Cape Town road traffic accident analysis: Utilising supervised learning techniques and discussing their effectiveness

by Christo du Toit (DTTCHR015)

Supervised by

Dr. Sebnem Er

Mr. Sulaiman Salau

A Thesis Submitted for the Degree of Mphil in Data Science at
University of Cape Town in 2022
Department of Statistical Sciences

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. I have used the Harvard convention for citation and referencing. Each contribution to, and quotation in, this essay/report/project from the work(s) of other people has been attributed, and has been cited and referenced. Any section taken from an internet source has been referenced to that source.

3. This essay/report/project is my own work, and is in my own words (except where I have attributed it to others).

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Signed by candidate

Print Name: Christo du Toit (DTTCHR015)

Date: 14 February 2022

Signature:

Abstract

Road traffic accidents (RTA) are a major cause of death and injury around the world and in South Africa. Methods to understand and reduce the frequency and injury-severity of RTAs are of utmost importance. There is limited South African literature on modelling RTA injury-severity using supervised learning (SL) methods that fit a model that relates a target variable to a set of predictor variables. In this thesis, multinomial logistic regression, classification trees (CT), random forests (RF), gradient boosted machines (GBM) and artificial neural networks (ANN) are used to model the potentially non-linear relationships between accident-related factors and injury-severity. Data on RTAs that occurred in the city of Cape Town during the period 2015-2017 are used for this study. The data contain the injury-severity of the RTAs as well as several accident-related variables. The injury-severity categories of RTAs are classified as: "no injury", "slight", "serious" and "fatal" injury. Additional locational and situational variables were added to the dataset. The exploratory analysis revealed that the vast majority of alleged causes (as deduced by the data capturers from the accident report) of RTAs are related to driver/human error, accidents with pedestrians make up only 5.86% of all RTAs yet account for 58.56% of "fatal" accidents and 55.37% of "serious" accidents, the majority of "fatal" and "serious" RTAs occur on the weekend and involve only one vehicle. It was also identified that the RTA data was severely imbalanced with regards to injury-severity. Imbalanced data occur when the number of observations belonging to each of the classification categories are not approximately equal and can negatively affect the performance of classification methods. This paper employed three common approaches to address class imbalance namely (i) undersampling of the majority class, (ii) oversampling of the minority class and (iii) the synthetic minority oversampling technique (SMOTE). The RTA data was split into training, validation and test sets keeping the proportions of the injury-severity category consistent. Four training datasets were analysed: the original imbalanced data, data with the minority class over-sampled, data with the majority class under-sampled and data with synthetically created observations. The performance of the SL methods trained on these four different datasets were compared using accuracy, recall, precision and F1 score as evaluation metrics. All three data sampling methods improved the CT, RF and GBM model's average recall and ability to identify observations belonging to the minority class ("fatal" RTAs). With regards to maximising average recall, the SMOTE technique was the most effective data sampling method to address class imbalance. Further analysis was done to determine whether simple SL methods such as multinomial logistic regression are sufficient to model RTA injury-severity or if more complex SL methods such as ANNs are required. The ANN model achieved a higher average recall and correctly identified more observations belonging to the minority class, "fatal" RTAs, than the multinomial logistic regression

model. Using average recall as the main evaluation metric, the ANN was selected as the "best" performing model on the validation data. The ANN model correctly identified a large number of "fatal" RTAs while also resulting in a high number of false positives. The ANN model was very effective at correctly identifying "no injury" RTAs as evidenced by the high recall and precision scores, but performed poorly at correctly identifying "slight" and "serious" RTAs. Finally, the variable importance of the CT, RF and GBM models trained on the SMOTE data revealed the geographical location of an RTA, crash type as well as the number of vehicles involved in an accident to be significant risk factors associated with RTA injury-severity. The CT and RF models both determined the alleged cause of an accident to be significant, while the RF and GBM models determined several weather-related variables to be significant risk factors associated with RTA injury-severity. Future road safety policies should focus on reducing human/driver error, reducing pedestrian-related RTAs and increasing policing efforts over weekends and during poor weather conditions. Road safety policies should take the geographical location of RTAs into account in order to identify high-risk areas for "serious" and "fatal" RTAs.

Acknowledgements

I would like to thank both my supervisors, **Dr Sebnem Er** and **Mr. Sulaiman Salau**, for all their effort, patience and time spent guiding me during the research process.

The author would like to thank the City of Cape Town for providing the road traffic accident data used in this paper.

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number 130479)

Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za

Contents

List of Figures

List of Tables

Abbreviations

ART	Adaptive Resonance Theory
ANN	Artificial Neural Networks
CT	Classification Trees
CART	Classification and Regression Trees
GBM	Gradient Boosting Machine
MARS	Multivariate Adaptive Regression Splines
MAE	Mean Absolute Error
MLE	Maximum Likelihood Estimator
MLP	Multilayer Perceptron
OOB	Out-Of-Bag
PTW	Powered Two-Wheelers
RF	Random Forest
RTA	Road Traffic Accident
RTMC	Road Traffic Management Corporation
SAPS	South African Police Service
SMOTE	Synthetic Minority Oversampling Technique
SL	Supervised Learning
SVM	Support Vector Machines
TAR	Traffic Accident Records
USA	United States of America
XGBoost	Extreme Gradient Boosting Tree

Chapter 1

Introduction

A road traffic accident (RTA) can be defined as a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment (ROSPA 2002). RTAs are a major cause of death and injury around the world. According to the World Health Organisation, RTA injuries are the leading cause of death of people between the ages of 5 to 29 years old (Organisation 2021). This is especially true for South Africa, where RTAs are a big public health concern. In 2018, there were 12 921 fatalities recorded in South Africa as a result of RTAs. In addition to the social cost, RTAs also have significant economic costs for South Africa. In a 2017 report from the Organisation for Economic Co-operation and Development, it was reported that the estimated total cost of RTAs on South African roads was R142.95 billion (South African Rand), which is roughly equal to 3.4% of South Africa's Gross Domestic Product (OECD 2017).

In order to effectively reduce the number and injury-severity of RTAs in South Africa, a better understanding of the relationship between RTA injury-severity and accident-related factors is needed.

Various methods exist to analyse large amounts of data. Two popular methods include unsupervised learning and supervised learning (SL) methods. Unsupervised learning methods are used to identify patterns in data consisting of multiple variables but no associated target variable (Gareth et al. 2013). SL methods fit a statistical model that relates a target variable to a set of predictor variables using a training set to tune the model. This model can then be used to either predict the target variable for a new set

of data, or to gain a deeper understanding of the relationship between the predictor variables and the target variable (Gareth et al. 2013). Popular SL methods include logistic regression, classification trees (CT), random forests (RF) as well as artificial neural networks (ANN). Each method has its own strengths and weaknesses depending on the problem or data that it is being applied to. These SL methods allow researchers to predict the injury-severity of RTAs more accurately than in the past and thus providing a better understanding of the relationship between accident-related variables and RTA injury-severity.

There are very few studies conducted on RTAs in South Africa and even fewer studies predicting RTA injury-severity. The use of SL methods can be very useful for informing future road safety campaigns and potentially reduce the frequency and injury-severity of RTAs in South Africa. The focus of this study is on RTAs that occurred in Cape Town in the Western Cape province of South Africa during the 2015-2017 period. The choice of the study area and year was guided by data availability and the fact that the Western Cape Province experienced 1 064 RTA fatalities in 2018.

1.1 Research Aim and Objectives

The aim of this study is to evaluate different SL methods in the prediction of injury-severity of Cape Town RTAs and consequently identify the most significant risk factors associated with RTA injury-severity. The main objectives of this study are:

- Review both international and South African literature on RTAs and RTA injury-severity prediction,
- Determine the best data sampling method to address the issue of class imbalance in RTA data,
- Utilise several different SL classification methods in order to predict the injury-severity of Cape Town RTAs and compare their effectiveness using suitable evaluation metrics,
- Identify the most significant risk factors associated with the injury-severity of Cape Town RTAs.

1.2 Significance/Contributions of the Study

This study contributes to the field by analysing and predicting the injury-severity of Cape Town RTAs through the use of SL methods. This study also presents the most effective data sampling method to address class imbalance which is prevalent in RTA datasets. A better understanding of the relationship between injury-severity of RTAs and other accident-related factors gained through SL methods and data corrected for class imbalance makes it possible to identify the most significant risk factors associated with RTA injury-severity. This could potentially inform governments on how to more effectively implement road safety policies and regulations thus potentially reducing the frequency and injury-severity of RTAs.

1.3 Structure of Dissertation

Chapter Two contains a review of previous RTA research conducted internationally as well as in South Africa. Chapter Three contains the description, processing and exploration of the Cape Town RTA dataset used for this study. The SL methods used for RTA injury-severity prediction are discussed in Chapter Four along with the evaluation metrics to measure model performance. Results are presented and discussed in Chapter Five. Finally, a summary of the main findings of this study as well as recommendations, limitations and suggestions for future work are presented in Chapter Six.

Chapter 2

Literature Review

The following sections look into previous research regarding the classification/prediction of RTA injury-severity outcomes. Classification problems use SL methods that are tuned using a training set to predict a categorical target variable related to a set of predictor variables for a new set of data (Gareth et al. 2013). Most of the literature on RTA injury-severity focus on developing statistical models that can accurately predict/classify RTA injury-severity based on accident-related factors such as driver, vehicle, roadway and environmental factors. While earlier research mainly applied relatively simple statistical models such as logistic regression, the latest literature includes the use of more modern and complex statistical tools such as classification and regression trees (CART), RF and ANN due to the availability of higher quality and larger volumes of RTA data. The first section discusses the research that applied various logistic regression models to RTA injury-severity data. The second section summarises research that made use of CART as well as other, more complex SL methods to model RTA injury-severity outcomes. The last section focuses on the South African literature on RTAs.

2.1 Logistic Regression in RTA Research

Regression methods are a very useful tool for analysing relationships between one or more predictor variables and a response variable. Logistic regression is used when the response variable is binary or categorical in nature and has been a widely used tool in RTA research. Logistic regression provides a powerful tool to test the relationships between

several predictor variables and injury-severity outcomes. Additionally, the odds ratios are a convenient method of measuring the magnitude of the effect that the predictor variables have on injury-severity outcomes in RTAs (Kim et al. 1995). When it comes to predicting injury-severity outcomes of RTAs, several statistical models have been effectively employed in previous research. This includes binary logistic regression (Al-Ghamdi 2002), multinomial logistic regression (Kong & Yang 2010), ordinal logistic regression (Mercier et al. 1997) as well as mixed logistic regression models (Milton et al. 2008). Poisson and Poisson-gamma regression models have been commonly used to predict RTA frequencies (Joshua & Garber 1990, Lord et al. 2005).

Al-Ghamdi (2002) investigated the influence of accident-related factors on accident severity in Saudi Arabia. The injury-severity of each accident was categorised as either fatal or non-fatal. Since the response variable was discrete and of a binary nature, the author made use of a binary logistic regression model to determine the influence of accident factors on the injury-severity. The data consisted of nine predictor variables related to the accident. The author found that the two accident-related factors most significantly associated with injury-severity were the location of the accident and the cause of the accident. The study determined that the odds of an accident taking place at an intersection being fatal were lower than for a non-intersection accident.

Similarly, Valent et al. (2002) made use of binary logistic regression to identify factors that contribute to RTA injury-severity in Italy. The study mainly focused on the injury-severity of the driver of the vehicle and was classified as either fatal or non-fatal. The main goal of the study was to determine which accident-related factors were associated with accident injury-severity. To determine this, the authors calculated odds ratios to estimate the likelihood of a more severe injury occurring compared to a lesser injury. A multivariate logistic regression model was used for the analysis and included several predictor variables related to the accident. The study found that men had higher odds of being involved in a fatal accident compared to women and that victims over the age of 65 had higher odds of sustaining a fatal injury compared to victims under the age of 30. Accidents occurring at night had higher odds of resulting in fatalities compared to accidents occurring during the day. Additionally, accidents that occur during fall and winter were found to be more likely to result in injury or death compared to accidents that occur during summer. Finally, fatal injury was found to be strongly associated with lack of seatbelt use.

Kong & Yang (2010) studied pedestrian-vehicle collisions in China with the aim of determining the association between vehicle impact speed and pedestrian age on the injury-severity of the pedestrian. The authors developed both univariate logistic regression models as well as multivariate logistic regression models for pedestrian fatality risk and pedestrian injury risk. Vehicle impact speed and pedestrian age were the two predictor variables of interest. The results showed that vehicle impact speed had a statistically significant relationship with both pedestrian injury risk and pedestrian fatality risk, while pedestrian age had a statistically significant relationship with pedestrian injury risk but not pedestrian fatality risk.

Nassar et al. (1994) made use of sequential binary logistic regression models to investigate RTA injury-severity outcomes in Ontario, Canada. According to the authors, previous studies have failed to take into account the interdependence among different injury-severities. The approach followed by the authors assumed that a sequential relationship exists between the different injury-severities. This means that in order to reach a higher level of injury-severity, the victim of a RTA must first exceed all the lower injury-severities. The authors considered five categories of injury-severity, namely: no injury, minimal injury, minor injury, major injury and fatality. In order to take the sequential relationship between injury-severities into account, the authors developed four binary logistic regression models to separate the injury-severity categories. The first model classified injury-severity as either no injury or at least a minimal injury. The second model classified injury-severity as either minimal injury or at least a minor injury. The third model classified injury-severity as either minor injury or a major injury. The fourth and final model classified injury-severity as either major injury or fatality. This approach allowed the conditional probabilities for each injury-severity category to be calculated given that the previous injury-severity category has been exceeded. Their results determined that seating position, seat belt use and the vehicle condition were the most important variables for classifying injury-severity outcomes.

Lukongo (2020) applied unordered multinomial logistic regression models to investigate contributing factors to RTA injury-severity in Louisiana, USA. Multinomial logistic regression allows the response variable to consist of more than two categories and can be considered an extension of the binary logistic regression model. Injury-severity was represented by five categories, namely fatal, severe injury, moderate injury, complaint injury and no injury. The data contained variables regarding driver characteristics, road and

environmental factors as well as temporal factors that were used as predictor variables in the multinomial logistic regression model. The choice of unordered classification of crash injury-severity was motivated by the threat of underreporting of no-injury RTAs by police officers and the flexibility of the model structure since there is no restriction on the classification of the injury-severity. Previous researchers have also recommended the use of unordered classification of injury-severities (Eustace et al. 2011). The author calculated the estimated response probability for each predictor variable and injury-severity class combination. The estimated response probability was calculated by subtracting 1 from the odds ratio. The results showed that drivers' gender and age had a significant effect on injury-severity. Specifically, male and older drivers had high risk of being involved in severe and fatal accidents. The results also found that the geometry of the road, major roads, weekdays and dry road surfaces increased the risk of fatal road accidents.

Kim et al. (1994) investigated the relationship between crash type and injury-severity of RTAs in Hawaii in the United States of America (USA) and emphasised the advantages of using log-linear analysis and logistic regression modelling to analyse injury-severity outcomes. Data for RTAs that occurred in the year 1990 was used with a focus on the crash type, seatbelt use and the injury level sustained. The authors used log-linear analysis to determine underlying relationships then converted the log-linear equations into a multinomial logistic regression function to estimate the model parameters. The multinomial logistic regression model was then used to calculate odds ratios which were used to compare the odds of sustaining a specific injury-severity for a given crash type. The odds ratio is very useful for interpretation purposes, since it estimates how much a predictor variable increases or decreases the odds of sustaining a particular injury-severity category. The results of the study indicated that rollover and head-on crash types had the highest odds of resulting in fatal injury. The authors also found that the use of interaction terms could be used to account for behavioural factors not included in the dataset.

Kim et al. (1995) went on to investigate the relationship between driver characteristics, behaviours and injury-severity in Hawaii, USA using a log-linear model. They found that driver intoxication as well as not wearing a seatbelt significantly increased the odds of sustaining severe injuries from a RTA. The results also indicated that driver demographics such as age and gender did not significantly affect the injury-severity.

Mercier et al. (1997) investigated the influence of gender and age on the injury-severity of head-on RTA collisions that occurred in Iowa, USA, while controlling for several accident-related factors. The factors controlled for were: collision type, vehicle speed, position in vehicle and safety restraint use. The study also attempted to compare, by age of the victim, the rates at which severe injuries occur in head-on collisions. To carry out the statistical analysis, the authors made use of ordinal logistic regression. Ordinal logistic regression is similar to multinomial logistic regression, except that the response variable is treated as an ordinal variable instead of a nominal variable. Injury severity was classified as either fatal, major or minor and was the target ordered categorical variable. The predictor variables consisted of both individual and interaction variables. The authors also made use of hierarchical and principal component logistic regression models in order to verify their findings from the ordered logistic regression model. Principal components logistic regression was used in order to overcome the issue of multi-collinearity between the high number of predictor variables. The results of the analysis determined that injuries sustained by older drivers/passengers are more severe than injuries sustained by younger drivers/passengers. A possible explanation is the loss of bone density and bone mass as people age. The authors also found variations in results occurred when the dataset was split by gender, with improvements in the results when data was split by gender compared to when the total population was used. This indicated that gender had an effect on the injury-severity of head-on collisions. The results showed that age was the most important predictor variable for both males and females, but that the use of a seatbelt was more beneficial for males compared to females. Islam & Mannering (2006) found similar results where separate models for male and female victims were found to be more accurate compared to the model that used the full sample.

Mixed logistic regression allows the model parameters to vary across observations, unlike the multinomial logistic regression model that only has fixed model parameters. It also relaxes several assumptions of the multinomial regression model such as the independence of irrelevant alternatives and the assumption of independent and identically distributed errors. It also accounts for potential unobserved heterogeneity in data (Jones & Hensher 2007).

Milton et al. (2008) made use of mixed logistic regression to investigate the injury-severity distributions of RTAs on different highway segments and how weather, traffic

and highway characteristics affected these distributions. The study made use of data on RTAs that occurred on highways in Washington State, USA. Contrary to previous RTA injury-severity research that modelled the injury-severity outcome, the authors aimed to develop a model that would estimate the proportions of the various injury-severities on different highway road segments. This study assumed that the reported accident frequencies for specific highway road segments are known before estimating the proportion of accidents resulting in each injury-severity outcome. The study did not make use of driver and vehicle characteristics. Instead, the study used roadway characteristics as well as environmental and weather factors. According to the authors, this allowed for a more general, non-event-specific interpretation of the factors that can affect injury-severity of RTAs. The authors considered three injury-severity categories, namely: property damage only, possible injury and injury. The injury category combined the original categories of evident injury, disabling injury and fatality since they made up such a small number of the total accidents in the dataset. The injury-severity of an accident was defined as the injury-severity of the worst-injured victim. The authors stated that it was important to use a model that allows for the possibility that the influence of variables that affect injury-severity can vary across different highway road segments. Due to variation in driver behaviour, the authors believed it was unrealistic to assume that the influence of roadway, weather and environmental characteristics on injury-severity remain the same across all highway segments. This was the main motivation for the use of a mixed logistic regression model in this study. The mixed logistic regression model accounts for possible unobserved heterogeneity between different highway segments that can arise due to human, vehicle or road and environmental factors. In other words, the mixed logistic regression model allows model parameters to vary randomly across different highway segments in order to account for variations in the influence that accident-related factors have on injury-severity. The study found that volume-related factors, such as daily traffic per lane, average daily truck traffic as well as weather conditions were best modelled as random parameters that can vary over different road segments. Conversely, the study found that roadway characteristics such as pavement friction and number of horizontal curves were best modelled as fixed parameters that remain constant across all road segments.

Islam & Hernandez (2013) applied mixed logistic regression to model injury-severity outcomes resulting from RTAs involving heavy vehicles on highways in Texas, USA.

The main objective of the study was to determine the significant contributory factors to injury-severity outcomes. The study used five categories of injury-severity and developed a separate mixed logistic regression model for each category of injury-severity. The authors stated that by not grouping together all the injury-severity categories, greater insights can be obtained regarding the influence that accident-related factors have on each injury-severity outcome. Similar to Milton et al. (2008), the authors emphasise that mixed logistic regression offers flexibility since it accounts for possible unobserved heterogeneity between observations/accidents related to human, vehicle, road and environmental factors not necessarily captured in the dataset. The mixed logistic regression model used in this study was conditioned on a crash already having occurred. The results of the study showed that driver demographics, roadway characteristics, weather, lighting conditions as well as the time of day were significant contributory factors for injury-severity outcomes. The authors also noted that while some predictor variables such as spatial and roadway characteristics were best modelled as fixed parameters, others like driver and traffic factors were best modelled as random parameters that vary across observations.

Kim et al. (2013) investigated single-vehicle RTAs in California and the factors that influence injury-severity while accounting for possible driver-specific heterogeneity due to age and gender. This was done by using a mixed logistic regression model with a driver-specific heterogeneous mean that is a function of the age and gender of the driver. The results showed that both age and gender cause heterogeneity in the single-vehicle crash population. The results showed that for drivers over 65 years old, half of the population had a higher risk of sustaining a fatal injury given that an accident occurred compared to the age group 24-65 years old. The other half of the 65+ population had a lower risk of fatal injury compared to the younger age group. This indicates that the parameter for old drivers is best modelled as a random parameter and supports the use of mixed logistic regression to model injury-severity outcomes. Similar to previous studies, the results also showed that male drivers had a higher risk of fatal injury given a single-vehicle crash occurred compared to female drivers. The authors went on to compare the mixed logistic regression model with a multinomial logistic regression model and found that most parameters were similar between the two models. The authors mention that heterogeneity could be captured by the multinomial logistic regression model by adding interaction variables. This could be used to avoid the complexity

of the random parameters used in the mixed logistic regression model. The authors conclude that the random parameters of the mixed logistic regression model does provide greater interpretive power compared to the fixed parameters of the multinomial logistic regression model.

Ye & Lord (2011) investigated the effects that underreporting of RTAs had on commonly used RTA injury-severity models. The models investigated included the multinomial logistic regression model as well as the mixed logistic regression model. Underreporting of crash data is a very common issue affecting RTA datasets around the world. It is therefore important to determine how it affects different models that are commonly used to model injury-severity of RTAs. In particular, crashes that result in low to no injury-severity levels are most often underreported. This can lead to underrepresentation of low and no injury-severity crashes in RTA datasets, while serious and fatal injury-severity levels are overrepresented. This could potentially lead to biased estimations for injury-severity classification models. The authors made use of a Monte Carlo approach that used observed as well as simulated RTA datasets. The results showed that both the multinomial and mixed logistic regression models were both affected by underreporting. The authors went on to show that by setting fatal injury-severity as the baseline severity for both the multinomial and mixed logistic regression models, the effect of underreporting on the two models could be minimised, the bias of the model can be minimised and the variability reduced.

Ye & Lord (2014) went on to investigate the effects of sample size on the multinomial and mixed logistic regression models. To investigate this, the authors once again made use of a Monte Carlo approach that used both an observed and simulated RTA dataset. The results of the study were consistent with prior research in that RTA injury-severity models were significantly influenced by the sample size from which they were estimated. This was expected, since logistic regression uses the maximum likelihood estimator (MLE) to estimate model parameters. The standard errors of parameter estimates are reduced as the sample size increases when using the MLE. Specifically, the results found that the minimum sample size to estimate a multinomial logistic regression model was 2 000 observations while the mixed logistic regression model required a minimum sample size of 5 000 observations. However, these are only the minimum sample size requirements and it would be beneficial to seek out larger datasets. The authors further noted that the

mixed logistic regression model was more interpretive and had a significantly better fit on the RTA injury-severity data compared to the multinomial logistic regression model.

The literature discussed above demonstrates that several logistic regression models can be applied to model RTA injury-severity. There is no consensus thus far on which logistic regression model performs the best. The selection of a specific logistic regression model is usually influenced by the availability and characteristics of the dataset (Savolainen & Quddus 2011). Injury-severity levels are inherently ordered, hence ordinal models have been widely used to model RTA injury-severity. However, ordinal logistic regression models have some significant limitations. They use the same coefficients/parameters for a variable for all of the response variable categories, which could potentially limit the performance of the model. Additionally, ordinal logistic regression models are significantly affected by underreporting of RTA injury data which leads to biased model parameters (Savolainen & Quddus 2011). This could be corrected if the true rate of underreporting in a population is known, but it is most often not known making corrections difficult. Due to these limitations of the ordinal logistic regression model, nominal models such as the multinomial or mixed logistic regression models are often preferred by researchers.

CART are a non-parametric classification method that do not require any pre-defined underlying relationship between the predictor variables and the target variables, unlike logistic regression models. This is advantageous, since if the model assumptions and underlying relationship between the predictor variables and target variables was violated in a parametric model, it could lead to misleading results. CART offers several advantages as a SL method such as its interpretability, computation speed and reasonable prediction accuracy (Loh 2014). Previous studies have indicated that CART methods outperform logistic regression when dealing with large sample sizes (Perlich et al. 2003). The following section discusses the use of CART in previous RTA research as well as other SL methods such as ANNs.

2.2 CART and Other SL Methods in RTA Research

About 57 years ago, Morgan & Sonquist (1963) published the first regression tree algorithm. Since then, CART methods have increasingly been used in research and applied

as a SL method. Breiman et al. (1984) published a book titled "Classification and Regression Trees" that sparked a renewed interest in decision tree methods. This included several improvements on previous tree methods, such as growing a large tree and then pruning it according to a specified cost-complexity parameter instead of using stopping rules when growing the tree. This approach solved the problem of under- and overfitting that previous decision tree methods suffered from. It also included a new approach to handling missing data. CART handles missing data by using a series of "surrogate" splits, which are splits on alternate variables that substitute for the split on the preferred variable when the preferred variable contains a missing value (Loh 2014). These surrogates can also be used to calculate a variable significance/importance score for the predictor variables. Since then, CART methods have become even more sophisticated. Modern CART methods can fit a variety of statistical models. For instance, modern classification trees can fit nearest neighbour, kernel density and several other models for the data partitions. They can also partition data with linear splits on subsets of variables. Similarly, modern regression trees can fit least-squares, logistic, Poisson and several other statistical models (Loh 2014).

All of these developments have led to an increase in the capabilities and prediction accuracy of CART methods. The availability and affordability of software that implements CART methods has also lead to greater adoption of CART methods by the research community. As a result, CART has become a popular data modelling tool for researchers investigating RTA injury-severity outcomes.

Chang & Wang (2006) conducted a study to examine the effectiveness of the CART model in identifying the risk factors that contribute to the injury severity of RTAs. The study made use of accident data containing accidents that occurred during 2001 in Taipei, Taiwan. The data included driver/vehicle characteristics, road and environmental factors as well as other accident-related factors. The injury-level of the worst-injured victim was used as the injury-severity of the accident. Similar to other studies, the road accident dataset was imbalanced with regards to the injury-severity classes. The injury-severity variable consisted of three classes and accounted for the following proportion of accidents in the dataset: No-injury (39.7%), Injury (59.9%) and Fatality (0.4%). Using the CART model, the overall model prediction accuracy for the training data was 90.3% and for the test data 91.7%. However, due to the very small number

of fatal accidents in the dataset, the CART model failed to classify accidents that resulted in a fatality. This emphasises the importance of having a balanced dataset with regards to the target variable when it comes to classification tasks. The results of the CART model identified vehicle type as the most important variable that contributed to injury-severity of RTAs. It also determined that pedestrians, bicyclists and motorcycle riders as having higher risk of being injured in a RTA compared to other types of vehicle drivers. The results of the study indicated that the CART model was an effective model to use when analysing the injury-severity of RTAs. The authors further discussed the main theoretical and practical advantages and disadvantages of the CART model. The main theoretical advantages of the CART method were identified as the following: not necessary to specify the functional form, can effectively handle multi-collinearity issues and not sensitive to outliers. The practical advantages include the graphic display of results that are easily interpretable as well as being effective at handling large datasets that include many predictor variables. The authors also noted that the CART model had some disadvantages. The CART model does not provide a probability level for the predictor variables and its predictions, it is difficult to conduct a sensitivity analysis to determine the marginal impact that different predictor variables have on the target variable and it is a high variance procedure. The authors recommended bagging procedures to assist in getting more stable/reliable predictions from tree-based methods, especially if the goal is to compare it to different supervised learning methods.

Kuhnert et al. (2000) attempted to demonstrate that sophisticated, non-parametric modelling methods have advantages over more traditional logistic regression models. In addition, they also showed that non-parametric methods such as CART and multivariate adaptive regression splines (MARS) could also be used as exploratory tools to improve logistic regression modelling. The authors made use of injury data resulting from RTAs in Brisbane, Australia. In order to determine whether risk-taking was a significant contributory factor to RTAs that result in serious injury or fatalities. The dataset contained variables such as driver aggression, age, gender, number of years driving experience as well as other demographic variables. Both the CART and MARS models achieved a higher classification accuracy than the logistic regression model. The authors further emphasised that each of the three modelling techniques had their own unique advantages and disadvantages. Even though each of the three techniques can be used as a modelling tool on their own, it can be advantageous to use the results of one tool to inform the

other. For instance, the authors demonstrate how the CART and MARS techniques can be used to identify the most important variables in the data. This information can then be used to inform the logistic regression model. By combining the strengths of each technique, a more informative and effective predictive model can be created. Additionally, by investigating the results of all three modelling techniques, a more complete analysis and understanding of the data can be achieved. The authors combined the results of all three modelling techniques to identify driving experience, seatbelts, gender and age as the most significant contributory factors of crashes that result in serious injury or fatalities.

Montella et al. (2012) analysed RTAs involving powered two-wheelers (PTW) in Italy. The aim was to determine interdependence and dissimilarities between the accident-related factors. The analysis was performed using both decision trees and rules discovery. Rules discovery is a method that is used to identify sets of items that occur together in a given event more often than they would if they were independent of each other. Using classification trees and rules discovery, the authors assessed several target variables including injury-severity, crash type, vehicles involved and alignment of the road. The results showed that rural provincial and national roads, curves in the road, PTW with greater cylinder capacity and night-time were associated with greater injury-severity. The results also identified head-on, run-off-the-road and hit pedestrian crash types as being associated with greater injury-severity. The authors also noted that PTW accidents were significantly associated with several combinations of driver, road and environmental factors. For instance, the combination of run-off-the-road crash type and curve alignment of the road were strongly associated with fatal injury-severity. These results further emphasise that RTAs are rarely caused by a single factor, but rather a combination of contributory factors. The authors conclude that both decision trees and rules discovery were effective in gaining a better understanding of RTAs and accident-related factors and that the two methods could be complementary.

Shanthi & Ramani (2012) compared the accuracies of different classification algorithms and the impact of feature selection on the performance of the different algorithms. Among the classification algorithms used in the study to model the injury-severity of RTAs were CART, naïve bayes and RF. Data for USA road accidents were used for the study. The data consisted of 33 accident-related variables as well as the target variable, the injury-severity of each accident. Using all 33 predictor variables included in

the dataset, the RF performed the best with a resultant misclassification rate of 14.2%. In order to reduce the error rates, the authors attempted to select only the most relevant predictor variables by applying several feature selection algorithms. After applying the feature selection algorithms, the performance of all the classification algorithms improved. The RF using only the relevant predictor variables as chosen by the Feature Ranking algorithm led to the best performance of all the classification algorithms in modelling injury-severity of RTAs. The results of the RF indicated that the manner of collision, age, drug involvement and seating position were the most important factors that contribute to injury-severity of RTAs. This study emphasised the importance of feature selection and its impact on the performance of classification algorithms. The results of this study support previous findings that tree ensemble methods such as RF generally outperforms single-tree methods.

ANNs are another SL method that has been used to classify RTA injury-severity outcomes. ANNs are a non-linear data modelling tool and are most often used to model complex relationships between predictor variables and a target variable (Akin & Akbaç 2010). One of the main benefits of ANN as a supervised learning method is that they do not require any pre-specified information and that they are effective at modelling non-linear and complex relationships between variables (Abdelwahab & Abdel-Aty 2001).

Abdelwahab & Abdel-Aty (2001) investigated the use of ANNs to predict driver injury-severity in RTAs that occurred in Florida, USA. According to the authors, the use of ANN could lead to a better understanding of the relationship between accident-related factors and injury-severity of RTAs. The authors made use of two popular neural network paradigms, namely the multilayer perceptron (MLP) and fuzzy adaptive resonance theory (ART) neural networks. The study specifically focused on two-vehicle accidents that occurred at signalised intersections. The data consisted of driver, vehicle, road and environmental characteristics related to the accident. The target variable, driver injury-severity, was divided into three classes namely: no injury, possible/evident injury and disabling injury/fatality. The authors performed variable selection using a Chi-squared test on the nominal/ordinal variables to determine the independence and significance of the variables. The results of the study indicated that the MLP neural network performed better than the fuzzy ARTMAP neural network. The authors then compared the performance of the MLP neural network to an ordered logistic regression model and found that the MLP neural network also performed better. The results of the

study also showed that rural intersections were more dangerous than urban intersections, female drivers had a higher risk of severe injuries compared to male drivers and that drivers not at fault were more likely to sustain serious injuries compared to drivers at fault. The authors concluded that ANNs had potential in modelling injury-severity of RTAs, especially MLP neural networks.

A study conducted in South Korea by Sohn & Shin (2001) applied three different classification algorithms to RTA injury-severity data. The algorithms included logistic regression, CART and ANN. Their aim was to compare the classification accuracy of the three methods as well as to identify the most important contributory factors of RTA injury-severity in South Korea. The authors used data captured by Traffic Accident Records (TAR) by the South Korean police. TAR forms contain data on RTA injury severity as well as several accident-related factors. The authors performed variable selection by performing a chi-squared test on the variables to determine their significance. The authors were able to successfully reduce the total number of predictor variables without significantly reducing classification accuracy. The results of the three classification algorithms identified the proper use of a protective device, either a safety belt or helmet, as the most important factor for classifying RTA injury-severity. However, the authors found that there was no significant difference in the classification accuracy between the logistic regression, decision tree or neural network models.

Chong et al. (2005) compared the performance of four supervised learning methods applied to predicting the injury-severity level of RTAs in the USA. The four supervised learning methods used in the study were ANN, classification trees, support vector machines (SVM) as well as a hybrid decision tree - neural network model. The study used road accident data for the years 1995-2000. The dataset only contained information regarding the driver of the vehicle. While most research in the past focused mainly on classifying injury-severity of RTAs as either fatal or no injury, the authors of the study added three more injury-severity classes. The target variable, injury-severity, therefore consisted of five classes namely: no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. The authors further subset the dataset to include only head-on front impact point road traffic accidents. This was done to ensure that the final dataset was relatively balanced between the different output classes. Head-on front impact point accidents had the highest proportion of fatal accidents. The features included in the final dataset were gender, alcohol usage, driver's age, road surface and

light conditions, roll-over and several vehicle characteristics. The authors then separated each target class and used a one-against-all approach. This involved the selection of one target class to be the positive class and assigned the value 1, while the rest of the target classes were combined into the negative class and assigned the value 0. Essentially, a separate classifier was constructed and trained for each injury-severity class in the dataset. The results from the study indicated that the classification tree and hybrid decision tree-neural network model outperformed both the SVM and ANN for every injury-severity class. The classification tree also outperformed the hybrid model for the no-injury and possible injury classes. Conversely, the hybrid model performed better than the classification tree for the non-incapacitating injury, incapacitating injury, and fatal injury classes. Most of the SL methods applied in the study performed the best when modelling the no injury and fatal injury-severity classes. The fatal injury-severity category is the most important to identify, since road fatalities carry the highest social and economic cost to society.

Olutayo & Eludire (2014) analysed data on road accidents that occurred on the first 40 kilometres of one of Nigeria's busiest roads, the Ibadan-Lagos Express. The authors made use of both CART and ANN to model the data. The study made use of a dataset obtained from the Nigeria Road Safety Corps that contained records of road accidents that occurred during a two-year period from January 2002 to December 2003. The dataset contained the following categorical variables: vehicle type, time of the day, season and the cause of the accident. The target variable of the study was the location of the accident. The 40 kilometre stretch of road was split into three separate regions, namely Region A, Region B and Region C. The authors made use of both CART and ANN to attempt to predict the location of an accident based on the available variables. Using the mean absolute error (MAE) as the main performance metric, the study determined that CART performed better than the ANN when predicting the location of a road accident. The study also concluded that the three major causes of road accident on Nigerian roads were tyre burst, broken shaft and loss of control.

The literature discussed above demonstrates that the use of non-parametric, sophisticated SL methods such as CART, RFs and ANNs can also successfully be applied to model RTA injury-severity. These non-parametric SL methods were shown to regularly outperform logistic regression methods. However, there is no consensus as to which of these non-parametric SL methods perform the best. Additionally, it has been shown

that the use of multiple SL methods can lead to a more complete analysis and understanding of the data. The following section discusses South African research regarding RTAs.

2.3 South African RTA Research

The Road Traffic Management Corporation (RTMC) is the main source of South African RTA data used by researchers. The South African Police Service (SAPS) collects fatal road traffic accident data (using the Culpable Homicide Crash Observation Report form) which is then sent to the RTMC. The RTMC is responsible for capturing, processing and verifying the data. RTA research in South Africa are descriptive in nature and largely consist of the identification of contributory factors, using RTA data sourced from the RTMC or from surveys. There are very few papers that make use of predictive modelling.

Accidents usually occur due to a combination of factors. If it is possible to identify these factors, efforts can be made to reduce/combat these factors and as a result reduce the frequency and injury-severity of road traffic accidents. According to the RTMC's State of road safety report (Department of Transport 2018), human factors were the main cause of 89.3% of fatal RTAs in South Africa in 2018. The most frequent causes were speeding, jay-walking and hit-and-run. Road and environmental factors were found to be the main cause of 6.5% of fatal RTAs in 2018 where poor visibility, wet road surfaces and sharp bends in the road were found to be the most frequent causes. Vehicle factors were the main cause of 4.2% of fatal RTAs with burst tyres the most frequent cause. Among all the victims of fatal RTAs in 2018, 26% were drivers, 33% were passengers, 38% were pedestrians and only 2% were cyclists. The report also determined that among fatal RTAs, the most frequent major crash types were head-on collisions, multiple vehicle crashes as well as single vehicles overturned. In total, South Africa recorded 12 921 fatalities resulting from RTAs in 2018.

Human behaviour is the cause of the majority of road accidents in South Africa compared to vehicle defects which account for far fewer road traffic accidents (Department of Transport 2018). However, vehicle defects are a factor that is very much in the control of the driver. Developing countries such as South Africa are more likely to have older, less reliable cars on the roads compared to developed countries (van Schoor et al. 2001).

It is therefore important to determine the impact that mechanical defects of vehicles have on RTAs.

In a study conducted by van Schoor et al. (2001), mechanical failures as a cause of motor vehicle accidents was investigated. The study used data on motor vehicle accidents in the Pretoria region of South Africa. The authors goal was to determine how much mechanical failures of motor vehicles contributed to RTAs and then compare it to international trends. The study determined that the two main mechanical defects that caused motor vehicle accidents were issues with tyres and brakes. However, the data showed that mechanical failures were only responsible for 3% of motor vehicle accidents, which is similar to international trends. Furthermore, the study collected information on the mechanical condition of road-going motor vehicles on both suburban roads as well as on highways in Pretoria and found that 40% of vehicles on suburban roads and 29% of vehicles on the highway had significant mechanical defects that violated South African road and traffic regulations.

Similarly, Moodley & Allopi (2008) aimed to determine the extent of vehicle defects in South Africa and their contribution to RTAs. The data used in the study was obtained from a pilot survey that collected data at randomly selected shopping centres in the eThekwini Metropolitan area in Durban, KwaZulu-Natal province. The result of the pilot survey determined that tyre defects were the most common vehicle defect. The authors also mention that the age and condition of the vehicles inspected differed between shopping centres located in high- and low-income areas. If these vehicle defects can be identified and managed, it could lead to a decrease in the number of RTAs on South African roads.

Botha & van der Walt (2006) discussed the results and findings of a road traffic offence survey obtained from the South African Department of Transport that was conducted on South African roads in 2005. The author used the information from the survey to determine the general level of lawlessness on South African roads and to identify contributory factors to RTAs. The authors classified contributory factors into the following four categories: road users, vehicles, roadway and environmental factors. According to the authors, human behaviour accounts for the first three categories, while only environmental factors are considered to be beyond the control of the driver of a vehicle. This suggests that the human element is a very significant contributory factor to RTAs.

The authors state that it is generally assumed that 90% of road accidents occur as a direct consequence of a traffic offence or violating prescribed norms and standards. The authors identified the most significant contributory factors to road accidents for each category. High speeds and pedestrian jay-walking were found to be the road user factors that contributed the most to fatal road crashes. With regards to vehicle factors, tyre bursts and faulty brakes were found to be the most common. Finally, sharp bends in the road as well as poor visibility and road conditions were the most reported road and environmental factors. The results of the survey also indicated that the level of lawlessness on South African roads were high, with speeding and intoxicated driving occurring at alarming rates. The authors concluded by making several recommendations for improving road safety in South Africa. The main recommendations were to improve traffic law enforcement, to improve road safety education and to identify hazardous pedestrian locations.

As previously mentioned, road user behaviour is the main contributory factor for the majority of all RTAs in South Africa. This indicates that in order to reduce the frequency and injury severity of RTAs, it is essential to change the behaviour of road users. According to Linu et al. (2013), 83.2% of the people who violated road rules were already aware of them. This makes changing road user behaviour a challenging task, since it is not as simple as informing road users about road safety rules. Moyana & Chibira (2016) determined that from the various possible interventions to improve road safety, interventions that focus on changing the behaviour of road users have a significant impact on improving overall road safety.

In order to accurately predict RTA injury-severity and inform intervention programs, comprehensive and good quality RTA data needs to be available/collected. In a study conducted by Chokotho et al. (2013), the quality of existing data sources on road traffic injuries in the Western Cape Province was investigated. In the Western Cape Province, several institutions collect RTA-related data. These include hospitals, mortuaries, the police as well as traffic authorities. The authors investigated the completeness of the police dataset by comparing it to the mortuary dataset. The mortuary dataset is used as the gold standard for the purpose of this study, since the authors assumed that mortuaries have comprehensive records of all non-natural deaths occurring in the Western Cape Province. In order to determine the completeness of the police dataset, the authors made use of the capture-recapture method which evaluates the degree of overlap

between two datasets to obtain a measure of data completeness. The authors identified several quality issues with the police RTA data namely underreporting, duplication as well as missing values. The authors concludes that in order to effectively inform road traffic prevention programs, the data quality issues identified had to be addressed.

The lack of good quality RTA data in South Africa is concerning and limits the effectiveness of research attempting to classify RTA injury-severity outcomes. Previous studies have shown that missing values in datasets used to train and test classification algorithms can negatively affect their prediction accuracy (Saar-Tsechansky & Provost 2007).

Twala (2013) investigated RTAs in the Gauteng Province in South Africa. The study aimed to determine the effect of incomplete RTA data on the predictive accuracy of different SL methods. The SL methods used in this study included ANN, SVM, CART, k-nearest neighbour, naïve Bayes classifier as well as a grey relational classifier which can be useful when only incomplete data are available. The study made use of RTA data obtained from the RTMC. However, like most real-life RTA datasets, the data contained a large number of missing values. The injury-severity in the data was classified into four classes, namely: fatal injury, serious injury, slight injury and property damage. In order to determine the effect that incomplete data had on the different SL methods, the author made use of two datasets. The first dataset was the original, incomplete dataset. The second dataset was created by imputing all the missing values contained in the original dataset in order to obtain a complete dataset. The author then calculated a measure called the excess error for each method. The excess error for each method was calculated as the classification error using incomplete data and then subtracting the classification error when using the complete data. This allowed the author to determine the effect on the predictive accuracy of each method when using complete versus incomplete data. The study found that there were significant differences in the predictive accuracy of all the SL classifiers when complete versus incomplete RTA data was used. The difference for all methods was significant at the 5% level. Overall, the grey relational classifier was found to be the most robust classifier when dealing with incomplete data. Additionally, the CART method was found to be more robust than the ANN. The study also conducted a variable importance test for the different SL methods and calculated the overall average for each variable. The most important determinants of RTAs and injury-severity in

Gauteng were found to be the age of the driver, driving while intoxicated, traffic control device, the road condition as well as the speed of the vehicle before the accident.

Mokoatle et al. (2019) investigated the performance of different classification methods in order to predict the injury-severity of drivers involved in RTAs. The study used RTA data for the period 2015 to 2017 that occurred in Soshanguve, Pretoria. The data included several features including road characteristics, injury severity of the driver, lighting- and weather-conditions as well as details about the driver and the vehicle. The authors decided to add a distance feature to the road accident data that represented the distance to the nearest place of interest from where the accident occurred. For the purposes of the study, places of interest included malls, bars, schools and similar locations. The purpose of adding the distance feature to the data was to determine if the distance to places of interest had an impact on the injury-severity of drivers involved in RTAs. The data used by the authors contained a lot of missing data, with nearly all observations containing a missing value for at least one predictor variable. They determined that the data was missing at random. In order to solve this issue, the authors made use of k-Nearest Neighbour multiple imputation to fill in the missing values in the dataset. The dataset used by the authors, like most real-world datasets, had a class-imbalance of the target variable/injury-severity of the driver. In order to balance the data, the authors made use of the Synthetic Minority Oversampling Technique (SMOTE). This technique fixed the imbalance problem by oversampling the minority class in the data. In this study the minority classes were the killed, serious and slight injury-severity classes. The majority class was the no injury class. The classification models used to predict the injury severity were multivariate logistic regression and the Extreme Gradient Boosting Tree (XGBoost). XGBoost is an ensemble learning method that is an implementation of gradient-boosted decision trees. It is a method that is well known for its scalability and speed (Chen & Guestrin 2016). Each of the two classification models was trained using three different datasets. The first dataset contained only the distance feature. The second dataset contained all the information captured on the accident report form obtained from the SAPS. The third dataset combined the accident report form data with the distance feature. The authors first made use of a relatively simple classification method, namely multinomial logistic regression. Then a more complex model, XGBoost, was used classify the injury severity. The authors made use of accuracy, precision, recall and the F1-score to evaluate the performance of the

different classification models. All three of the multinomial logistic regression classifiers performed poorly on the data, regardless of the dataset used to train it. Even the classifier that was trained using both the distance feature as well as the accident report form data failed to outperform the other two classifiers. Since multinomial logistic regression performed poorly on the data, the authors decided to use a more complex/powerful classification model, namely XGBoost. The XGBoost classifier using the combined dataset containing both the distance feature and the accident report data outperformed the other two XGBoost classifiers that were trained on the separate datasets. This result indicated that data related to the location of an accident combined with the usual traffic accident data such as vehicle type, road characteristics and light and weather conditions can improve the performance of RTA injury-severity prediction models. Additionally, the study determined that changing the multi-class label of injury-severity into a binary variable had a negative impact on classification performance. It also showed that model performance could be improved by ensuring that all classes of the target variable were almost equally distributed in the dataset. Using the XGBoost model with the combined data, the authors determined that truck license code C1, vehicle type, light motor duty license code EB, vehicle manoeuvre as well as the distance to the closest building had a significant impact on injury-severity of drivers involved in RTAs.

In 2020, the RTMC published a study in which driver intoxication and fatal crashes were investigated using frequency and proportion analysis as well as logistic regression (Govender et al. 2020). The study made use of data on 13 074 fatal RTAs during the period 2016-2018. The study found that over three quarters of the fatal RTA victims in the dataset were male and that speeding was most frequently found to be the main cause of fatal RTAs. It was also found that driver intoxication significantly increased the risk of RTAs that involved both the driver as well as other road users. The study also made some notable findings regarding temporal, vehicle and spatial characteristics of fatal RTAs in South Africa. It was found that the majority of fatal RTAs occurred at night (55%), over weekends (64%) and during non-vacation periods (70%). It also found that RTAs involving minibuses (public transport) were significantly more likely to be fatal for both passengers and pedestrians. Regarding the spatial characteristics of fatal RTAs, the study found that 72% of fatal RTAs occurred in the jurisdiction of local municipalities as opposed to metropolitan municipalities.

It is evident that South African literature on RTAs mostly consist of identifying significant contributors to RTAs. The literature has shown that the quality of South African RTA data is generally poor due to issues such as underreporting, duplication as well as missing values in the data. There are limited studies modelling RTA injury-severity using SL methods, with the focus area mainly in the province of Gauteng. There is a definite need for more research focusing on South African RTA injury-severity prediction using SL methods, especially in the Western Cape province.

2.4 Summary

This chapter contained a review of both international and South African literature regarding RTAs. Various supervised learning methods have been used to model RTA injury-severity in international literature. These methods range from simpler models such as logistic regression to more complex methods such as ANNs. South African literature mostly consists of identifying significant contributors of RTAs. There is a need for research focusing on South African RTA injury-severity prediction using SL methods. This will lead to a better understanding of the relationship between risk factors and RTA injury-severity, which could be used to inform future rules and regulations regarding road safety in South Africa. The following chapter describes and explores the dataset used in this study.

Chapter 3

Data

3.1 Description

Data on RTAs in Cape Town were sourced from the City of Cape Town. The dataset contains records of more than 82 000 RTAs that occurred during the 2015-2017 period. The choice of the study area and years was guided by data availability and the fact that the Western Cape Province experienced 1 064 RTA fatalities in 2018. The dataset contains several variables related to the accident. These variables include: street name, crash date, weekday, time of day, alleged cause, crash type, vehicle type, number of vehicles, number of passengers, number of pedestrians involved in the accident as well as the worst injury-severity sustained during the accident.

This dataset was expanded by adding weather-related variables such as temperature, precipitation, wind speed, visibility and cloud cover for Cape Town on the day the RTA occurred (*Visual Crossing Weather API* 2021). Several other variables such as those relating to whether an accident occurred on a public holiday, on a weekend, the season the accident occurred, the number of vehicles involved, whether the accident occurred during peak traffic times as well as whether an accident occurred at an intersection or non-intersection were also added.

Since this dissertation aims to determine the effect that the location of an accident (amongst other variables) has on predicting the injury-severity, the street addresses of the accidents were geocoded in order to obtain geographical coordinates for each accident. After inspecting that valid coordinates were returned for each accident's street

address, the longitude and latitude coordinates were added as variables to the dataset. The full set of variables and their descriptions can be found in Table B.1 in Appendix B. Figure 3.1 shows a map of the accident locations in Cape Town.

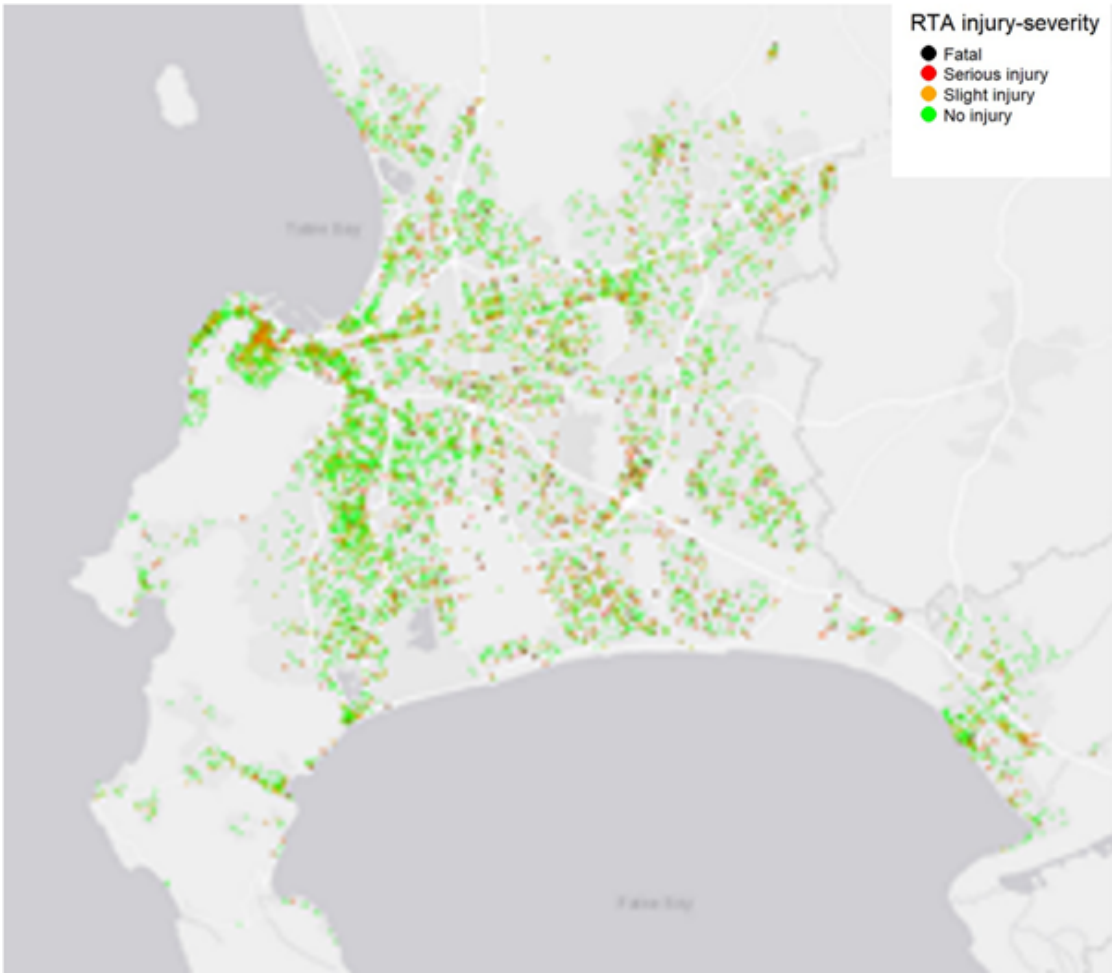

FIGURE 3.1: Geographic locations of RTAs in Cape Town 2015-2017.

A common issue with RTA datasets is that they are imbalanced with regards to injury-severity. The target variable for this study is the worst injury-severity sustained by a person during a RTA. The target variable consists of four injury classes, namely: "fatal", "serious", "slight" and "no injury". The data used in this study is severely imbalanced with regards to injury-severity as can be seen in Figure 3.2. According to Chawla et al. (2002), a dataset is imbalanced if the number of observations belonging to each of the classification categories are not approximately equal. Imbalanced data can negatively affect the performance of certain classification methods, especially with regards to predicting the minority class (Weiss & Provost 2001). This is an issue since the minority class is often the class researchers are most interested in predicting correctly.

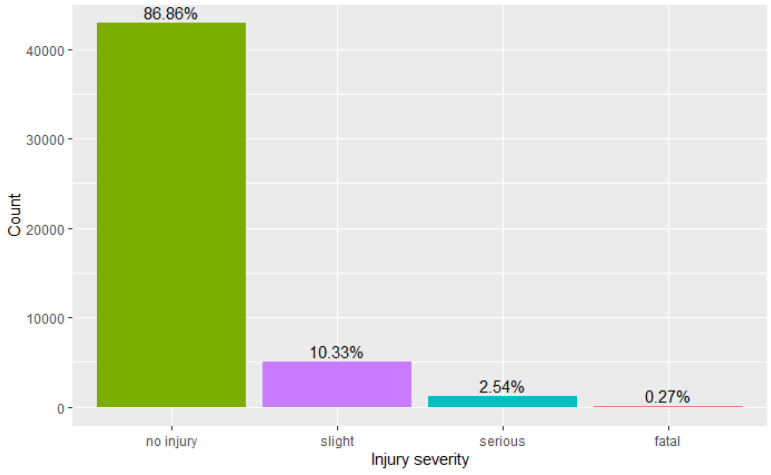

FIGURE 3.2: The distribution of injury-severity of RTAs in Cape Town 2015-2017 (N= 82 363).

Since it is important to be able to accurately predict RTAs that result in some sort of injury, it is necessary to address the data imbalance issue before beginning the process of building classification models. Several methods exist to address the issue of imbalanced data. The two most common approaches used by researchers to address imbalanced data are cost sensitive learning and data sampling techniques. Cost sensitive learning methods address the class imbalance issue by assigning higher costs/penalties for misclassifying the minority class compared to misclassifying the majority class. Sampling methods involve modifying the original, imbalanced dataset to achieve a new dataset with a more equal distribution of the target variable's classes.

Three common data sampling approaches used by researchers to address imbalanced data are utilised in this study, namely (i) undersampling of the majority class, (ii) oversampling of the minority class and (iii) SMOTE. SMOTE is a popular over-sampling method developed by Chawla et al. (2002) that creates artificial data examples of the minority class in order to improve the imbalanced distribution of the target variable. While random over-sampling methods simply duplicate existing minority class examples, SMOTE creates artificial minority examples by extrapolating between existing minority examples. It does this by by finding the k-nearest neighbours of the minority class for each minority example and then generating artificial examples in the feature space of the nearest neighbours.

3.2 Exploratory Data Analysis

After expanding the dataset by adding new variables, the data was then prepared and cleaned. This process involved identifying missing values, removing duplicate observations, transforming variables, grouping of categories for variables with a large number of levels as well as identifying outliers and removing invalid values for variables. The RTA dataset received from the City of Cape Town originally contained 219 052 RTAs, but after cleaning the data and only selecting RTAs whose street addresses could be correctly geocoded, the final dataset contained 82 363 RTAs. The final, cleaned RTA dataset had similar proportions of injury-severity to the original RTA dataset. The next step after cleaning the data is to explore the dataset. This is done in order to find interesting patterns in the data, identify significant variables and detect outliers and/or errors in the data. The data cleaning, processing and exploratory analysis for this study was conducted using R (R Core Team 2019) and various packages within R, all of which can be found in Appendix C.

Figure 3.3 shows the distribution of the different crash types in the cleaned RTA dataset. It can be seen that head-on/rear-end and sideswipe crashes are the most common crash types in the dataset. From previous research it is known that RTAs usually result from either human, vehicle, road or environmental factors. The distribution of the alleged cause (as deduced by the data capturers from the accident report) of the RTAs are shown in Figure 3.4. Insufficient following distance as well as other driver-related causes, which contain several driver-related causes, account for more than half of the RTAs in the dataset. Figure 3.4 shows that the vast majority of alleged causes of RTAs are related to driver/human error, similar to the findings of the Department of Transport (2018), while vehicle, road and environmental factors only account for about 4% of RTAs in the dataset. It is clear that in order to reduce RTAs, focus should be placed on reducing driver-related errors and causes. This is further supported by Table 3.1 which shows the distribution of alleged cause for each injury-severity. Pedestrian and other driver related causes account for the vast majority of both “fatal” (91.9%) and “serious”(81.6%) RTAs.

In order to effectively reduce the number of RTAs that result in “fatal” and “serious” injuries, it is important to identify the crash types most associated with them. The distribution of crash types for each injury-severity category is examined next. While accidents with pedestrians constitute only 5.86% of all RTAs, Table 3.2 shows accidents

TABLE 3.1: Distribution of alleged cause by injury-severity.

		Injury-severity			
		fatal	no injury	serious	slight
Alleged cause	Bypass distance too close	1.40%	6.10%	1.20%	1.60%
	Change lane while unsafe	0.00%	8.10%	1.40%	3.00%
	Entered traffic while unsafe	2.70%	9.40%	5.70%	8.20%
	Insuff. following distance	2.30%	27.30%	6.50%	20.40%
	Other driver related	47.30%	33.70%	46.30%	35.70%
	Pedestrian related	44.60%	1.80%	35.30%	26.20%
	Roadway and Environment related	0.50%	3.20%	1.50%	2.20%
	Vehicle related	0.50%	1.00%	1.10%	1.10%
	Vehicle reversed	0.90%	9.50%	0.90%	1.70%
	Total	100.00% (222)	100.00% (71541)	100.00% (2095)	100.00% (8505)

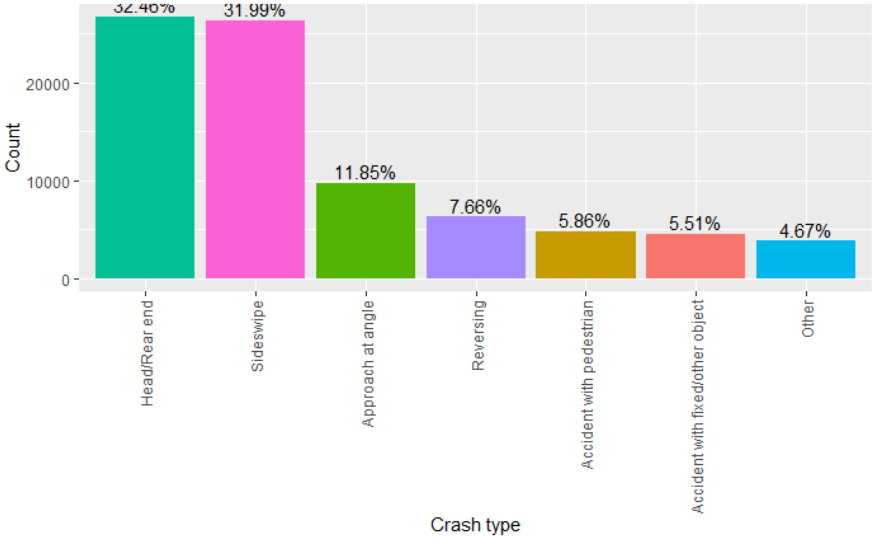

FIGURE 3.3: Distribution of crash type.

with pedestrians make up 58.56% of "fatal" accidents and 55.37% of "serious" accidents. In order to reduce the number of RTAs that result in "fatal" and "serious" injuries, focus should thus be placed on reducing pedestrian accidents.

TABLE 3.2: Distribution of crash type by injury-severity.

		Injury-severity			
		fatal	no injury	serious	slight
Crash type	Accident with fixed/other object	7.21%	5.71%	4.34%	4.15%
	Accident with pedestrian	58.56%	1.30%	55.37%	30.65%
	Approach at angle	6.31%	11.64%	10.36%	14.13%
	Head/Rear end	4.50%	34.11%	8.11%	25.26%
	Other	14.86%	4.10%	9.45%	7.96%
	Reversing	0.00%	8.68%	0.38%	1.11%
	Sideswipe	8.56%	34.46%	11.98%	16.74%
	Total	100.00% (222)	100.00% (71541)	100.00% (2095)	100.00% (8505)

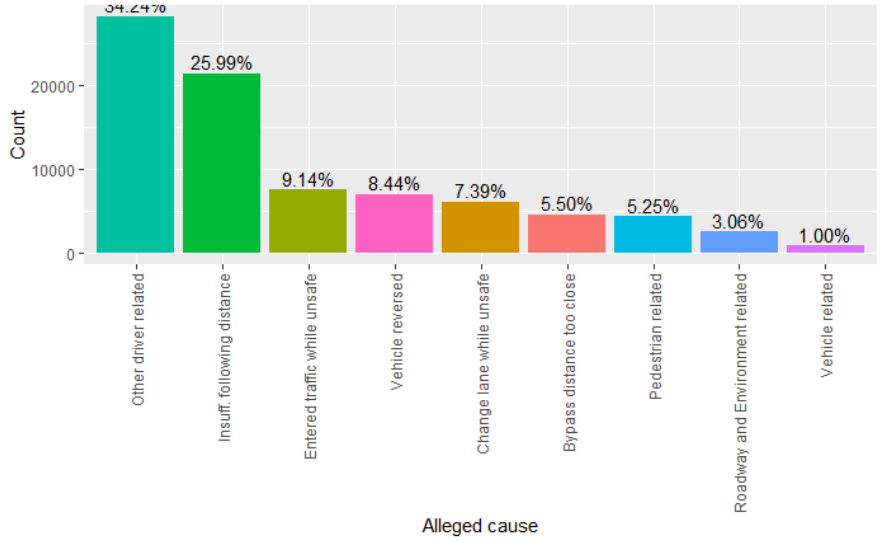

FIGURE 3.4: Distribution of alleged cause.

Table 3.3 shows the distribution of RTAs that occur on weekends for each injury-severity category. According to Yau (2004), RTAs occurring on a Friday follow a similar pattern to RTAs occurring on the weekend. As a result, an accident occurring on Friday, Saturday or Sunday is classified as occurring on the weekend. The table shows that the majority of RTAs that result in "fatal" or "serious" injuries occur on the weekend, similar to the findings of Govender et al. (2020). In contrast, the majority of RTAs that result in "slight" or "no injuries" occur during weekdays (non-weekend).

TABLE 3.3: Distribution of accidents occurring on weekends by injury-severity.

		Injury-severity			
		fatal	no injury	serious	slight
Weekend	FALSE	45.05%	60.71%	49.16%	57.52%
	TRUE	54.95%	39.29%	50.84%	42.48%
	Total	100.00% (222)	100.00% (71541)	100.00% (2095)	100.00% (8505)

TABLE 3.4: Distribution of the number of vehicles involved by injury-severity.

		Injury-severity			
		fatal	no injury	serious	slight
Number of vehicles involved	1	72.97%	8.62%	63.53%	38.12%
	2	23.87%	88.55%	32.46%	55.37%
	3+	3.15%	2.83%	4.01%	6.51%
	Total	100.00% (222)	100.00% (71541)	100.00% (2095)	100.00% (8505)

An interesting relationship between injury-severity and the number of vehicles involved in an accident is shown in Table 3.4. The table shows that 72.97% of "fatal" and 63.53% of "serious" injury RTAs involve only one vehicle. In contrast, the majority of RTAs that result in "slight" or "no injuries" involve two vehicles. This seems to indicate that reducing single vehicle accidents would be an effective method of reducing RTAs that result in "fatal" or "serious" injuries.

Next, the Cape Town RTA dataset is split into training, validation and test sets containing 60%, 20% and 20% of the RTAs in the full dataset. It is important for the validation and test data to be sufficiently large in order to obtain reliable performance measures of the SL methods. The proportions of the injury-severity category are kept consistent between all datasets to ensure that the training, validation and test data are representative of the original data. Four separate training datasets are then constructed which include the original imbalanced data, data with the minority class oversampled, data with the majority class undersampled and the SMOTE data containing artificially created observations. Each of these four datasets are then used to train different SL models. The following section discusses the methodology and different SL methods used to predict the injury-severity of Cape Town RTAs.

Chapter 4

Research Methodology

4.1 Supervised Learning

Supervised learning methods fit a model that relates a target variable, either numerical or categorical, to a set of predictor variables using a training dataset. It is assumed that there exists a relationship between the predictor variables and the target variable that can be written as (Gareth et al. 2013, p. 17):

$$\mathrm{Y} = \mathrm{f(X)} + \varepsilon \tag{4.1}$$

, where f represents an unknown, fixed function of how the predictor variables relate to the target variable and ϵ represents the random error that has a mean of 0 and is independent of the predictor variables (X). Since the function, f, is usually unknown it must be estimated using existing observations/training data. Different SL methods, both parametric and non-parametric, can be used to estimate f. Since the expected value of the error term is 0, the target variable can be estimated as:

$$\hat{Y} = \hat{f}(\mathrm{X}) \tag{4.2}$$

, where $\hat{f}$ is the estimated function, f, and $\hat{Y}$ the predicted value of the target variable. SL methods estimate the function by optimizing a cost, error or loss function during the training process in order to obtain the parameters that minimise the prediction error

of the estimated function. By estimating f, the target variable can be predicted using the predictor variables. It also provides insights regarding the relationship between the target variable and each predictor variable.

The target variable in this study is the worst injury-severity caused by a RTA, which is a categorical variable. Due to the categorical nature of the target variable, SL classification methods will be utilised. The goal is to apply different SL methods to RTA injury-severity training data in order to accurately estimate the function/relationship between RTA injury-severity and accident-related factors.

This analysis will make use of CT, RF and Gradient Boosted Machine (GBM) methods. These methods have been chosen because they do not have the limitations of logistic regression and are more effective at handling large datasets with many predictor variables. The performance of these models are used to determine the most effective (leading to greatest improvement in a specified evaluation metric) data sampling method to address class imbalance in RTA data. Additionally, a multinomial logistic regression model as well as an ANN model are fitted to the RTA data. This is done in order to determine whether the relationship between RTA injury-severity and accident-related factors can be modelled sufficiently using a simple SL method such as multinomial logistic regression or whether more complex methods such as ANNs are required.

4.1.1 Multinomial Logistic Regression

As mentioned in Chapter Two, several logistic regression models have been employed in previous RTA research such as binary logistic regression, multinomial logistic regression, ordinal logistic regression as well as mixed logistic regression models. Multinomial logistic regression is a SL method that is used for classification problems. It is an extension of binary logistic regression that is used when the target variable consists of more than two categories. Multinomial logistic regression is often used in research to model RTA injury-severity since it does not require the assumptions of normality, linearity or homoscedasticity (Abdulhafedh 2017).

Multinomial logistic regression uses one category of the target variable as the reference category for the other categories of the target variable to be contrasted/compared to. For this analysis, the “no injury” category was selected as the reference category since

the majority of RTAs in the data belong to this category. The multinomial logistic regression model calculates the probability of an RTA belonging to each injury-severity category relative to "no injury" as follows (Yasmin & Eluru 2013):

$$P_i(j) = \frac{\exp(\beta_j X_{ij})}{\sum_{j=1}^{J} \exp(\beta_j X_{ij})} \tag{4.3}$$

, where $P_i(j)$ is the probability of RTA i resulting in injury-severity category j, X_{ij} is a vector of predictor variables and β_j is a vector of coefficients to be estimated for injury-severity category j. Essentially, a separate logistic regression model and a set of coefficients is estimated for each injury-severity category relative to the reference category "no injury".

4.1.2 Classification Tree (CT)

Classification trees (CT) are a non-parametric classification method that do not require any pre-defined underlying relationship between the predictor variables and the target variables to be specified. CTs use a tree-like structure in order to model the relationship between the predictor variables and the target variable. CTs offer several advantages as a SL method such as its interpretability, computation speed and reasonable prediction accuracy (Loh 2014). CTs can also effectively handle multi-collinearity issues, are resistant to outliers and effective at handling large datasets that include many predictor variables.

CTs are grown using recursive binary splitting to create a hierarchy of univariate binary decisions. A split point on a single variable is determined at each internal node of a CT. A diagram illustrating the structure of a basic CT can be seen in Figure 4.1. At each node a single variable, X, is evaluated to create a split point. The final nodes of the CT, also known as terminal nodes, represent the predicted category R of the target variable. The size of a CT is determined by the number of terminal nodes, which in the case of the CT illustrated in Figure 4.1 is five terminal nodes.

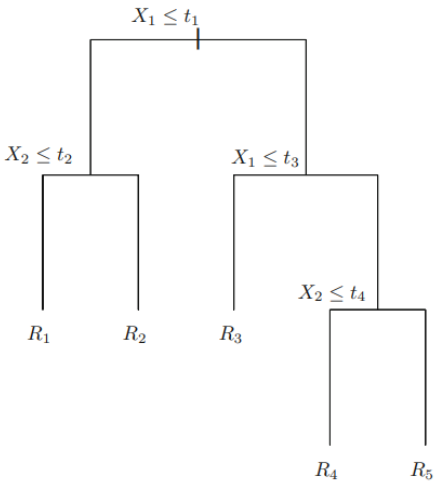

FIGURE 4.1: Diagram of a decision tree [Adapted from Gareth et al. (2013)].

The three most commonly used splitting criteria for CTs are the Gini index, entropy and deviance. The "tree" package (Ripley 2019) used in this dissertation uses deviance as the splitting criteria. Deviance is calculated as:

$$\text{Deviance} = -\sum_{k=1}^{K} \hat{p}_{\text{mk}} \log \hat{p}_{\text{mk}} \tag{4.4}$$

, where k=1,...,K represents the classification category, m the tree node and $\hat{p}_{\text{mk}}$ the proportion of class k observations in node m. When using deviance as a splitting criterion, the CT is grown by choosing the split that reduces deviance the most at each step. Once a fully grown CT has been constructed, it must be pruned in order to avoid overfitting the model. Pruning can be done using a technique known as cost complexity pruning. Cross-validation, which is a resampling method used to estimate the test error using training data (Gareth et al. 2013), can then be used to estimate the misclassification rate (percentage of predictions that were wrong) of each of the different sized trees created by the pruning process. Finally, the tree resulting in the lowest cross-validation error is chosen as the optimal CT (Loh 2014).

4.1.3 Random Forest (RF)

While a CT might be easily interpretable, it comes at the expense of predictive accuracy as well as high sampling variability (Chang & Wang 2006). RFs are used to reduce the variance of CTs. This method is built on the idea that averaging a set of predictions reduces the variance. RF is an ensemble learning method, meaning that many CTs are combined/ensembled into one, better model. A RF model is built by growing a multiple number of trees, B, on B bootstrapped samples (random sub-samples of data with replacement, $B>0$).

As an example, if there are B training sets, it would be possible to build a CT on each one and then average the predictions of all the trees on the test set. The average predictions for the test set, $\hat{y}_{\text{ave}}$, would be calculated as:

$$\hat{y}_{\text{ave}} = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_{\text{b}} \tag{4.5}$$

Since predicting injury-severity of RTAs is a classification problem, the prediction for an observation is the injury category that is the most commonly occurring among the B predictions. This process is illustrated in Figure 4.2.

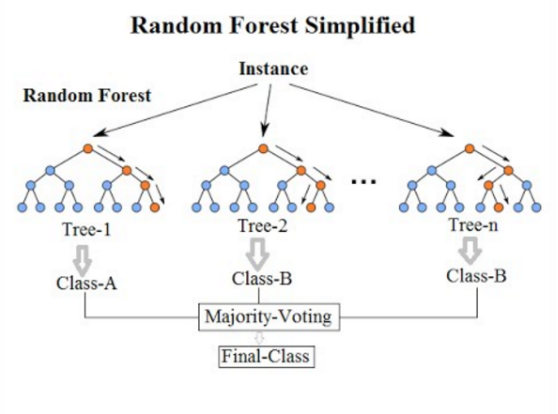

FIGURE 4.2: Diagram of a Random Forest [Adapted from Koehrsen (2020)].

Each time a split in a tree is considered, a random sample of m predictor variables are chosen as possible split candidates where $m < p$, with p representing the total number of predictors in the data. This decorrelates the trees since all the trees will look different. This also results in reduced sampling variability of the predictions. The number of

predictors to be considered at each split, m, for RF models are usually calculated as (Gareth et al. 2013, p. 320) :

$$m \approx \sqrt{p} \tag{4.6}$$

This dissertation will use cross-validation to find the optimal number of predictors to be considered at each split as well as the optimal number of trees to be grown for the RF model. The test error of the RF model can be estimated by calculating the Out-Of-Bag (OOB) error (Gareth et al. 2013, p. 318). The OOB error is calculated from the OOB observations which are the observations that were not used to grow a specific CT. To obtain an estimate of the test error, the response for an observation is predicted using each of the trees for which that observation was OOB (not included in the bootstrapped sample). The predictions for that single observation are then combined into one prediction per observation. From these predictions the out-of-sample classification error can be calculated.

4.1.4 Gradient Boosting Machine (GBM)

Gradient boosting machines (GBM), similar to RF, is an ensemble learning method. Unlike RF, which grows trees independently, GBM grows trees sequentially. This means each tree can learn from the mistakes made by the previous trees. In a GBM model, the separate trees are correlated and dependent. The residuals of the first tree informs the building of the next tree, and the residuals of the second tree informs the building of the third tree. This process continues until a sequence of B trees are built, each one accounting for some variation in the target variable that was not explained by the previous trees. The first few trees will identify the strongest patterns present in the data and as a result will provide the biggest reductions in the classification error.

A GBM model has the following three tuning parameters (Gareth et al. 2013, p. 322) :

- The number of trees, B
- The shrinkage parameter or learning rate, λ
- The number of splits in each tree, d

Cross-validation can be used to compare the effect of different values of these tuning parameters on the predictive accuracy of the model. By using cross-validation to evaluate a range of different parameter values, the optimal parameter values can be obtained to fit the GBM model. In general, methods that learn slowly/build more trees tend to perform better. The learning rate and number of splits parameters, λ and d, affects the rate at which the cross-validation error decreases.

4.1.5 Artificial Neural Networks (ANN)

Artificial neural networks (ANN) are a SL method that can be used for both regression and classification problems. ANNs are especially useful when one does not need interpretable results and when there are non-linear relationships present in the data (Trevor et al. 2009). ANNs model the relationship between predictor variables and the target variable by using a model that is derived from how a biological brain reacts to stimuli from sensory inputs. ANNs mimick this by using a network of artificial neurons in order to solve statistical learning problems.

In this dissertation, a MLP which is a type of feed-forward ANN will be used as previous research has shown it to outperform other SL methods (Abdelwahab & Abdel-Aty 2001). ANNs consist of layers, where a simple ANN would consist of only an input and an output layer, while a complex ANN would contain several hidden layers as well. Each hidden layer in an ANN receives multi-dimensional input and then processes it using weighted summation and an activation function. The output of each hidden layer is then fed forward to the next layer. The connections between layers of an ANN are weighted and these weights are optimised in order to minimise the error function. The error function in this case is Cross-entropy, which is calculated as (Trevor et al. 2009, p. 395) :

$$\text{Cross-entropy} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{\text{ik}} \log f_{\text{k}}(x_{\text{i}}) \tag{4.7}$$

, where $f_{\text{k}}(x_{\text{i}})$ is the predicted target value, y_{ik} is the actual target value, K is the number of categories in the target variable and N is the number of observations.

When developing an ANN model, initial values for weights are usually chosen as random values close to zero (Trevor et al. 2009). Using the training data, if the ANN correctly

predicts the class, the weights remain unchanged. However, if the ANN wrongly predicts the class, the weights of the ANN are updated proportional to the value of the input. This process of minimising the error function is known as backward propogation (Trevor et al. 2009). The optimal weights of the ANN model should be chosen in such a manner that it minimises the error function.

The output layer of the ANN will use the "Softmax" activation function, which normalises the output of an ANN to a probability distribution over the predicted output classes. The observation is classified as the class of the target variable that has the highest probability value. In this dissertation, the output layer for the ANN model will have 4 neurons since the target variable has 4 categories. An illustration of an ANN with a single hidden layer can be seen in Figure 4.3.

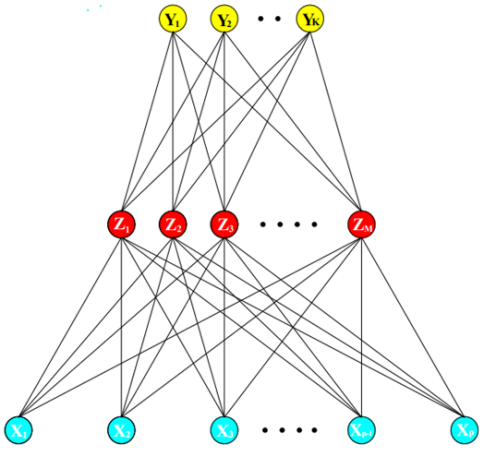

FIGURE 4.3: Diagram of a single layer, feed-forward ANN [Adapted from Trevor et al. (2009)].

The complexity of ANNs can be controlled by specifying the number of hidden layers as well as the number of neurons in each hidden layer in the model. Weight decay will be used to avoid overfitting the ANN model. Weight decay is a regularisation method which is applied to the weights of an ANN. Weight decay adds a penalty term to the error function being minimised in order to decrease the magnitude of the weights and avoid overfitting (Trevor et al. 2009).

Since there are several parameters that can influence the performance of the model, cross validation will be used. Different values for the following parameters are used in order to find the optimal model:

- Number of neurons in hidden layer 1,
- Number of neurons in hidden layer 2,
- Number of neurons in hidden layer 3,
- Weight decay

A range of values between zero and one will be evaluated for the weight decay parameter and the cross-validation accuracy will be calculated to identify the best performing model. This metric represents the percentage of observations in the validation data that the model classified correctly.

Next, the evaluation metrics that will be used to compare the effectiveness of the different SL methods will be discussed.

4.2 Evaluation Metrics

To identify the "best" performing model with regards to predicting RTA injury-severity, evaluation metrics are needed to compare the performance of the different models. While accuracy might be the most simple metric to understand and calculate, it can be misleading especially when dealing with imbalanced data. A model could classify all observations as the majority class and still achieve a relatively high accuracy despite it failing to identify any observations belonging to the minority class. For this reason, several other metrics in addition to accuracy will be used to evaluate model performance.

The validation data performance of the multinomial logistic regression, CT, RF, GBM and ANN models will be evaluated using the average (i) accuracy, (ii) recall, (iii) precision and (iv) F1 score of each model. These evaluation metrics were chosen due to their popularity and widespread use among researchers (Gareth et al. 2013, p. 149).

Accuracy is simply calculated as the percentage of observations correctly predicted by the model. Recall, also known as sensitivity, is defined (Gareth et al. 2013, p. 149) as:

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4.8}$$

Recall is the percentage of actual positive cases that is correctly identified by the model. Recall is an appropriate measure to evaluate model performance when there is a high cost associated with false negatives and relatively low cost associated with false positives, as is the case with RTA injury-severity prediction. If a "fatal" accident (actual positive) is incorrectly classified as a lesser injury, the consequences are severe. As a result, recall is a very important metric when selecting the best performing model since correctly identifying accidents resulting in some sort of injury or fatality is of utmost importance.

Precision is defined (Gareth et al. 2013, p. 149) as:

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positive} \tag{4.9}$$

Precision is the percentage of predicted positive cases that actually are positive. In other words it measures how precise a model is at identifying positive observations. Precision is an appropriate measure to evaluate model performance when there is a high cost associated with false positives and relatively low cost associated with a false negative. Although this is not the case with RTA injury-severity prediction, precision is still an important metric to determine how precisely the model can identify RTAs that result in some sort of injury or fatality.

In order to select the model that can correctly predict most positive cases with high precision, both recall and precision measures should be maximised. This can be an issue, since an increase in recall can often lead to a decrease in precision and vice versa (Haibo & Yunqian 2013). The F1 score is a metric that combines both recall and precision measures and can be used to identify a model with the best balance between recall and precision.

The F1 score is the harmonic mean of recall and precision and is calculated as:

$$\text{F1 score} = 2{*}\frac{Recall * Precision}{Recall + Precision} \tag{4.10}$$

In the next chapter, the results of different SL methods and comparisons using the various evaluation metrics are discussed.

Chapter 5

Results and Discussion

This chapter presents the validation data performance of the CT, RF and GBM methods trained on the four different datasets. The results of each model are compared using the evaluation metrics previously discussed. The effectiveness of the different sampling methods to address class imbalance is also discussed. The classification performance of the multinomial logistic regression and ANN models trained on the data with the most effective data sampling method are then compared. This is done in order to determine whether simple SL methods, such as multinomial logistic regression, are sufficient to model RTA injury-severity or if more complex SL methods such as ANNs are required. The model with the best performance on the validation data, with regards to average recall and recall for "fatal" RTAs, is used to predict the injury-severity of the RTA test dataset. Additionally, the variable importance of the CT, RF and GBM models are investigated in order to identify the most significant variables associated with injury-severity of RTAs, including the geographical location of an RTA. The analysis was conducted using R (R Core Team 2019) and all plots were created using the "ggplot2" package (Wickham 2016).

5.1 Classification Tree

The CT models were fitted using the "tree" package (Ripley 2019). The number of terminal nodes remaining after the tree pruning process for each CT model is shown in Table 5.1.

TABLE 5.1: Terminal nodes of each CT model.

Training data	# of terminal nodes
Original	7
Oversampled	14
SMOTE	14
Undersampled	18

5.1.1 CT Performance Summary

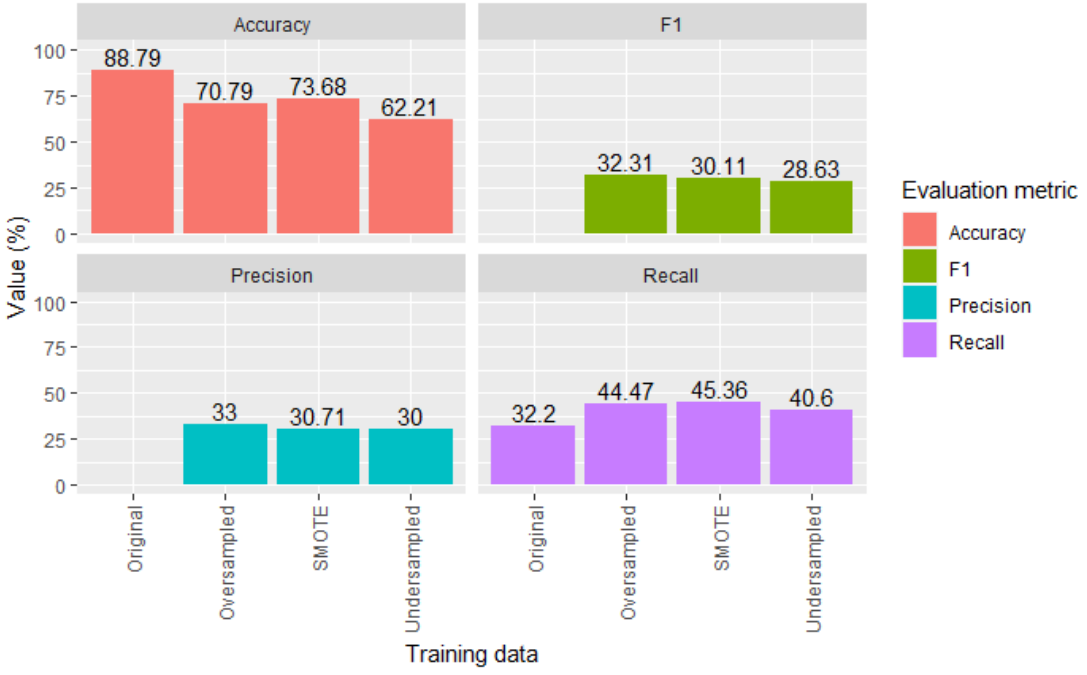

FIGURE 5.1: CT performance summary graph.

The average validation data accuracy, recall, precision and F1 score achieved by the CT models trained on the four different training datasets are shown in Figure 5.1. The CT trained on the original, imbalanced data achieved the highest accuracy, followed by the CT trained on the SMOTE data. However, as mentioned previously, accuracy can be a misleading evaluation metric when dealing with imbalanced data. In Table 5.2 which

contains the confusion matrices of all the CT models, the CT model trained on the original imbalanced data failed to identify any RTAs resulting in "fatal" or "serious" injuries. This is a major issue since RTAs that result in "fatal" or "serious" injuries carry the highest social and economic impact. The CTs trained on the three datasets where data sampling methods were applied to address the imbalanced data all managed to identify some "fatal" and "serious" RTAs. This emphasises the need to address the issue of imbalanced data when dealing with classification problems as well as the need for alternative evaluation metrics.

Since the CT model trained on the original imbalanced data failed to identify any "fatal" or "serious" RTAs, the average precision and F1 score could not be calculated. The oversampled CT achieved the highest precision and F1 score, followed by the SMOTE CT and finally the undersampled CT respectively (Figure 5.1). Recall is used to evaluate model performance when there is a high cost associated with false negatives and relatively low cost associated with false positives, as is the case with RTA injury-severity prediction. For this reason, average recall will be the main metric used to evaluate model performance. The CT model trained on the SMOTE data achieved the highest average recall (45.36%), followed by the oversampled CT (44.47%), undersampled CT (40.6%) and finally the original CT (32.2%) respectively (Figure 5.1). The SMOTE CT correctly identified the highest number of "fatal" RTAs compared to the other CT models (Table 5.2). With regards to maximising average recall, it is clear that the SMOTE technique was the most effective data sampling method to address class imbalance. The CTs trained on the oversampled and undersampled data also achieved higher average recall compared to the CT trained on the original imbalanced data, indicating that any of the three methods to address class imbalance could improve the average recall of a CT model.

Next, the results of the RF models are presented and discussed.

Table 5.2: Confusion matrix of all CT models.

Training data	**Confusion matrix**					
Original			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**0**	0	0	0
		no injury	22	**14113**	172	1188
		serious	0	0	**0**	0
		slight	22	195	247	**513**
Oversampled			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**21**	1370	196	602
		no injury	1	**11184**	38	511
		serious	13	361	**140**	272
		slight	9	1393	45	**316**
SMOTE			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**29**	761	248	596
		no injury	4	**11757**	62	632
		serious	4	441	**71**	194
		slight	7	1349	38	**279**
Undersampled			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**17**	896	165	465
		no injury	1	**9642**	46	462
		serious	18	628	**121**	306
		slight	8	3142	87	**468**

5.2 Random Forest

The RF models were fitted using the "randomForest" package (Liaw & Wiener 2002). The optimal number of predictors considered at each split for each RF model is shown in Table 5.3.

TABLE 5.3: Number of predictors considered at each split for each RF model.

Training data	Predictors considered at each split
Original	3
Oversampled	7
SMOTE	7
Undersampled	7

5.2.1 RF Performance Summary

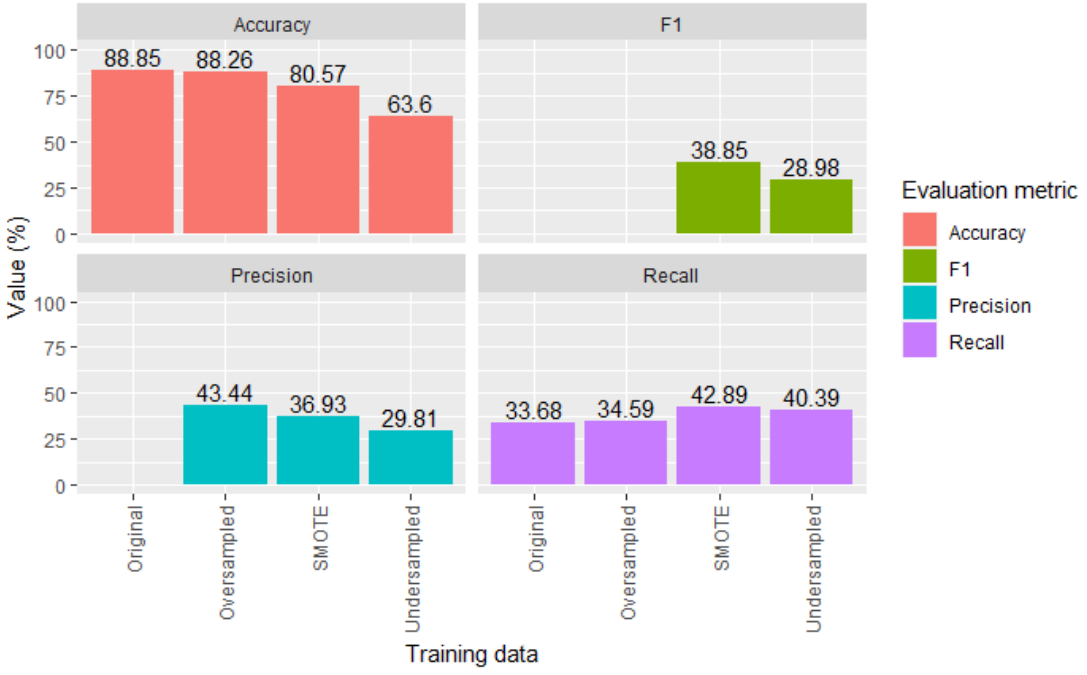

FIGURE 5.2: RF performance summary graph.

Similar to the CT results, the RF model trained on the original, imbalanced data achieved the highest accuracy followed closely by the oversampled RF, SMOTE RF and undersampled RF respectively. Despite their high accuracy, both the RFs trained on the original data as well as the oversampled data failed to correctly identify any "fatal" RTAs as shown in Table 5.4. Since the RF model trained on the original data failed

to identify any "fatal" RTAs, either correctly or incorrectly, the average precision and F1 score could not be calculated. Similarly, the average F1 score for the oversampled RF could not be calculated since it failed to correctly identify any "fatal" RTAs. The oversampled RF achieved the highest average precision followed by the SMOTE RF and undersampled RF respectively, which is similar to the results found with the CT models. The SMOTE RF achieved a higher F1 score than the undersampled RF.

With regards to average recall, the SMOTE RF performed the best (42.89%) followed by the undersampled RF (40.39%), oversampled RF (34.59%) and original RF (33.68%) respectively. Interestingly, the undersampled RF managed to correctly identify the highest number of "fatal" RTAs as shown in Table 5.4. Once again, using average recall as the main evaluation metric, the SMOTE technique seems to be the most effective method of addressing imbalanced data. The results also show that all three data sampling methods resulted in RF models with a higher average recall compared to the RF trained on the original, imbalanced data.

Next, the results of the GBM models are presented and discussed.

Table 5.4: Confusion matrix of all RF models.

Training data	**Confusion matrix**					
Original			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**0**	0	0	0
		no injury	21	**14111**	173	1188
		serious	2	9	**29**	17
		slight	21	188	217	**496**
Oversampled			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**0**	1	1	1
		no injury	19	**13974**	164	1130
		serious	5	33	**42**	48
		slight	20	300	212	**522**
SMOTE			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**3**	56	28	62
		no injury	4	**12395**	80	650
		serious	15	224	**149**	265
		slight	22	1633	162	**724**
Undersampled			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**15**	713	149	424
		no injury	3	**9934**	61	556
		serious	22	634	**150**	343
		slight	4	3027	59	**378**

5.3 Gradient Boosting Machine

The GBM models were fitted using the "gbm" package (Greenwell et al. 2019) and the "caret" package (Kuhn 2020). Using cross-validation, the optimal values found for the various tuning parameters for each GBM model are shown in Table 5.5.

TABLE 5.5: Parameters used for each GBM model.

Training data	Number of trees	Learning rate	Number of splits in each tree
Original	100	0.01	10
Oversampled	100	0.05	20
SMOTE	150	0.05	20
Undersampled	500	0.01	2

5.3.1 GBM Performance Summary

FIGURE 5.3: GBM performance summary graph.

The GBM model trained on the original data, similar to the CT and RF models, achieved the highest average accuracy followed by the SMOTE GBM, oversampled GBM and then undersampled GBM respectively (Figure 5.3). However, the GBM trained on the original, imbalanced data suffers from the same issues as both the CT and RF models

where it fails to identify any “fatal” RTAs. Once again, the average precision and F1 score could not be calculated for the GBM model trained on the original data. The GBM model trained on SMOTE data achieved the highest average precision and F1 score, followed by the oversampled GBM and finally the undersampled GBM.

In contrast to the CT and RF results, the GBM trained on the oversampled data achieved the highest average recall (45.39%), followed closely by the SMOTE GBM (44.88%) and undersampled GBM (44.19%). Once again, the GBM trained on the original data achieved the lowest average recall (34.34%). The GBM trained on the undersampled data managed to correctly identify the highest number of “fatal” RTAs. Looking at average recall, it seems that oversampling of the minority class was slightly more effective than SMOTE and undersampling of the majority class. The results once again indicate that any of the three data sampling methods improve model performance as measured by average recall when compared to the model trained on the original, imbalanced data.

Table 5.6: Confusion matrix of all GBM models.

Training data	**Confusion matrix**					
Original			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**0**	0	0	0
		no injury	22	**14113**	172	1188
		serious	6	11	**43**	29
		slight	16	184	204	**484**
Oversampled			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**4**	95	25	68
		no injury	2	**11988**	54	565
		serious	20	382	**197**	359
		slight	18	1843	143	**709**
SMOTE			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**6**	99	35	87
		no injury	2	**12361**	71	634
		serious	17	245	**155**	257
		slight	19	1603	158	**723**
Undersampled			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**20**	662	146	413
		no injury	4	**9779**	56	514
		serious	15	852	**164**	369
		slight	5	3015	53	**405**

5.4 Effectiveness of Data Sampling Methods

The effectiveness of the data sampling methods compared to the imbalanced data is determined by the improvement in the evaluation metrics, specifically average recall, of the SL methods. This dissertation employed three different data sampling methods to address the issue of imbalanced data. These methods include (i) undersampling of the majority class, (ii) oversampling of the minority class and the SMOTE technique. The results of the CT, RF and GBM models trained on the four different datasets show that all three data sampling methods improved the model's average recall and ability to identify observations belonging to the minority class, which in this case is "fatal" RTAs. None of the models trained on the original imbalanced data were able to identify "fatal" RTAs, similar to the findings of Chang & Wang (2006).

While all three data sampling methods have successfully been used by previous researchers, the SMOTE technique is widely considered as one of the most influential and popular data sampling methods in SL in recent years (García et al. 2016). For this analysis, the main evaluation metric used to measure model performance is average recall. The results of the CT, RF and GBM models seem to support that the SMOTE technique to address class imbalance is the most effective method with regards to maximising average recall. The CT and RF models that achieved the highest average recall were the models trained on the SMOTE data. The GBM trained on the SMOTE data achieved the second highest average recall, just slightly less than the GBM trained on the oversampled data.

The results indicate that oversampling is the next most effective method after SMOTE, with both the CT and GBM models trained on the oversampled data achieving a higher average recall compared to the models trained on the undersampled data.

Next, the performance of the multinomial logistic regression and ANN models on the validation data are presented and compared.

5.5 Simple vs Complex SL Methods

The performance of both a multinomial logistic regression and ANN model on the validation data are presented in this section. Since the SMOTE technique was determined to be the most effective data sampling method for the CT, RF and GBM models, both the multinomial logistic regression and ANN models are only trained on the SMOTE data. These models are compared in order to determine whether simple SL methods, such as multinomial logistic regression, are sufficient to model RTA injury-severity or if more complex SL methods such as ANNs are required.

The assumptions of the multinomial logistic regression model are checked before fitting the model. The data contains no extreme outliers, no severe multi-collinearity is present and the observations are independent. The target variable, "worst injury-severity", is measured at a nominal level for this dissertation. Since the assumptions of the model are satisfied, the multinomial logistic regression model was fitted using the "nnet" package (Venables & Ripley 2002).

The ANN model was fitted using the "RSNNS" package (Bergmeir & Benítez 2012) and the "caret" package (Kuhn 2020). Using cross-validation, different model architectures containing a range of different values for each tuning parameter were evaluated. The optimal values for the various tuning parameters of the ANN model found through cross-validation are shown in Table 5.7. The ANN model consists of one hidden layer with 30 neurons and a weight decay value of 0.0001.

TABLE 5.7: Parameters used for ANN model.

Parameter	**Value**
# of neurons in hidden layer 1	30
# of neurons in hidden layer 2	0
# of neurons in hidden layer 3	0
Weight decay	0.0001

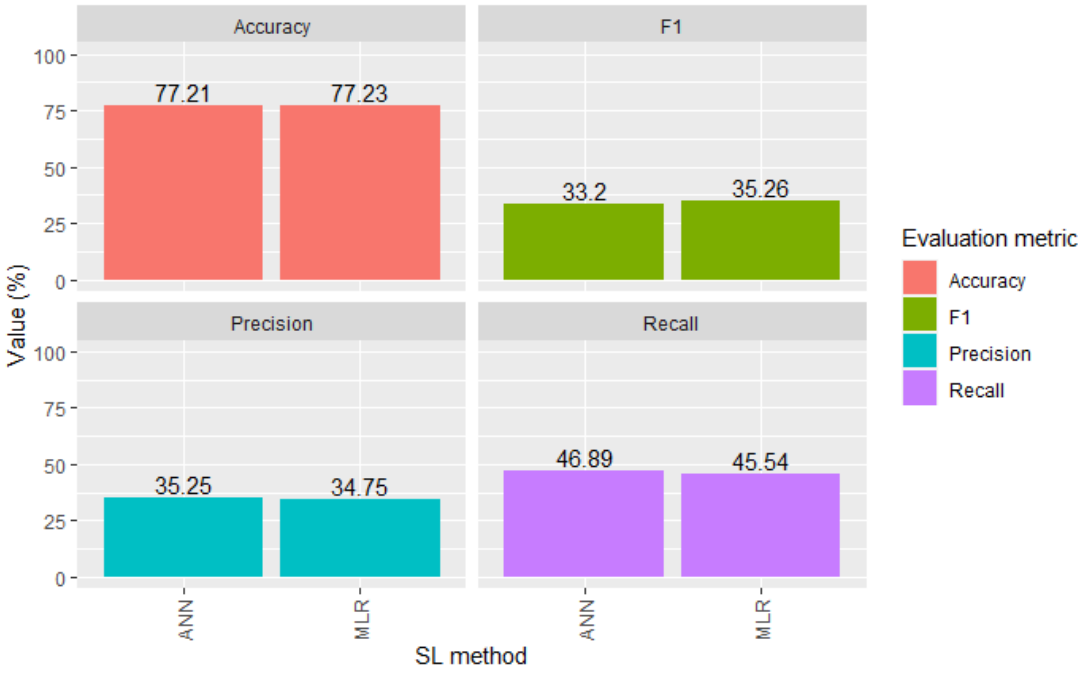

FIGURE 5.4: ANN vs multinomial logistic regression performance summary.

The performance of both the multinomial logistic regression and ANN models are provided in Figure 5.4, while the confusion matrix for each model is shown in Table 5.8. The accuracy achieved by the multinomial logistic regression and ANN models are nearly identical. As shown in Table 5.8, both the multinomial logistic regression and ANN models managed to correctly identify at least some RTAs belonging to each injury-severity category. The ANN model correctly identified more "fatal" and "no injury" RTAs than the multinomial logistic regression model, while the multinomial logistic regression model correctly identified more "slight" and "serious" injury RTAs.

Figure 5.4 shows that the ANN model achieved a higher average precision and average recall compared to the multinomial logistic regression model, while the multinomial logistic regression model achieved a slightly higher F1 score. With regards to average recall, the ANN model performs slightly better (46.89%) than the multinomial logistic regression model (45.54%). The ANN model also correctly identified a far greater amount of "fatal" RTAs compared to the multinomial logistic regression model. The results indicate that a more complex SL method, such as ANN, does achieve a higher average recall and more correctly identified observations belonging to the minority category than a simpler SL method, such as multinomial logistic regression. This supports the

findings of Abdelwahab & Abdel-Aty (2001) who found that a MLP ANN outperforms logistic regression when predicting RTA injury-severity.

TABLE 5.8: Confusion matrix of multinomial logistic regression and ANN models trained on SMOTE data.

SL method	**Confusion matrix**					
Multinomial logistic regression			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**17**	590	144	407
		no injury	2	**12169**	64	620
		serious	19	273	**150**	288
		slight	6	1276	61	**386**
ANN			Actual			
	Prediction		fatal	no injury	serious	slight
		fatal	**29**	958	214	579
		no injury	4	**12358**	70	732
		serious	5	115	**88**	147
		slight	6	877	47	**243**

Next, the model with the best performance on the validation data is identified based on average recall.

5.6 Model with the Highest Average Recall

When dealing with an imbalanced dataset such as the RTA dataset used in this study, it is important to evaluate model performance using metrics other than accuracy. Thus, the average recall was chosen as the main evaluation metric to select the best performing model. As previously mentioned, recall is a very appropriate measure to evaluate model performance when there is a high cost associated with false negatives and relatively low cost associated with false positives, as is the case with RTA injury-severity prediction. If a "fatal" accident (actual positive) is incorrectly classified as a lesser injury, the consequences are severe.

The average recall of all the different SL methods utilised during this analysis are compared in order to identify the best model based on its performance on the validation data. The comparisons of the average recall of all SL methods are shown in Figure 5.5.

FIGURE 5.5: Comparison of average recall.

There is a noticeable difference in average recall between the different SL methods as well as the training data the models were trained on. As shown in Figure 5.5, the model with the highest average recall is the ANN trained on the SMOTE training data (46.89%). The multinomial logistic regression model trained on the SMOTE data achieved the second highest average recall. The GBM model trained on the oversampled data and the CT trained on the SMOTE data have the next highest average recall. It seems that the RF models, regardless of the training data used, achieved a lower average recall compared to most of the other SL methods.

A comparison of the recall for "fatal" RTAs of all the different SL methods is shown in Figure 5.6.

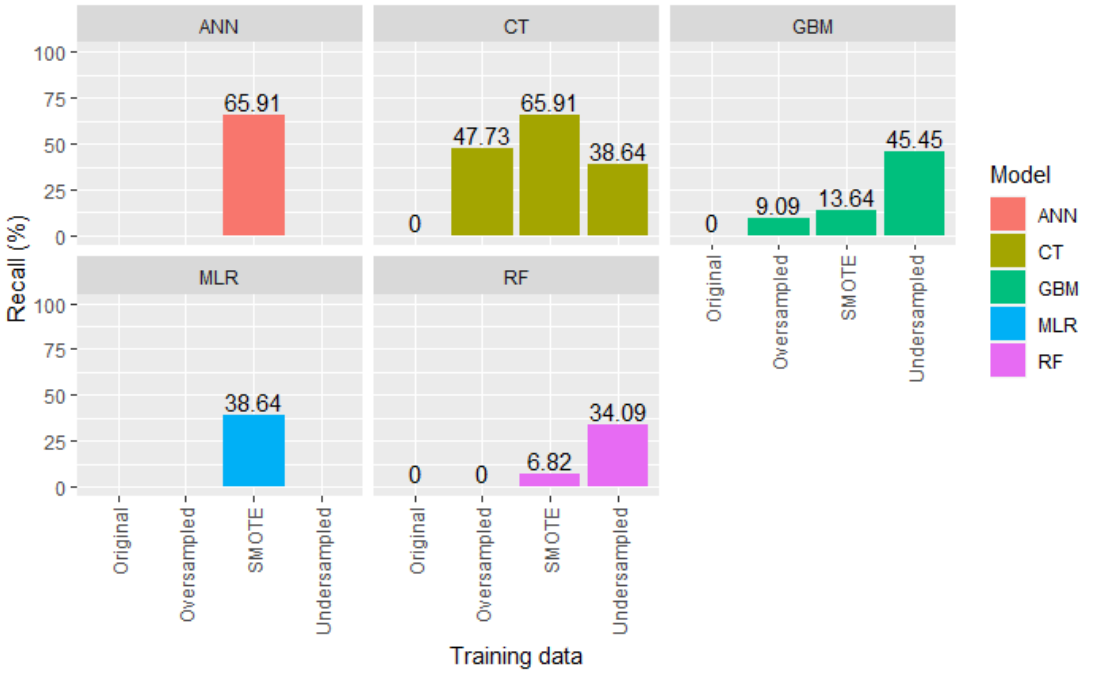

FIGURE 5.6: Comparison of recall for "fatal" RTAs.

Figure 5.6 shows that the ANN and CT trained on the SMOTE data achieved by far the highest recall for "fatal" RTAs compared to the other models. The results show that the ANN and CT models seem to achieve the highest recall for "fatal" RTAs overall, followed by the GBM, multinomial logistic regression and finally the RF models respectively. The CT, RF and GBM models trained on the original data could not identify any "fatal" RTAs. The RF model trained on the oversampled data also failed to identify any "fatal" RTAs. The evaluation metrics by class for all the SL methods can be found in Appendix A.

Based on the comparison of average recall and recall for "fatal" RTAs, it has been determined that the ANN model trained on the SMOTE data is the best performing model. The ANN model trained on the SMOTE data achieved the highest average recall as well as the tied highest recall for "fatal" RTAs. Next, the performance of the ANN model on the test data, otherwise known as "unseen" data, is presented.

5.7 Performance of the Best Model on Test Data

The ANN model trained on the SMOTE data achieved the highest average recall as well as the highest recall for "fatal" RTAs on the validation data. This model has therefore been selected as the "best" performing model and its performance on the test data, also known as "unseen" data, is now presented. The confusion matrix of the ANN's performance on the test data is shown in Table 5.9 while the evaluation metrics are shown in Table 5.10.

TABLE 5.9: Confusion matrix of ANN on test data.

		Actual			
		fatal	**no injury**	**serious**	**slight**
Prediction	fatal	38	1011	223	574
	no injury	2	12257	64	698
	serious	1	132	80	157
	slight	3	908	52	272

TABLE 5.10: Evaluation metrics of ANN on test data by class.

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	86.36	2.06	4.02	
No Injury	85.67	94.13	89.70	
Serious	19.09	21.62	20.28	
Slight	15.99	22.02	18.53	
Average	51.78	34.96	33.13	76.78

The model has an accuracy of 76.78% while being able to correctly identify at least some RTAs belonging to each of the four different injury-severity categories (Table 5.10). The model has a higher average recall (51.78%) than any of the SL methods manage to achieve on the validation data. As shown in Table 5.9, the model also managed to correctly identify a large number of "fatal" and "no injury" RTAs, in contrast to fewer correctly identified "slight" and "serious" RTAs. Table 5.10 also shows that the ANN model has a very high recall for both "fatal" (86.36%) and "no injury" (85.67%) RTAs and a comparatively low recall for "slight" (15.99%) and "serious" (19.09%) RTAs. This is consistent with the findings of Chong et al. (2005), who found that several SL methods applied during their study predicted "no injury" and "fatal" RTAs most accurately out

of all the injury-severity categories. The model also has a high precision for “no injury” RTAs, indicating that it is very precise at correctly identifying “no injury” RTAs. This is in contrast with “fatal” RTAs, for which the model has a low precision score. This suggests that although the model correctly identifies the vast majority of “fatal” RTAs, it also results in a large number of false positives for “fatal” RTAs.

With regards to RTA injury-severity prediction, there is a high cost associated with false negatives and a relatively low cost to false positives. Considering this, the fact that the model has low precision for “fatal” RTAs is less of a concern due to it having a very high recall for “fatal” RTAs. For both “slight” and “serious” RTAs, the model has low recall and low precision scores. This indicates that the model is not effective at identifying RTAs belonging to those injury-severity categories.

In summary, the results presented in this chapter show that the ANN model, which achieved the best results on the validation data, can correctly identify a large number of “fatal” RTAs while also resulting in a high number of false positives. The ANN model is very effective at correctly identifying “no injury” RTAs as evidenced by its high recall and precision score. Finally, the ANN model performs poorly at correctly identifying “slight” and “serious” RTAs as evidenced by the low recall and precision scores.

5.8 Variable Importance

As previously stated, SL methods model the relationship between predictor variables and a target variable. This allows researchers to gain insights regarding the relationship between the target variable and individual predictor variables. The accident-related variables that are the most significantly associated with RTA injury-severity according to the CT, RF and GBM models trained on the SMOTE data are investigated. While the ANN model was selected as the "best" performing model, ANNs are considered a "black-box" model which provides little insight into the importance of the predictor variables during the prediction process (Olden et al. 2004). For this reason, the variable importance of the ANN model is not included in this section.

In order to determine the most important predictor variables according to the CT, the predictor variables used as split variables are examined. The "randomForest" package (Liaw & Wiener 2002) that was used to construct the RF model contains a built-in function to calculate variable importance. The variable importance is determined by the mean decrease in the Gini-index (a measure of node purity) for each predictor variable. The "gbm" package used to construct the GBM model also has a built-in function for variable importance that calculates the relative influence of each predictor variable. The relative influence of each predictor is represented by the increase in the misclassification rate when the predictor variable is randomly permuted (Breiman 2001). A comparison of the 10 most important variables for each of the CT, RF and GBM models can be seen in Table 5.11.

TABLE 5.11: Comparison of variable importance.

Rank	**CT**	**RF**	**GBM**
1	crash_type	lng	num_vehicles_involved
2	alleged_cause	temperature	lng
3	lng	lat	wind_speed
4	pedestrian_involved	cloud_cover	multi_vehicle_accident
5	num_people_involved	relative_humidity	lat
6	passengers_involved	wind_speed	temperature
7	m_cycle	visibility	cloud_cover
8	multi_vehicle_accident	crash_type	visibility
9	num_vehicles_involved	alleged_cause	relative_humidity
10		num_vehicles_involved	crash_type

The variable importance of the CT, RF and GBM models differ slightly. The CT model used the crash type and the alleged cause of the RTA as its first two split variables. The location of the RTA, the longitude coordinates specifically, was also used as a split variable. The results show that pedestrians and passengers being involved in an RTA is a significant risk factor as well as the number of vehicles involved in the accident. RTAs involving a motorcycle is also determined to be an important risk factor. The results of the CT variable importance support the findings of Chang & Wang (2006) that the involvement of pedestrians and motorcycles in RTAs are significantly associated with injury-severity. The results also support the findings of the exploratory data analysis where it was shown that pedestrian-related RTAs accounted for the vast majority of "serious" and "fatal" RTAs.

Interestingly, both the RF and GBM have several weather-related variables in their 10 most important variables, in contrast to the CT which did not include a single weather-related variable. Both the RF and GBM models determined that the temperature, cloud cover, relative humidity, visibility and wind speed were important. This suggests that the weather can be a significant risk factor associated with RTA injury-severity.

Similar to the CT model, both the RF and GBM models determined the longitude coordinates to be important. However, the RF and GBM models also determined the

latitude coordinates to be important. This indicates that the geographical location of an RTA can be a significant risk factor associated with RTA injury-severity. This is similar to the results found by Mokoatle et al. (2019), who determined that geographical features such as the distance to the nearest points of interest from where an RTA occurred were significantly associated with RTA injury-severity.

Similar to the CT model, the RF and GBM models also determined the crash type and the number of vehicles involved in a RTA to be important. This is consistent with the results of the exploratory data analysis, which found that the majority of "serious" and "fatal" RTAs involved only one vehicle while the majority of "no injury" and "slight" RTAs involved two vehicles. The exploratory data analysis also showed that crash types involving pedestrians most often resulted in serious injuries or fatalities.

The variable importance determined by the CT, RF and GBM models provides further insight into the most important risk factors associated with RTA injury-severity. The results indicate that the geographical location of a RTA to be a significant risk factor associated with RTA injury-severity, especially the longitude coordinates. The crash type was deemed a significant factor across all three models, confirming the findings of previous studies by Kim et al. (1994) and Shanthi & Ramani (2012). Additionally, all three models determined the number of vehicles involved in an accident to be important. Similar to the findings of Al-Ghamdi (2002), the CT and RF models both determined the alleged cause of an accident to be significant. The RF and GBM models determined several weather-related variables to be significantly associated with RTA injury-severity as well, similar to the findings of Islam & Hernandez (2013).

Chapter 6

Conclusion, Recommendations and Future Work

6.1 Conclusion

The aim of this dissertation was to analyse RTAs that occurred in the City of Cape Town in order to gain a better understanding of the relationship between accident-related factors and the injury-severity of RTAs through the use of SL methods. Data was collected on RTAs that occurred in Cape Town during the period 2015-2017. The study focused on achieving four main objectives: Review both international and South African literature on RTAs and RTA injury-severity prediction; determine the best data sampling method to address the issue of class imbalance in RTA data; utilise several different SL classification methods to predict the injury-severity of Cape Town RTAs and compare their effectiveness and finally to identify the most significant risk factors associated with RTA injury-severity.

Prior to the analysis of the data, a review of both international and South African literature on RTAs and RTA injury-severity prediction was conducted. This included a review of previous international literature that made use of logistic regression, CART as well as other SL classification methods in modelling RTA injury-severity. Next, the South African literature on RTAs was reviewed. It was found that South African literature mostly consists of identifying significant contributors of RTAs, with limited literature on modelling RTA injury-severity as a classification problem using SL methods. It was

clear that there is a need for research focusing on South African RTA injury-severity prediction using SL methods.

After the literature review was completed, the RTA data used for the study was presented. This included the source of the data, the cleaning and processing of the data and the addition of new variables to the dataset. Variables added to the dataset included weather-related variables; whether an RTA occurred on the weekend or on a public holiday; whether an RTA occurred during peak traffic times and more. In order to determine the importance of geographical location as a predictor of RTA injury-severity, the street addresses of the RTAs were geocoded in order to obtain the geographical coordinates of each RTA. After the data preparation, an exploratory data analysis was performed to find interesting patterns in the data. The findings of the exploratory data analysis included the following: the vast majority of alleged causes of RTAs are related to driver/human error, accidents with pedestrians make up only 5.86% of all RTAs yet account for 58.56% of "fatal" accidents and 55.37% of "serious" accidents, the majority of "fatal" and "serious" RTAs occur on the weekend and involve only one vehicle.

It was also identified that the RTA data was severely imbalanced with regards to injury-severity. Imbalanced data can negatively affect the performance of certain classification methods, especially with regards to predicting the minority class. The RTA was split into training, validation and test sets containing 60%, 20% and 20% of the RTAs in the full dataset respectively. The proportions of the injury-severity category were kept consistent between all datasets. This dissertation employed three common data sampling approaches used by researchers to address imbalanced data, namely (i) undersampling of the majority class, (ii) oversampling of the minority class and (iii) SMOTE. Four separate training datasets were then constructed which included the original imbalanced data, data with the minority class oversampled, data with the majority class undersampled and the SMOTE data containing artificially created observations.

Chapter Four of this dissertation presented a brief overview of SL methods, followed by the different SL methods that were utilised in the study, namely: multinomial logistic regression, CT, RF, GBM and ANN methods. The different metrics used to evaluate model performance such as accuracy, recall, precision and F1 score were also presented and discussed in this chapter. Chapter Five presented and discussed the results of this study. The validation data performance of the CT, RF and GBM models trained on

the four different datasets were presented and discussed. The results showed that all three data sampling methods improved the CT, RF and GBM model's average recall and ability to identify observations belonging to the minority class, which in this case is "fatal" RTAs. With regards to maximising average recall, it was determined that the SMOTE technique was the most effective data sampling method to address imbalanced data. It was also found that the CT, RF and GBM model trained on the imbalanced data all failed to identify any "fatal" RTAs, further emphasising the need to account for class imbalance before training classification models. Next, the validation data performance of the multinomial logistic regression and ANN models trained on the SMOTE data was presented and compared. This was done in order to determine whether simple SL methods such as multinomial logistic regression are sufficient to model RTA injury-severity or if more complex SL methods such as ANNs are required. The results showed that the ANN model achieved a higher average recall and correctly identified more observations belonging to the minority category, "fatal" RTAs, than the multinomial logistic regression model. Using average recall as the main evaluation metric, the ANN was selected as the best performing model on the validation data. The ANN trained on the SMOTE data achieved the highest average recall as well as the highest recall for "fatal" RTAs out of all the models. The performance of the ANN model on the test data, also known as "unseen" data, showed that the ANN model correctly identified a large number of "fatal" RTAs while also resulting in a high number of false positives. The model was very effective at correctly identifying "no injury" RTAs as evidenced by its high recall and precision scores. However, the model performed poorly at correctly identifying "slight" and "serious" RTAs as evidenced by the low recall and precision scores.

Finally, the variable importance of the CT, RF and GBM models trained on the SMOTE data were investigated to identify the most significant risk factors associated with RTA injury-severity. The results indicate the geographical location of a RTA to be a significant risk factor associated with RTA injury-severity, especially the longitude coordinates. The crash type as well as the number of vehicles involved in an accident was deemed to be significant risk factors across all three models. The CT and RF models both determined the alleged cause of an accident to be significant, while the RF and GBM models determined several weather-related variables such as temperature, cloud cover, relative humidity, visibility and wind speed to be significant risk factors associated with

RTA injury-severity as well.

6.2 Recommendations

Due to the high cardinality of the predictor variables, it is highly difficult to provide guidance on road safety policy. As a result, recommendations on policy will come from the exploratory data analysis informed by the variable importance results. The results of the exploratory analysis indicated human error as being the alleged cause of the vast majority of RTAs analysed during this study, while road and environmental factors only accounted for a small portion of the RTAs. This suggests that in order to effectively reduce RTAs, the emphasis should be placed on reducing human/driver error. The exploratory analysis also showed that pedestrian-related RTAs account for the majority of RTAs that result in serious injuries and fatalities. Improving the safety of pedestrian crossings, either through better infrastructure or better policing of pedestrian jaywalking, should be of utmost importance in order to reduce "serious" and "fatal" RTAs. The results of the exploratory analysis also showed that the majority of "serious" and "fatal" RTAs occurred over the weekend. Increased traffic law enforcement during weekends could potentially help to reduce the number of RTAs resulting in serious injuries and fatalities.

Imbalanced data is a common issue found with RTA data. The comparison of the CT, RF and GBM models trained on the four different training datasets indicate that the best data sampling technique to address class imbalance in RTA datasets is the SMOTE technique with regards to maximising average recall. It is therefore recommended that future researchers use the SMOTE technique to address imbalanced RTA datasets when predicting RTA injury-severity. The results of the different SL methods indicate that more complex SL methods such as ANN can offer improvements in predicting RTA injury-severity compared to simpler SL methods, especially with regards to improving average recall and recall of the minority class. It is recommended that more complex SL methods be utilised during RTA injury-severity prediction as they can offer improved performance depending on the evaluation metric being maximised. While ANNs may offer improved classification performance, decision tree models such as CTs, RFs and GBMs offer greater interpretability of results. Depending on the problem being addressed, policy creators should choose the SL method used accordingly.

The results of the variable importance suggest that the geographic location of a RTA is a significant risk factor associated with RTA injury-severity. It is recommended that the City of Cape Town start recording and collecting the geographic coordinates of future RTAs. This will allow high-risk areas of "serious" and "fatal" RTAs to be more accurately identified and help focus safety efforts and policing more effectively. The results also indicated that several weather-related variables were significant risk factors associated with RTA injury-severity. Increased policing and warning signs during poor weather conditions could potentially help reduce RTA injury-severity.

Finally, it is recommended that the City of Cape Town expand and improve the quality of their RTA data. This study added several new predictor variables to the dataset obtained from the City of Cape Town, several of which were found to be significantly associated with RTA injury-severity. This will allow future researchers to analyse and model RTA injury-severity more comprehensively and hopefully reduce the frequency and injury-severity of RTAs.

6.3 Limitations

The data used for this study was sourced from the City of Cape Town, who collected and processed the data from the SAPS. The data contained several accident related variables along with the RTA injury-severity. However, compared to RTA data used in similar international studies, the data used for this study contains relatively few accident-related variables. This can negatively affect the performance of the SL methods as the models will be trained on data that is potentially missing some important accident-related variables. These include variables that were found to be significantly associated with RTA injury-severity in previous studies, such as seat belt use (Nassar et al. 1994), vehicle impact speed (Kong & Yang 2010), driver's gender and age (Lukongo 2020), driver experience (Kuhnert et al. 2000), drug involvement (Shanthi & Ramani 2012) and more.

The data used for this study was limited to accidents where the location could be correctly geocoded. This limited the number of RTAs included in the final dataset used in this study. While the geocoding was able to return a sizeable number of coordinates, the accuracy of geocoding algorithms must always be assessed. During the cross-validation

process to find optimal values for the tuning parameters of the various SL methods, accuracy was used as the evaluation metric. While the cross-validation accuracy is the default metric used to select the optimal parameter values, it may be beneficial to use a different metric such as recall or F1 score, especially when dealing with imbalanced data.

Finally, only RTAs occurring in Cape Town was analysed during this study. The SL methods fitted during this study might not perform as well on new RTA data from other geographical regions or cities in South Africa since the models were only trained on Cape Town RTAs. The variables found to be most significantly associated with RTA injury-severity in Cape Town may not hold true for other regions in South Africa.

6.4 Future Work

This study only used RTA data for a specific time window (2015-2017) and the geographical coordinates of an RTA were added as predictor variables to the data. The use of models that explicitly take the spatio-temporal nature of RTA data into account could be beneficial since this study determined that the geographical location of an accident is significantly associated with RTA injury-severity. Reducing the cardinality of predictor variables in RTA data may result in more interpretable variable importance results. Splitting the data into training, validation and test datasets reduces the data available to train models. The use of cross-validation is recommended for future work instead, in addition to it being a better estimate of the test-error. Obtaining RTA data for multiple cities in South Africa would allow researchers to develop more generalised SL methods to model RTA injury-severity in South Africa, which would help inform road safety regulations that could be implemented across the country. It would also allow researchers to identify any differences between cities or geographic regions in South Africa with regards to the most significant risk factors associated with RTA injury-severity.

Appendix A

The validation data evaluation metrics (by class) for all the SL methods trained on the different training datasets are presented in this appendix.

A.1 Classification Tree

A.1.1 Original Data

TABLE A.1: Evaluation metrics of CT trained on original data by class on validation data

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	0.00	NA	NA	
No Injury	98.64	91.08	94.71	
Serious	0.00	NA	NA	
Slight	30.16	52.51	38.31	
Average	32.20	NA	NA	88.79

A.1.2 Undersampled Data

Table A.2: Evaluation metrics of CT trained on undersampled data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	38.64	1.10	2.14	
No Injury	67.39	94.99	78.84	
Serious	28.88	11.28	16.22	
Slight	27.51	12.63	17.31	
Average	40.60	30.00	28.63	62.21

A.1.3 Oversampled Data

Table A.3: Evaluation metrics of CT trained on oversampled data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	47.73	0.96	1.88	
No Injury	78.17	95.31	85.89	
Serious	33.41	17.81	23.24	
Slight	18.58	17.92	18.24	
Average	44.47	33.00	32.31	70.79

A.1.4 SMOTE Data

Table A.4: Evaluation metrics of CT trained on SMOTE data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	65.91	1.77	3.46	
No Injury	82.17	94.40	87.86	
Serious	16.95	10.00	12.58	
Slight	16.40	16.68	16.54	
Average	45.36	30.71	30.11	73.68

A.2 Random Forest

A.2.1 Original Data

Table A.5: Evaluation metrics of RF trained on original data by class on validation data

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	0	NA	NA	
No Injury	98.62	91.08	94.70	
Serious	6.92	50.88	12.18	
Slight	29.16	53.80	37.82	
Average	33.68	NA	NA	88.85

A.2.2 Undersampled Data

Table A.6: Evaluation metrics of RF trained on undersampled data by class on validation data

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	34.09	1.15	2.23	
No Injury	69.43	94.13	79.91	
Serious	35.80	13.05	19.13	
Slight	22.22	10.90	14.63	
Average	40.39	29.81	28.98	63.60

A.2.3 Oversampled Data

Table A.7: Evaluation metrics of RF trained on oversampled data by class on validation data

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	0.00	0.00	NA	
No Injury	97.67	91.41	94.43	
Serious	10.02	32.81	15.36	
Slight	30.69	49.53	37.89	
Average	34.59	43.44	NA	88.26

A.2.4 SMOTE Data

Table A.8: Evaluation metrics of RF trained on SMOTE data by class on validation data

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	6.82	2.01	3.11	
No Injury	86.63	94.41	90.35	
Serious	35.56	22.82	27.80	
Slight	42.56	28.49	34.13	
Average	42.89	36.93	38.85	80.57

A.3 Gradient Boosting Machine

A.3.1 Original Data

Table A.9: Evaluation metrics of GBM trained on original data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	0.00	NA	NA	
No Injury	98.64	91.08	94.71	
Serious	10.26	48.31	16.93	
Slight	28.45	54.50	37.39	
Average	34.34	NA	NA	88.88

A.3.2 Undersampled Data

Table A.10: Evaluation metrics of GBM trained on undersampled data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	45.45	1.61	3.11	
No Injury	68.35	94.46	79.31	
Serious	39.14	11.71	18.03	
Slight	23.81	11.64	15.64	
Average	44.19	29.86	29.02	62.94

A.3.3 Oversampled Data

Table A.11: Evaluation metrics of GBM trained on oversampled data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	9.09	2.08	3.39	
No Injury	83.79	95.07	89.07	
Serious	47.02	20.56	28.61	
Slight	41.68	26.13	32.13	
Average	45.39	35.96	38.30	78.30

A.3.4 SMOTE Data

Table A.12: Evaluation metrics of GBM trained on SMOTE data by class on validation data

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	13.64	2.64	4.43	
No Injury	86.39	94.59	90.31	
Serious	36.99	23.00	28.36	
Slight	42.50	28.89	34.40	
Average	44.88	37.28	39.37	80.41

A.4 Multinomial Logistic Regression

A.4.1 SMOTE Data

Table A.13: Evaluation metrics of multinomial logistic regression trained on SMOTE data by class

Class	Recall (%)	Precision (%)	F1 (%)	Overall Accuracy (%)
Fatal	38.64	1.47	2.83	
No Injury	85.05	94.66	89.60	
Serious	35.80	20.55	26.11	
Slight	22.69	22.33	22.51	
Average	45.54	34.75	35.26	77.23

A.5 Artificial Neural Networks

A.5.1 SMOTE Data

Table A.14: Evaluation metrics of ANN trained on SMOTE data by class on validation data

Class	**Recall (%)**	**Precision (%)**	**F1 (%)**	**Overall Accuracy (%)**
Fatal	65.91	1.63	3.18	
No Injury	86.37	93.88	89.97	
Serious	21.00	24.79	22.74	
Slight	14.29	20.72	16.91	
Average	46.89	35.25	33.20	77.21

Appendix B

B.1 Data Dictionary

Table B.1: Data Dictionary

Variable	Description
accident_location_type	Categorical: Accident occurred at either an intersection or non-intersection • intersection • non_intersection
year	Categorical: The year the accident occurred • 2015 • 2016 • 2017
crash_type	Categorical: The type of crash that occurred • Sideswipe • Head/Rear end • Other • Approach at angle • Accident with pedestrian • Reversing • Accident with fixed/other object

Table B.1 continued from previous page

Variable	Description
alleged_cause	Categorical: The alleged cause of the crash is deduced by the data capturers when capturing the accident report at the City of Cape Town. They look at all the factors indicated on the accident form and then deduct the alleged cause of the crash. • Other driver related • Insuff. following distance • Roadway and Environment related • Pedestrian related • Vehicle reversed • Entered traffic while unsafe • Change lane while unsafe • Bypass distance too close • Vehicle related
num_vehicles_involved	Categorical: The number of vehicles involved in the accident • 1 • 2 • 3+
num_people_involved	Categorical: The number of people involved in the accident • 1 • 2 • 3 • 4+
lng	Numerical: The longitude (geographic coordinate) of the accident's location
lat	Numerical: The latitude (geographic coordinate) of the accident's location
weekend	Categorical: Accident occurred on weekend • TRUE • FALSE

Table B.1 continued from previous page

Variable	Description
time_category	Categorical: Time of the day when accident occurred • Morning-peak (6-9am) • Off-peak • Afternoon-peak (4-7pm)
public_holiday	Categorical: Accident occurred on a public holiday • TRUE • FALSE
season	Categorical: Season that accident occurred • Summer • Autumn • Winter • Spring
passengers_involved	Categorical: Passengers were involved in the accident • TRUE • FALSE
multi_vehicle_accident	Categorical: More than one vehicle was involved in the accident • TRUE • FALSE
pedestrian_involved	Categorical: A pedestrian was involved in the accident • TRUE • FALSE
motor_car	Categorical: A motor car was involved in the accident • TRUE • FALSE
light_deliv_vehicle	Categorical: A light delivery vehicle was involved in the accident • TRUE • FALSE

Table B.1 continued from previous page

Variable	Description
mini_bus	Categorical: A mini-bus was involved in the accident • TRUE • FALSE
large_deliv_vehicle	Categorical: A large delivery vehicle was involved in the accident • TRUE • FALSE
m_cycle	Categorical: A motorcycle was involved in the accident • TRUE • FALSE
bus	Categorical: A bus was involved in the accident • TRUE • FALSE
truck_articulated	Categorical: A truck was involved in the accident • TRUE • FALSE
other_vehicle	Categorical: A vehicle not belonging to any of the main vehicle categories was involved in the accident • TRUE • FALSE
bicycle	Categorical: A bicycle was involved in the accident • TRUE • FALSE
panel_van	Categorical: A panelvan was involved in the accident • TRUE • FALSE
temperature	Numerical: Temperature on the day the accident occurred measured in degrees Celsius (C)

Table B.1 continued from previous page

Variable	Description
precipitation	Categorical: Precipitation on the day the accident occurred measured in millimetres (mm) • 0 mm • 0-1 mm • 1+ mm
wind_speed	Numerical: Wind speed on the day the accident occurred measured in kilometres per hour (kph)
visibility	Numerical: Visibility on the day the accident occurred measured in kilometres
cloud_cover	Numerical: Cloud cover on the day the accident occurred measured as a percentage (%) of maximum cloud cover (no visible sky)
relative_humidity	Numerical: The relative humidity on the day the accident occurred measured as a percentage (%). Relative humidity is defined as the amount of atmospheric moisture present relative to the amount needed for saturation at the same temperature.
conditions	Categorical: The weather conditions on the day the accident occurred • Clear • Partially cloudy • Heavy rain • Rain
worst_injury_severity	Categorical: The worst injury sustained by a person during the accident • no injury • slight • serious • fatal

Appendix C

C.1 R packages used for data cleaning, processing and exploratory analysis

- "janitor" (Firke 2021)
- "leaflet" (Cheng et al. 2019)
- "leafgl" (Appelhans 2020)
- "lubridate" (Grolemund & Wickham 2011)
- "readxl" (Wickham & Bryan 2019)
- "stringr" (Wickham 2019)
- "sf" (Pebesma 2018)
- "tidyverse" (Wickham et al. 2019)
- "UBL" (Branco et al. 2016)

Bibliography

Abdelwahab, H. & Abdel-Aty, M. (2001), 'Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections', *Transportation Research Record: Journal of the Transportation Research Board* **1746**(1), 6–13.

Abdulhafedh, A. (2017), 'Incorporating the multinomial logistic regression in vehicle crash severity modeling: A detailed overview', *Journal of Transportation Technologies* **7**, 279–303.

Akin, D. & Akbaç, B. (2010), 'A neural network (nn) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics', *Scientific Research and Essays* **5**(19), 2837–2847.

Al-Ghamdi, A. S. (2002), 'Using logistic regression to estimate the influence of accident factors on accident severity', *Accident Analysis & Prevention* **34**(6), 729–741.

Appelhans, T. (2020), *leafgl: High-Performance 'WebGl' Rendering for Package 'leaflet'*. R package version 0.1.1.
URL: *https://CRAN.R-project.org/package=leafgl*

Bergmeir, C. & Benítez, J. M. (2012), 'Neural networks in R using the stuttgart neural network simulator: RSNNS', *Journal of Statistical Software* **46**(7), 1–26.
URL: *https://www.jstatsoft.org/v46/i07/*

Botha, G. & van der Walt, H. (2006), Fatal road crashes, contributory factors and the level of lawlessness, *in* 'Proceedings of the 25th Southern African Transport Conference (SATC 2006)', 10, pp. 377–387.

Branco, P., Ribeiro, R. P. & Torgo, L. (2016), 'UBL: an r package for utility-based learning', *CoRR* **abs/1604.08079**.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and regression trees. 1st ed*, Chapman and Hall/CRC, Wadsworth.

Chang, L.-Y. & Wang, H.-W. (2006), 'Analysis of traffic injury severity: An application of non-parametric classification tree techniques', *Accident Analysis & Prevention* **38**(5), 1019–1027.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research* **16**, 321–357.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (August 2016),', KDD, pp. 785–794.

Cheng, J., Karambelkar, B. & Xie, Y. (2019), *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.0.3.
URL: *https://CRAN.R-project.org/package=leaflet*

Chokotho, L., Matzopoulos, R. & Myers, J. (2013), 'Assessing quality of existing data sources on road traffic injuries (rtis) and their utility in informing injury prevention in the western cape province, south africa', *Traffic Injury Prevention* **14**(3), 267–273.

Chong, M., Abraham, A. & Paprzycki, M. (2005), 'Traffic accident analysis using machine learning paradigms', *Informatica* **29**, 89–98.

Department of Transport, R. (2018), 'State of road safety report: Calendar january-december 2018', Available at https://www.rtmc.co.za/index.php/publications/reports/traffic-reports (2021/07/19).

Eustace, D., Indupuru, V. & Hovey, P. (2011), 'Identification of risk factors associated with motorcycle-related fatalities in ohio', *Journal of Transportation Engineering* **137**(7), 474–480.

Firke, S. (2021), *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 2.1.0.
URL: *https://CRAN.R-project.org/package=janitor*

García, S., Luengo, J. & Herrera, F. (2016), 'Tutorial on practical tips of the most influential data preprocessing algorithms in data mining', *Knowledge-Based Systems* **98**, 1–29.

Gareth, J., Daniela, W., Trevor, H. & Robert, T. (2013), *An introduction to statistical learning: with applications in R*, Spinger.

Govender, R., Sukhai, A. & van Niekerk, A. (2020), *Driver intoxication and fatal crashes*, Road Traffic Management Corporation Research and Development, [Online]. Available from.
URL: *https://www.rtmc.co.za/images/rtmc/docs/research$_d$ev$_r$ep/Driver*

Greenwell, B., Boehmke, B., Cunningham, J. & Developers, G. (2019), *gbm: Generalized Boosted Regression Models*. R package version 2.1.5.
URL: *https://CRAN.R-project.org/package=gbm*

Grolemund, G. & Wickham, H. (2011), 'Dates and times made easy with lubridate', *Journal of Statistical Software* **40**(3), 1–25.
URL: *https://www.jstatsoft.org/v40/i03/*

Haibo, H. & Yunqian, M. (2013), *Imbalanced Learning: Foundations, Algorithms, and Applications*, Wiley.

Islam, M. & Hernandez, S. (2013), 'Modeling injury outcomes of crashes involving heavy vehicles on texas highways', *Transportation Research Record: Journal of the Transportation Research Board* **2388**(1), 28–36.

Islam, S. & Mannering, F. (2006), 'Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence', *Journal of Safety Research* **37**(3), 267–276.

Jones, S. & Hensher, D. (2007), 'Evaluating the behavioural performance of alternative logit models: An application to corporate takeovers research', *Journal of Business Finance Accounting* **34**(7-8), 1193–1220.

Joshua, S. & Garber, N. (1990), 'Estimating truck accident rate and involvements using linear and poisson regression models', *Transportation Planning and Technology* **15**(1), 41–58.

Kim, J., Ulfarsson, G., Kim, S. & Shankar, V. (2013), 'Driver-injury severity in single-vehicle crashes in california: A mixed logit analysis of heterogeneity due to age and gender', *Accident Analysis Prevention* **50**, 1073–1081.

Kim, K., Nitz, L., Richardson, J. & Li, L. (1994), Analyzing the relationship between crash types and injury severity in motor vehicle collisions in hawaii, *in* 'Transportation Research Record 1467, TRB, National Research Council, , DC', pp. 9–13.

Kim, K., Nitz, L., Richardson, J. & Li, L. (1995), 'Personal and behavioral predictors of automobile crash and injury severity', *Accident Analysis Prevention* **27**(4), 469–481.

Koehrsen, W. (2020), 'Random forest simple explanation'.
URL: *https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d*

Kong, C. & Yang, J. (2010), 'Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in china', *Accident Analysis & Prevention* **42**(4), 987–993.

Kuhn, M. (2020), *caret: Classification and Regression Training*. R package version 6.0-86.
URL: *https://CRAN.R-project.org/package=caret*

Kuhnert, P. M., Do, K.-A. & McClure, R. (2000), 'Combining non-parametric models with logistic regression: an application to motor vehicle injury data', *Computational Statistics & Data Analysis* **34**(3), 371–386.

Liaw, A. & Wiener, M. (2002), 'Classification and regression by randomforest', *R News* **2**(3), 18–22.
URL: *https://CRAN.R-project.org/doc/Rnews/*

Linu, K. J., Minu, P. K., Nithya, N. P., Sasidharan, S. & Sreeshma, K. P. (2013), Road safety awareness index road user behavior – a case study at kazhakkoottam, *in* 'Proceedings of International Conference on Energy and Environment', 2(1, pp. 270–276.

Loh, W.-Y. (2014), 'Fifty years of classification and regression trees', *International Statistical Review* **82**(3), 329–348.

Lord, D., Washington, S. & Ivan, J. (2005), 'Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory', *Accident Analysis Prevention* **37**(1), 35–46.

Lukongo, O. (2020), 'Examining prominent causes of traffic injury severity in louisiana with multinomial logistic models', *Transportation Research Record: Journal of the Transportation Research Board* **2675**(1), 245–257.

Mercier, C. R., Shelley, M. C., Rimkus, J. B. & Mercier, J. M. (1997), 'Age and gender as predictors of injury severity in head-on highway vehicular collisions', *Transportation Research Record* **1581**(1), 37–46.

Milton, J. C., Shankar, V. N. & Mannering, F. L. (2008), 'Highway accident severities and the mixed logit model: an exploratory empirical analysis', *Accident Analysis & Prevention* **40**(1), 260–266.

Mokoatle, M., Vukosi Marivate, D. & Michael Esiefarienrhe Bukohwo, P. (2019), Predicting road traffic accident severity using accident report data in south africa, *in* 'Proceedings of the 20th Annual International Conference on Digital Government Research', pp. 11–17.

Montella, A., Aria, M., D'Ambrosio, A. & Mauriello, F. (2012), 'Analysis of powered two-wheeler crashes in italy by classification trees and rules discovery', *Accident Analysis Prevention* **49**, 58–72.

Moodley, S. & Allopi, D. (2008), An analytical study of vehicle defects and their contribution to road accidents, *in* 'Proceedings of the 27th Southern African Transport Conference (SATC 2008', pp. 469–479.

Morgan, J. & Sonquist, J. (1963), 'Problems in the analysis of survey data, and a proposal', *Journal of the American Statistical Association* **58**(302), 415–434.

Moyana, H. & Chibira, E. (2016), Improving safety in the road transport sector through road user behaviour changing interventions: a look at challenges and prospects, *in* 'Proceedings of the 35th Southern African Transport Conference (SATC 2016', pp. 516–528.

Nassar, S., Saccomanno, F. & Shortreed, J. (1994), 'Road accident severity analysis: a micro level approach', *Canadian Journal of Civil Engineering* **21**(5), 847–855.

OECD (2017), *South Africa*.
URL: *https://www.oecd-ilibrary.org/content/component/irtad-2017-38-en*

Olden, J. D., Joy, M. K. & Death, R. G. (2004), 'An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data', *Ecological Modelling* **178**(3-4), 389–397.

Olutayo, V. A. & Eludire, A. A. (2014), 'Traffic accident analysis using decision trees and neural networks', *International Journal of Information Technology and Computer Science* **6**(2), 22–28.

Organisation, W. H. (2021), 'Road traffic injuries'.
URL: *https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries*

Pebesma, E. (2018), 'Simple Features for R: Standardized Support for Spatial Vector Data', *The R Journal* **10**(1), 439–446.
URL: *https://doi.org/10.32614/RJ-2018-009*

Perlich, C., Provost, F. & Simonoff, J. (2003), 'Tree induction vs. logistic regression: A learning-curve analysis'.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ripley, B. (2019), *tree: Classification and Regression Trees*. R package version 1.0-40.
URL: *https://CRAN.R-project.org/package=tree*

ROSPA (2002), 'The royal society for prevention of accidents (rospa) road safety engineering manual', Available at hhttps://trid.trb.org/view/730321 (2021/07/19).

Saar-Tsechansky, M. & Provost, F. (2007), 'Handling missing values when applying classification models', *Journal of Machine Learning Research* **8**, 1625–1657.

Savolainen, P., M. F. L. D. & Quddus, M. (2011), 'The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives', *Accident Analysis & Prevention* **43**(5), 1666–1676.

Shanthi, S. & Ramani, R. G. (2012), Feature relevance analysis and classification of road traffic accident data through data mining techniques, *in* 'Proceedings of the World Congress on Engineering and Computer Science', Vol. 1, sn, pp. 24–26.

Sohn, S. & Shin, H. (2001), 'Pattern recognition for road traffic accident severity in korea', *Ergonomics* **44**(1), 107–117.

Trevor, H., Robert, T. & Jerome, F. (2009), *The Elements of Statistical Learning*, Spinger.

Twala, B. (2013), 'Extracting grey relational systems from incomplete road traffic accidents data: the case of gauteng province in south africa', *Expert Systems* **31**(3), 220–231.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S. & Barbone, F. (2002), 'Risk factors for fatal road traffic accidents in udine, italy', *Accident Analysis Prevention* **34**(1), 71–84.

van Schoor, O., van Niekerk, J. & Grobbelaar, B. (2001), 'Mechanical failures as a contributing cause to motor vehicle accidents — south africa', *Accident Analysis Prevention* **33**(6), 713–721.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Visual Crossing Weather API (2021).
URL: *https://www.visualcrossing.com/weather-api*

Weiss, G. M. & Provost, F. (2001), 'The effect of class distribution on classifier learning'.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wickham, H. (2019), *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.
URL: *https://CRAN.R-project.org/package=stringr*

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686.

Wickham, H. & Bryan, J. (2019), *readxl: Read Excel Files*. R package version 1.3.1. **URL:** *https://CRAN.R-project.org/package=readxl*

Yasmin, S. & Eluru, N. (2013), 'Evaluating alternate discrete outcome frameworks for modeling crash injury severity', *Accident Analysis and Prevention* **59**, 506–521.

Yau, K. K. (2004), 'Risk factors affecting the severity of single vehicle traffic accidents in hong kong', *Accident Analysis amp; Prevention* **36**(3), 333–340.

Ye, F. & Lord, D. (2011), 'Investigation of effects of underreporting crash data on three commonly used traffic crash severity models', *Transportation Research Record: Journal of the Transportation Research Board* **2241**(1), 51–58.

Ye, F. & Lord, D. (2014), 'Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models', *Analytic Methods in Accident Research* **1**, 72–85.