importance	**Variable**
0.58	Pedestrian_Location
0.51	Element_Type
0.44	LightCondition
0.38	HumanPreCrashFactor1
0.37	Location_Type
0.34	AlcoholTest
0.33	Hospitalized
0.33	Weather
0.31	SpeedLimitPosted
0.20	Driver_Pedestrian_Gender

Table 4-3 VIM for Injury severity

Attribute Importance	**Variable**
0.70	Day_of_Week
0.67	Element_Type
0.65	Pedestrian_Location
0.62	TimeNew
0.59	LightCondition
0.54	AlcoholTest
0.52	HumanPreCrashFactor1
0.48	Weather
0.46	Hospitalized
0.46	Road_Surface
0.41	SpeedLimitPosted

Table 4-4 VIM for Accident Severity

6.2.4. Classification Rule Extraction

Finally, the significant rules were identified by using FURIA a novel fuzzy rule-based classification method. “weka.classifiers.rules.FURIA” is a class for generating a fuzzy rule-based sets. FURIA built upon RIPPER algorithm and extends the futures of it. RIPPER is a rule learning algorithm which generate simple and comprehensible rule sets. No default rule is used and also the order of the class is not considered by FURIA. Certainty factor (CF) of a fuzzy rule specifies the percentage of confidence that an extracted rule is correct. The tables 4-5 and 4-6 show the rule sets generated by using FURIA algorithm.

FURIA Rule				
Rule No	**Antecedent**		**Consequent**	**Certainty Factor (CF)**
1	(Element_Type = 00)	=>	Severity=1	0.71
2	(Element_Type = 09) and (Pedestrian_Location = 2) and (LightCondition = 1) and (HumanPreCrashFactor1 = 02)	=>	Severity=1	0.71
3	(Pedestrian_Location = 3) and (AlcoholTest = 1) and (LightCondition = 1) and (Element_Type = 05) and (HumanPreCrashFactor1 = 02)	=>	Severity=1	0.69
4	(LightCondition = 4) and (Pedestrian_Location = 1) and (HumanPreCrashFactor1 = 02) and (Location_Type = 1)	=>	Severity=1	0.69
5	(Element_Type = 07)	=>	Severity=1	0.70
6	(Hospitalised = 1) and (HumanPreCrashFactor1 = 01) and (AlcoholTest = 1) and (Pedestrian_Location = 1)	=>	Severity=2	0.69
7	(Hospitalised = 1) and (HumanPreCrashFactor1 = 01) and (Element_Type = 02)	=>	Severity=2	0.60
8	(Hospitalised = 2)	=>	Severity=3	0.90
9	(LightCondition = 1)	=>	Severity=3	0.61

Table 4-5 FURIA rule generation for injury severity

FURIA Rule				
Rule No	**Antecedent**		**Consequent**	**Certainty Factor (CF)**
1	(Element_Type = 00)	=>	Highest_Severity=1	0.73
2	(Element_Type = 05) and (Day_of_Week = 4) and (Pedestrian_Location = 3) and (HumanPreCrashFactor1 = 02)	=>	Highest_Severity=1	0.70
3	(Element_Type = 09) and (Pedestrian_Location = 2) and (TimeNew = 5)	=>	Highest_Severity=1	0.67
4	(LightCondition = 4) and (Day_of_Week = 2) and (Pedestrian_Location = 2)	=>	Highest_Severity=1	0.85
5	(Element_Type = 09) and (Day_of_Week = 1)	=>	Highest_Severity=1	0.63
6	(AlcoholTest = 1) and (Pedestrian_Location = 3) and (Day_of_Week = 2) and (TimeNew = 5)	=>	Highest_Severity=1	0.73
7	(Day_of_Week = 3) and (TimeNew = 2) and (SpeedLimitPosted = 1)	=>	Highest_Severity=1	0.72
8	(Hospitalised = 1) and (HumanPreCrashFactor1 = 01) and (Pedestrian_Location = 1) and (AlcoholTest = 1)	=>	Highest_Severity=2	0.71
9	(Hospitalised = 1) and (Day_of_Week = 1) and (Element_Type = 02)	=>	Highest_Severity=2	0.58
10	(Hospitalised = 1) and (Element_Type = 05) and (LightCondition = 5) and (Pedestrian_Location = 3)	=>	Highest_Severity=2	0.83
11	(Hospitalised = 1) and (Day_of_Week = 7) and (Pedestrian_Location = 2)	=>	Highest_Severity=2	0.54
12	(Hospitalised = 2)	=>	Highest_Severity=3	0.91
13	(LightCondition = 1)	=>	Highest_Severity=3	0.61

Table 4-6 FURIA rule generation for Accident severity

6.3 Summary

This chapter provided the implementation details of each research module of proposed solution. Moreover, it mentioned, software algorithms, different parameters and resulting models. Next chapter evaluates the solution implemented.

Chapter 7

Evaluation

7.1 Introduction

The previous chapter discussed the details on implementation of the model of the solution. This chapter focuses on how testing strategies are carried out for the research objectives in the terms of evaluation measurements for the selected data mining techniques such as percentage of accuracy, TP rate and ROC area for the classification.

7.2 Evaluation for classification

We have used confusion matrix for evaluating a classifier quality, with the help of confusion matrix we can find various evaluation measures. For instance, accuracy, recall and precision to evaluate the data miming classifiers.

These measurements and their definitions are given in the following table.

Measure	**Formula**	**Meaning**
Precision/Positive Predictive Value	$PPV = \frac{TP}{(TP + FP)}$	"The fraction of relevant instances among the retrieved instances"
Recall / Sensitivity/True Positive Rate	$TPR = \frac{TP}{(TP + FN)}$	"The fraction of relevant instances that have been retrieved over the total amount of relevant instances"
Specificity/True Negative rate	$TNR = \frac{TN}{(TN + FP)}$	"Measures the proportion of actual negatives that are correctly identified"
Accuracy	$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$	"The percentage of predictions those are correct"
AUC (Area Under the Curve)		"The probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one"

Table 4-7 Evaluation measures for classifiers

7.3 Evaluation of Injury severity

The performance evaluation of each algorithm in predicting the injury severity was conducted in three different ways.

Initially, with the training data set. Then, the performance was evaluated using 10-fold cross-validation. Finally, testing data set was used to evaluate the prediction performance. The performance evaluation results summary of the J48 algorithm is presented in Table 4-8.

J48 Tree based classifier						
Test Mode	Correctly Classified Instances	Incorrectly Classified instances	Accuracy	Mean absolute error	Root Mean Squared Error	AUC
Training Set	847	427	66.4835 %	0.3026	0.389	0.765
10 Fold CV	744	530	58.3987 %	0.3432	0.4318	0.663
Test Set	335	393	46.0165 %	0.3906	0.4831	0.575

Table 4-8 J48 Tree Classifier evaluation summary for injury severity model

As depicted in Table 4-9, the prediction accuracy of J48 algorithm using test data set for Fatal, Grievous, and Non-Grievous accidents was 23.1%, 25.3%, and 66.3%, as per the order.

However, the prediction efficiency cannot completely described by using accuracy only. Accuracy is merely a single aspect of others, when describing the prediction efficiency. Therefore it is necessary to evaluate the perdition model by other ways.

Further evaluation of the model was done based on ROC curves. The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test.

The test set has an AUC of > 0.5, AUC explains the capability of the model to distinguish between classes. A model can predict more precisely, as the AUC value increases.

Classification Model	Class Values	TP rate	FP rate	Precision	Recall	F-Measur	AUC
J48	Fatal	0.231	0.178	0.178	0.231	0.201	0.525
	Grievous	0.253	0.166	0.448	0.253	0.323	0.587
	Non Grievous	0.666	0.569	0.549	0.666	0.602	0.581

Table 4-9 The prediction accuracy of J48 by using the test data set

The evaluation results summary for the variable importance measure with Random Forest Tree based classifier is presented in Table 4-10.

VIM with Random Forest Tree based classifier						
Test Mode	Correctly Classified Instances	Incorrectly Classified instances	Accuracy	Mean absolute error	Root Mean Squared Error	AUC
Training Set	1025	249	80.4553 %	0.21	0.3003	0.932
10 Fold CV	774	500	60.7535 %	0.3295	0.4333	0.676
Test Set	354	374	48.6264 %	0.3768	0.4747	0.593

Table 4-10 Random Forest evaluation summary for injury severity model

The evaluation results summary for the classification rule extraction with FURIA rule based classifier is presented in Table 4-11.

Classification Rule Extraction with FURIA rule based classifier						
Test Mode	Correctly Classified Instances	Incorrectly Classified instances	Accuracy	Mean absolute error	Root Mean Squared Error	AUC
Training Set	782	492	61.3815 %	0.2795	0.4618	0.651
10 Fold CV	735	539	57.6923 %	0.297	0.4808	0.617
Test Set	367	487	42.9742 %	0.3877	0.5428	0.538

Table 4-11 FURIA evaluation summary for injury severity model

In each test mode the AUC is greater than 0.5.

7.4 Evaluation of Accident Severity

The same approach was used to perform the evaluation of each algorithm in predicting the injury severity. Initially, with the training data set. Then, the performance was evaluated using 10-fold cross-validation. Finally, testing data set was used to evaluate the prediction performance. The performance evaluation results summary of the J48 algorithm is presented in Table 4-12.

J48 Tree based classifier						
Test Mode	Correctly Classified Instances	Incorrectly Classified instances	Accuracy	Mean absolute error	Root Mean Squared Error	AUC
Training Set	926	414	69.1045 %	0.2838	0.3767	0.794
10 Fold CV	780	560	58.209 %	0.3372	0.4404	0.642
Test Set	371	405	47.8093 %	0.3803	0.4912	0.567

Table 4-12 J48 Classifier Evaluation Summary for accident severity model

The prediction accuracy of J48 algorithm using test data set for Fatal, Grievous, and Non-Grievous accidents was 16.2%, 27.8%, and 71.1%, as per the order. The test set has an AUC of > 0.5, The prediction accuracy of J48 by using test data set is presented in Table 4-13.

Classification Model	Class Values	TP rate	FP rate	Precision	Recall	F-Measure	AUC
J48	Fatal	0.162	0.104	0.207	0.162	0.182	0.523
	Grievous	0.278	0.208	0.425	0.278	0.336	0.570
	Non Grievous	0.711	0.598	0.543	0.711	0.616	0.577

Table 4-13 The prediction accuracy of J48 by using test data set

The performance evaluation results summary of the Random Forest algorithm is presented Table 4-14.

Random Forest Tree based classifier						
Test Mode	Correctly Classified Instances	Incorrectly Classified instances	Accuracy	Mean absolute error	Root Mean Squared Error	AUC
Training Set	1262	78	94.1791 %	0.1403	0.2001	0.989
10 Fold CV	768	572	57.3134 %	0.3364	0.4342	0.663
Test Set	373	403	48.067 %	0.3747	0.4679	0.588

Table 4-14 Random Forest evaluation summary of accident severity model

The evaluation results summary for the classification rule extraction with FURIA rule based classifier is presented in Table 4-15.

Classification Rule Extraction with FURIA rule based classifier						
Test Mode	Correctly Classified Instances	Incorrectly Classified instances	Accuracy	Mean absolute error	Root Mean Squared Error	AUC
Training Set	830	510	61.9403 %	0.2755	0.4573	0.659
10 Fold CV	770	570	57.4627 %	0.2961	0.479	0.621
Test Set	764	788	49.2268 %	0.3512	0.5179	0.554

Table 4-15 FURIA evaluation summary of accident severity model

In each test mode the AUC is greater than 0.5.

7.5 Summary

This chapter concluded with the test result used to evaluate the data model. Final chapter will summarize the overall research and highlight the significant findings of the research and further improvements for the proposed solution

Chapter 8

Conclusion and Further Works

8.1 Introduction

All the previous chapters discussed the problem identified and proposed solution. This chapter provides an overview of the research and how we provide the solution to address the problem of analyzing the crash patterns of Sri Lankan road accident data. Furthermore this chapter focuses on limitations and further works of this research.

8.2 Overview of the Research

In the area of road traffic accident analysis, data mining technique has been recognized as reliable technique which can be used beyond the conventional techniques. When analyzing road traffic accidents, different models were developed to identify factors affecting the severity of a traffic accident.

In analyzing road accidents, basically two types of models have been developed. The frequency based models and severity based models. This research is focusing on developing a solution based on severity model. Classification and Clustering are the two directions of road accident analysis with data mining techniques. This research study is focusing on the classification technique.

To build the solution model, RTA Data is cleansed through data pre-processing. Since the original dataset has large set of attributes, but only few of them are relevant, next step was to reduce the dataset by attribute subset selection/feature selection. For the attribute/feature subset selection, the wrapper method of Weka was used with J48 algorithm as the classifier. Random forest was applied to measure the variable importance. Finally, FURIA a fuzzy based rule extraction algorithm was applied on dataset.

8.3 Key Findings

This research basically studies the application of J48, Random Forest and FURIA data mining algorithms on road accident data to predict the underlying factors which influence on severity of an accident.

The degree of importance of each attribute to the accident severity, varies based on the algorithm used. As per the results, the 'Pedestrian_Location' and 'Element_Type' are frequent factors which influence the severity of a pedestrian accident. It is also revealed that 'Day_of_Week', substantially influences on overall accident severity.

Importance	**Random Forest**	**FURIA**
1	Pedestrian_Location	Element _Type
2	Element_Type	HumanPreCrashFactor1
3	LightCondition	Pedestrian_Location
4	HumanPreCrashFactor1	LightCondition
5	Location_Type	Hospitalised
6	AlcoholTest	AlcoholTest
7	Hospitalised	Location_Type
8	Weather	Weather
9	SpeedLimitPosted	SpeedLimitPosted
10	Driver_Pedestrian_Gender	Driver_Pedestrian_Gender

Table 4-16 Most influential factors of injury severity based on algorithm

Importance	**Random Forest**	**FURIA**
1	Day_of_Week	Day_of_Week
2	Element_Type	Pedestrian_Location
3	Pedestrian_Location	Element _Type
4	TimeNew	Hospitalised
5	LightCondition	LightCondition
6	AlcoholTest	Time
7	HumanPreCrashFactor1	AlcoholTest
8	Weather	HumanPreCrashFactor1
9	Hospitalised	SpeedLimitPosted
10	Road_Surface	Road_Surface
11	SpeedLimitPosted	Weather

Table 4-17 Most influential factors of accident severity based on algorithm

8.4 Problems Encountered and Limitations

The proposed solution model only address the single vehicle accident with pedestrian involvement due to the time limit and availability of data.

In RTA data set, there are missing values because of two reasons. Missing and unknown values due to not reporting the incident to the police. For an example ‘Damage only’ accidents are not reported to the police as a result of the interferences of insurance agents and settling without reporting to police. Missing and unknown values, due to not recording it by the police at the incident point, which limits the study of analyzing contributory factors.

8.5 Further Works

This research only addresses two sub research questions attached with road accident data. This can be further extended to analyses multi-vehicle accidents, injury servility of drivers and passengers, etc. The data mining techniques which was applied to the dataset is classification and used the classification rule extraction. But clustering technique can also be used to form natural groups and association rule mining can be applied for rule extraction. Therefore this research could be extended to address different crash patterns of road accidents in Sri Lanka.

8.6 Summary

This chapter concluded the thesis by describing the solution given with data mining to analyze road traffic accident data, the limitations and problems encountered and how it can be further extended to analyze different crash patterns of road accidents in Sri Lanka.

References

[1] World Health Organization, *Global status report on road safety 2015*. 2015.

[2] The Island, "Roads in Sri Lanka - A matter of life & death”.

[3] F. Moradkhani, S. Ebrahimkhani, and B. S. Begham, “Road Accident Data Analysis: A Data Mining Approach,” May 2014.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[5] S. Kumar and D. Toshniwal, “A data mining approach to characterize road accident locations,” *J. Mod. Transp.*, vol. 24, no. 1, pp. 62–72, Mar. 2016.

[6] World Health Organization, “Research Framework for Road Safety in the South-East Asia Region.” WHO publications, 2015.

[7] “WHO | Road safety training manual,” *WHO*. [Online]. Available: http://www.who.int/violence_injury_prevention/road_traffic/activities/training_manuals/en/. [Accessed: 07-Jul-2018].

[8] S. Kumar and D. Toshniwal, “A data mining framework to analyze road accident data,” *J. Big Data*, vol. 2, no. 1, Dec. 2015.

[9] M. Makaiwi, “Modeling crash frequency and severity using historical traffic and weather data: truck involved crashes on I-80 in Iowa,” p. 70.

[10] C. Lee, F. Saccomanno, and B. Hellinga, “Analysis of Crash Precursors on Instrumented Freeways,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1784, pp. 1–8, Jan. 2002.

[11] W.-H. Chen and P. Jovanis, “Method for Identifying Factors Contributing to Driver-Injury Severity in Traffic Crashes,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1717, pp. 1–9, Jan. 2000.

[12] M. Gupta, V. K. Solanki, and V. K. Singh, “Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review,” 2017, pp. 197–199.

[13] Y. Castro and Y. J. Kim, “Data mining on road safety: factor assessment on vehicle accidents using classification models,” *Int. J. Crashworthiness*, vol. 21, no. 2, pp. 104–111, Mar. 2016.

[14] T. Beshah and S. Hill, “Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia.,” in *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

[15] A. Pakgohar, R. S. Tabrizi, M. Khalili, and A. Esmaeili, “The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach,” *Procedia Comput. Sci.*, vol. 3, pp. 764–769, 2011.

[16] M. Singh and A. Kaur, “A Review on Road Accident in Traffic System using Data Mining Techniques.”

[17] A. Tavakoli Kashani, A. Shariat-Mohaymany, and A. Ranjbari, “A data mining approach to identify key factors of traffic injury severity,” *PROMET-TrafficTransportation*, vol. 23, no. 1, pp. 11–17, 2011.

[18] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edition. Morgan Kaufmann, 2012.

[19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[20] K. B. Agyapong, D. J. B. Hayfron-Acquah, and D. Michael, “An Overview of Data Mining Models (Descriptive and Predictive),” vol. 4, no. 5, p. 8, 2016.

[21] M. H. Dunham, *Data Mining, Introductory and Advanced Topics*. Pearson Education.

[22] E. A. Devi and E. J. Kaur, “A Survey on Data Mining and Its Current Research Directions,” *Int. J. Adv. Res. Comput. Sci.*, p. 5, 2017.

[23] “Weka 3 - Data Mining with Open Source Machine Learning Software in Java.” [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 10-Nov-2017].

[24] “R: What is R?” [Online]. Available: https://www.r-project.org/about.html. [Accessed: 13-Jul-2018].

[25] D. Bhosale, R. Ade, and P. R. Deshmukh, “Feature Selection based Classification using Naive Bayes, J48 and Support Vector Machine,” *Int. J. Comput. Appl.*, vol. 99, no. 16, pp. 14–18, Aug. 2014.

[26] F. A. S. Borges and R. A. S. Fernandes, “Comparison Between Random Forest Algorithm and J48 Decision Trees Applied to the Classification of Power Quality Disturbances,” p. 2.

[27] J. Hühn and E. Hüllermeier, “FURIA: an algorithm for unordered fuzzy rule induction,” *Data Min. Knowl. Discov.*, vol. 19, no. 3, pp. 293–319, Dec. 2009.

[28] “Scientific method,” *Wikipedia*. 04-Feb-2019.

[29] “Experimental Research - A Guide to Scientific Experiments.” [Online]. Available: https://explorable.com/experimental-research. [Accessed: 09-Feb-2019].

[30] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Feature Subset Selection Problem using Wrapper Approach in Supervised Learning,” *Int. J. Comput. Appl.*, vol. 1, no. 7, pp. 13–17, Feb. 2010.

[31] L. Breiman, “Random Forests, Machine Learning,” 2001.

[32] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, “Understanding variable importances in forests of randomized trees,” p. 9.

[33] E. Frank and I. H. Witten, “Generating A curate Rule Sets Without Global Optimization,” p. 8.

Appendix A: Table Details of Accident Database

Table1:AttendantCircumstances	
1	Accident_Key
2	Number_of_Vehicles
3	Number_of_Casualties
4	DSDivision
5	StationNo
6	Date
7	Time
8	TimeNew
9	Serial_No
10	UrbanRural
11	WorkDay_Holiday
12	Day_of_Week
13	RoadNumber
14	RoadStreetName
15	NearestLowerKmPost
16	DistanceLowerKmPost
17	Node_Number
18	Link_ Number
19	DistanceFromNode
20	East_coordinate
21	North_coordinate
22	Collision_type
23	Second_Collision
24	Road_Surface
25	Weather
26	LightCondition
27	Location_Type
28	Pedestrian_Location
29	Traffic_Control
30	SpeedLimitPosted
31	SpeedLimit_LightVeh
32	SpeedLimit_HeavyVeh
33	PoliceAction
34	CaseNumber
35	BReport
36	Research_Purpose
37	Export_Status
38	Highest_Severity

Table2:VehicleDetails	
1	Accident Key
2	Vehicle Reference Number
3	Element Type
4	Vehicle Registration Number
5	Vehicle Year of Manufacture
6	Age of Vehicle
7	VehicleOwnership
8	Direction of Moving
9	Driver/Pedestrian Gender
10	Driver/Pedestrian Age
11	Driver License Number
12	ValidityofLicence
13	Licence Year of Issue
14	Number of Years Since Issue
15	HumanPreCrashFactor1
16	HumanPreCrashFactor2
17	PedPreCrashFactor
18	RoadPreCrashFactor
19	VehiclePreCrashFactor
20	CrashFactorforSeverity
21	OtherCrashFactor
22	AlcoholTest
23	DriverRideratFault
24	Research Purpose

Table3: CasualtyDetails	
1	AccidentKey
2	CasualtyReferenceNumber
3	TrafficElementNumber
4	Severity
5	Category
6	CasualtyGender
7	Age
8	Protection
9	Hospitalised

Appendix B: Initial Attributes Sets for Sample Set 01

```
Relation:     Colombo Division Pedestrian Accidents 32 Attributes
Instances:    1168
Attributes:   32
              Category
              CasualtyGender
              Age
              Hospitalised
              Date
              TimeNew
              UrbanRural
              WorkDay_Holiday
              Day_of_Week
              Collision_type
              Second_Collision
              Road_Surface
              Weather
              LightCondition
              Pedestrian_Location
              Location_Type
              Traffic_Control
              SpeedLimitPosted
              Element _Type
              Direction_of_Moving
              Driver_Pedestrian_Gender
              Driver_Pedestrian_Age
              ValidityofLicence
              HumanPreCrashFactor1
              HumanPreCrashFactor2
              PedPreCrashFactor
              RoadPreCrashFactor
              VehiclePreCrashFactor
              AlcoholTest
              DriverRideratFault
              Age_of_Vehicle
              Severity
```

Appendix C: Selected Attributes Sets for Sample Set 01

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

number of folds (%)  attribute
           0(  0 %)     1 Category
           6( 60 %)     2 CasualtyGender
           7( 70 %)     3 Age
          10(100 %)     4 Hospitalised
           6( 60 %)     5 Date
           7( 70 %)     6 TimeNew
           6( 60 %)     7 UrbanRural
           4( 40 %)     8 WorkDay_Holiday
           5( 50 %)     9 Day_of_Week
           0(  0 %)    10 Collision_type
           5( 50 %)    11 Second_Collision
           7( 70 %)    12 Road_Surface
           8( 80 %)    13 Weather
           8( 80 %)    14 LightCondition
           9( 90 %)    15 Pedestrian_Location
           8( 80 %)    16 Location_Type
           5( 50 %)    17 Traffic_Control
           8( 80 %)    18 SpeedLimitPosted
           8( 80 %)    19 Element _Type
           7( 70 %)    20 Direction_of_Moving
           9( 90 %)    21 Driver_Pedestrian_Gender
           3( 30 %)    22 Driver_Pedestrian_Age
           7( 70 %)    23 ValidityofLicence
           8( 80 %)    24 HumanPreCrashFactor1
           4( 40 %)    25 HumanPreCrashFactor2
           5( 50 %)    26 PedPreCrashFactor
           4( 40 %)    27 RoadPreCrashFactor
           2( 20 %)    28 VehiclePreCrashFactor
           8( 80 %)    29 AlcoholTest
           6( 60 %)    30 DriverRideratFault
           7( 70 %)    31 Age_of_Vehicle
```

Appendix D Initial Attributes Sets for Sample Set 02

```
Relation:     Colombo Division Accident Severity
Instances:    1228
Attributes:   26
              Number_of_Vehicles
              Number_of_Casualties
              Date
              TimeNew
              UrbanRural
              WorkDay_Holiday
              Day_of_Week
              Collision_type
              Second_Collision
              Road_Surface
              Weather
              LightCondition
              Location_Type
              Pedestrian_Location
              Traffic_Control
              SpeedLimitPosted
              Protection
              Hospitalised
              Element _Type
              DriverRideratFault
              AlcoholTest
              HumanPreCrashFactor1
              HumanPreCrashFactor2
              ValidityofLicence
              Age_of_Vehicle
              Highest_Severity
```

Appendix E: Selected Attributes Sets for Sample Set 02

```
Attribute selection 10 fold cross-validation (stratified), seed: 1

number of folds (%)  attribute
           0(  0 %)     1 Number_of_Vehicles
           5( 50 %)     2 Number_of_Casualties
           3( 30 %)     3 Date
           6( 60 %)     4 TimeNew
           3( 30 %)     5 UrbanRural
           5( 50 %)     6 WorkDay_Holiday
           6( 60 %)     7 Day_of_Week
           5( 50 %)     8 Collision_type
           5( 50 %)     9 Second_Collision
           7( 70 %)    10 Road_Surface
           6( 60 %)    11 Weather
          10(100 %)    12 LightCondition
           4( 40 %)    13 Location_Type
           9( 90 %)    14 Pedestrian_Location
           2( 20 %)    15 Traffic_Control
           6( 60 %)    16 SpeedLimitPosted
           0(  0 %)    17 Protection
          10(100 %)    18 Hospitalised
          10(100 %)    19 Element _Type
           5( 50 %)    20 DriverRideratFault
           6( 60 %)    21 AlcoholTest
           9( 90 %)    22 HumanPreCrashFactor1
           5( 50 %)    23 HumanPreCrashFactor2
           3( 30 %)    24 ValidityofLicence
           2( 20 %)    25 Age_of_Vehicle
```

Appendix F: Measuring Attribute Importance

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -attribute-
importance -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     Colombo Division Pedestrian Accidents
Instances:    1274
Attributes:   11
              Hospitalised
              Weather
              LightCondition
              Pedestrian_Location
              Location_Type
              SpeedLimitPosted
              Element_Type
              Driver_Pedestrian_Gender
              HumanPreCrashFactor1
              AlcoholTest
              Severity
Test mode:    evaluate on training data

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-
check-capabilities

Attribute importance based on average impurity decrease (and number
of nodes using that attribute)

      0.58 (  3017)  Pedestrian_Location
      0.51 (  1187)  Element_Type
      0.44 (  3106)  LightCondition
      0.38 (  3879)  HumanPreCrashFactor1
      0.37 (  3108)  Location_Type
      0.34 (  2719)  AlcoholTest
      0.33 (  1853)  Hospitalised
      0.33 (   627)  Weather
      0.31 (  1218)  SpeedLimitPosted
      0.2  (   495)  Driver_Pedestrian_Gender


Time taken to build model: 0.22 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===
Correctly Classified Instances        1025               80.4553 %
Incorrectly Classified Instances       249               19.5447 %
Kappa statistic                          0.6556
Mean absolute error                      0.21
Root mean squared error                  0.3003
Relative absolute error                 53.8434 %
Root relative squared error             68.0121 %
Total Number of Instances             1274

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
MCC      ROC Area  PRC Area  Class
                 0.679    0.035    0.796      0.679    0.733
0.687    0.962     0.848     1
                 0.681    0.082    0.759      0.681    0.718
0.620    0.920     0.840     2
                 0.903    0.242    0.825      0.903    0.862
0.674    0.929     0.944     3
Weighted Avg.    0.805    0.163    0.802      0.805    0.801
0.661    0.932     0.899

=== Confusion Matrix ===

   a   b   c   <-- classified as
 144  26  42 |   a = 1
  18 239  94 |   b = 2
  19  50 642 |   c = 3
```

Appendix G: FURIA Rule Extraction

```
=== Run information ===

Scheme:       weka.classifiers.rules.FURIA -F 3 -N 2.0 -O 2 -S 1 -p
0 -s 0
Relation:     Colombo Division Pedestrian Accidents
Instances:    1274
Attributes:   11
              Hospitalised
              Weather
              LightCondition
              Pedestrian_Location
              Location_Type
              SpeedLimitPosted
              Element_Type
              Driver_Pedestrian_Gender
              HumanPreCrashFactor1
              AlcoholTest
              Severity
Test mode:    10-fold cross-validation

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

FURIA rules:
===========

(Element_Type = 00) => Severity=1 (CF = 0.71)
(Element_Type = 09) and (Pedestrian_Location = 2) and
(LightCondition = 1) and (HumanPreCrashFactor1 = 02) => Severity=1
(CF = 0.71)
(Pedestrian_Location = 3) and (AlcoholTest = 1) and (LightCondition
= 1) and (Element_Type = 05) and (HumanPreCrashFactor1 = 02) =>
Severity=1 (CF = 0.69)
(LightCondition = 4) and (Pedestrian_Location = 1) and
(HumanPreCrashFactor1 = 02) and (Location_Type = 1) => Severity=1
(CF = 0.69)
(Element_Type = 07) => Severity=1 (CF = 0.7)
(Hospitalised = 1) and (HumanPreCrashFactor1 = 01) and (AlcoholTest
= 1) and (Pedestrian_Location = 1) => Severity=2 (CF = 0.69)
(Hospitalised = 1) and (HumanPreCrashFactor1 = 01) and (Element_Type
= 02) => Severity=2 (CF = 0.6)
(Hospitalised = 2) => Severity=3 (CF = 0.9)
(LightCondition = 1) => Severity=3 (CF = 0.61)

Number of Rules : 9
```