Road Traffic Accident Analysis Using Machine Learning Techniques for Soshanguve, Pretoria

Mpho Mokoatle

orcid.org NWU- **00392-18-A9**

Dissertation submitted in fulfillment of the requirements for the degree *Master of Computer Science* at the North West University

Supervisor: Prof BM Esiefarienrhe

Co-supervisor: Dr VN Marivate

Co-supervisor: Mrs TL Letlonkane

Graduation ceremony: July 2019

Student number: 24037958

Acknowledgments

I would like to express my appreciation to everyone who played a role in ensuring that the proposed outputs in this work materialize:

- To Prof. Michael Esiefarienrhe Bukohwo in his capacity as a supervisor, thank you for all your hard work, support, advice, and wisdom.
- To Dr. Vukosi Marivate in his capacity as a co-supervisor, thank you for grooming me and guiding me throughout this study.
- To the CSIR, thank you for providing me with all the resources I needed to complete this work.

Table of Contents

Abstract

Road traffic accidents (RTAs) in South Africa reached the highest road death toll in 2017, in spite of road safety campaigns and initiatives. " Ongoing campaigns are simply not sufficient'', said a representative from the South African Automobile Association (AA). RTA data is usually collected at accident scenes and those who collect this data lack sufficient knowledge and skill to translate the data into knowledge that can be used to gain a better understanding of the root causes and factors associated with the occurrence of an accident. The South African literature on RTAs has shown a limitation in using advanced methods such as machine learning, to extract insight from RTA data or trace patterns and trends that are associated with the occurrence of an accident. In this work, machine learning methods were deployed to study the relationships that exist in data that is captured on South African Accident Report (AR) forms. An AR form is a form that is completed for all RTAs that occur on a public road where a motor vehicle was involved [1]. In order to increase the data, distances to the nearest places of interest such as bars, malls, schools, restaurants, and buildings were extracted through a geospatial database and added to the AR data. The reason behind this was to determine if distances to the nearest places of interest have an impact on the injury severity of drivers in RTAs. First, the main characteristics in the data were summarized by performing the exploratory data analysis. Upon completion of the exploratory data analysis, it was found that truck license holders particularly code ***C1***, used light vehicles such as motor cars, in comparison to heavy motor vehicles. The results from the exploratory data analysis also revealed that these drivers sustained the most severe injuries in RTAs, different from light motor vehicle drivers. In South Africa, duty licenses are granted with various codes that indicate the kind of vehicle that may be used with that duty license; the codes are shown in **Appendix A**. It should also be noted that the tests for each license code are conducted differently using different vehicles. For example, when testing for a heavy duty license code ***C1***, the test is conducted on a vehicle with a Gross vehicle mass (GVM) of 3500 kg and less than 16000 kg, and when testing for a light motor vehicle duty license code ***B***, the test is conducted on a vehicle with a GVM of $\leq$ 3500 kg. To determine if truck licenses code ***C1*** and the distances to the nearest places of interests such as malls, bars, schools, restaurants, and buildings have a high importance on the injury severity of drivers in RTA, three classifiers were created by using parametric and non-parametric machine learning algorithms namely;

Multivariate Logistic Regression (MLR) and the Extreme Gradient Boosting Tree (XGBoost), where XGBoost outran MLR. The first classifier was created using the extracted distance features and the target class (injury severity), the second classifier was created using the initial data that is collected on AR forms and the third classifier was created by integrating engineered distance features and data that is collected on AR forms. This model achieved an accuracy of 83.14%±3.34 %, and a precision, recall, and an F1 score of 82.83%±3.18 %, 82.66 %±3.16 %, and 82.35%±3.25 %, respectively. Also, the most significant predictors of injury severity of drivers in RTAs were found to be truck licenses code C1, light motor duty license code EB, vehicle type (motor car or station wagon), single vehicle: overturned accident type, vehicle maneuver and the distance to the closest building. There are several mitigation strategies that can arise as a result of confirming whether or not the distances to the nearest places of interest, or if the kind of duty license that a motorist has have a high importance on injury severity. For example, if the type of duty license and the distances to the nearest places of interest have a significant impact on the level of injury in RTAs, than it may be useful to explore how adjusting the current status quo will improve safety on the road which states that motorists with heavy duty licenses are allowed to use light motor vehicles. Moreover, if closest places of interests also play a role in the injury severity of drivers in RTAs, this information can direct policymakers to areas of high accident occurrences and proactive measures can then be taken such as to create a road safety awareness down the affected line i.e, N1, N14; allocate more funds to improve the road, ensure that medical services are close by to provide optimal treatment of rehabilitation following the injury such as effective first aid and appropriate care, and also increase traffic personnel in the affected area.

The study also searched for frequent attributes that co-exist in the incidence of a RTA by applying the association rule mining technique. Before searching for frequent items that co-exist in the data, the clustering was performed as a preliminary step. This resulted into having two clusters and the support, confidence, and lift of the rules found in the first cluster were 0.21, 0.71 and 2.05 respectively. Similary, the support, confidence, and lift of the rules found in the second cluster were 0.22, 0.71 and 2.35 respectively

List of Tables:

List of Figures

List of abbreviations

AA	Automobile Association
ANN	Artificial Neural Network
CNN	Convolutional neural network
CHAID	Chi-square Automatic Interaction Detector
DM	Data Mining
DS	Data Science
FARS	Fatality Analysis Reporting System
KDD	knowledge discovery in databases
K-NN	K-nearest neighbor
NNs	Neural Network(s)
MAR	Missing at random
MNAR	Missing not at random
MCAR	Missing completely at random
MI	Multiple imputation
SADC	South African Development Community
LR	Logistic Regression
AARTO	Administrative Adjudication of Road Traffic Offences Act
RTMC	Road Traffic Management Corporation
CART	Classification and Regression Tree
PDO	Property Damage Only
PAM	Partitioning around medoids
MCA	Mobility Centre for Africa
MPVs	Multi-Purpose vehicles
MLR	Multivariate logistic regression
GVM	Gross vehicle mass
GCM	Gross combination mass
RTAs	Road traffic accidents

GIS	Geographic information system
GPS	Geographic positioning system
RTMC	Road Traffic Management Corporation
SUV	Sport Utility vehicle
SMOTE	Synthetic Minority Oversampling Technique
SAPS	South African Police Services
TW	Tare weight
ROC	Receiving operating characteristic
FP	False positive
TP	True positive
OSA	Obstructive sleep apnea
XGBOOST	Extreme gradient-boosting tree

List of publication(s) and poster(s)

1. International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD 2018), paper title: ‘’ Collision Course: Challenges with Road Traffic Accident Data in South Africa’’
2. Poster presentation at the Black in AI workshop 2017, California USA.
3. Poster presentation at the Deep Learning Indaba 2018, Stellenbosch SA.

Chapter 1: Introduction

Road accidents in South Africa have reached epidemic proportions. A former police chairman, Howard Dembovsky, said the injuries and fatalities of road crashes are a national catastrophe that the government and citizens are not taking seriously. " More people die every day on South African roads than in the Middle East in war zone countries, per day!" he ridiculed [2]. The challenge of RTAs is significant because it costs the South African economy a fortune. For example, in 2015, the expenses of crashes were approximated to be R143 billion [3]. When an individual is involved in an accident, if serious injuries are sustained, it is an expense to manage the individuals in hospitals. After a thorough literature read of RTAs in South Africa, it was gathered that RTAs have mostly been studied or investigated using statistical methods. However, this is not sufficient if the relationship that is being investigated is too intricate or if the data is too cumbersome. Moreover, RTA data is commonly collected at accident scenes yet it wastes away in data warehouses due to insufficient training and skill. For example, law enforcement personnel such as police officers are only trained on how to collect AR data but are not trained to perform any data analysis on the data. The RTA data includes information about the individual(s) that were involved in the RTA, the extent of the injury, description of the accident, vehicle characteristics, and light, weather and road conditions. The full list of the description of the data is found in **Appendix B**. Collecting such data is important to ensure that the root causes of accidents and where they happen are explained and understood. As such, planning mitigation strategies at national and local levels to address the outcomes of injuries sustained in RTAs is heavily dependent on this data [4]. With the recent developments in the internet, data collection, data storage, retrieval, need for computational power and business insight, this gave birth to the evolution of Data Science (DS). DS has its origin in Statistics and has been applied to various disciplines such as medicine, engineering and social sciences [5]. DS is an interdisciplinary combination of inference, algorithm advancement, and technology that is meant to solve complex tasks and strives on using data in creative ways to generate useful insight [6]. One of the most frequently used DS methods is modern neural networks (NNs) and decision trees.

NNs are non-linear statistical data modeling techniques that are usually used to model complicated relationships between inputs and outputs. NN have an advantage over regression

analyses in that in regression analysis, the analyst has to choose a model to fit the data while in NN, pre-selection of models are not required. Moreover, in NN you are given the ability to scale hidden layers to provide more desired accuracy [7]. Decision trees are also one of the DS models. A decision tree is a predictive model which can be used to imitate both classifiers and regression models whereas, in operations research, decision trees invoke a hierarchical model of decisions and their consequences. When a decision tree is used for classification problems, it is called a classification tree and when it is applied in regression problems, it is called a regression tree [8]. In this work, some DS methods were used to discover insight from road traffic accident data that was granted by the South African Police Services (SAPS) in a small township in South Africa, Soshanguve. Attempts were made to get more data from multiple cities/towns across South Africa, but this was not successful. Future works would involve reusing the model on different cities/towns, to determine if the same behavior will be observed as in Soshanguve. Next, the problem statement is introduced.

1.1 Problem Statement

RTAs are one of the most contributing factors of death and disability on the continent. Predictions suggest that RTAs will surpass malaria and HIV/AIDS as sources of death in 2020. As reported by Road Traffic Management Corporation (RTMC), 90% of road accidents in South Africa are the end-product of anarchy, they are predictable and preventable. South African minister of transport, Ms. Dipuo Peters announced that South Africa had a road death rate of 23.5 per 100 000 people in 2014 although the global average is 17.4 fatalities per 100 000 people [9]. The minister also added that road deaths have not decreased at the rate required to meet international goals as set out by the United Nations Decade of Action for Road Safety 2011-2020. The South African Road Safety Strategy 2011-2020 report stated that former road safety strategies have not accomplished their set goals. This is because of a number of factors namely: the strategy was too extensive in its focus; preference was not given to easily fixable problems and adequate resources were not set aside to realize the application of the identified strategies within the three spheres of Government.

The South African Department of Transport has put forward measures to combat road carnage. Amongst the interventions includes:

- The Railway Level Crossing Unit: This program promotes and ensures safety railway level crossings.
- Enhancement of compliance: The inauguration of the Administrative Adjudication of Road Traffic Offences Act (AARTO) that improves road traffic quality by producing a scheme to discourage road traffic violations and also to assist the progress of adjudication of road traffic violations.
- Fighting Fraud and Corruption: The formation of the National Traffic AntiFraud and Corruption Unit within the RTMC to fight acts of fraud and dishonesty by working together with other law enforcement agencies has developed into various prosecutions for illegal acts across the traffic setting.
- Law Enforcement in SADC (South African Development Community): The Cross-Border Road Transport Agency is administered to facilitate the unlimited flow of passengers and goods within the SADC region [9].

The above interventions do not include addressing or evaluating the RTA challenge using modern machine learning interventions. In this work, machine learning techniques are applied to RTA data that is captured on AR forms. The aim is to use these techniques to extract insight that may be used to implement enhanced road safety strategies that are meant to improve safety on the road. Next, research questions and objectives will be outlined.

1.2 Research Questions

RQ1: What are the main features in the given accident data? What can be asked from the data preceding modeling or hypothesis testing?

RQ2: Are there any co-occurring frequent item sets prior to the incidence of a road traffic accident?

RQ3*:* What are the significant predictors and/or spatial features that contribute to the injury severity of drivers in RTAs?

1.3 Research Aim and Objectives

1.3.1 Aim

The aim of this research is to use machine learning algorithms to discover patterns and trends that are associated with the incidence of an accident using data that is obtained from Soshanguve, Gauteng Province of South Africa.

1.3.2 Objectives

RO1: Explore and describe interesting features of the data by performing exploratory data analysis.

RO2: Perform clustering as a preliminary task for data segmentation to remove data heterogeneity thereafter, apply association rule mining to each cluster to identify frequent itemsets that are associated with the incidence of an accident.

RO3: Extract spatial features using a geographic information system (GIS) and consolidate it with the original accident data to create a classification model that predicts injury severity of drivers in RTAs.

1.4 The Significance of the study

The findings of this study will reveal how machine learning techniques to address the challenge of RTAs in South Africa. Since this research focuses on a specific area-Soshanguve, it will reveal co-occurring factors associated with the incidence of accidents in this particular area and extract useful predictors that have a high influence on injury severity.

1.5 The limitations of the study

This study makes use of a small sample size and as such, the results may not be a general representation of other parts of the country. However, the study aims to increase the sample size by extracting more spatial data using a geospatial database. Another limitation is that the study is executed on secondary data (data was collected by other people and not the researchers.) and only focuses on drivers or motorists on the road and excludes other road users, i.e passengers, pedestrians or cyclists. The reason for this is that data about the other road users was not

consistently measured or recorded during the data collection processes. This might have been caused by a procedural issue during the data collection process at the police station and/or accident scene(s).

1.6 Ethical Clearance

The application for ethical clearance was submitted and approved the North-West University, as the data included sensitive information about human subjects. The ethical clearance certificate was issued and the ethical clearance number is **NWU-00392-18-A9**. Permission to use the data was granted by the Soshanguve SAPS through an approval letter dated 22 May 2017.

1.7 Summary of the chapter

In this chapter, the application area and the research argument was introduced, followed by research questions and objectives. In the following chapter, the state of the art and how the study aims to add to the body of knowledge in the RTA domain will be discussed.

Division of chapters

The general subject of the research will be explained in the ***Introduction*** chapter, then narrowed down to the main argument of the study. In this chapter, the problem, context, limitations, and importance of the study will be presented. Following this, the current state of knowledge in the field will be explained in the ***Literature Review*** chapter, where important topics such as the application of logistic regression (LR), decision trees, and NNs in RTAs will be discussed. The next chapter will involve giving a description of the methods that will be used to conduct the research. Specifically, the chapter will discuss themes such as the chosen research design and ML algorithms that will be applied. This chapter is discussed in the ***Methodology*** chapter. Then, the research findings will be reported and interpreted in the ***Results and Discussion*** chapter. The study will then conclude and summarize the research by outlining the research recommendations and present potential avenues for future research in the ***Conclusion, Recommendations and Future work*** chapter.

Chapter 2: Literature review

RTAs continue to be a national challenge. In 2016, statistics from the RTMC show that road fatality increased by 9% from the previous year 2015. That is a total of 14 071 road fatalities. The RTMC is the biggest South African road safety entity that is responsible for generating, combining, reporting and analyzing RTA data across South Africa [10], [11]. In South Africa, statistical tools have been used to achieve a better understanding of the factors and causes that are associated with the occurrence of an accident. However, these tools do not reveal enough knowledge if the dependency that is analyzed is too intricate or when the dataset is too large [12]. In preceding years, researchers around the world have investigated RTAs using DS techniques. DS is connected to Big Data as it is an interdisciplinary field that combines various fields including but not limited to software engineering, data engineering, business intelligence, computer science, statistics and so on [13]. Pattern recognition is one of the DS application areas. Pattern recognition embroils assuming or predicting the occurrence or non-occurrence of a phenomenon or natural event, for example classifying an observation as either blue or white, true or false, real or fake [14]. This is the exact task that a pattern recognition algorithm does: classifying a set of elements on the basis of a specific condition [15]. While pattern recognition takes its basis from engineering, machine learning is an approach of data analysis that uses algorithms that learn from data to discover hidden knowledge [16] and takes its basis from computer science. Nonetheless, these concepts can be merged into a single field [17].

Now that the origin of the field applied in the study and the problem area has been introduced, prior work that has been carried out to address RTAs will now be discussed. First, international literature is discussed by reviewing applied research in RTAs using logistic regression (LR) methods, association rule mining, artificial intelligence, and classification and regression trees (CART) methods. Then, South African literature on the RTA domain will be discussed where a review will be given on how RTA data is collected. The data quality problems that arise in this way of storing or manipulating data will then be discussed.

2.1 RTA research using Logistic Regression (LR)

To have a better understanding of the risk of pedestrian fatality in public traffic in China, earlier publications [18] focused on investigating the link between impact speed and risk of pedestrians in passenger vehicle collisions. The sampling population consisted of independent variables such as the accident type, the age of the pedestrians, severity, vehicle type (Sport Utility vehicle (SUVs) or Multi-Purpose vehicles (MPVs)); and the impact speed was then treated as the target class. The results from the LR models showed that the risk of pedestrian fatality increases as the impact speed increases. To study the strength of the correlations amongst features, the author(s) [18] used the Wald Chi-Square test. To have more substantial correlation findings, the above-mentioned author(s) could have used more robust statistical tests such as t-test or Phi. Also, the data that was used in the study had a variety of features such as consultations that were carried out between the victims that were involved in the accident and data gathered from emergency rooms. It would have been useful to explore how these features contribute to the risk of pedestrian fatality in the urban public traffic in China because usually, using more features to build statistical models or machine learning models help to improve model performance and accuracy. To correctly define crash-types in the United States, a ‘’speeding-related’’ framework or method is used. This framework is divided into two subgroups: ‘’exceeding the recommended speed limit’’ or ‘’driving too fast for conditions’’. However, the quality or effectiveness of this framework has not been evaluated before. As such, the author(s) [19] studied the quality of this framework by creating LR models. The models were then able to label crash-types that were not initially defined as ‘’speeding-related’’ but had crash evidence that suggested “speeding” as a causative factor. The author(s) also found that the ‘’driving too fast for conditions’’ code was used in three distinct situations and instead of using the ‘’exceeding the recommended speed limit’’ code, the ‘’driving too fast for conditions’’ code was also used. In addition to validating the models, a Hosmer-Lemeshow test was used. A Hosmer-Lemeshow is a test used to evaluate how well a model fits a given set of observations. It investigates if the real measurements of observations are the same to the predicted probabilities of observations in groups of the dataset by employing a Pearson’s chi-square test. The author(s) also took the models through a blind-reviewed process consisting of crash narratives.

Accidents do not just occur in random places at random times, meaning that the incidence of an accident is also influenced by a spatial setting of an environment [20], where the time of the incidence of an accident is also a factor. In order to improve road safety and identify areas that have a high accident concentration in 10 roads in Beijing city, the author(s) [21] built LR models that could predict accident hotspots. Moreover, the author(s) also ran statistical significance tests to determine important attributes that led to a traffic accident. To evaluate the performance of the prediction model, the author(s) created a plot of correctly and incorrectly predicted or observed hotspot locations. However, there are more improved ways of evaluating the performance of a prediction model. One of the most well-known methods is the ROC (receiving operating characteristic) curve. ROC curve plots the True Positive (TP) rate against the False Positive rate (FP) for several cut-off points of an element. Each element on the ROC curve is a representation of the TP/FP pair associated with a specific decision threshold. The area under the ROC curve is an indication of how well an element differentiates between two groups [22].

Prior work has suggested that driving under the influence of drugs and/or alcohol increases the likelihood of drivers sustaining serious injuries or being killed in RTAs. To further support or validate this, a study [23] was carried out to investigate the impact of alcohol and drug use in motor vehicle crashes. The study used a fatally-injured dataset from three Australian states. To determine if the motorists were guilty or not guilty, a LR model was created and the independent variables were age, gender, crash type, and alcohol and drug use; where ‘’guilty’’ was the target class. The author(s) then found that motorists who tested positive for psychotropic drugs had a higher likelihood of being ‘’guilty’’ than motorists who tested negative. The author(s) also deducted that motorists who had a high ‘’guilty’’ rate were under the 25 and greater than the 65 age group. Furthermore, the author(s) concluded that the top three drugs that contributed to driver severity were tetrahydrocannabinol, amphetamines and a mixture of psychoactive drugs.

To clear out RTA scenes, law enforcement agencies such as the police or towing companies need to be alerted. To add to this challenge, vehicles near the area of the accident scene end up using the affected accident route assuming that the bottleneck is caused by traffic congestion. Having said this, work in [24] studied accident data to detect RTAs and automatically send

alerts to law enforcement agencies notifying them of the accident. To achieve this objective, the author(s) supplied the data to machine learning algorithms namely: LR, bagging classifier, AdaBoost classifier, voting classifier, and trivial classifier. Following this, the author(s) then evaluated the performance of each machine learning algorithm to accurately detect RTAs. The author(s) then concluded that the classifier that outperformed other classifiers was the AdaBoost classifier with an accuracy of 85%. The key difference between work in [23] and [24] is that the author(s) in [24], first split their data into training and test sets before creating LR models. To evaluate the performance and accuracy of a machine learning model, it is important to test the model on a test dataset.

The author(s) [25] attempted to identify risk factors that are associated crimes with accidents and the characteristics of these individuals. To achieve this task, the author(s) used descriptive analysis and created LR models to identify the risk factors. Similar to [23], the author(s) also identified alcohol consumption as one of the leading causes of road crashes. The author(s) also discovered that in road crashes, seat belts were 50% less likely to be worn and that the most significant causes of death were drunk driving, high speed, vehicle failure, and motorist's negligence. Human factors are responsible for a large portion of RTAs. Previous studies have discovered that mental disorders [26], [27], [28] increase the rate at which road accidents occur. The author(s) [28] investigated the impact of predictors such as personality traits, driving behavior, mental illness and their contribution to RTAs. The study population included bus and truck motorists that were involved in accidents and those that were not involved in accidents. To collate their findings, the author(s) performed descriptive analysis and MLR models. The author(s) discovered that depression, anxiety, and neuroticism were among the most significant predictors in road accidents. In [25], the main causes of death in road safety are due to drunk driving, high speed, vehicle failure, and negligence. In [28], the main causes of death in road safety are depression, anxiety, and neuroticism.

Author(s) [29] investigated the relationship between the use of cellphones and traffic accidents in motor vehicles. To carry out their investigation, author(s) used two groups of individuals: those who were involved in an accident and those who were not involved in an accident. Through the use of a LR model, the author(s) were able to discover that talking on the cell phone while driving increases RTA risk. The author(s) also discovered that cognitive activities

such as thinking of problems and watching scenery also increased RTA risk. Before building a model, it is important to discard all the insignificant attributes that will only add noise to a model. As such, the author(s) here removed all the insignificant attributes to arrive at a more robust and reliable model that yielded important results.

Predicting the driving ability in patients with obstructive sleep apnea (OSA) has been very difficult because it is unknown if computer-based driving simulators can predict patients that are at more risk. The author(s) [30] investigated if using data that is derived from a computer-based simulator generated more information compared to information obtained from history and if whether or not their study might be useful to advise OSA patients about driving. The results from the first LR model revealed that older patients, that are female and alcohol consumption had an influence on the patient's performance on the simulator. The second LR model was created to investigate if clinical history, sleep study results, and data obtained from the computer-based driving simulator were helpful in predicting individuals with OSA as having had an RTA. The LR model could only classify 100% of individuals who did not have an RTA and only 10% of the individuals who had the RTA could be classified. This was clearly caused by an imbalance in the data. Class imbalance occurs when the labels in the target class are not represented equally [31]. When class imbalance exists in the data, the machine learning model gets biased and cannot correctly classify the minority class thus leading to inaccurate results. It is therefore important to correct class imbalance prior to model building.

Udine, Italy, has the highest fatality rate in RTAs across Italy. The author(s) [32] used LR models to study the factors that are associated with RTAs. The findings from the study revealed that the risk of fatal accidents is lower in females than in males. Compared to individuals who are aged < 30, individuals aged >= 65 had a higher fatal injury risk as pedestrians, motorists, moped riders, and cyclists. In RTAs that happened between 1:00 hrs and 5:00 hrs, fatality risk was higher than from 6:00 hrs and 11:00 hrs between pedestrians, motorists, cyclists and moped riders. The study also found that the risk of death is significantly higher on roads outside the urban area. Also, the injury of motorists was closely connected with seatbelts that are not worn.

To investigate the consequences after the occurrence of a RTA, a study [33] was carried out to investigate the psychological and social outcome at three months and one year after the occurrence of a RTA. Through the use of a LR model, the findings showed that after one year,

most people reported severe physical problems and a small portion of the study population reported psychiatric consequences. The author(s) then concluded that there is an important need to make changes in the medical care and socio-legal policy to recognize and remedy chronic problems. Similar to this work [33], work in [34] also investigated the consequences after having suffered an RTA. Some people reported psychiatric challenges as in [33] and moderate or severe pains after three years of having suffered a RTA. The study population here also reported physical problems as in [33]. Again, the author(s) [33] also maintained that there is a need for changes in clinical care and socio-legal procedures.

2.2 RTA research using Association Rule Mining

When investigating frequent items that co-occur in datasets, analysts often apply association rule mining. Assume you are a manager in a grocery store and want to determine which items are bought in-conjunction by customers. To answer this question, you would use an association rule mining technique. This technique was introduced in (Agrawal et al. 1993). Its objective is to illuminate significant correlations, frequent patterns, and but not limited to, associations among observations within a transaction database [35].

Kumar and Toshniwal [36] used association rule mining to investigate frequent itemsets that occur together in the incidence of a RTA. The author(s) hypothesized that in order to find more significant rules within a dataset; data segmentation must be performed as a preliminary step. Decreasing data heterogeneity is an important concept in DS and statistical methods. Heterogeneity and its opposite, homogeneity, is how constant a variable relationship is. Removing heterogeneity removes noise and as such, sensitivity for the target class is then improved [37]. As such, prior to applying the association rule algorithm; the author(s) first segmented the data into clusters by using K-modes clustering. K-modes clustering is different from K-means clustering as it is used to cluster categorical data. K-modes clustering was first introduced in 1998 by Huang. K-modes use a matching dissimilarity measure for categorical data. Instead of using means, it uses modes for clusters and a frequency-based method to renew modes in the clustering procedure to reduce the cost of the clustering function with the matching dissimilarity measure for categorical data. These adjustments to the K-means algorithm enable k-modes to be able to cluster large categorical data [38]. Nonetheless, the

results showed that executing clustering prior to applying the association rule mining algorithm yielded more important rules that would have remained undisclosed if data segmentation was not executed as a preliminary step. The author(s) also enforced the algorithm onto the un-segmented data and thus were able to further validate and prove their hypothesis. In addition to their findings, the author(s) performed trend analysis on monthly and hourly road accident counts for each cluster.

The author(s) [39] investigated high-frequency accident locations by using association rule mining. By using this technique, the author(s) were able to discover frequent itemsets that occur together in high-frequency accident locations. This procedure was then repeated on low-frequency traffic accident locations. Following this, the results from the high-frequency and low-frequency traffic accident locations were then compared. The author(s) found that human and behavioral factors are significant to analyze frequent item sets occurring in RTAs. The key difference between the low-frequency and high-frequency accident locations was mainly with the infrastructure and location of the traffic accident location.

The key similarities between prior publications [36], [39]; are that the author(s) executed data segmentation as a preliminary step before applying the association rule algorithm. The rules found were significant and had lift values greater than 1. A lift value is a measure that tells us about the goodness of a rule. With a rule of lift value greater than 1, this indicates that the appearance of A and B together is more expected whereas a lift value that is less than 1 suggests the contradictory concept.

Rail safety policymakers need to ensure that freight and passengers arrive safely to their destinations. Developing rail safety standards require a strong knowledge base that can be used to support decision making and improve safety standards. To help rail policymakers achieve this objective, the author(s) [40], focused their study in the rail safety domain by analyzing rail data of past accidents in Iran. The main similarity between [36], [39] is that the author(s) here also applied association rule mining to discover patterns within the rail data. The author(s) were then able to discover that the most contributing factors to the incidence of an accident were human-error, wagon, and track. To evaluate the ‘’interestingness’’ of a rule, the author(s) only considered the support and confidence of a rule. The support of a rule is the portion of the transactions that include all elements that are in A and B. The confidence of a rule is the

conditional likelihood, calculated as the portion of transactions including both A and B. One of the challenges in applying the association rule mining technique to real datasets is knowing which values (high or low) to set for the support constraint. A high support constraint bypasses combinational explosion in unearthing frequent item sets, but at the cost of excluding important patterns of low support. Nevertheless, rules with high support are evident and well-known and it is rules with low support that produce new knowledge [41]. Used independently, these two parameters are not sufficient to indicate that the rule is significant. It is important to also consider the lift of the rules. Another way to measure the goodness of a rule is by calculating its strength.

Preceding work [42] evaluated the relationship between RTA attributes and the causes of RTAs. Two research approaches were used; the rough set theory and association rule mining. To understand the causes of death in road safety, four aspects were studied: driver factors (fatigue, alcohol/drug effects, mood etc), vehicle factors (steering system, brake system, electrical system etc), road factors (geometry, road conditions etc), environmental factors (climate conditions, land-use and so on). If a study used the association rule mining technique, it is expected that the description of the rules found are explained or described along with their support, confidence and lift values. In this study, this information was missing. No knowledge was deduced (if any) from the rules and the results were too abstract. However, the two theories that were used in the study were well discussed.

As mentioned earlier on that accidents tend to cluster at specific areas, a study [43] investigated which attributes frequently appear together in high and low accident zones through the use of the association rule mining concept. These frequent itemsets found in high and low accident zones were then comparatively analyzed. It was found that in high accident zones, items that frequently occurred together were “left turns at signalized intersections, collisions with pedestrians, loss of vehicle control and rainy weather’’. Frequent itemsets that occurred in low accident zones were “head-on collisions and drunken road users’’. The rules had high lift values which emphasized their significance.

If the comparison is done between [42],[43]; in [42], there was no information about the evaluation metrics that were gathered from the rules such as support, confidence and lift values.

Also, the rules were not explicitly explained as in [43]. As a result, it was difficult to deduce significant knowledge about RTAs.

A study [44], similar to [43], investigated the frequent items that occur together in low and high accident zones. The main difference between [43] and [44] is that in [44], clustering was performed on the dataset as a preliminary task before frequent item sets could be mined. K-means clustering was used to find the clusters within the data and categorized the data in three groups: high-frequency, moderate frequency and low-frequency accident zones. Rules generated for high-frequency accident zones showed that intersections on highways are more unsafe for all types of accidents. Also, high-frequency accident zones typically involved two-wheeler accidents in regions with hills. In moderate frequency accident zones, colonies surrounding local roads and intersections on highways are unsafe for pedestrian-hit accidents. Low-frequency accidents are spread across the district and most of these accidents are not critical.

In a similar analysis [45] as in [44], frequent items that occurred together in RTAs were also mined. To overcome data heterogeneity, the author(s) here also applied K-means clustering as in [44] and partitioned the data into clusters. The key difference between the two publications [44] and [45] is that the author(s) [45], applied the association rule mining on the entire un-partitioned dataset and also on the clusters. The author(s) then stated that combining K-means clustering and association rules revealed significant knowledge. In cluster analysis, choosing the optimal number of clusters $\boldsymbol{K}$ to be used to partition the data is a very critical process. Choosing the number of $\boldsymbol{K}$ will have a direct impact on the strength of the clustering results. In this paper [45], the author(s) chose $\boldsymbol{K}$=5 as the optimal number of clusters. However, according to their "within-cluster sum of squares" plot, $\boldsymbol{K}$=2 should have been the optimal number of clusters since it exhibited less variability (the observations are closest to each other when $\boldsymbol{K}$=2, in comparison to when the cluster size $\boldsymbol{K}$ is equal to 5.). Generally speaking, a cluster that has the least sum of squares is better than a cluster that has a large sum of squares. Clusters that have larger values have greater differences in the observations within the clusters [46].

To define key role players that have a direct influence on accidents and to also identify which factor is more dangerous in megacities, India, the author(s) [47] used the Info Gain Evaluator technique and association rule mining. The author(s) used the first technique to rank attribute significance or to rank which attribute is more prone to accidents. However, the results of this

technique were highly summarized without further explanation of what the values from the rank entail or which attributes highly impact accidents. As a result, it became difficult to learn from the given results because they were not thoroughly discussed. The same trend is observed in the second technique (association rule mining) that was in the study. Here, rules that were found were not discussed as in prior work [36] or [48]. If symbols are used in the association rule analysis, it is important that they are well labeled and defined so that readers know what each symbol entail and thus infer meaning from the rules. Also, to assess the significance of the rules, the only mentioned performance evaluation matrices were the support and confidence values. These are not sufficient to indicate the significance of a rule. For example, if the lift measure was also used, this would have indicated whether the antecedent influenced the consequent negatively or positively [49].

The author(s) in [50], analyzed frequent itemsets or trends utilizing the accidents, casualties, and vehicles of a real data sample from the United Kingdom (UK). The author(s) applied two techniques namely: descriptive analytics and predictive analytics. In descriptive analytics, data is analyzed to understand what already happened, to understand the past. The outcome of this kind of analysis is not used to make predictions or forecast the future. In predictive analytics, historical data is analyzed to make future predictions by using statistical tools, data mining (DM) tools, and machine learning algorithms [51]. To achieve the predictive task, the author(s) [50] used three classification algorithms namely: Random Forest, Gradient Boosted Classifier and Random Forest Big Data to classify accidents as fatal or nonfatal. The latter model had the least classification error. For the descriptive analysis (association rule mining), there were some interesting rules that were found, e.g. accidents that occurred on Sundays had fatal consequences with the highest likelihood of occurring excluding the fact that this day had the least number of accident occurrences. The main difference between [45], [47] and [50] is that work in [50] did not involve any preliminary clustering prior to searching for frequent itemsets.

2.3 RTA research using Artificial Neural Networks (ANNs)

Traffic sign recognition systems have been created to reduce road carnage in some parts of the world [52]. These systems are usually embedded in some vehicles (i.e BMW 7-Series 2008), BMW 5- Series, and Mercedes-Benz E-Class), to recognize traffic signs on the roads. This technology is used in Image Processing. However, not much work has been done to develop accurate road sign recognition systems. Having said this, [53] developed an algorithm that will be used to classify the shape of traffic signs and their recognition to produce a driver alert system. The proposed algorithm is composed of two stages: shape classification stage and content classification stage. The algorithm is first fed a list of Bounding Boxes to classify. The shapes of the bounding boxes are classified by an artificial neural network (ANN). These shapes are triangular, squared or circular. To classify the content or meaning of these shapes, both the color and shape of traffic signs are classified as danger, information, obligation or prohibition. After classifying the shapes of the traffic signs, they are then fed as input to the second ANN. These shapes then label the icon of the road sign thus the result of the second ANN gives the full classification of the road sign.

More work in the identification of road signs was proposed by [54]. In this work, the author(s) proposed a driving alert system that will assist drivers and reduce the occurrence of accidents. Same as in [53], the author(s) here also proposed a technique to detect and recognize road signs. Their methodology consists of two subsystems: the detection subsystem and the recognition subsystem. In the detection subsystem, two techniques are used: the color filter and the selective search. Both techniques are used to remove noise on an image. In the recognition subsystem, a convolutional neural network (CNN) is applied to classify road signs.

So far, it has been observed that publications in [53] and [54] have had two phases in the implementation of creating driver assistance systems. The two phases are the detection phase and the recognition phase. The detection phase in [53] consisted of classifying shapes of road signs whereas the detection phase in [54] included classifying the content of the shapes where both the color and the shape of road signs are taken into account and classified as danger, information, obligation or prohibition; were the final output in the recognition phase is the

classification of the road sign. The recognition phase in [54] consisted of feeding input from the detection to classify road signs. The similarity between [53] and [54] is that the output achieved from both works is the identification of road signs.

A slightly different study from [53] and [54], the author(s) [55] used convolutional neural networks to recognize vehicles on two-lane roads on images or videos, and then classify the vehicle that was erroneous in the RTA. The key similarity gathered from [53], [54] and [55]; is that all works apply a form of artificial neural networks in image processing to improve road safety.

2.4 RTA research using CART and other classification algorithms

Decision trees are an essential part of predictive modeling in Machine Learning. A classification and Regression Tree (CART) is a kind of a decision tree that can be used for classification and regression tasks. Provided an input, the tree is searched by evaluating the specific input started at the root node of a tree [56]. CART models have been applied in the road safety domain to make predictions on injury severity and understand the main factors in RTAs.

To model RTA data, a study [57] used a Taiwan accident dataset. The CART model was used to model the relationship between injury severity, driver/vehicle characteristics, environmental attributes, and accident attributes. The results showed that the most significant contributing factor in injury severity according to the Taiwan dataset is vehicle type and the most vulnerable groups that have a higher risk of being injured are pedestrians, cyclists and bicycle riders. To evaluate the performance of the CART model, the author(s) tested the performance of the model on unseen data (testing data). The model had three target classes: no injury, injury and fatal. However, the model completely failed to classify fatality observations. This might have been caused by an imbalance that existed in the data. Although the author(s) argued that their findings should not be judged solely on this flaw, class imbalance should be detected at the early stages of model building or model testing and be corrected. Methods that are used to correct class imbalance include under-sampling, over-sampling, under and oversampling and

many more. A reliable model should be able to make predictions for all specified labels in the target class.

After analyzing injury severity on two-lane two-way rural roads in Iran, the author(s) [58] discovered that incorrectly overtaking and not using seat belts had a great impact on the intensity of injury severity. As observed in [57], the author(s) here also encountered a class imbalance problem. However, the class imbalance problem here was corrected prior model building. For example, the author(s) combined the serious injuries and the light injuries into a single class and the fatality class was placed in a separate class. This resulted in having a fairly balanced target class. In the analysis, the author(s) also performed variable importance to discover variables that had a significant impact on the target label. The common thing to note between [57] and [58], is that the author(s) all built classification predictive models to classify injury severity levels (fatal, non-fatal, injury). Both works experienced class imbalance problems and addressed it differently.

Studies in [57] and [58] investigated factors that affect injury severity in RTAs. In [59], the author(s) investigated the impact of the human element, the occurrence of RTAs and injury severity of RTAs in Iran. Three techniques were used namely; descriptive analysis, LR classification, and regression trees. As observed in [57] and [58], the target class here had 3 classes: fatal, non-fatal and injury. Interesting enough, no class imbalance was experienced as in previous publications [57] and [58]. It is very rare to have a real-world dataset were both positive and negative classes are fairly balanced. Nonetheless, the results from the classification using LR and CART model showed that CART exhibited better prediction accuracy than the LR model. The author(s) found that the driving license, safety belt, and age were among the most significant attributes that had a high influence on injury severity in Iran. For passengers, injury severity increased when seat belts were not used. Injury severity decreased for those who had driving licenses, the gender of the motorists had no influence on injury severity. To sum it up, the human factor had a great impact on all accidents followed by environmental factors and vehicle factors accordingly.

The main argument in [60], was to investigate the role of road users (drivers, passengers, pedestrians) on RTA injury risks. To investigate, the author(s) employed two techniques namely; CART and Random forest classification. The results from the CART model showed

that attributes that contributed sufficiently to whether or not a collision would result with an injury were: the movement of pedestrians and vehicles, the profession of the victims, the age of the victims, the code of the driving license and driving experience and the health condition of victims. The author(s) used 10-fold cross-validation and the ROC curve to validate their findings and to make their model reliable. Given the fact that class imbalance problems are highly likely in classification problems, the author(s) used a technique called priors equal, which generates an equal probability for all groups defined in the target class. Although class imbalance was corrected, the CART model performed well in classifying non-injury (majority class) observations over injury (minority class) observations. The Random forest model adopted in the study also classified the majority classes better than the minority class. Comparing the results of the two classification methods, the models both performed satisfactorily. However, it is worthwhile to note that the Random forest model was more robust and had better model performance than the CART model.

A slightly different approach from the classification problems that were discussed in the preceding sections, the author(s) [61] investigated the duration of an accident taking into consideration the detection time, the response time, the clearance time and the recovery time. The author(s) used three techniques on a Beijing incidence record dataset to study the duration time namely: CART, Chi-square Automatic Interaction Detector (CHAID), and Exhaustive CHAID. It was then found that the CART model outperformed the other two models as it performed better and had a better model accuracy. However, the author(s) eluded that the CART algorithm does well at dealing with missing values. From what is learned in the literature, the other purpose of data preprocessing or data cleaning in the early stages of knowledge discovery is to handle missing data so that it is not carried over in the model building. Therefore, missing values should have been removed before incorporating them in the model building.

2.5 The South African literature on road safety

The growing numbers of RTAs propose that the existing approach for addressing RTAs in South Africa is inadequate. Given that in 2016, RTA fatality grew by 9% [2] from the year 2015. The expenses of crashes were approximated to be R143 Billion [3] in the year 2015.

Having said that, road safety continues to remain a national health challenge. Figure 1 shows the fatal RTAs in South Africa's provinces between the years 2004 and 2015. Comparing to other countries, South Africa has a higher fatality rate with Gauteng and Kwa Zulu Natal at the lead.

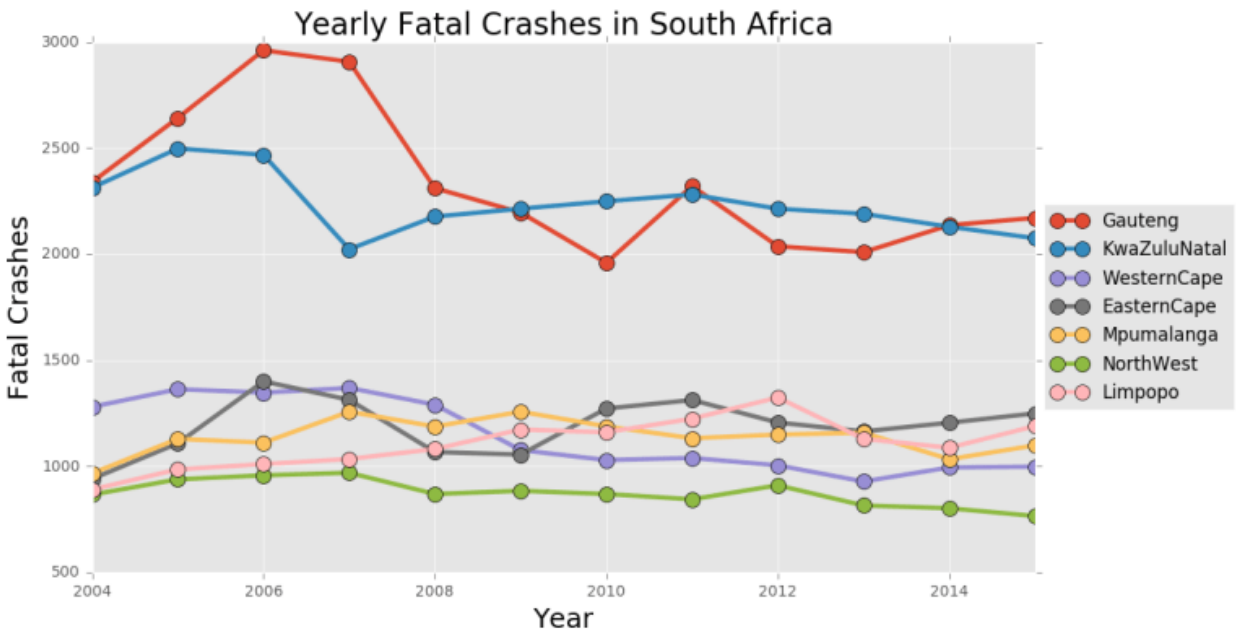

Figure 1: This figure displays fatal crashes of the seven provinces of South Africa (two provinces were omitted for ease of reading) – Source: [48: 2].

An exploratory analysis [62] investigation on fatal RTAs identified contributing elements and the level of lawlessness in RTAs. In DS, researchers often perform exploratory data analysis to compile the main features of a dataset. Steps include descriptive statistics, data visualization, and inferential statistics and so on. This kind of analysis does not provide an in-depth understanding of a specific challenge data question. For instance, the author(s) were able to find the total number of crashes for the years 2003 and 2005, the total counts of un-roadworthy vehicles and the number of motorists that violated the legal breath alcohol limit. Although the author(s) performed descriptive data analysis on the data, they were able to give recommendations to combat road safety challenges such as increased traffic law intervention, educating motorists and the general public about road safety, and ways to identify dangerous pedestrian locations.

A study [63], focused on interventions that can be used to change the behavior of road users to improve road safety. The author(s) found that influencing the behavior of road users will have a positive effect in addressing road safety concerns in South Africa. The author(s) also found that low and middle-income countries account for over 90% of road fatalities, even though they have a low vehicle population. The key difference between [62] and [63] is that the author(s) [63] used a qualitative research methodology while the author(s) [62] used a quantitative research methodology. The similarity between the publications [62], [63] is that the author(s) gave recommendations to curb road safety challenges and one of the common recommendations from both publications was road safety education. Educating scholars at an early age about road safety and preparing them to be skilled motorists.

Dissimilar to publications [62], [63]; the author(s) [64] tackled key challenges in RTAs by investigating the rate of traffic blockage of roads in Kimberley, a city in South Africa. The author(s) employed two methods in their study namely; Level of Service and Percent Traffic Diversion. The Level of Service model is a qualitative method that is applied to measure or explain the aspect of traffic service and the latter model is a technique that is applied to swerve traffic blockage by using alternative routes or diversion curves. Although the author(s) did not provide advanced recommendations to remedy traffic congestion, they were able to direct the attention of policymakers to specific road segments that needed attention.

When addressing road safety challenges, an all-inclusive approach needs to be applied and all parameters that contribute to RTAs need to be evaluated. In order to holistically address road safety challenges, there is a necessity to advise drivers who are in the process of purchasing new vehicles to choose safe vehicles. To reach this objective, it is important to understand the underlying factors that play a role in how new car buyers in South Africa prioritize buying safe vehicles. The author(s) [65] investigated significant factors that new car buyers consider or prioritize when purchasing vehicles by conducting surveys between car buyers and salespersons in two South African cities, Stellenbosch and Mthatha. The study found that new car buyers did not fully understand the safety features and crashworthiness of vehicles. The buyers prioritized reliability, cost, and comfort more than safety features. In order to make informed purchasing decisions, the new car buyers were heavily dependent on the information they receive from

dealerships. Although car dealerships had more information about the vehicles, they did not always convey them to new car buyers.

As eluded by the Mobility Centre for Africa (MCA) 2017, vehicle population in South Africa increases by 3% on a yearly basis. This increase in vehicle population translates to more cars on road networks that drive at unapproved speeds. As such, safe vehicle maneuvers have to be executed and this heavily depends on the mechanical state of a vehicle. In developing countries like South Africa, the state of the economy puts pressure a large percentage of the population to drive older and less trustworthy cars. As such, motor vehicle accidents that occur as a result of mechanical failure increases, with the fatality rate being amongst the highest cause of death in RTAs. The author(s) [66] investigated the role that mechanical failure plays in motor vehicle accidents and compared it to international trends. The study findings showed that tares, brakes, and overloading contributed the most to mechanical failures.

Summary

In summary of this chapter, discussions about using statistical and machine learning approaches in the road safety domain was discussed. First, a review was given on how LR was used to tackle road safety challenges followed by association rule mining, artificial neural networks, and CART models. Then, a discussion on the South African literature on RTAs was given. After a thorough literature search, there is a need for new computational methods, such as machine learning, to help in uncovering hidden knowledge from RTA data that would assist in planning effective strategies to improve road safety. These strategies would be implemented by road safety initiatives such as the Transport department, RTMC, and AA, so as to better road carnage on road. Next, the research methodology that will be used in the study is introduced.

Chapter 3: Research Methodology

This section of the work provides an understanding of the methods that will be used in the study. The chosen research design, which is the Cross-Industry Process for Data Mining (CRISP-DM) methodology will be explained. Then, the focus is placed on data quality problems that were encountered in the study and methods that were used to remedy these challenges are explained. Specifically, methods such as data that is either Missing Completely At Random (MCAR), Missing At Random (MAR), Missing Not At random (MNAR) and multiple imputation that will be used to prepare the data collected before data modeling. Following this, clustering techniques such as K-means and Partitioning Around Medoids (PAM) clustering algorithms will be discussed and briefly compared. To look for frequent itemsets in the data, the association rule mining technique is used and methods that were used to execute this task are explained. Finally, two machine learning modeling methods namely; MLR and XGBoost are explained.

3.1 Research Design

The CRISP-DM methodology will be applied in this research. This is a form of another data science methodology that is meant to aid in the execution of knowledge discovery processes. It is meant to make DM projects less expensive, more reliable, repetitive and faster. The CRISP-DM framework consists of six stages as depicted in Figure 2 and explained as follows:

- Business understanding: This phase consists of discerning research questions and objectives from a business point of view then translating this knowledge into a DM task and an initial plan structured to achieve set goals.
- Data understanding: This phase includes data collection followed by activities in order to get accustomed with the data and detecting problems to reveal initial meaning from the data.
- Data preparation: This stage comprises of activities that are carried out to create the final dataset that will be used in the modeling stage.

- Modeling: Here, modeling techniques are chosen and applied in order to answer data questions and the parameters of these models are adjusted to suit the data so as to generate accurate findings.
- Evaluation: After creating the model, it is crucial to review its performance and ensure that it answers set research questions and objectives.
- Deployment: Construction of the model is usually not the completion of the project. Typically, the knowledge discovered will need to be sorted and presented in a manner that the client can use or understand it. This phase can be as plain as producing a report or as intricate as implementing an iterative DM process [67]:

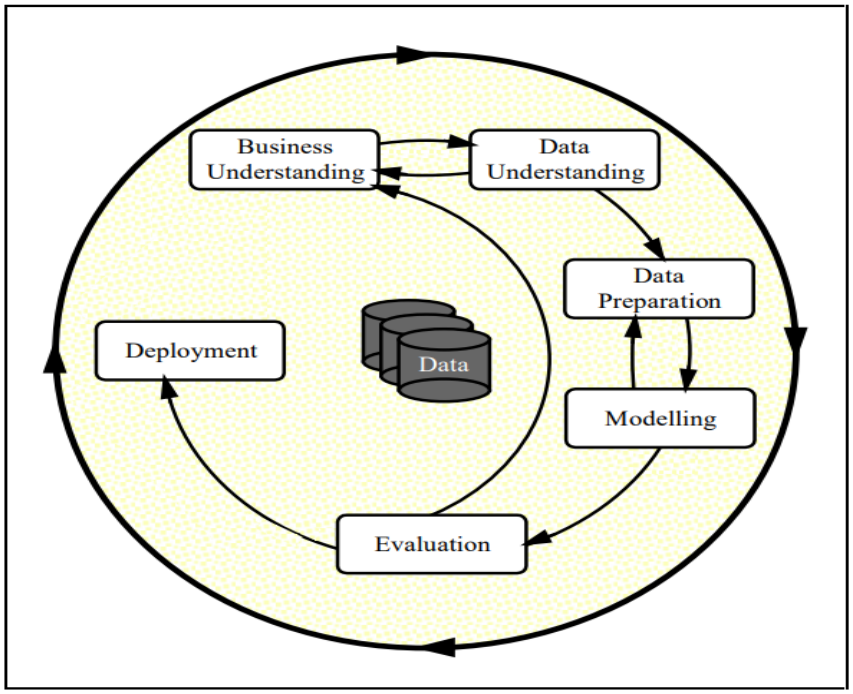

Figure 2: A CRISP-DM framework - Source: [67: 5].

3.2 Missing data

Sooner or later (typically sooner), data analysts or anyone that does any statistical analysis stumbles into missing data problems. Missing data occurs when no data value is captured for an observation in a given dataset [68]. In a conventional data set, information is missing for some variables for different reasons [69]. Missing data may be due to human error, equipment malfunction, the noise of transmission, etc [70]. To make a decision as to how missing data will be handled, it is important to understand the mechanism or pattern of the 'missingness' in the

data. In this study, 'missingness' will be defined as the degree to which data is missing where three general missing data patterns or mechanisms are considered:

a. ***Missing completely at random (MCAR):*** the tendency for an observation to be missing is completely at random, this means that the missing observation is not dependent on the observed data. The missing data points are just a random subset of the dataset [68].

b. ***Missing at random (MAR):*** means there is an association between the propensity of the missing data points and the observed data. When an observation is missing, this 'missingness' is not connected to the missing values but is connected to the values of a subject's observed attributes [71].

c. ***Missing not completely at random (MNCAR):*** if the possibility that the data is missing depends on non-observed information then the data is missing not at random. If data is MNCAR, there are no general methods of handling the missing data the correct way [72]. The method that will be used to handle missing data points in the study is discussed next.

3.3 Multiple imputation (MI)

Multiple imputation (MI) is a method that replaces each missing data point or deficient value with acceptable values that represent a distribution of possibilities. MI was chosen because simple imputation methods cause an under-estimate or non-randomness in data. In simple imputation methods, the same value is filled in every time there is no variance [73]. For a statistical conclusion to be valid, it is important to introduce the correct degree of randomness during imputation and to include that uncertainty when calculating standard errors and confidence [74]. Figure 3 shows a multi-imputed dataset where each missing value is replaced by an estimate to a vector of ***m*** values. The newly imputed values are stored in a subordinate matrix with one row for missing data point and ***m*** columns.

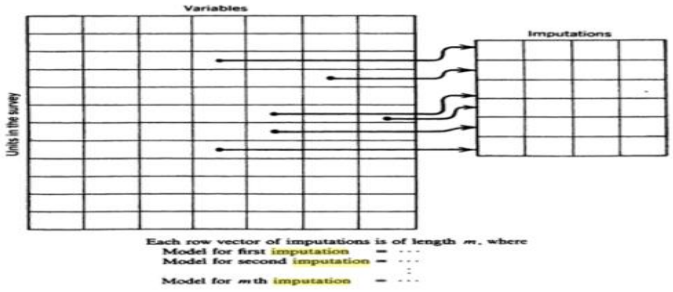

Figure 3: Dataset with **m** imputations for each missing datum [70: 3].

The first step in MI is to create multiple copies of the data, where the missing values are replaced by the newly imputed ones. These are generated from the distribution in the complete data so MI is based on a Bayesian approach. The second stage is to fit a statistical model to the imputed dataset. In this study, k-Nearest Neighbor (kNN) imputation was the statistical method that was chosen for imputation because it handles a small dataset easily, is non-parametric and it matches a missing data point precisely closest to its neighbor. The algorithm detects samples that are nearest to each other by using the Euclidean distance measures and the values for the missing data are determined from their closest observed neighbors. This method gives the best approximation of missing values [75]. The other major advantage of kNN is that it achieves great precision in imputing missing data and decreases errors in inferential statistics. It is important to note that before applying the kNN algorithm to impute data, the optimal number of ***K*** should be chosen so as to avoid impairing the original instability of the data [76]. To find the optimal ***K***, data will first be clustered and silhouette analysis will be used. Silhouette analysis is used to evaluate the separation distance between clusters by giving a score of how well matched a data point in one cluster is to data points in nearby clusters. Silhouette scores that are closer to +1 suggest that the data point is at a great distance from nearby clusters. A score of 0 suggests that the data point is on or very near to the decision boundary and -1 score indicates that the data points might have been assigned an incorrect cluster [77]. The flowchart in Figure 4 shows how the data will be imputed:

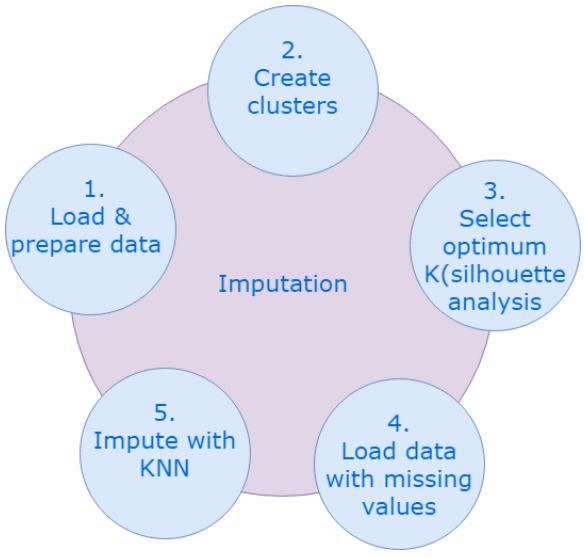

Figure 4: kNN Imputation.

3.4 Clustering

Categorizing data into sensible groups is one of the most significant concepts of understanding and learning. For instance, a popular system of scientific grouping puts organisms into a system of ordered taxa such as class, kingdom, and phylum. Cluster analysis is the study of grouping or clustering elements in accordance with interesting features or similarities [78]. In this study, K-Means clustering and PAM algorithms were used to cluster data. The K-means algorithm clusters data into ***k*** folds such that all elements of the same group belong to the same center. The elements in the same group are closer to that center compared to any other center. The algorithm keeps track of these centers called centroids and executes in simple iterations. The following steps are executed in the iterations:

i. cluster the elements based on the centers C(i), i.e, determine the centers to which each of the elements in the data belongs. The elements are divided based on the Euclidean Distance from the centers.

ii. set a new center $C(i+1)$to be the average of all elements that are closest to $C(i)$The new position of the center in a particular group is called the new position of the old center [72].

The algorithm seeks to minimize an objective function (sum squared error). The objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} || \, x_i^j - \, C_j||^2 \tag{1}$$

where $|| \, x_i^j - \, C_j||^2$ is the distance measure between an element x_i^j and the center C_j is the distance of the elements from their centroids [79].

K-means algorithm is simple and executes fast. In various application domains, the technique proves itself to be efficient and produces reliable clustering results. It takes as input the selected number of ***K*** and a dataset $\{D = d_{1,} d_2 \ldots\ldots d_n\}$, and outputs **K** clusters [80].

PAM clustering is based on the **K**-medoids clustering algorithm. Instead of using average/ centroids to represent the clusters, medoids are used. The medoid is a fixed object that represents data points whose mean dissimilarity compared to all other data points is small. That is, the medoid is also a data point of the dataset different to the mean that is used in K-means.

The algorithm takes as input $\boldsymbol{k}$ (cluster number) and $\boldsymbol{D}$ (dataset) and outputs a set of clusters [81]. The K-medoids algorithm is an improvement of the K-means algorithm because a large value in K-means can heavily change the distribution of data. The PAM algorithm pioneered the K-medoids algorithm. Preceding a random selection of medoids, PAM iteratively attempts to make a better selection of medoids [82]. Different from K-means, PAM can be applied to cluster categorical data. The main drawback of PAM clustering is that it cannot scale well on large datasets, but this worked in well in this research project because the dataset was small. The following are the steps of the PAM algorithm:

i. Select $\boldsymbol{k}$ objects in $\boldsymbol{D}$ summarily as the starting representative objects.
ii. Rerun the following steps until no change:
    i. Allocate each remaining object to the medoid with the closest representative object $\boldsymbol{O_{random}}$.
    ii. Choose a non-representative object randomly $\boldsymbol{O_j}$
    iii. Calculate the cost function (S) of switching representative objects $\boldsymbol{O_j}$ with non-representative objects $\boldsymbol{O_{random}}$
    iv. If the total cost is less than zero S<0, switch $\boldsymbol{O_j}$with $\boldsymbol{O_{random}}$ [103].

3.5 Association rule mining

Association rule mining involves taking a group of transactions called basket data and obtaining association rules in them [83]. Association rules imply $X \rightarrow I_J$, where X is a set of items in I and I_J is an individual item in I that is not available in $\boldsymbol{X}$. The rule $X \rightarrow I_J$ is achieved in the group of transactions T with the confidence component $0 \leq c \leq 1$ if the partial amount $c\%$ of transactions in T that satisfy X also satisfy I_J. Provided a group of transactions T, the goal is to produce all rules that satisfy specific additional restraints of two non-identical forms [84]. Association rule mining can be applied in various areas such as medical diagnosis, protein sequences, and retail and so on. In association rule mining, the support, confidence and lift are matrices that are used to evaluate the significance of a rule. The support illustrates the frequency of the patterns within a rule. The confidence is the significance of the implication of a rule and

the lift is a measure that tells us about the increase in the possibility of the consequent given the antecedent [85]. The Apriori algorithm will be used to search for frequent item sets in the data.

3.6 Classification

Classification is the mechanism of predicting a class (usually called target or label) for a given dataset. Classification predictive modeling consists of estimating a mapping function ***f*** from predictors ***x*** to discrete dependent variables *y* [86]. Common classification problems may include the following:

- Predicting whether a certain credit card is fraudulent;
- Determining the weather forecast if it will be sunny, windy or cold;
- Assessing if whether an individual is a good or bad credit risk;
- Assessing which environmental areas will suffer drought in the future;

Classification was applied in this work because the target class is a discrete variable ‘’**driver severity**’’ with four values: 1= killed, 2= Serious, 3= Slight, = No injury. The classification algorithms are discussed next.

3.6.1 Multinomial Logistic Regression (MLR)

MLR, also named Softmax Regression, is a supervised machine learning algorithm that can be applied in various domains such as text classification. MLR is a form of a regression model that extends LR to classification tasks where the output can host more than two values [87]. The algorithm consists of training and a testing phase. Training data of size L, with target labels, are expressed as $D_L = \{(X_1, Y_1), \ldots., (X_L, Y_L)\}$. The accepted MLR model is defined as:

$$P(y_1 = k|x_i, w) = \frac{e(^{w_k}x_i)}{\sum_{k=1}^{K} e(w^{(k)}x_i)} \quad (2)$$

where,

- w_k is logistic regressors for target *k*
- *w* is described as $(w_1^T, \ldots., w_K^T)$ [88].

MLR has several advantages over other multivariate techniques. For instance, it has different data distribution expectations which suggest that it provides satisfactory and accurate results when it comes to model fit and model accuracy. MLR is also sturdy to breaching of assumptions of multiple normalities and equivalent variance-covariance scores across classes and it also does not expect a linear relationship between the predictors and target class [89].

3.6.2 Extreme Gradient Boosting Tree (XGBoost)

XGBoost is a supervised machine learning algorithm that is used on training data $\boldsymbol{x_i x_j}$ to predict a class variant $\boldsymbol{y_i y_j}$ [90]. In this work, the XGBoost algorithm was chosen for the classification problem because of its efficiency and scalability. XGBoost supports several other objective functions including regression and ranking. It also has the potential of executing parallel computations and is ten times faster than other gradient boosting trees. The XGBoost package accepts several input formats such as a dense matrix, sparse matrix, local data files or its own xgb.DMatrix class [91]. In order to train the supervised machine learning algorithm, an objective function must be defined and optimized, and the goal is to find the best parameter that optimizes this function. This function takes as input the training loss L term and the regularization Ω. The training loss is the error on the train set and the regularization term is used to control the model from over-fitting. During the training phase, multiple learners or trees are created. The goal is to add a learner that optimizes the objective function by using the additive training method. Put plainly, let $\check{y}_i^t$ be the predicted output of the i^{th} observation at the t^{th} iteration, then, to minimize the objective function, an individual tree structure q and leaf weights w needs to be added at each step. This means that a learner that greedily improves the objective function will added [92].

Summary

This chapter discussed the methods that will be used in the study. The CRISP-DM research method will be followed to generate research findings. Before searching for co-occurring features in the data, data clustering will be performed as a preliminary task. Then, MLR and XGBoost will be applied for data modeling. The next chapter presents the results of the study.

Chapter 4: Results and Discussion

This chapter aims to explain results uncovered from the study. First, a brief discussion on how RTA data is collected after the occurrence of an RTA will be given. Then, data quality challenges will be discussed using two RTA datasets from two case studies used in the study. In the second section of this chapter, the description of the data used in the study will be reviewed. This section will explain elements such as data features and data collection and collation. The third section of this chapter will focus on how data was prepared for modeling and will cover concepts of data preparation such as one hot encoding, multiple imputation etc. The fourth section will cover results from the exploratory data analysis where the purpose was to understand the main characteristics in the data. The fifth section of this chapter will focus on results from data clustering and mining frequent item sets. Then, the outcomes from the predictive modeling will be explained by first applying Multinomial Logistic Regression followed by the application of XGBoost.

4.1 RTA data collection and collation

In South Africa, when a road traffic accident occurs, the person or people involved in the accident need to come to an immediate halt and if able, give assistance to those that sustained injuries. In event of serious injuries or deaths, the law enforcement agencies and or medical services need to be alerted to the accident scene for support or assistance. Upon arrival of the medical services, the driver(s) involved in the accident, which are not injured, or bystanders need to give a description of the accident and explain how the accident unfolded. Then, if the driver(s) involved in the RTA are still alive, they need to report the accident within 24 hours at the SAPS then fill out an AR form. In some cases, the AR form can be filled at the accident scene by the law enforcement officials. The filled out AR form can then be used for vehicle insurance claims or, the law enforcement department that collects the forms can be doing so to have a data pool that they can explore, analyze or study when making road safety standards or procedures. Data qualities challenges encountered in two datasets that are used in the study are discussed next.

4.1.1 Data quality challenges using two case studies

Table 1: This table gives a description of a RTA dataset obtained from North West, South Africa; the table also highlights some data quality challenges found in the data.

Data	July 2015-December 2016
Data presentation	Excel spreadsheet
Included features	Date, time, location, accident type, number of slight, serious, killed drivers, passengers, pedestrians, cyclists, light/weather conditions, road surface type, quality of road surface, road surface, road marking visibility, obstructions, overtaking control, traffic control type, road signs clearly visible, condition of road signs, the direction of the road, flat or sloped, position of the vehicle before the accident, vehicle maneuver vehicle damage
Excluded features	GPS coordinates, demographic attributes, tributes, driver/learner license details, road type, junction type, light/weather conditions, seat belt fitted etc.
Comments	The data had a large number of missing values, it was poorly presented and lacked useful attributes.

The data described in Table 1 lacked substantial attributes that would yield a strong data product (knowledge). For example, the data lacked details about the road infrastructure on which accidents occurred. It is of critical importance to include road infrastructure details in the data collection process because it is believed that road characteristics somehow contribute to accident severity, basing this assumption on a notion that accidents do not just happen anyhow at any place [93]. It is also important to note that environmental features were not recorded. Collecting environmental features data is important in order to analyze the role that an environmental setting plays on the severity of injuries [94] as it has been proven that RTAs gravitate towards areas that have a lot of facilities surrounding them, facilities such as; bars, shopping malls, restaurants and so on. The location was often also ill-defined. For example, one location would be defined using several different names. The manner in which locations were recorded was inconsistent. Now, because locations were described using inconsistent names or labels, just before applying any geoinformatics technique; these addresses or locations need to be decoded first into geographic coordinates. Having to perform the decoding of addresses before applying machine learning algorithms slows down knowledge discovery as it is a gradual process. One of the reasons why it is recommended to use a geographic positioning system

(GPS) when describing RTA location, amongst other things, is because visualizing these accidents on a regional map such as a heat map displays RTA hotspots, so that policymakers may be able to identify different levels of intensity of accidents and how and where they are mostly clustered. In addition to this, the dataset lacked information about the person(s) involved in the accident such as demographic attributes. Demographic attributes allow researchers to investigate which proportion of the population is most vulnerable to RTAs; which ethnicity and age group so as to correctly categorize or organize accident features. There was also no information about the level of compliance on the road of the motorist(s) involved in the accident, i.e seatbelt or helmet used, liquor or drug use suspected etc. Having this kind of information allows investigators to allocate driver responsibility levels in RTAs [95].

Table 2: Soshanguve, South Africa dataset. The data (paper forms) were collected at the Soshanguve police station, and analyzed at the research institution.

Data	Jan 2015-Dec 2017
How was it presented?	Paper forms
Included attributes	Date, time, location, accident type, demographics, accident severity, vehicle type/model, environmental conditions, etc
Lacking attributes	Pedestrian/cyclist information such as position, location, pedestrian/ cyclist action etc. Special observations such as tare condition, the condition of the vehicle, length of skidmarks etc. GPS coordinates and more.
General comments	The data had a large number of missing values. Old traditional method of handling data is still used. Overall poor data representation.

In the second case study, a dataset obtained from the police station of accidents that occurred in Soshanguve, Gauteng is described. The dataset is explained in Table 2 and the same procedure as in Table 1 is followed by briefly discussing some of the challenges that were found in the data. The first challenge experienced was that paper forms are still used to record data instead of digital forms. This tradition of capturing data through paper forms is observed in most places or organizations across South Africa were data is usually collected through questionnaires or paper surveys. The use of paper forms to collect data is a challenge and one of the reasons for this is that the data must first be computerized before performing any advanced data analytics. One valuable advantage about using digital forms to collect data instead of traditional paper forms is

that the ratio of omission on the forms is reduced thus resulting in having more filled out forms or complete datasets. As discussed in [36], instead of collecting data through paper forms, there are other advanced solutions such using a website that is connected to a secure server, in which its task is to store, retrieve and manage data. This solution is relevant in the South African RTA domain because it is also implemented in other parts of the world and not only does it provide transparency but, it also makes for less difficult reproducibility of data analysis results carried out by researchers. An example would be the United States Fatality Analysis Reporting System (FARS). FARS collects RTA data online and makes it available in several machine readable formats. This solution also addresses the problem of having incomplete and/or incorrect datasets more especially if data were collected from the drivers. Moreover, if the dataset is digitized, this also improves the accuracy of the data analysis process and also allows for easier collaborations among researchers. Another thing to note about this dataset is that the speed limit was also typically under-reported or omitted. The speed limit attribute allows for the investigation of associations between high-speed driving and the motorist's visual acuity or vision. In prior publication [96], the visual acuity of motorists changes when they are driving in high speed compared to when they are resting in motion or driving at low speeds. This also means that the capability to judge far and near objects also changes. Although the speed limit was omitted or missing at most times, if the law department that collects the accident data had access to digital resources and training, collecting geographic data through a geoinformatics system enables the fetching of attributes such as the speed limit with less effort. A great number of accidents are associated with the motorist's behavior and road infrastructure. Nevertheless, according to literature, causes resulting from mechanical failure such as faults with the vehicle tires, brakes, and steering wheel are less studied. If more information about the vehicle mechanical features were provided; this would have enabled the study to evaluate how these mechanical features contribute to accident severity.

At the MCA 2017, it was found that the growth rate of the South African population decreased by 1.25% [48] in 2017 and the vehicle population of registered cars increases by 3% on a yearly basis. It would have been interesting to see how the population and the number of registered vehicles on the road affect safety on the road.

As previously discussed, using text or label descriptions to define accident locations makes data analysis a needlessly intricate process. In this second case study, a similar trend as in the first case study is observed. If an interest occurred to investigate how time and space contribute to the incidence of RTAs, this process would have been slow because the semantic address locations require to be translated into geographic coordinates before applying any spatiotemporal techniques [97]. Also, if an attempt was made to translate these text-based addresses to geographic coordinates, it is very likely that this would result into having wrong addresses due to the incorrect spelling of the text-based descriptions or labels or capture error.

4.2 Data Description

In this study, data that is captured on AR forms were used and a sample is shown in Figure 5 below. The dataset was provided by the Soshanguve SAPS in Gauteng, province of South Africa and ethical clearance to use the data was granted. The data was collected from Jan 2015-Dec 2017, and comprised of features such as road type, junction type, accident type, driver age, vehicle type and so on. The data had 45 attributes and 1525 observations. A detailed description of the data is provided in **Appendix B**. It is important to note that all other information related to the identity of the driver such as driver name, surname, residential address, vehicle registration and so on, were removed for privacy preservation reasons and to protect driver anonymity.

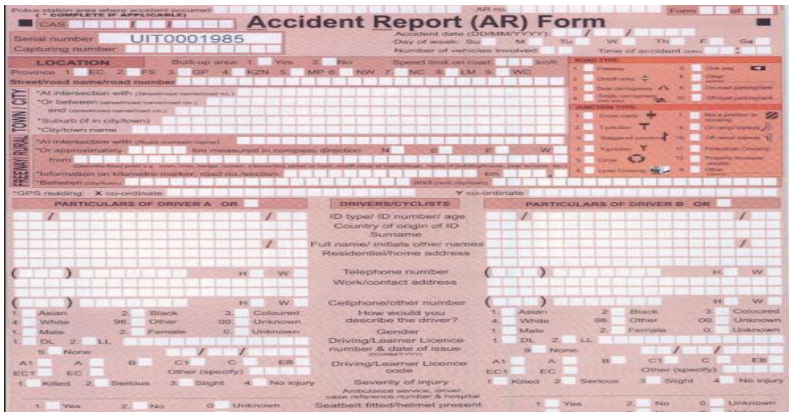

Figure 5: South African AR form

In order to augment the data, distance features were extracted such as the distance to the nearest bar, mall, building, school, and point of interest using a spatial database. The intention behind

this was to determine if distances to nearest places of interest have an impact on injury severity of drivers in RTAs. If it is found that the closest places of interests also play a role in the injury severity of drivers in RTAs, this information can direct policymakers to areas of high accident occurrences and proactive measures can then be taken such as to increase traffic personnel in the affected area, introduce speed cameras, enhance pedestrian walkways, provide tactical allocation of traffic signals, and ensure that medical services are close by to provide optimal treatment of rehabilitation following the injury such as effective first aid and appropriate care. Next, discussion on how data was prepared for modeling is provided.

4.3 Data preparation

To draw meaning from data in order to provide services and support decision making has become a focal point in all data management tasks and guaranteeing data quality has become more significant than ever. Errors in data tend to arise for several reasons. Some of these reasons consist of faults during data collection and errors during data consolidation from several databases. As a result, errors in the data result in false reports and can impact decision making negatively [98],[99]. The process of ensuring outstanding data quality is known as data cleaning. Different datasets need different kinds of cleaning. In this work, the first data cleaning step involved getting rid of duplicate and irrelevant values. An example of an irrelevant value may include mistakenly reporting '**age**' of the driver in a '**day of the week'** column. The second step in our data cleaning process had to do with checking for typos or inconsistent labeling of features. For instance, the name of the street 'Mandela Avenue' might have been labeled 'mandela avenue' in another row. Following this, the categorical representations of some variables in the dataset were converted to binary vectors through one hot encoding. The reason for choosing one hot encoding over integer encoding was to prevent the machine learning model from thinking that there exists some kind of sequential relationship between the variables. The third step of arranging the data consumed the most time as it consisted of handling missing values. After a careful literature exploration and uncovering that data is missing at least at random, the missing values were then imputed using the K-Nearest Neighbor (kNN) imputation. Different from model-based methods, kNN searches all complete observations and selects k observations that closely match the missing observation. kNN is well known for its simplicity, easy-interpretability, and high accuracy [100]. Before imputation, it

was critical to determine the optimum number of clusters ***k***. To partition the data, K-means clustering was chosen. K-means clustering is a popular unsupervised machine learning algorithm for dividing data into a set of k clusters where k is the number of groups defined by the analyst. It categorizes objects in several clusters in such a way that objects within the same group are similar, and objects from different groups are dissimilar [101]. To evaluate the 'goodness' of these clusters, the average silhouette method was used. This method can be used to evaluate the separation distance between the returned clusters. The silhouette graph represents a measure of how well matched each point in one cluster is in comparison to points in neighboring clusters [77]. Although there were several data qualiy challenges in the study, attemps were made to improve the quality of the data.

Before clustering, all rows that had missing values were discarded as the clustering algorithm is not good at missing values. However, the dataset had an extensive amount of missing values at a total of 31434. When all rows that included null values were cleared out all at once, this resulted in having only 3 rows only and 45 columns. This was clearly not a desirable option and as a result, an alternative way of clustering the data had to be implemented. This was resolved by clustering a pair of attributes, one at a time, and also imputing each pair one at a time. The steps that were followed to execute this procedure are as follows:

- Select a pair of features
- Find the total number of missing values in the selected pair
- Remove all missing values from the selected pair
- Find the optimum k (n_cluster)
- Impute, given optimum k

Table 3: The first column represents the selected pair of features; the second column reveals the number of missing values given the selected pair of features. The third column shows the number of rows that were left after clearing out all missing values from the pair of features before clustering. The fourth column shows the average silhouette score that was returned from the silhouette analysis and the fifth column shows the corresponding optimum ***k.***

Pair of attributes	Number of missing values	Rows left after removing missing values	Average silhouette width	Corresponding n_cluster
"accident.type", "closest.school"	1383	427	0.595	4
''weather.condition'',"road.surface.type"	140	1439	0.943	5
"distance.of.closest.bar.in.meters", "closest.restaurant"	1886	582	0.672	5
''driving.or.learners.licence.code'',''severity.driver''	610	1064	0.794	5
"gender.driver", "driving.or.learners.licence"	736	951	0.997	4
"liquor.drug.use.suspected", "liquor.drug.evidentiary.tested"	1116	943	0.993	5
"overtaking.control", "traffic.control.type"	1032	741	0.773	5
"position.of.vehicle.before.accident", "vehicle.manoeuvre"	475	1176	0.842	5
"speed.limit.km/h", "Built.up.area"	2421	140	0.878	5
"any.passengers.or.pedestrians", "carried.passengers.for.reward"	1442	582	0.964	5
"age.of.driver", "race.of.driver"	387	1237	0.961	2
"quality.of.road.surface", "road.surface"	147	1438	0.967	5
"road.type", "junction.type"	2200	204	0.768	5
"road.marking.visibility", "obstructions"	371	1237	0.941	5
"seatbelt.fitted.or.helment.present", "seatbelt.or.helmet.definitely.used"	769	1113	0.996	5
"type.of.building", "distance.of.closest.building.in.meters"	2121	3346	0.874	5
"vehicle.type", "light.condition"	151	1414	0.770	5

"road.signs.clearly.visible", "condition.of.road.signs"	304	1333	0.974	4
"distance.of.closest.school.in.meters", "closest.bar"	1886	582	0.662	5
"distance.of.closest.restaurant.in.meters", "closest.building"	1885	582	0.741	5
"closest.point", "type.of.point", "distance.of.closest.point.in.meters"	2859	566	0.697	5

What was deduced from the above table is that if the sample size is too small after removing rows that included missing values, the average silhouette score diminishes and consequently, the kNN imputation algorithm will have a high number of data points that failed to impute. The original data had 1525 observations and 64 columns. 18738 values failed to impute out of 59467 values, that is, a total of 32% of the data failed to impute and was then filled with zeros by the imputation algorithm. What was further learned was that kNN imputation does really well in imputing discrete values in comparison to continuous values.

The next step in the data preparation phase dealt with correcting the class imbalance. Just to recap, the target class had four labels (1: killed, 2: serious, 3: slight, 4: no injury). From the data, 76.4% of drivers were not injured, 12.1% were slightly injured, 10.4% sustained serious injuries and only 11% of motorists were killed. The goal was to arrive at a robust model that had good sensitivity and specificity and as such, it was crucial to correct the class imbalance before model building. Class imbalance occurs when negative classes outweigh positive classes. Class imbalance is associated with cost-sensitive learning; this means that incorrectly classifying a positive class instance is generally more significant than incorrectly classifying a negative class instance [102]. The Synthetic Minority Oversampling Technique (SMOTE) was selected to rectify the class imbalance. SMOTE works by oversampling the minority classes through creating artificial instances as opposed to oversampling with substitution [103]. Next, results from the exploratory data analysis are explained.

4.4 Exploratory Data Analysis (EDA)

Usually, the subsequent step prior to data modeling is the exploratory data analysis. The reason for performing EDA was to expand insight, discover underlying data structure, outliers and abnormality; and discover important variables. What was noticed from Figure 6 is that most motorists in this case study are in possession of heavy duty license code **C1**. The reason for this might be that **C1** tests are less complicated compared to light duty license code **B**. In code **B** tests, the learner is tested on movements such as parallel parking, emergency stop, and alley docking while in **C1** tests, the learner is only tested on engaging corners, intersections and large vehicle reverse [48].

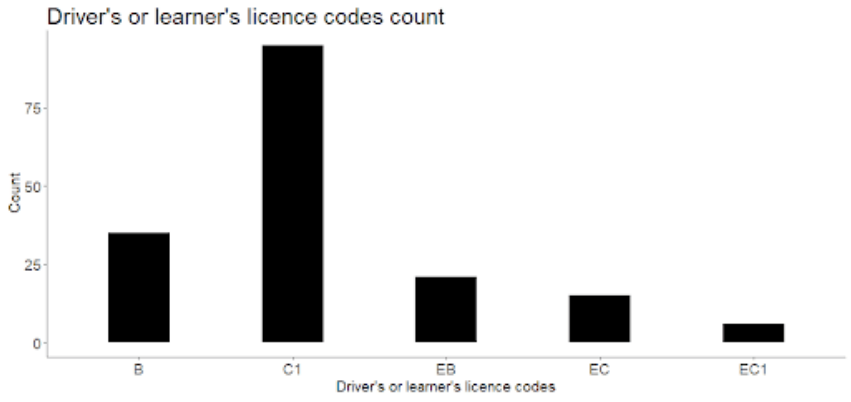

Figure 6: Plot showing the frequency of each duty license

In Figure 7, the truck drivers particularly code **C1** drivers, use light motor vehicles (vehicles with a tare weight of <=3500 kg or <) such as motor cars or station wagons more than heavy motor vehicles (vehicles with a tare weight of range 3500 kg and 16000 k).

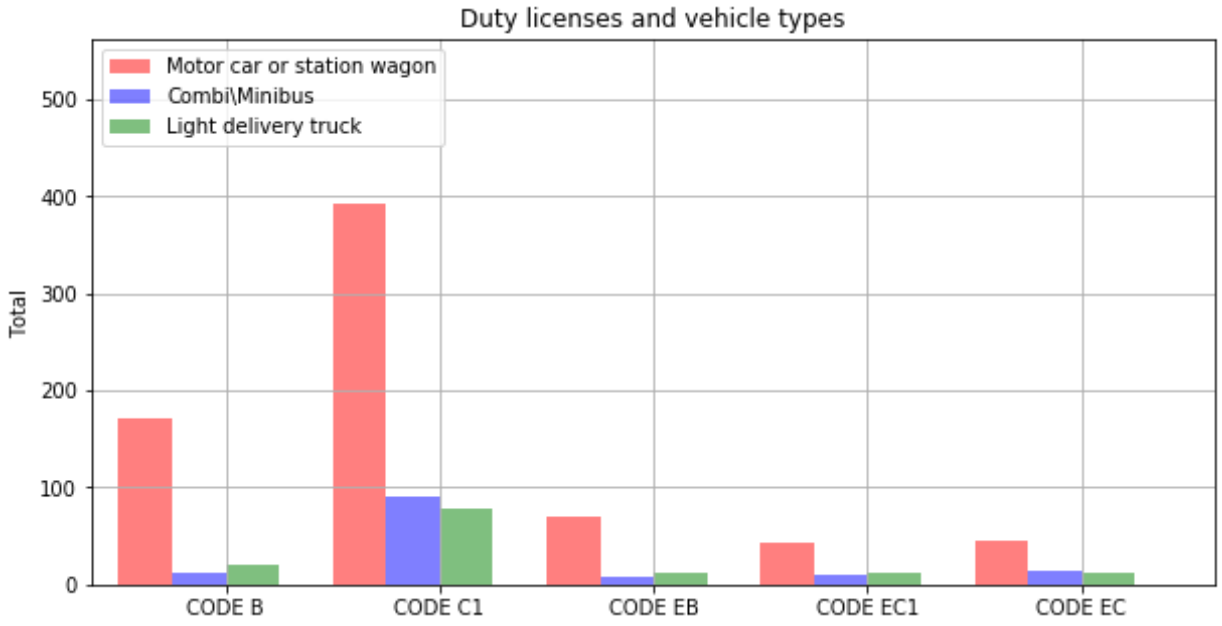

Figure 7: Plot showing vehicle types found per duty license

Figure 8 below reveals that truck license code **C1** motorists sustain the most serious injuries in RTAs compared other duty licenses such as code ***B***, ***EB***, **EC1**, and ***EC***.

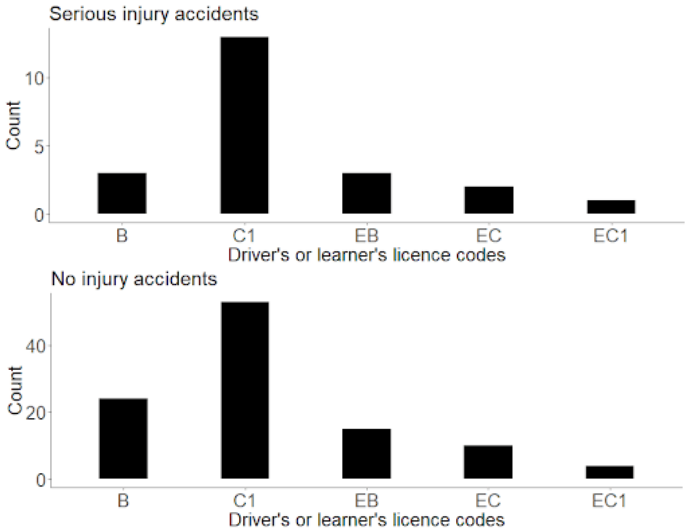

Figure 8: Illustration of serious injuries and no injury incidents found in each duty license.

The results from the EDA revealed a rather interesting finding against heavy duty license code ***C1*** motorists. In the upcoming section, amongst other things, the impact that heavy duty licenses have on driver injury severity will be investigated. Next, the results from mining frequent items are presented.

4.5 Association rules

Before searching for co-occurring features, the data was first divided into clusters in order to find meaningful rules following the work of the author(s) in [19]. To perform clustering, the PAM [20] clustering technique was selected because it supports heterogeneous data types. The optimum **k** that corresponded to the average silhouette score was **2**, which translated to having two clusters. The number of observations found in each cluster is shown below:

Table 4: % of observations found in each cluster

Cluster number	% of observations
1	30.43
2	69.56

Cluster 1 comprised of head/rear end accident types that typically took place on Saturdays in uncontrolled junctions. Drivers in this cluster did not attain any injuries. ***Cluster 2*** included Sideswipe: opposite directions accidents that mostly appeared on Saturdays and Sundays. Drivers would either be slightly injured or not injured. Upon completion of clustering, the Apriori algorithm was applied to search for co-occurring item sets in the clusters.

Rules for cluster 1 revealed that head/rear end accidents usually occurred during the day on flat, straight tarmac roads that did not have any obstructions. The road marking visibility and quality of the road appeared to be in good condition. Drivers in this cluster were in possession of truck licenses code ***C1***, although typically got involved in RTAs using light motor vehicles. Rules also showed that before the occurrence of the accident, some drivers would be traveling straight while others would be making a right turn. Motorists in this cluster were not under the influence of liquor or drugs and did not sustain any injuries. The support, confidence, and lift of these rules were 0.21, 0.71 and 2.05 respectively.

Rules for cluster 2 revealed that Sideswipe: opposite directions accidents that appeared in this cluster typically occurred during daylight on flat tarmac roads that had satisfactory road signs. Motorists would be traveling on the correct side of the road lane and in a straight direction. Even though seat belts were fitted, it was not known if they were used. The motorists here were

also ***C1*** truck drivers but were involved in accidents using light motor cars. However, these motorists would either sustain slight injuries or no injuries at all. The support, confidence, and lift of these rules were 0.22, 0.71 and 2.35 respectively. The next section discusses the results of creating classification models with Multinomial Logistic Regression.

4.6 Classification with Multinomial Logistic Regression (MLR)

Three classifiers were created using the multinomial logistic regression algorithm. The first classifier was created with distance features only and the target class. To recall, the target class had four labels (1: killed, 2: serious, 3: slight, 4: no injury) and the predictors of this classifier were:

- Distance to the closest point (in meters)
- Distance to the closest bar (in meters)
- Distance to the closest restaurant (in meters)
- Distance to the closest school (in meters)
- Distance to the closest building (in meters)
- Type of building
- Type of point

The second classifier was created with the original data that is captured on AR forms without distance features. In no particular order, the list of predictors is given below with driver injury severity as the dependent variable:

- accident_type
- age_of_driver
- race_of_driver
- any_passengers_or_pedestrians
- direction_of_road
- flat_or_sloped
- gender_driver
- driving_or_learners_licence

- liquor_drug_use_suspected
- liquor_drug_evidentiary_tested
- road_marking_visibility
- obstructions
- overtaking.control
- traffic_control_type
- position_of_vehicle_before_accident
- vehicle_manoeuvre_code
- quality_of_road_surface
- road_surface_code
- road_signs_clearly_visible
- condition_of_road_signs
- junction_type_code
- seatbelt_fitted_or_helment_present
- seatbelt_or_helmet_definitely_used
- speed_limit_km/h
- Built_up_area
- vehicle_type
- light_condition
- weather_condition_code
- weather_condition
- road.surface.type
- driving_or_learners_licence_code

The third and final classifier was created with the distance features fused with data captured on AR forms. All three classifiers were cross-validated where 80 % of the training data was divided into 5 folds and 20% was kept for testing. To evaluate the classification system the F1, recall and precision scores were used and the performances of the classifiers are shown in Table 5. The recall is defined as the measure of total amount of true positive predictions divided by the number of positive examples plus the number of false negatives and the precision is defined as the number of correct positive predictions divided by the total number of true positives plus

the total amount of false positives [104]. The F1 is the harmonic mean of the precision and recall. The accuracy, precision, recall and F1 score of all three classifiers were not plausible despite the balanced class distribution. What was important to us was to magnify the number of true positives, however, this would also result in the high rate of ‘’false alarms’’ but, the cost of incorrectly classifying injury severity as negligible is higher than the reverse classification. In the attempt to create more reliable models, a non-parametric form of gradient boosting tree was chosen and the results are discussed in the following section.

Table 5: Performance evaluation matrices for the MLR classifiers

	Distance only classifier	AR data classifier	Distance features & AR data classifier
Accuracy	52.04 %±2.67 %	71.05 %±3.71 %	59.10 %±4.15 %
Precision	52.05 %±3.68 %	71.36 %±1.18 %	58.50 %±4.72 %
Recall	52.03 %±3.30 %	72.48 %±0.99 %	59.46 %±4.39 %
F1-score	49.86 %±3.36 %	71.28 %±1.10 %	58.06 %±4.66 %

4.7 Classification with the Extreme Gradient Boosting Tree (XGBoost)

The results concerning predictive modeling using XGBoost will now be explained. Three classifiers were created with a focus of predicting injury severity in RTAs. XGBoost has been widely employed by data scientists to achieve outstanding results on many machine learning algorithms. The purpose behind the algorithm’s popularity is due to its scalability in all events. It executes 10 times faster than existing boosting trees on a single machine. This scalability is due to a novel tree learning algorithm and parallel/distributed computing [92]. Before running XGBoost, 3 kinds of parameters were set namely: general parameters, booster parameters and lastly; task parameters. General parameters have to do with deciding which booster to use i.e, tree or linear model. Booster parameters rely on the chosen booster and task parameters depend on the learning event, i.e regression problems may employ different parameters. To choose optimum parameters during the construction of the estimators, the Exhaustive Grid Search method was used. This method applies the ‘fit’ and ‘predict’ method like all classifiers, the only difference is that the parameters that are used to predict are improved or optimized by cross-validation [77]. All three classifiers were created using the same data that was used in creating the three MLR models. The best parameters returned from the grid search were:

Table 6: Optimum grid search parameters for the XGBoost classifiers

	Distance only classifier	AR data classifier	Distance features & AR data classifier
Learning rate	0.2	0.2	0.2
n_estimator	100	100	100
max_depth	6	6	6
min_child_weight	2	2	3
max_delta_step	0	1	1
subsample	0.9	0.9	0.8

Since this was a multiclass problem, the learning task that was selected was the multi: softmax, the number of classes (num_class) was set to 4 and the number of boosted trees was 100. After running XGBoost, the performance of the distance only classifier is shown in Table 7 and the confusion matrix is shown in Table 8.

Table 7: Performance evaluation matrix of the distance only classifier

Accuracy	Precision	Recall	F-score
41.51 %±4.52 %	52.51 %±2.30 %	4.00 %±3.01 %	37.52 %±3.81 %

Table 8: Confusion matrix for the distance only classifier

	Predicted label			
Actual label	**1**	**2**	**3**	**4**
1	**205**	0	0	0
2	**113**	61	23	9
3	**126**	16	64	6
4	**129**	19	33	32

From these results, it can be inferred that the model performance of this classifier is rather poor and there could be several reasons for this. One, it might be that more data features need to be added. Two, it might be that there is no association between the target class and predictors. To test these two theories, this led to the creation of the second classifier (AR data classifier). The performance of this classifier is shown in Table 9 and Table 10.

Table 9: Performance evaluation matrix for the AR data classifier

Accuracy	Precision	Recall	F-score
77.75 %±1.48 %	76.00 %±2.51 %	76.32 %±2.35 %	75.99 %±2.35 %

Table 10: Confusion matrix for the AR data classifier

	Predicted label			
Actual label	**1**	**2**	**3**	**4**
1	**205**	0	0	0
2	1	**176**	13	14
3	3	25	**166**	27
4	2	17	9	**177**

The results from the second classifier certainly prove that adding more data improves the performance of a classifier and there is an association between the target class and the predictor variables. However, to determine if nearest places of interest such as malls, bars, restaurants, and schools have an impact on the severity of driver injury in RTAs, this led to the creation of the third classifier and the results are shown in Table 11 and Table 12.

Table 11: Performance evaluation matrix for the distance and AR data classifier

Accuracy	Precision	Recall	F-score
83.14 %±3.34 %	82.83 %±3.18 %	82.66 %±3.16 %	82.35 %±3.25 %

Table 12: Confusion matrix for the distance & AR data classifier

	Predicted label			
Actual label	**1**	**2**	**3**	**4**
1	**221**	0	0	3
2	0	**178**	1	23
3	0	1	**166**	21
4	0	11	6	**214**

When the results from the three classifiers are compared, two theories can be confirmed. A 6% increase in the accuracy of the third classifier that was created using AR forms data and distance features was gained. Results also show that XGBoost produced better results than MLR. XGBoost has several advantages of MLR because its boosting procedure creates interactions

between variables which in turn provide numerous combinations to detailed event clarification. It also generates better flexibility, speed, stability and improved accuracy [105].

A question was raised from the exploratory data analysis- does the heavy duty license code **C1** have an impact on the severity of drivers in RTAs and, are distances to the nearest places of interest have an impact on injury severity of drivers in RTAs? To answer these questions, the 'plot_importance' library was used where the third classifier was given as an argument. The table of feature importance is shown in Figure 9, but only highlight the top 10 features will be highlighted. The results showed that truck licenses code C1, light motor duty license code EB, vehicle type (motor car or station wagon), single vehicle: overturned accident type, vehicle maneuver, and the distance to the closest building were among the top predictors that contributed to injury severity of drivers in RTAs. In a study [59] that was carried out in Iran, it was also found that the driving license code had an impact on injury severity of drivers in RTAs. To be specific, drivers who owned a second-grade license were typically more involved in RTAs compared to drivers that owned a first-grade license. In a pattern recognition problem that was executed in Ethiopia, the author(s) [60] also discovered that duty licenses have a significant impact on the severity of drivers in RTAs. These findings from the literature confirm the findings in this study that suggests that duty license codes have an impact on injury severity of drivers in RTAs. One of the significant features that contribute to injury of drivers in RTAs was the distance to the nearest building. A study [106] carried out in Hong Kong assessed the risk of the incidence of an accident while taking into account the environmental setting of an area. This study found that accidents had a likelihood of occurring near commercial areas, market stalls, and building material storage areas.

INDEPENDENT VARIABLES	F1-SCORE
DRIVER'S OR LEARNER'S LICENSE CODE: C1	168.5
DRIVER'S OR LEARNER'S LICENSE CODE: EB	159.33
CARRIED PASSENGERS FOR REWARD: NO	137.33
VEHICLE TYPE: MOTOR CAR OR STATION WAGON	129.25
ACCIDENT TYPE: SINGLE VEHICLE OVERTURNED	124
VEHICLE MANOEUVRE: TRAVELLING STRAIGHT	121
OVERTAKING CONTROL: NONE	116
DISTANCE TO CLOSEST BUILDING	112.75
OVERTAKING CONTROL:BARRIER LINE	110
LIGHT CONDITION: NIGHT, LIT BY STREET LIGHTS	107.67

Figure 9: The XGBoost variable importance of the distance & AR data classifier.

4.8 Binary and Multi-label classification using XGBoost

During the cleaning process, there was a lot of data manipulation that was done to ensure that the data is ready for modeling and will also produce reliable classifiers. For instance, the target label (injury severity) consisted of more negative classes than positive classes i.e, killed, serious and slight injuries were unlikely to occur as a large share of RTAs belonged to the ''no injury'' class as shown in Figure 10. As a result, the SMOTE oversampling technique was used to oversample the minority classes as mentioned in the data preparation section of the chapter.

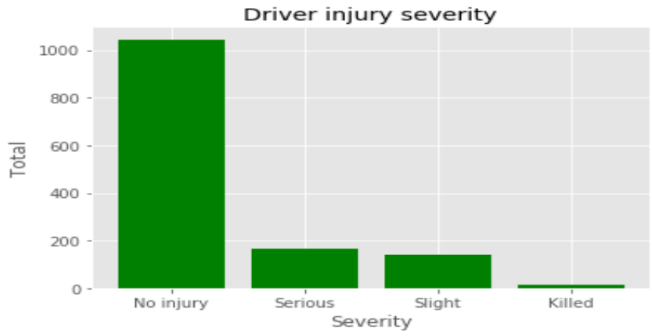

Figure 10: This bar plot shows the unbalanced amount of instances that were found in each target class

An attempt was then made to re-build the third classifier which was created using the AR data and the distance features. The only difference with this classifier is that the classes will be left as unbalanced as they are and no sampling technique will be carried out. The reason for this was

to determine if class imbalance will have an effect on the performance of the classifier. The performance of this classifier is shown in Table 13 and Table 14:

Table 13: The performance evaluation matrix for the distance and AR data classifier (where no sampling technique was performed).

Accuracy	Precision	Recall	F-score
70.09 %±5.14 %	61.20 %±1.74 %	71.56 %±2.55 %	65.78 %±1.69 %

Table 14: Confusion matrix for the multi-class XGBoost classifier without SMOTE.

	Predicted label			
Actual label	**1**	**2**	**3**	**4**
1	**0**	0	0	3
2	0	**6**	3	19
3	0	1	**2**	29
4	0	4	3	**23**

As shown in the above table, this classifier does really poorly in predicting a negative label as a positive label. As said earlier on, it is rather desirable to have a true negative class labeled as positive, than to have a true positive class labeled as negative. The recall is also not gratifying as this score means that the classifier is not exceptional in predicting positive samples. This output than ultimately led to a poor harmonic mean, F1-score. Thus, it was necessary to correct class imbalance before creating classification models.

Most publications in literature that dealt with multi-class tasks often restored to reducing their multiclass label into a binary classification task, such as reducing the four label class presented in this work, as a binary representation where 1 represents ‘’***serious***’’ and ‘’***killed***’’ injuries and 0 represents ‘’***slight***’’ and ‘’***no injury’’*** instances. This was mainly done because of a large class imbalance that existed. In this work, a similar approach was adopted where ‘’***serious***’’ and ‘’***killed***’’ injuries were merged into a single positive class 1 and ‘’***slight***’’ and ‘’***no injury’’*** were fused into a negative class 0. This classifier achieved an accuracy of 91.61%±2.17%. However, the F1 score was zero which meant that the classifier predicted all positive instances as negative. This was clearly a terrible classifier. This classifier was then re-created; the only difference now was that the binary classes were oversampled where both classes were equal.

This classifier achieved an accuracy of $83.8\% \pm 2.7\%$, a precision of $84.5\% \pm 3.7\%$, a recall of $84.2\% \pm 3.8\%$, and finally an F1 score of $85\% \pm 3.7\%$. This result shows that even if the classes are reduced into fewer classes, it is still necessary to have a balanced class distribution so as to improve overall classification performance. Chang and Chien [107] investigated factors that have a high importance on injury severity of drivers in RTAs. However, there was a large class imbalance in the data; i.e "***no injury"*** levels occupied 60% of the data while "***injury"*** and "***fatal"*** levels occupied 35% and 4.75% of the data respectively. This led to the classification model incorrectly classifying "***injury"*** accidents as "***no injury"*** accidents and ultimately, the classification system was rather poor. The author(s) [108] investigated severity in roll-over crashes. The author(s) first applied the CART algorithm to identify the best predictors for the classification model, then used the support vector machine algorithm to build the prediction model. To remedy the class imbalance problem that existed in the data, the author(s) reduced their five-level response variable to a three-level response variable and this resulted in a better performance prediction model. Sharaf et al. [109] also encountered a class imbalance problem when investigating factors that affect RTAs. The negative class accounted for 59% of the data, while "***moderate injuries***", "***severe injuries***", and "***death***" cases accounted for 31%, 7%, and 3% of the data respectively. As anticipated, the classification model performed really poorly on the test dataset with an accuracy of 0%. This then brings us to the conclusion that; when addressing multi-class problems that have a class imbalance, it is important to correct rectify the data before endeavoring to build prediction models.

Chapter 5: Conclusion, Recommendations & Future Work

5.1 Conclusion

This work focused on RTAs using a small community called Soshanguve in South Africa. The study used data captured on AR forms that were collected over a period of two years (Jan 2015-Dec 2017), and the main goal of the study was to evaluate how machine learning methods can be applied to extract meaning or knowledge from the given data. The study sets out three objectives: the first goal was to describe or summarize the main characteristics in the data by performing exploratory data analysis; the second goal was to discover co-occurring factors that appear at the incidence of a RTA by applying association rule mining and the third objective dealt with how to best create a classification model. Predictors that have a high importance on the injury severity of drivers in RTAs were also discussed. Before attempting to address the set objectives, it was important to identify the significant gaps in the literature. First, international publications were reviewed that involved using machine learning techniques such as LR methods, association rule mining, artificial neural networks, and CART models in RTAs. Then, the literature search was filtered down to South African literature where publications concerning RTAs and the methods used to address RTA challenges were discussed. Afterward, it was gathered that RTA data in South Africa has mostly been studied using low-level statistical analysis and thus there is a need for new computational methods such as machine learning to help extract knowledge from RTA data floods. After careful analysis of the literature, methods that were used in the study were then discussed in chapter 3. The chapter first introduced the CRISP-DM methodology as the chosen research design. Thereafter, the focus was placed on methods that were used in the study, namely: kNN imputation, K-means and PAM clustering technique, association rules, and classification algorithms namely; MLR and XGBoost. The subsequent chapter which was chapter 4, involved the implementation of the proposed research objectives. First, the chapter provided a brief explanation of how RTA data are collected at accident scenes or police stations in South Africa. Thereafter, data quality challenges encountered in the two datasets used in the study were then discussed where the overall finding was that both datasets lacked substantial attributes that hinder the process of discovering knowledge that would be useful in supporting road safety strategies and decision-making. Both datasets contained a large amount of missing data, ill-defined, and inconsistent labeling of

variables and one of the datasets was collected using paper forms instead of electronic forms and as a result, the rate of omission on the forms was high. The next section of this chapter described the data used in the study and also explained how the data cleaning process was carried out. The exact steps that were executed in preparation of the data were:

- The removal of duplicate and irrelevant values, typos and the correction of inconsistent labeling.
- The binarization of categorical variables using one-hot encoding.
- The replacement of missing values using kNN imputation where 32% of the data failed to impute.
- Equalizing class distribution by using the SMOTE technique.

Upon completion of the data cleaning process, the exploratory data analysis was performed where the main intention was to understand and summarize the main characteristics in the data. The key findings in this section of the chapter were as follows:

- Most motorists in Soshanguve hold truck licenses code ***C1***.
- These motorists use light motor vehicles more than heavy motor vehicles.
- These motorists also sustained the most serious injuries in RTAs in comparison to other duty licenses such as code ***B***, ***EB***, ***EC1***, and ***EC***.

The next section of this chapter focused on searching for frequent co-occurring item sets in the data by using the Apriori algorithm. The rules showed that head/rear end accidents mostly occur when some motorists are traveling straight while others would be making a right turn. Sideswipe: opposite directions accidents typically occur when both motorists are traveling in a straight correction on correct road lanes. These accidents occur on good road conditions with clear road markings. The final section of the results chapter involved creating classification models that predict injury severity of drivers in RTAs. To augment the data, the distances to the nearest places of interest were extracted using a spatial database. The reason behind this was to determine if nearest places of interest had an influence on injury severity of drivers in RTAs. First, three MLR classifiers were created. The first classifier was created with distance features only and the target class (1: killed, 2: serious, 3: slight, 4: no injury). The second classifier was created using the original data that was captured on AR forms without distance features, and the

third and final classifier was created with the distance features fused with data captured on AR forms. All classifiers were cross-validated were 80% of the training data was divided into 5 folds and 20% was reserved for testing. The MLR classifiers showed very poor performance, this output then led to the re-creation of the three classifiers using the XGBoost algorithm. The third classifier that was created using AR data and distance features achieved an accuracy of 83.14 %±3.34 % with an F1-score of 82.35 %±3.25 %. This result led to the conclusion that the augmentation of data that is captured on AR forms with distances to nearest places of interest data yields better classification models that predict injury severity of drivers in RTAs. Also, truck licenses code ***C1***, light motor duty license code EB, vehicle type (motor car or station wagon), single vehicle: overturned accident type, vehicle maneuver, and the distance to the closest building were among the top predictors that contribute to injury severity of drivers in RTAs. Moreover, the third classifier was re-created where no class imbalance was corrected; this classifier produced a poor model where the F1-score was not satisfactory. Another classifier was created where the four label target was reduced to a two label binary representation, where no oversampling was performed. This classifier achieved an accuracy of 87.58 %±1.85 % with an F1 score of zero. This meant that the classifier predicted all instances as negative This classifier was then re-created only this time, the two label binary class was oversampled which resulted in having two balanced class distribution. This classifier achieved an accuracy of 83.8%±2.7%, a precision of 84.5%±3.7%, a recall of 84.2%±3.8%, and finally an F1 score of 85%±3.7%. This result shows that even if the classes are reduced into fewer classes, it is still necessary to have a balanced class distribution so as to improve overall classification performance. What can be deduced from the latter two points is that; it is never favorable to create prediction models without correcting class imbalance and it is also not sufficient to reduce a multi-class task into a binary class task. The results obtained in the study were also compared to results found in previous work. Next, the recommendations are provided.

5.2 Recommendations

First, road traffic safety officials and all other personnel that will be involved with data collection and handling should be thoroughly trained on how to collect valid, consistent and accurate RTA data. It was found that some law enforcement departments are not trained on how to collect reliable RTA data. If the quality of data is reliable, this will help in finding more in-

depth understanding or knowledge that can be used to support RTA strategies and decision making. Results from the predictive modeling revealed that truck drivers ***C1***, have a high importance on the injury severity of drivers in RTAs. In 2011, a new law prohibiting the use of truck licenses by motorists using light motor vehicles was tabled but not signed into the Road Traffic Act [110], after the announcement of this modification, the new law was abrogated and the status quo restored because the review process arrived at the conclusion that this might not be in the best interest of South Africans. As a result, motorists that are granted truck licenses will still be allowed to use the licenses to drive light motor vehicles. In light of this, it may be useful to see how amending the current status quo will have an impact on injury severity of drivers in RTAs. The results from the study suggest geographic features are important in improving severity prediction accuracy. To know that closest places of interests also play a role in the injury severity of drivers in RTAs can enable road safety policymakers to take proactive measures in the areas of high accident concentration before the occurrence of an accident. For example, if the location of high accident severity is caused by the lack of a traffic control system, decision makers can apply an informed decision as to where to initiate a robot, yield sign, or stop sign. Also, if the area of high accident concentration is caused by road-related factors, decision-makers can decide to create a road safety awareness down the affected line i.e, N1, N14; allocate more funds to improve the road, ensure that medical services are close by to provide optimal treatment of rehabilitation following the injury such as effective first aid and appropriate care, and also increase traffic personnel in the affected area.

5.3 Future Work

In the future, it would be worthwhile to have more data that spans multiple cities or towns, this would then allow us to create a generic model that can then be applied to new areas and reviewed on the basis of overall predictive power. Another avenue of future work would be to collaborate with road safety entities such as AA, RTMC, and Statistics South Africa, in order to work together towards unearthing impactful knowledge that can be used to design improved strategies to curb RTAs.

If more RTA data is available, another interesting area to tap into would be accident prediction. Predicting the occurrence of an accident before it happens by harnessing the power of artificial

intelligence. For instance, in Dubai, the Smart Dubai office partnered with the Transport Authority and a local AI start-up to improve road safety by designing an early warning accident detector that would warn motorists of dangerous and unpredictable motorists around them [111].

Having more data would also give rise to the investigation of the role of driver fatigue in RTAs, knowledge discovered from this investigation would then enable the enhancement of safety on the road by introducing advanced driver assistance systems i.e, designing an intelligent tool that can detect if the motorist is falling asleep while driving then alert them to take a stop to refresh or rest [112].

Moreover, another area to investigate if more data is granted would be the role of driver distraction in RTAs. Distractions can be as trivial as changing a radio station or as big as answering a call or other activities behind the steering wheel. Intelligent cell phone applications are currently being tested that are created to automatically disable phone calls while the motorist is driving to assist in reducing driver distraction [113].

Appendix A

In this appendix, a table [114] that describes the types of license codes that are used in South Africa and the type of vehicles that can be used with each license code is provided.

Code	Vehicle type	Includes
	Motorcycles	
A1	Motorcycles with an engine of 125 cc or less	
A	Motorcycles with an engine that is $>$ 125 cc	Code A1
	Light motor vehicles	
B	Vehicles with TW of $\leq$ 3500 kg, and buses, minibes, and goods vehicles with GVM of $\leq$ 3500 kg	
EB	Vehicles with GCM $\leq$ 3500 kg, and vehicles allowed by Code B	Code B
	Heavy motor vehicles	
C1	Vehicles with TW between 3500 & 16000 kg; buses, minibuses, and goods vehicles with GVM between 3500 kg & 16000 kg	Code B
C	Buses and goods vehicles with GVM $>$ 16000 kg	Codes B, C1
EC1	Articulated vehicles between 3500 & 16000 kg, and vehicles allowed by code C1 with a trailer of GVM $>$ 750 kg	Codes B, EB, C1
EC	Articulated vehicles with GCM $>$ 18000 kg, and vehicles allowed by Code C with a trailer of GVM $>$ 750 kg	Codes B, EB, C1, C, EC1.

Appendix B

In this appendix, a table that describes the data detail with all the features and possible values is provided.

Table 15: Data Description

Feature	**Description**
speed.limit.km/h	Integer: Speed limit on the road
Built.up.area	Integer: 1= Yes, 2= No
road.type	Integer: 1= Freeway, 2= On/off ramp, 3= Dual carriageway,4= Single carriageway,5= One way,8= Other,9= On road parking,10= Off road parking
junction.type	Integer:1= Cross roads,2=T-junction,3= Staggered junction,4= Y-junction,5= Circle,6= Level crossing,7= Not junction or crossing,9= On ramp/ slipway,11= Pedestrian crossing,12= Property driveway,8= Other
age.of.driver	Integer
race.of.driver	Integer: 1= Asian, 2= Black, 3= Coloured, 4= White,98= Other,00= Unknown
gender.driver	Integer: 1= Male, 2= Female, 0= Unknown
driving.or.learners.licence	Integer: 1= Driver's license, 2= Learner's license, 9= None
driving.or.learners.licence.code	Categorical: A1, A, B, C1, C, EB, EC1, EC
Carried.passengers.for.reward	Integer: 1= Yes, 2= No, 0= Unknown
seatbelt.fitted.or.helment.present	Integer: 1=Yes, 2= No, 3= Unknown
seatbelt.or.helmet.definitely.used	Integer: 1=Yes, 2= No, 3= Unknown
liquor.drug.use.suspected	Integer: 1=Yes, 2= No
liquor.drug.evidentiary.tested	Integer: 1=Yes, 2= No
any.passengers.or.pedestrians	Categorical: No/Yes
vehicle.type	Integer: 1-5= Passenger vehicles, 6-10= Goods vehicles,

	11-14= Motor cycles, 15-19= Other vehicles
light.condition	Integer: 1= Daylight, 2= Night: lit by street lights, 3= Night (unlit), 4= Dawn/ dusk, 8= Other
weather.condition	Integer: 1= Clear, 2= Overcast, 3= Rain, 4= Mist/ Fog, 5= Hail/ Snow, 6= Dust, 7= Fire/smoke, 9= Severe wind, 0= Unknown
road.surface.type	Integer: 1= Concrete, 2= Tarmac, 3= Gravel, 4= Dirt, 8= Other
quality.of.road.surface	Integer: 1= Good, 2= Bumpy, 3= Pothole, 4= Cracks, 5= Corrugated,8= Other
road.surface	Integer: 1= Dry, 2= Wet, 3= Wet in areas, 4= Ice, 5= Snow, 6= Loose gravel, 7= Slippery, 8= Other, 9= Water: standing or moving
road.marking.visibility	Integer: 1= Unknown, 1= Good,2= Not good, 7= N/A
obstructions	Integer: 1= Accident site, 2= Road works, 3= Road block, 8= Other, 9= None
overtaking.control	Integer: 1= Barrier line, 2= Road sign, 7= N/A, 9= None
traffic.control.type	Integer: 1=, Robot, 2= Stop sign, 3= Yield sign, 4= Officer, 5= Officer + robot, 6= Uncontrolled junction, 77= Not at junction, crossing or barrier line, 8= All robots out of order, 9= Some robots out of order, 10= Flashing robots, 11: Boom, 12= Pedestrian crossing, 13= Barrier line
road.signs.clearly.visible	Integer: 1= Yes, 2= No, 7= N/A
condition.of.road.signs	Integer: 1= Good, 2= Not good, 3= Damaged or missing, 7= N/A
direction.of.road	Integer: 1= Straight, 2= Curving, 3= Sharp curve
flat.or.sloped	Integer: 1= Flat, 2= Uphill, 3= Downhill, 4= Steep uphill, 5= Steep downhill
position.of.vehicle.before.accident	Integer: 1= Correct road lane, 2= Wrong road lane, 3= Wrong side of road, 4= Road shoulder, 5= On road parking bay, 6= Off road parking bay
vehicle.manoeuvre	Integer: 1= Turning right, 2= Turning left, 3= U-turn, 4= Enter traffic flow, 5= Merging, 6= Diverging, 7=

	Overtaking: pass right, 8= Overtaking: pass left, 9= Travelling straight, 10= Reversing, 11= Sudden start, 12= Sudden stop, 13= Busy parking, 15= Changing lane,16=, Swerving, 17= Slowing down, 18= Avoiding object, 19= Stationary, 20=Parked, 98= Other
accident.type	Integer: 1= Head/rear end, 2= Head on, 3= Sideswipe: opposite directions, 4= Sideswipe: same directions, 8= Approach at angle, 16= Single vehicle: left of the road, 11= Single vehicle: overturned, 12= Accident with pedestrian, 13= Accident with animal, 14= Accident with train, 15=Accident with fixed/ other object
closest.school	Categorical: Name of closest school
distance.of.closest.school.in.meters	Numerical: distance to closest school (in meters)
closest.bar	Categorical: Name of closest bar
distance.of.closest.bar.in.meters	Numerical: distance to closest bar (in meters)
closest.restaurant	Categorical: Name of closest restaurant
distance.of.closest.restaurant.in.meters	Numerical: distance to closest restaurant (in meters)
closest.building	Categorical: Name of closest building
type.of.building	Categorical: Type of closest building
distance.of.closest.building.in.meters	Numerical: distance to building (in meters)
closest.point	Categorical: Name of closest point
type.of.point	Categorical: Type of closest point
distance.of.closest.point.in.meters	Numerical: distance to closest point (in meters)
severity.driver	Integer: 1= Killed, 2= Serious, 3= Slight, 4= No injury

References

[1] F. Bauchot, G. Marmigere, C. Truntschka, and F. Tressols, “Vehicle fed accident report.” Google Patents, 2010.

[2] S. Bachoo, A. Bhagwanjee, and K. Govender, “The influence of anger, impulsivity, sensation seeking and driver attitudes on risky driving behaviour among post-graduate university students in Durban, South Africa,” *Accid. Anal. Prev.*, vol. 55, pp. 67–76, 2013.

[3] F. J. . Labuschagne, “Cost of Crashes in South Africa,” *Road Traffic Management Corporation Research and Development*, 2016. [Online]. Available: https://www.arrivealive.co.za/documents/Cost-of-Crashes-in-South-Africa-RTMC-September-2016.pdf, 2016. [Accessed: 29-Jan-2018].

[4] A. F. Dove, J. C. Pearson, and P. A. Weston, “Data collection from road traffic accidents.,” *Emerg. Med. J.*, vol. 3, no. 3, pp. 193–198, 1986.

[5] M. A. Waller and S. E. Fawcett, “Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management,” *J. Bus. Logist.*, vol. 34, no. 2, pp. 77–84, 2013.

[6] W. v. d. Aalst and E. Damiani, “Processes Meet Big Data: Connecting Data Science with Process Science,” *IEEE Trans. Serv. Comput.*, vol. 8, no. 6, pp. 810–819, Nov. 2015.

[7] D. Akin and B. Akba, “A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics,” *Sci. Res. Essays*, vol. 5, no. 19, pp. 2837–2847, 2010.

[8] L. Rokach and O. Z. Maimon, *Data mining with decision trees: theory and applications*, vol. 69. World scientific, 2008.

[9] Dipuo Peters, “Minister Dipuo Peters: 2016 Africa Road Safety,” *South African Government*, 2016. [Online]. Available: https://www.gov.za/speeches/africa-road-safety-2016-31-oct-2016-0000-0. [Accessed: 01-May-2017].

[10] P. Mitra and S. K. Pal, *Pattern recognition algorithms for data mining*. Chapman and Hall/CRC, 2004.

[11] "South Africa, Road Safety Annual Report 2017," *OECD Publ. Paris*, 2017.

[12] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using machine learning paradigms," *Informatica*, vol. 29, no. 1, 2005.

[13] T. H. Davenport and D. J. Patil, "Data scientist," *Harv. Bus. Rev.*, vol. 90, no. 5, pp. 70–76, 2012.

[14] M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper.," in *FLAIRS conference*, 1999, vol. 1999, pp. 235–239.

[15] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media, 2013.

[16] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[17] K. C. Fu, *Sequential methods in pattern recognition and machine learning*, vol. 52. Academic press, 1968.

[18] C. Kong and J. Yang, "Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in china," *Accid. Anal. Prev.*, vol. 42, no. 4, pp. 987–993, 2010.

[19] C. D. Fitzpatrick, S. Rakasi, and M. A. Knodler Jr, "An investigation of the speeding-related crash designation through crash narrative reviews sampled via logistic regression," *Accid. Anal. Prev.*, vol. 98, pp. 57–63, 2017.

[20] R. B. Taylor and A. Harrell, *Physical environment and crime*. US Department of Justice, Office of Justice Programs, National Institute of Justice Washington, DC, 1996.

[21] T. Lu, Z. Dunyao, Y. Lixin, and Z. Pan, "The traffic accident hotspot prediction: Based on the logistic regression method," in *Transportation Information and Safety (ICTIS)*,

2015 International Conference on, 2015, pp. 107–110.

[22] N. A. Kramer, T. T. Lopez-Capp, E. Michel-Crosato, and M. G. H. Biazevic, "Sex estimation from the mastoid process using Micro-CT among Brazilians: Discriminant analysis and ROC curve analysis," *J. Forensic Radiol. Imaging*, vol. 14, pp. 1–7, 2018.

[23] O. H. Drummer *et al.*, "The involvement of drugs in drivers of motor vehicles killed in Australian road traffic crashes," *Accid. Anal. Prev.*, vol. 36, no. 2, pp. 239–248, 2004.

[24] R. W. Thomas and J. M. Vidal, "Toward detecting accidents with already available passive traffic information," in *Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual*, 2017, pp. 1–4.

[25] J. Demberelsuren, J. Oyunbileg, D. Uranchimeg, and M. Roine, "474 Road traffic injuries and deaths and their risk factors in Mongolia," *Inj. Prev.*, vol. 22, p. A172, 2016.

[26] P. Stallard, E. Salter, and R. Velleman, "Posttraumatic stress disorder following road traffic accidents," *Eur. Child Adolesc. Psychiatry*, vol. 13, no. 3, pp. 172–178, 2004.

[27] J. Kenardy, M. Heron-Delaney, J. Warren, and E. Brown, "The effect of mental health on long-term health-related quality of life following a road traffic crash: results from the UQ SuPPORT study," *Injury*, vol. 46, no. 5, pp. 883–890, 2015.

[28] S. S. Alavi, M. R. Mohammadi, H. Souri, S. M. Kalhori, F. Jannatifard, and G. Sepahbodi, "Personality, driving behavior and mental disorders factors as predictors of road traffic accidents based on logistic regression," *Iran. J. Med. Sci.*, vol. 42, no. 1, p. 24, 2017.

[29] J. M. Violanti and J. R. Marshall, "Cellular phones and traffic accidents: an epidemiological approach," *Accid. Anal. Prev.*, vol. 28, no. 2, pp. 265–270, 1996.

[30] P. M. Turkington, M. Sircar, V. Allgar, and M. W. Elliott, "Relationship between obstructive sleep apnoea, driving simulator performance, and risk of road traffic accidents," *Thorax*, vol. 56, no. 10, pp. 800–805, 2001.

[31] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance

problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[32] F. Valent, F. Schiava, C. Savonitto, T. Gallo, S. Brusaferro, and F. Barbone, "Risk factors for fatal road traffic accidents in Udine, Italy," *Accid. Anal. Prev.*, vol. 34, no. 1, pp. 71–84, 2002.

[33] R. Mayou and B. Bryant, "Outcome in consecutive emergency department attenders following a road traffic accident," *Br. J. Psychiatry*, vol. 179, no. 6, pp. 528–534, 2001.

[34] R. Mayou and B. Bryant, "Outcome 3 years after a road traffic accident," *Psychol. Med.*, vol. 32, no. 4, pp. 671–675, 2002.

[35] Q. Zhao and S. S. Bhowmick, "Association rule mining: A survey," *Nanyang Technol. Univ. Singapore*, 2003.

[36] S. Kumar and D. Toshniwal, "A data mining framework to analyze road accident data," *J. Big Data*, vol. 2, no. 1, p. 26, 2015.

[37] I. E. Allen and J. E. Seaman, "Is Heterogeneity Your Friend?," *Qual. Prog.*, vol. 50, no. 2, p. 44, 2017.

[38] Z. Huang and M. K. Ng, "A note on k-modes clustering," *J. Classif.*, vol. 20, no. 2, pp. 257–261, 2003.

[39] K. Geurts, G. Wets, T. Brijs, and K. Vanhoof, "Profiling of high-frequency accident locations by use of association rules," *Transp. Res. Rec. J. Transp. Res. Board*, no. 1840, pp. 123–130, 2003.

[40] A. Mirabadi and S. Sharifian, "Application of association rules in Iranian Railways (RAI) accident data analysis," *Saf. Sci.*, vol. 48, no. 10, pp. 1427–1435, 2010.

[41] W.-Y. Lin, M.-C. Tseng, and J.-H. Su, "A confidence-lift support specification for interesting associations mining," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002, pp. 148–158.

[42] R. Tian, Z. Yang, and M. Zhang, "Method of road traffic accidents causes analysis based

on data mining," in *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*, 2010, pp. 1–4.

[43] K. Geurts, I. Thomas, and G. Wets, "Understanding spatial concentrations of road accidents using frequent item sets," *Accid. Anal. Prev.*, vol. 37, no. 4, pp. 787–799, 2005.

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

[45] P. A. Nandurge and N. V Dharwadkar, "Analyzing road accident data using machine learning paradigms," in *I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017 International Conference on*, 2017, pp. 604–610.

[46] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm," *Procedia Comput. Sci.*, vol. 54, pp. 764–771, 2015.

[47] S. Gothane and M. V Sarode, "Analyzing Factors, Construction of Dataset, Estimating Importance of Factor, and Generation of Association Rules for Indian Road Accident," in *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*, 2016, pp. 15–18.

[48] M. Mokoatle and V. Marivate, "Collision Course: Challenges with Road Traffic Accident Data in South Africa," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, 2018, pp. 1–6.

[49] N. Hussein, A. Alashqur, and B. Sowan, "Using the interestingness measure lift to generate association rules," *J. Adv. Comput. Sci. Technol.*, vol. 4, no. 1, p. 156, 2015.

[50] F. Babič and K. Zuskáčová, "Descriptive and predictive mining on road accidents data," in *Applied Machine Intelligence and Informatics (SAMI), 2016 IEEE 14th International Symposium on*, 2016, pp. 87–92.

[51] S. Chawla and N. Chawla, *Data Analytics: A Problem-Solving Approach*. Chapman & Hall/CRC, 2016.

[52] F. Moutarde, A. Bargeton, A. Herbin, and L. Chanussot, “Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system,” in *Intelligent Vehicles Symposium, 2007 IEEE*, 2007, pp. 1122–1126.

[53] S. Hamdi, H. Faiedh, C. Souani, and K. Besbes, “Road signs classification by ANN for real-time implementation,” in *Control, Automation and Diagnosis (ICCAD), 2017 International Conference on*, 2017, pp. 328–332.

[54] S.-C. Huang, H.-Y. Lin, and C.-C. Chang, “An in-car camera system for traffic sign detection and recognition,” in *Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), 2017 Joint 17th World Congress of International*, 2017, pp. 1–6.

[55] E. Ba\cser and Y. Altun, “Classification of vehicles in traffic and detection faulty vehicles by using ANN techniques,” in *Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT), 2017*, 2017, pp. 1–4.

[56] W.-Y. Loh, “Fifty years of classification and regression trees,” *Int. Stat. Rev.*, vol. 82, no. 3, pp. 329–348, 2014.

[57] L.-Y. Chang and H.-W. Wang, “Analysis of traffic injury severity: An application of non-parametric classification tree techniques,” *Accid. Anal. Prev.*, vol. 38, no. 5, pp. 1019–1027, 2006.

[58] A. T. Kashani and A. S. Mohaymany, “Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models,” *Saf. Sci.*, vol. 49, no. 10, pp. 1314–1320, 2011.

[59] 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.

[60] T. Beshah, D. Ejigu, A. Abraham, V. Snasel, and P. Kromer, “Pattern recognition and knowledge discovery from road traffic accident data in ethiopia: Implications for

improving road safety," in *Information and Communication Technologies (WICT), 2011 World Congress on*, 2011, pp. 1241–1246.

[61] R. Li, X. Zhao, X. Yu, J. Li, N. Cheng, and J. Zhang, "Incident duration model on urban freeways using three different algorithms of decision tree," in *2010 International Conference on Intelligent Computation Technology and Automation*, 2010, pp. 526–528.

[62] G. Botha and H. der Walt, "Fatal road crashes, contributory factors and the level of lawlessness," in *Proceedings of the 25th Southern African Transport Conference (SATC 2006)*, 2006, vol. 10, p. 13.

[63] H. Moyana and E. Chibira, "Improving safety in the road transport sector through road user behaviour changing interventions: a look at challenges and prospects," 2016.

[64] D. Das and M. Keetse, "Assessment of traffic congestion in the central areas (CBD) of South African cities: a case study of Kimberly city."

[65] M. Khan and M. Sinclair, "Importance of safety features to new car buyers in South Africa."

[66] O. van Schoor, J. L. van Niekerk, and B. Grobbelaar, "Mechanical failures as a contributing cause to motor vehicle accidents—South Africa," *Accid. Anal. Prev.*, vol. 33, no. 6, pp. 713–721, 2001.

[67] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, pp. 29–39.

[68] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 333. John Wiley & Sons, 2014.

[69] P. D. Allison, *Missing data*, vol. 136. Sage publications, 2001.

[70] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 2004.

[71] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple

imputation be used for handling missing data in randomised clinical trials--a practical guide with flowcharts," *BMC Med. Res. Methodol.*, vol. 17, no. 1, p. 162, 2017.

[72] M. G. Kenward, G. Molenberghs, and others, "Likelihood based frequentist inference when data are missing at random," *Stat. Sci.*, vol. 13, no. 3, pp. 236–247, 1998.

[73] D. A. Bennett, "How can I deal with missing data in my study?," *Aust. N. Z. J. Public Health*, vol. 25, no. 5, pp. 464–469, 2001.

[74] P. Royston and others, "Multiple imputation of missing values," *Stata J.*, vol. 4, no. 3, pp. 227–241, 2004.

[75] E. G. Armitage, J. Godzien, V. Alonso-Herranz, Á. López-Gonzálvez, and C. Barbas, "Missing value imputation strategies for metabolomics data," *Electrophoresis*, vol. 36, no. 24, pp. 3050–3060, 2015.

[76] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: a critical evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 3, p. 74, 2016.

[77] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[78] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[79] V. K. Dehariya, S. K. Shrivastava, and R. C. Jain, "Clustering of image data set using k-means and fuzzy k-means algorithms," in *Computational Intelligence and Communication Networks (CICN), 2010 International Conference on*, 2010, pp. 386–391.

[80] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*, 2010, pp. 63–67.

[81] A. Bhat, "K-medoids clustering using partitioning around medoids for performing face recognition," *Int. J. Soft Comput. Math. Control*, vol. 3, no. 3, pp. 1–12, 2014.

[82] T. Velmurugan and T. Santhanam, “Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points,” *J. Comput. Sci.*, vol. 6, no. 3, p. 363, 2010.

[83] A. C. C. Coolen, “A beginner’s guide to the mathematics of neural networks,” in *Concepts for Neural Networks*, Springer, 1998, pp. 13–70.

[84] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in *Acm sigmod record*, 1993, vol. 22, no. 2, pp. 207–216.

[85] A. Rajak and M. K. Gupta, “Association rule mining: applications in various areas,” in *Proceedings of International Conference on Data Management, Ghaziabad, India*, 2008, pp. 3–7.

[86] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[87] W. P. Ramadhan, S. A. Novianty, and S. C. Setianingsih, “Sentiment analysis using multinomial logistic regression,” in *Control, Electronics, Renewable Energy and Communications (ICCREC), 2017 International Conference on*, 2017, pp. 46–49.

[88] O. Behadada, M. Trovati, M. A. Chikh, N. Bessis, and Y. Korkontzelos, “Logistic regression multinomial for arrhythmia detection,” in *Foundations and Applications of Self* Systems, IEEE International Workshops on*, 2016, pp. 133–137.

[89] C. Kwak and A. Clayton-Matthews, “Multinomial logistic regression,” *Nurs. Res.*, vol. 51, no. 6, pp. 404–410, 2002.

[90] M. Gumus and M. S. Kiran, “Crude oil price forecasting using XGBoost,” in *Computer Science and Engineering (UBMK), 2017 International Conference on*, 2017, pp. 1100–1103.

[91] T. Chen, T. He, M. Benesty, and others, “Xgboost: extreme gradient boosting,” *R Packag. version 0.4-2*, pp. 1–4, 2015.

[92] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of*

the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[93] M. D. Schluchter, "Mean Square Error," *Wiley StatsRef Stat. Ref. Online*, 2014.

[94] R. Good and H. J. Fletcher, "Reporting explained variance," *J. Res. Sci. Teach.*, vol. 18, no. 1, pp. 1–7, 1981.

[95] O. Siddiqui and M. W. Ali, "A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts," *J. Biopharm. Stat.*, vol. 8, no. 4, pp. 545–563, 1998.

[96] M. der Laan, K. Pollard, and J. Bryan, "A new partitioning around medoids algorithm," *J. Stat. Comput. Simul.*, vol. 73, no. 8, pp. 575–584, 2003.

[97] F. L. Mannering, V. Shankar, and C. R. Bhat, "Unobserved heterogeneity and the statistical analysis of highway accident data," *Anal. methods Accid. Res.*, vol. 11, pp. 1–16, 2016.

[98] M. Dallachiesa *et al.*, "NADEEF: a commodity data cleaning system," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 541–552.

[99] V. Ganti and A. Das Sarma, "Data cleaning: A practical perspective," *Synth. Lect. Data Manag.*, vol. 5, no. 3, pp. 1–85, 2013.

[100] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, 2012.

[101] A. Kassambara, "Practical Guide To Cluster Analysis in R," *Creat. North Charleston, SC, USA*, 2017.

[102] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst. Man, Cybern. Part B*, vol. 39, no. 2, pp. 539–550, 2009.

[103] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[104] S. Tan, “Neighbor-weighted k-nearest neighbor for unbalanced text corpus,” *Expert Syst. Appl.*, vol. 28, no. 4, pp. 667–671, 2005.

[105] S. E. Paulo Celio Di Cellio Dias, Serasa Experian, Melissa Forti,Bradesco, Marc Witarsa, “A Comparison of Gradient Boosting with Logistic Regression in Practical Cases,” *SAS Institute*, 2018. [Online]. Available: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1857-2018.pdf. [Accessed: 01-Nov-2018].

[106] K. Ng, W. Hung, and W. Wong, “An algorithm for assessing the risk of traffic accident,” *J. Safety Res.*, vol. 33, no. 3, pp. 387–410, 2002.

[107] L.-Y. Chang and J.-T. Chien, “Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model,” *Saf. Sci.*, vol. 51, no. 1, pp. 17–22, 2013.

[108] C. Chen, G. Zhang, Z. Qian, R. A. Tarefder, and Z. Tian, “Investigating driver injury severity patterns in rollover crashes using support vector machine models,” *Accid. Anal. Prev.*, vol. 90, pp. 128–139, 2016.

[109] S. Alkheder, M. Taamneh, and S. Taamneh, “Severity prediction of traffic accident using an artificial neural network,” *J. Forecast.*, vol. 36, no. 1, pp. 100–108, 2017.

[110] “Truck licences still valid for cars after review by Transport Department,” *Transport Department, South Africa*, Pretoria, 01-Feb-2011.

[111] C. Malek, “AI road safety system set for junctions in Dubai after agreement signed,” *The National UAE*, 2017. [Online]. Available: https://www.thenational.ae/uae/transport/ai-road-safety-system-set-for-junctions-in-dubai-after-agreement-signed-1.666832. [Accessed: 01-Oct-2018].

[112] “Artificial Intelligence to aid driving behavior and prevent accidents,” *The Economics Times*, 2017. [Online]. Available: https://auto.economictimes.indiatimes.com/news/auto-technology/artificial-intelligence-to-aid-driving-behavior-and-prevent-accidents/59013663. [Accessed: 30-Sep-2018].

[113] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, 2014.

[114] “National Road Traffic Act,” 93 of 1996, 2008.