TRAFFIC CRASHES PREDICTION USING MACHINE LEARNING MODELS,

CASE STUDY: RWANDA

by

James MUCYO NZABAMBARIRWA

Registration Number: 219013723

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

of

MASTER OF DATA SCIENCE IN ACTUARIAL SCIENCE

in the African Centre of Excellence in Data Science

College of Business and Economics

UNIVERSITY OF RWANDA

Supervisor: Dr. JMV HAKIZIMANA

September 2020

Declaration

I declare that this dissertation entitled **Traffic Crashes Prediction Using Machine Learning Models**, **Case Study: Rwanda** is the result of my work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Names: James MUCYO NZABAMBARIRWA

Signature

Approval Sheet

This dissertation entitled **Traffic Crashes Prediction Using Machine Learning Models, Case Study: Rwanda** was written and submitted by **James MUCYO NZABAMBARIRWA** in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in **Actuarial Science** is hereby accepted and approved.

Dr. Jean Marie Vianney HAKIZIMANA

Supervisor

17.8.2021

Dr Ignace KABANO
Head of Training

Dedication

I dedicate this work to my first family by blood as well as my second family by work (RDF Community).

Acknowledgements

I was able to complete this Dissertation with the help and active support of concerned bodies and authorities as well as several persons. I owe my debt and would like to express my feelings of gratitude to my lecturers and Dissertation supervisor, Dr JMV Hakizimana, Lecturer, College of Science and Technology, the University of Rwanda for his valuable guidance at different stages of this Dissertation research program. It would be impossible to carry on the research and make it to the completion phase without his able guidance and motivation. His sympathy for me will be fondly remembered.

I would be indebted to Dr Ignace Kabano, Head of Teaching, African Centre of Excellence in Data Science, College of Business and Economics, the University of Rwanda for his very useful guidance and counselling in overcoming difficulties during my study. I appreciate his continuous encouragement.

My thanks would be due to Senior Officers of Rwanda National Police, Traffic Department in particular who furnished all the relevant information and dataset for the study without which this dissertation would not be possible. I would express my gratitude to all those officers of RNP whose significant cooperation would be critical to me in compiling the sample from the RNP Database of road traffic accidents. To this end, I would also take the opportunity to thank colleagues: Mr Venant, Alfred, and Isaac whose transport to and from the centre was significant boosting morale when was needed most.

I gratefully acknowledge the financial support received from the African Centre of excellence in Data Science and the World Bank Grant. I would also like to thank the Ministry of Defense through RDF leadership for the flexibility for the whole period of my MSc. Studies.

Finally, my heartiest thanks go to my family, my wife Brigitte Niyonsaba, daughter Princess and my Son, Ivan Prince whose proximity, love provided me joy and relaxation. Once again, thank you, my parents, my Mother, Eline Mukankuyo with your old age, for your everyday prayers. You made my Dissertation!

Abstract

Predicting traffic crashes has become a significant emerging challenge yet Road traffic accidents in Rwanda accounts for 5000 crashes annually on average, claiming 54% of the lives of pedestrians and cyclists. And recent research attempts to tackle this area are limited to summary statistics. This study aims to build a predictive machine learning model so that road traffic accidents can be predicted and policymakers and road traffic users can make informed decisions. Based on the historical road traffic accidents dataset in Rwanda and python programming language, two supervised predictive machine learning models were trained and test on the sample accidents records for six years registered by RNP/ Traffic department across the country. The first model predicted the number of accidents. And the second model classified the accidents based on injury severity categories. The model results indicate the regression model correctly predicts the total number of accidents 100%. The model prediction results match the actual accidents records. The random forest model classifies accidents injury severity at the rate of 91%. Both models are recommended for use as a key to prediction to better prevent road traffic accidents. Moreover, it is recommended that the road traffic accidents database keep daily accidents records to improve the prediction. Further researchers may focus on building automated systems that integrated information from driver's behaviours, road, weather, vehicle data instantly.

Keywords: *Machine learning, road traffic accidents, prediction, linear regression, random forest, confusion matrix, classification, bootstrapping, random subspace.*

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AHP: Analytical Hierarchical Process.

CoK: City of Kigali

DBQ: Driver Behaviour Questionnaire

GES: General Estimate System

GIS: Geographic Information System.

GPS: Global Positioning System

IEEE: Institute of Electrical and Electronics Engineers

ML: Machine Learning

MSE: Mean Squared Error

NASS: National Automotive sampling system

RDF: Rwanda Defense Force

RF: Random Forest

RNP: Rwanda National Police

RTA: Road Traffic Accidents.

WHO: World Health Organisation

Chapter 1: General Introduction

1.1 Background of the Study

Nowadays machine learning has attracted the attention of scientists in this information age due to the exponential growth in data production made available every second through digital and smart devices. Scientists have the desire for extracting useful information and knowledge. The researchers in this field have shown that the information and knowledge gained in this process have a wider range of applications.

Notable fields where knowledge generated by machine learning techniques is used are marketing campaigns through customized and segmented marketing contents tailored for individuals. Next, machine learning is popular in the financial sector for fraud detection and customer retention. There is a range of machine learning classifiers to note: supervised tries to inferring a function from labelled training data, unsupervised are such that no labels are given to the learning algorithm and leaving it on its own to find structure in its input, and reinforced learning not like the two mentioned earlier, we give it an engine to learn the so-called utility function. Each category is subdivided into two categories as follows: classification, regression, clustering, dimensionality reduction, utility and Q-learning respectively.

The application of machine learning techniques in other fields including social sciences to solve social problems has gained popularity in the past two decades. It has gained the attention of researchers in road traffic accidents. Where the latter has been ranked number eight by World Health Organisation (WHO), Global status report on road safety 2018. The same report highlighted that road traffic accidents claimed 1.35 million lives each year (World Health Organisation, 2018). With those statistics, it becomes imperative to contribute to the existing knowledge to minimize this death rate by researching and answering the research question facing the world in the next section of the problem statement. Solving road traffic world accidents starts by countries level. That’s why Rwanda is my research location interest in this study.

1.2 Statement of the Problem

Traffic crashes have a significant impact on the economy in both the form of property damage and also in the form of lost time. The congestion likely to happen in busy areas will cause the waste of gas and air pollution. The worst cases are fatalities or severe injuries according to the 2018 WHO report, the world lost 1.35 million lives (World Health Organisation, 2018). Road safety remains a crucial topic in the transportation industry across the globe. Rwanda National police statistics show that in 2018, there were 5000 traffic crashes in Rwanda and the most vulnerable population in traffic crashes were pedestrians and cyclists with 54% (Rwanda National Police, 2019). The National Police, Traffic department officers have been keeping detailed records on roadways accidents, and built comprehensive traffic accidents information. However, according to the preliminary review of the available literature, the data have been only used to carry out rudimentary statistical analysis and the conclusions of which are mostly patterns and statistics. The more explanatory and predictive nature of the traffic accidents causation, characteristics, and factors analysis have been overlooked. Predicting traffic crashes has become a significant emerging challenge. The question is: How can road traffic crashes in Rwanda be predicted, for the country to be able to allocate the resources available effectively?

This Dissertation research will combine the machine learning model algorithms to construct the predictive model of traffic crashes, explores if the accidents reporting mechanism affects model prediction quality while bringing human behaviour, road network infrastructure conditions, car physical conditions, and geospatial characteristics conditions into the road traffic accidents analysis.

1.3 The objective of the Study

The main research objective of this study is to construct a predictive machine learning model on road traffic crashes in Rwanda.

1.4 Research Questions

- How does the road traffic data collection and reporting affect the quality of model prediction?
- Which of the attributes in the available dataset are most relevant for road traffic accidents prediction?
- Which of the evaluated Machine learning methods are best suitable to the problem at hand?

1.5 Research Hypothesis

The machine learning model can play a big role in road traffic accidents prediction in Rwanda.

1.6 Scope of the Study

This research Dissertation aims to construct a predictive machine learning on road traffic crashes in Rwanda. I will experiment with the road traffic accidents training dataset from Rwanda National Police, Traffic Department and relevant stakeholders where applicable. The data will consist of Rwanda road accidents information from the year 2010 to 2015. That is how the study will be limited both in location, time and size.

1.7 The organisation of the Study

This research Dissertation is organised into five chapters. The first chapter is entailing the introduction to the study and its sub-headings. The second chapter covers the literature review which encompasses the critical analysis of some existing works in this field made by previous researchers. The third chapter will cover the experimental design of this research through a methodical approach that will be applied to the dataset in use. The fourth chapter will cover the experimental setup, training dataset description, feature selection algorithms, model building and accurate measurement. This chapter will end up with an experimental results analysis. Finally, chapter five will conclude with a summary of the key research point of emphasis, suggested recommendations for the stakeholders who will use the research work and highlighting areas for future research.

Chapter 2: Literature Review

2.1 Overview of road traffic accidents and their determinants

Road traffic accidents are one of the major concerns for human beings. It threatens the life to the death of road users. Traffic accidents or collision occur when a vehicle collides with another vehicle, taxi moto, bicycle, pedestrian, animal, or any physical obstacle. Road traffic accidents can either be predicted or prevented.

Table 2.1 Regional road traffic death distribution by WHO, 2018 road and safety report

Period	**Africa**	**Americas**	**East Mediterranean**	**Europe**	**South East Asia**	**Western pacific**	**World**
2013	26.1	15.9	17.9	10.4	19.8	18	18.3
2016	26.6	15.6	18	9.3	20.7	16.9	18.2
Variation	**0.5**	**-0.3**	**0.1**	**-1.1**	**0.9**	**-1.1**	**-0.1**

The table shows how the rate of traffic death is distributed and keeps increasing in low and middle-income countries while it is decreasing in developed countries.

Since the road traffic accident is a result of any collusion, an accident can be interpreted as a function of its cause or determinants.

It can be mathematically expressed as follows:

$$\mathbf{RTA} = \mathbf{f}(\mathbf{HB}, \mathbf{RNI}, \mathbf{CPC}, \mathbf{GC})$$

Whereby:

RTA = Road Traffic Accidents in different forms (severity, medium, minor)

f = Function

HB = Human Behaviour

RNI = Road Network Infrastructure Condition

CPC = Car Physical Condition

GC = Geospatial Characteristic Condition

Briefly, those key road traffic accidents can be explained as follows:

- **Human behaviours**: This can be explained or interpreted as the way road users behave for an accident to happen. For drivers not putting on the seatbelts or violating the zebra crossing. Similarly, the pedestrian may not observe well traffic lights indications intentionally.
- **Road network infrastructure condition**: This may be explained as the general status of the country's road network in its different forms. Thus, asphalt road, one- or two-sided road, road signs, and lighting systems. All the above imply road user's vulnerability.
- **Car physical condition**: This is the status of the mechanical. For example, valid or faulty technical control checks.
- **Geospatial characteristic conditions**: This can be explained as the weather condition in general. For example, night vision may impair road user's visibility. In addition, day, night, rainy, cloudy weather may amplify the magnitude of road traffic accidents.

Within his context, let critically review what different researchers have written of this global challenge for an opportunity to contribute the existing knowledge.

2.2 Previous approaches to accidents prediction both on a global, regional level

According to the study made by (Stoop, 1995) on the in-depth investigation of the accidents, the study highlighted that road safety research has been very predominant in the USA and Canada before 25 years when they introduce this in-depth analysis approach. However, the researches made in this era have been focussing on the role of the operator or the conductor. The researchers were contaminated by the question of blame and reliability. In addition, the traditional statistical-oriented approach has proved itself limited due to the complex phenomenon involving traffic accident collisions.

Despite the traffic road accidents was ranked 8th globally claims lives of the World population and the limitation identified in the 70s research works on this topic, the literature is not large. Moreover, the most available attempts to carry out the machine learning process on road accidents data of a single city or very small area which may suffer population representation.

Furthermore, the same study (Stoop, 1995) suggested that there was a need by the researchers to introduce a variety of research techniques and the application system approaches enabled due to the introduction of information technology and modern management of transport logistics. Since then accidents and traffic processes have become a solid basic subject of research.

The dataset from Ethiopia was frequently used as basic data, the fact was that this country recorded a high number of road traffic accidents per capita. (Tibebe & Shawndra, March 2010) the study used 18,288 accidents data records from Addis Abba. In addition, the study used Naïve Bayes, decision trees, and K-nearest neighbour algorithms and cross-validation as a methodology to classify the data. The algorithm accuracy yield value is 80%. The algorithm's results are strong however, the author generalised the whole not taking into account the individual accidents. In addition, the algorithms may not yield the same accuracy values if the entire countries accidents records were used.

Another research conducted by (Tesema, Abraham, & Grosan, 2005) on rule mining and classification of road accidents using regression trees, used the data from Ethiopia. The dataset has many variables to the maximum of 36, however, 13 was used in model building. The researcher failed to show the feature selection process, this may render the study results to be not representative.

In 2012, a research work conducted by (Shanti & Geetha Ramani, October 24-26, 2012) on feature relevance analysis on the classification of Road Traffic Accidents in 56 states of U.S, used classification algorithms such C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree 457549 on the sample with 33 attributes within information records of 5 years. The study found that the misclassification errors were reduced in three-stage experiments performed at 84.62% accurate and 15.38% or error rates. The study was significantly well-performed since it passes all the phases of the model construction process. This study may have been supported by the strength of road networks infrastructure in the U.S.

Furthermore, a study conducted by (Quanjun, Xuan, Harutoshi, & Ryosuke, 2016) analyzed the effect of human mobility on the traffic accident risk. The same study used a sample of 300 thousand records of traffic accidents collected from Japan from January 1, 2013, to July 31, 2013, with 3 attributes. This dataset was combined with the human mobility data of 1.6 million users collected through GPS records both corresponding to the same period with traffic accidents records. The deep learning neuro-network was constructed and randomly trained 80% and 20% tested and evaluated to the dataset. The results of the model simulation found that high traffic accident risk is more intensive in the connection business activities in the regions than in other road network regions. This is significant because human mobility was often overlooked by many researchers as a key attribute to the increase in traffic accident level of risk.

In a study conducted by students from the computer science department from Oklahoma State University, in the USA in the partnership of Korean and Poland Universities tried to model the severity of injuries that occurred during accidents using four machine learning paradigms; artificial neural network, support vector machines, decision trees and a concurrent hybrid model of decision trees and neural network. The study used a sample dataset from the National Automotive sampling system (NASS) General Estimate System (GES). The dataset was assumed to be nationally representative. The total number of 417, 670 recorded cases from 1995 to 2000 was initially used. The study found that the hybrid decision trees-neural network model outperformed the rest of the individual models, with 90% promising results on fatal injuries (Miao, Ajith, & Marcin, December, 2004). This study importantly brings to the attention of researchers in the field that when different machine learning models are used, it is rare to obtain one that can perform better on all aspects of accidents classes.

In an exploratory postal survey conducted on decision-making style, driving style and self-reported involvement in road traffic accidents by (Davina, John, & James, 1993) decision-making and driving style questionnaires administered on 711 drivers across the UK was measured and assessed on 7 independent and internally coherent dimensions according to the principal component analysis and 6 independent dimensions of driving style. The results on multiple regression analysis indicated that drivers aged 60 years old and below scored low against thoroughness as the independent variable was at high risk of road traffic accidents mediated by faster driving. While drivers over 60 years old scored low at thoroughness, hesitancy was associated with their risk of road traffic accidents. The findings may have been applicable if the data used were enough to be worth a generalisation of the study. In addition, the scope of the study fails to include historical data insights. The responses from questionnaires may include subjectivity and individual bias of the survey participants.

Similarly, (Wahlberg, Dorn, & Kline, 2011) in their analysis of the Driver Behaviour Questionnaire as a predictor of road traffic accidents, the association between crashes and the violation and errors factors may be spuriously high due to the reporting bias, as a result, the DBQ may not be a successful predictor of accidents as claimed by (Davina, John, &

James, 1993) in their study on Decision-making style, driving style and self-reported involvement in road traffic accidents.

A multivariate study conducted by (Jianfeng, Zhonghao, Wei, & Quan, 2016) to establish a causational factor analysis of road traffic accidents using analytical Hierarchy Process (AHP)-Apriori algorithm logistic model on sample data of 10,000 accidents data and 20 types of accidents factors collected from eastern, north and northeastern regions of China. The Apriori algorithm was used was applied to analyse the degree of accidents or the level of influence. The study findings show that the method and AHP proved capable of determining the type and severity of accidents and their factors. The question is to know whether the method AHP-Apriori can or cannot be contextual or specific to a certain region, since the researchers did not mention its applicability of out the regions where the research was conducted, China.

2.3 Previous attempt studies to traffic accidents prediction in Rwanda

Apart from different policies formulated by the Ministry of the infrastructure of Rwanda and the Rwanda National Police and its various stakeholders to curb and prevent road traffic accidents, few studies have been conducted in this area in an attempt to bridge the gap. However, they were exploratory or epidemiology studies and they failed to employ the benefits the machine learning offers. In addition, they were limited in scope both in geographic location, time and sample size to the point the generalisation of the study outcome could be subjective and almost difficult to implement nationwide. The consulted previous studies on road traffic accidents were reviewed and can be critically analysed based on their sample data, methodology used, the outcome as follows:

The most recent epidemiology study of road traffic injury hotspots in Kigali (Patel, et al., 2016) has analysed the police data on traffic accidents recorded between January 1 and December 31, 2013, to identify the hotspot locations in Kigali. The study used descriptive statistics for visual representation and mapped the statistic using the GIS Software. The findings show that the study has succeeded in approximating the hotspots, however, the sample geographic study area was not representative to conclude based on the variables analysed, that the study can support national wide policy formulation. In addition, the study fails to predict in terms of a quantitative number of accidents that are likely to happen on the identified hotspots points shortly.

From the literature reviewed, the early 1975 studies on accidents analysis and prevention have proved their methodologies to be successful, however, with modern information edge, they are no longer practical. The modern transportation system with embedded sensing technologies, recent research studies have found it imperative to cope up with the new change in technology and employ modern tools capable of handling analyzing volume traffic flow information that is dumped in institution database systems. Furthermore, the researchers have identified unstandardized traffic accidents documentation resulting in incomplete and many outliers in the dataset used that could impact the prediction accuracies of their models

The disparities in the research gap on transportation processes are also evident between the developed countries and low and middle-income countries where most road traffic accidents are pertinent and a major health concern. The present Dissertation research will employ a machine learning model to predict the severity of the accidents in quantitative form, the data quality plays an important role. The next chapter will provide a detailed breakdown of the methodology to be adopted through the Dissertation research.

Chapter 3: Methodology

Data mining and training are two distinct phases in data compilation, model training and results from evaluation for model-based machine learning approaches. Data mining refers to the process of collecting, analyzing and consolidating the data from multiple sources. And training phase involves the use of the compiled data and fit the Machine learning model to the data. This later phase culminates with the evaluation of the outcome to see what inputs variables have influenced the model performance.

3.1 Machine learning

Machine learning is a subset of the Artificial intelligence field which is concerned with building automatic systems that learn from examples from different sources intending to extract actionable insights (Harrington, 2012). Furthermore, the rationale behind model-based machine learning is to form the problem domain-based assumptions into a model form (John Winn, 2013).

" A program is said to learn from experience E concerning some classes of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. $(\text{Mitchell.}, 1997)$. ''

In light of my research study, the experience E is the information from the road traffic accidents and other related information obtained from various sources. The task T is to predict the road traffic accidents and the performance measure P is related to the size of the error term in the prediction.

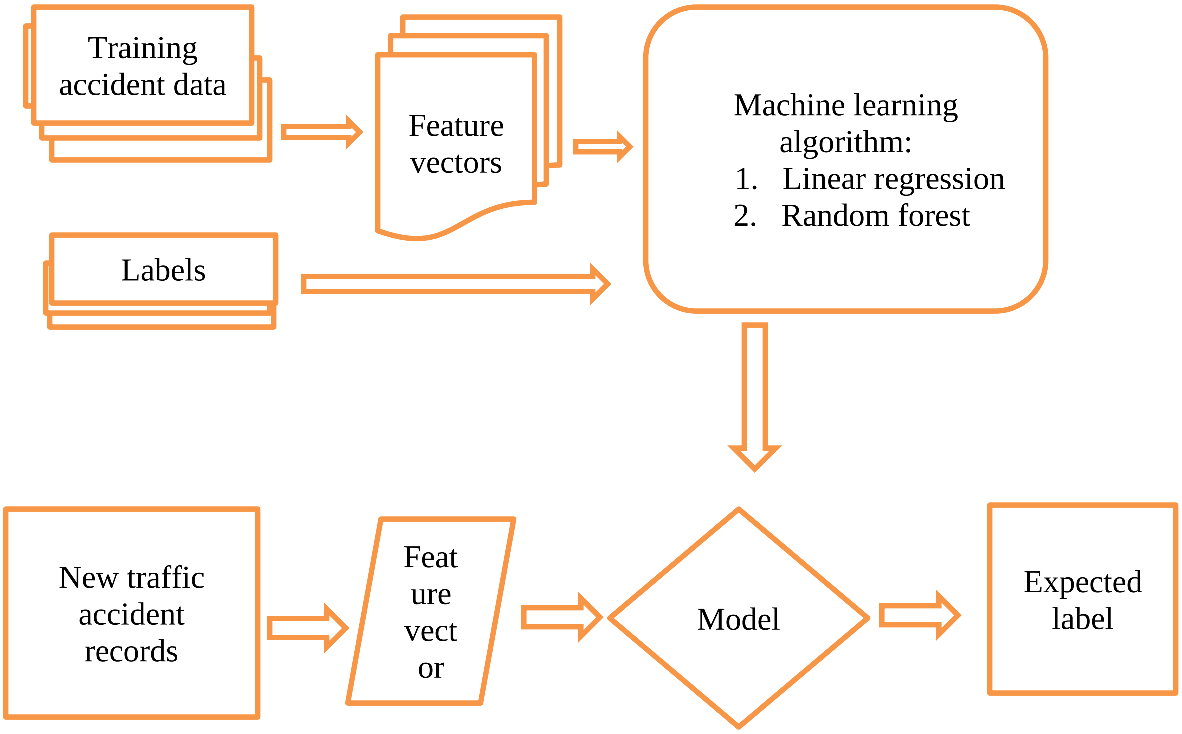


Figure 3.1 The Research modelling process design of this study

3.2 Linear regression method performance metrics

To evaluate the performance of the Machine learning regression model on mapping the inputs variables to the target variable, different metrics will be applied to the trained model. I intend to use the following metrics:

3.2.1 Bias and variance

In the machine learning regression task, the bias of the model is referred to as the loss incurred by model prediction from the optimal prediction. Thus the difference between the estimated value of a parameter and its true value (Domingos, June, 2000). Mathematically, the variance can be expressed as follows:

The bias of a model on an example x is $\boldsymbol{B(x) = L(y_{*}, y_m)}$ $\qquad$ $\mathbf{(3.1)}$

The variance can be referred to as the average loss incurred by predictions relative to the main prediction. This is a measure of the model stability in response to the new training dataset. It means variation estimations between model realizations. It does not measure model correctness but consistency. However, high variance is an indication of overfitting (Mannor, 2007). And it can be represented as follows:

$$\boldsymbol{V(x) = E_D[L(y_m, y)]} \qquad (\mathbf{3.2})$$

It can be deduced that the total loss is a function of $B(x)\ and\ V(x)$

3.2.2 Mean squared error

The mean squared error is the difference between estimates and the true value. The smaller the mean squared error the better. As previously shown the error comprises both variance and bias (Kong, 1995). It can be expressed as follows:

$$\boldsymbol{MSE = VarianceEstimate + Bias\ (Estimate, TrueValue)^2} \qquad (\mathbf{3.3})$$

3.2.3 Per cent error

The per cent is the relative error expressed as a percentage. It is used to complement the residual mean squared error for the model fitted prediction details. It can be mathematically expressed as follows: $\boldsymbol{\varepsilon = 100 \,.\left|\frac{(TrueValue - Estimate)}{True\,Value}\right|}$ $(\mathbf{3.4})$

3.2.4 The coefficient of determination (r^2)

The r^2 is the coefficient that measures the dependability of dependent variables to some inputs variables (Wang, 2017). It varies from a negative one to one (Alexander, 2015). The negative one indicates no relationship between the input variables and the dependent variables. Any value above 0.5 indicates a positive relationship while the one indicates a strong positive relationship between dependents and independent variables (Grömping, 2015).

3.3 Linear model regression

The linear model is a supervised learning method that predicts the value of one or more continuous target variables y given the value of a D-dimensional vector x of the input variable (Christopher, 2006). The simplest form of linear regression takes the following form: $y = w_1x + w_0$ where w_1 and w_0 are real values to be learned. The ultimate goal of the linear regression problem is to minimize the loss function (Christopher, 2006). Thus finding the weight (w^*) that guarantees the global minimum as expressed in the following formula:

$$\boldsymbol{w}^* = {}^{\boldsymbol{argmin}}{}_{\boldsymbol{w}}(\boldsymbol{h}_{\boldsymbol{w}}) \tag{3.5}$$

So that: $\boldsymbol{h}_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}_{\boldsymbol{1}}\boldsymbol{x} + \boldsymbol{w}_{\boldsymbol{0}}$ **3.6**

For multiple linear regression, there are more advanced computations that build on simple regression. The same principles apply but, in multiple linear regression each example say x_j is n-element vector and the goal is to find the hyper-plane that fits the out y based on some loss function. The squared loss function. The following functions give the hypothetical space (Mannor, 2007).

$$\boldsymbol{h}_{\boldsymbol{sw}}(\boldsymbol{x}_{\boldsymbol{j}}) = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{j}} = \sum_{\boldsymbol{i}} \boldsymbol{w}_{\boldsymbol{i}}\, \boldsymbol{x}_{\boldsymbol{j},\boldsymbol{i}} \tag{3.7}$$

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\overset{\boldsymbol{argmin}}{}}\sum_{\boldsymbol{j}} \boldsymbol{Loss}\left(\boldsymbol{y}_{\boldsymbol{j}}, \boldsymbol{h}_{\boldsymbol{sw}}\right) \tag{3.8}$$

In this research study, the linear regression model will be applied.

3.3.1 The Cook's distance

The Cook's distance (Cook’s D) is the term given to the metric that measures the distance between data points on the linear regression model (Mitchell., 1997). it is mostly used to measure the influence of data points when calculating the least-squares. To clear confusion, let interpret Cook’s D in a mathematical form:

$$\boldsymbol{D}_{\boldsymbol{i}} = \frac{\sum_{\boldsymbol{j}}^{\boldsymbol{n}} = \boldsymbol{1}(\tilde{\mathrm{Y}}_{\boldsymbol{j}} - \tilde{\mathrm{Y}}_{\boldsymbol{j}(\boldsymbol{i})})^{\boldsymbol{2}}}{\boldsymbol{pMSE}} \tag{3.9}$$

From this mathematical expression, $\tilde{\mathrm{Y}}_j$ is the model prediction on the observation j. And $\tilde{\mathrm{Y}}_{j(i)}$ is the model prediction on the same observation, where the trained model on data has missed the observation i. The parameters fitted to the model are expressed by p and the mean squared error by MSE (Bangalore, June, 2000). Therefore, Cook’s D will be used when analyzing the results of the regression linear model.

3.4 The random forest model design

The random forest is a regression machine learning method build on the decision tree architecture. But random forest differs from decision tree in that it assembles several decision trees regressed and trained. The decision in the random forest is based on the majority rule (Dogru, February, 2018).

The random forest learning classification method was invented to overcome the flaws of the overfitting of the decision tree method. The random forest is efficient on a large database and achieves high prediction accuracy even when there is missing data (Biau, 2012). Furthermore, the random forest has built-in functions that enable subspaces randomly on the training data to achieve high prediction accuracy (Breiman, 2001).

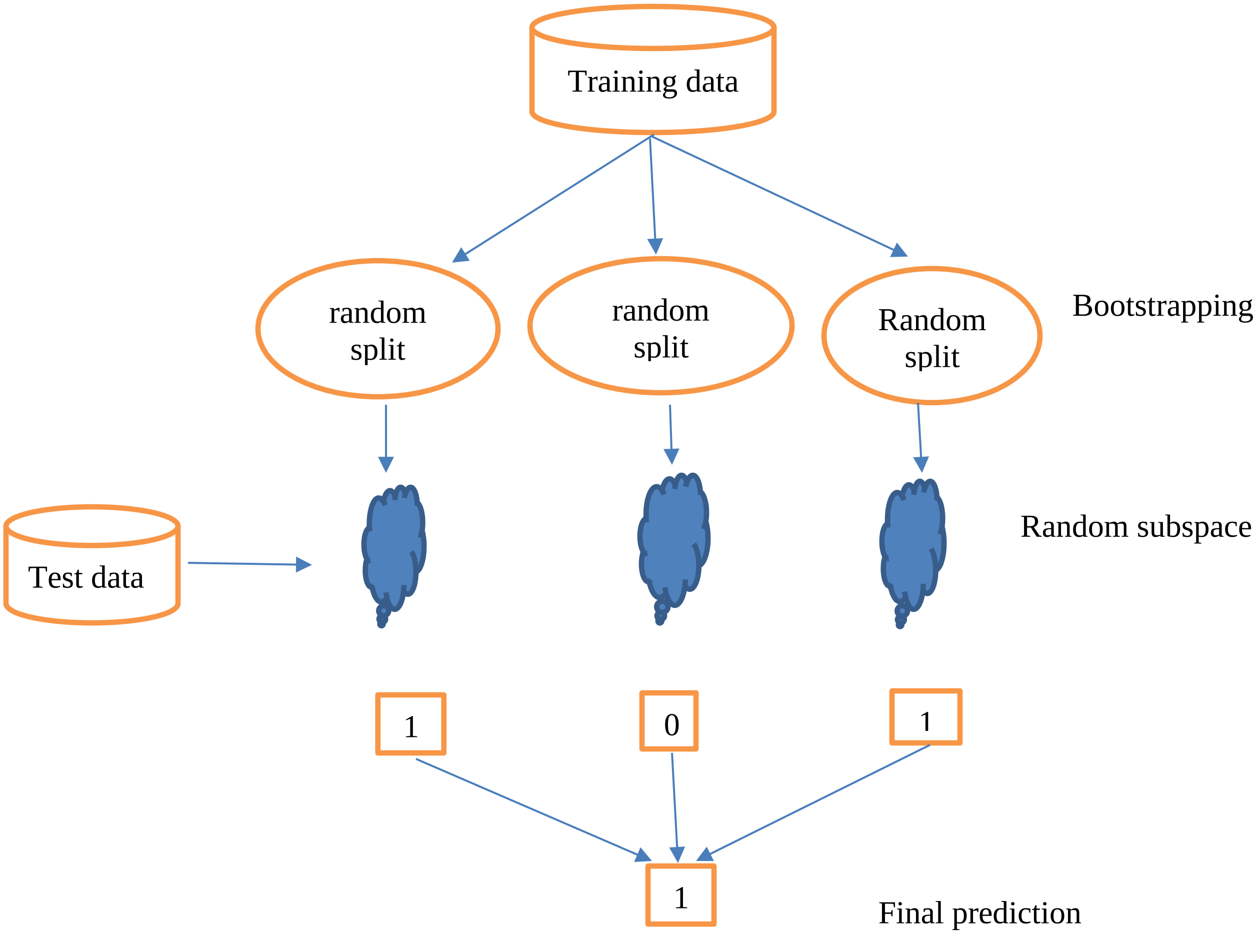


Figure 3.2 Random forest model development design

3.5 Evaluation

The error metrics on prediction test data will be computed to determine the best models suited to fit the data. Those metrics are, mean squared error, per cent error for a regression model, and precision and recall, F-score and confusion matrix for the random forest. Finally, to determine features significance, both linear and random forest models will be used to extract the ratings according to their importance and contribution in models.

3.5.1 Precision and recall

The precision and recall is an error metric for skewed analysis that gives the values between zero and one (Sajjadi, 2018). It is important when evaluating a model precision in predicting a value of each element in the test set, to know how often the model is causing a false alarm and by the recall, how sensitive our algorithm is making a correct guess in known correct elements (Morstatter, August, 2016). For both precision and recall, high good precision is closer to one.

3.5.2 F1 Score

The F-score is an error type metric skewed analysis that results in a unique value. It is like taking the mean of the precision and recall and assign the weight (Huang, 2015). It is deducted that if precision or recall is zero, the F-score also is zero, if precision or recall is one, the F-score is also one, else the remaining values are between zero and one (Chicco, 2020).

3.5.3 Confusion matrix

According to (IEEE, 2017), the confusion matrix is considered critical to the classification task or problem in that, the true positives and true negatives are identifiable in each class in question. Furthermore, this metric evaluates the performance of the model in a detailed manner. So that anyone interested in the cause of the model performance behaviour can investigate further and correct where the model fails to improve performance (Deng, 2016). The architecture of the confusion matrix by its name is $N X N$ matrix, where N is the number of predicted classes. In the present study this confusion metric will be used to evaluate the random forest model performance on the classification of road accidents severity injury classes:

Table 3.1 Confusion matrix architecture table

Confusion matrix		**Model prediction results**	
		Class 1	**Class 2**
Actual ground reality	**Class 1**	Value (positive guess)	Value (negative guess)
	Class 2	Value (negative guess)	Value (positive guess)

3.6 Practical implementation software

Throughout this research study, the python programming language will be used for both machine learning and statistical analysis. Python is a high-level scripting language convenient to human-readable and coding (Martins, June, 2016). In addition, python is an open-source scripting language and does not require a license to use, share and modify. It is reached in ML-enabled community libraries. In data collection and preprocessing, python will be used.

Chapter 4: Data Analysis and Results

In the present study, we used the secondary data source that was available to the office of the RNP/Traffic database. The data provided were a sample of road traffic accident records for the 30 District over six years starting from January 2010.

4.1 Road accidents management data

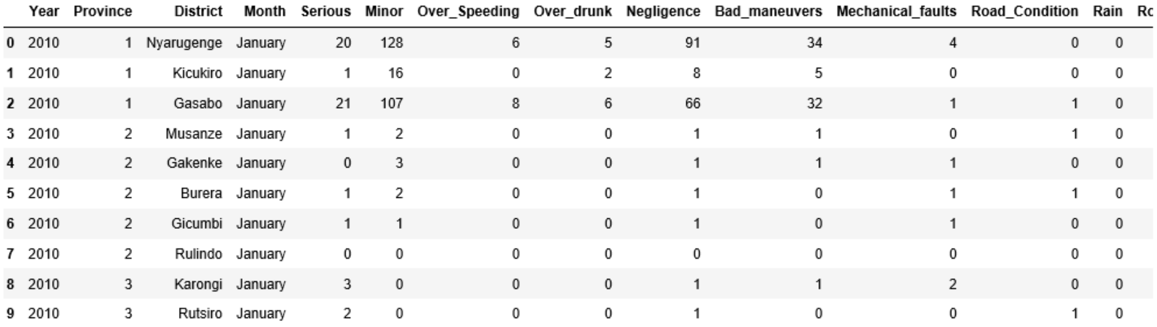
The road accidents database contains the information on registered accidents aggregated and reported every month on the country's road infrastructure network. The study will be limited to aggregated datasets reported monthly on these registered accidents. The limitation is due to missing individual-level daily accidents records and reporting which are widely used in developed countries. Furthermore, accidents were categorized into serious injury severities and minor injury severities. The data was given in the spreadsheet format file and contained the following information for the monthly aggregated injury severity per Month and District: Over-speeding, over-drunk, Negligence, Bad-maneuvers, Mechanical-faults, Road-condition, Rain, and Others. The table following table shows the variable of interest to our study.

Table 4.1 Variable of interest from road traffic accidents data

Variable name	**Description**
Year	When the accident happened
District	District where the accident took place
Month	Month (period) in which accident happened
Over-speeding	# Accident caused by driving out of speed control
Overdrunk	# Accident caused by drink and drive factor
Negligence	# Accident caused by the negligence of drivers
Bad manoeuvres	# Accidents caused by driver's bad decision
Mechanical faults	# Accidents caused by vehicle bad mechanical condition
Road condition	# Accident caused by road infrastructure condition
Rain	# Accidents caused by the weather factor condition
Road sign	# Accidents caused by either violation or inexistent of traffic signs

Other	# Accidents caused by the unspecified reasons
Serious	# Accident classified as Serious
Minor	# Accidents classified as Minor

Rwanda National Police operate across 30 District countrywide stations. These later report every accident in their area of operation and report back for consolidation to the RNP Traffic database. The majority of the road traffic accidents registered took place in Kigali City comprising of three districts, Nyarugenge, Gasabo and Kicukiro respectively. The considered accident records have been visualized in the figure. The crucial problem is whether the model will generalize and achieve accurate predictions for all districts.



	Year	Province	District	Month	Serious	Minor	Over_Speeding	Over_drunk	Negligence	Bad_maneuvers	Mechanical_faults	Road_Condition	Rain	Rc
0	2010	1	Nyarugenge	January	20	128	6	5	91	34	4	0	0	
1	2010	1	Kicukiro	January	1	16	0	2	8	5	0	0	0	
2	2010	1	Gasabo	January	21	107	8	6	66	32	1	1	0	
3	2010	2	Musanze	January	1	2	0	0	1	1	0	1	0	
4	2010	2	Gakenke	January	0	3	0	0	1	1	1	0	0	
5	2010	2	Burera	January	1	2	0	0	1	0	1	1	0	
6	2010	2	Gicumbi	January	1	1	0	0	1	0	1	0	0	
7	2010	2	Rulindo	January	0	0	0	0	0	0	0	0	0	
8	2010	3	Karongi	January	3	0	0	0	1	1	2	0	0	
9	2010	3	Rutsiro	January	2	0	0	0	1	0	0	1	0	

Figure 4.1 Variables of interest from road traffic accidents data

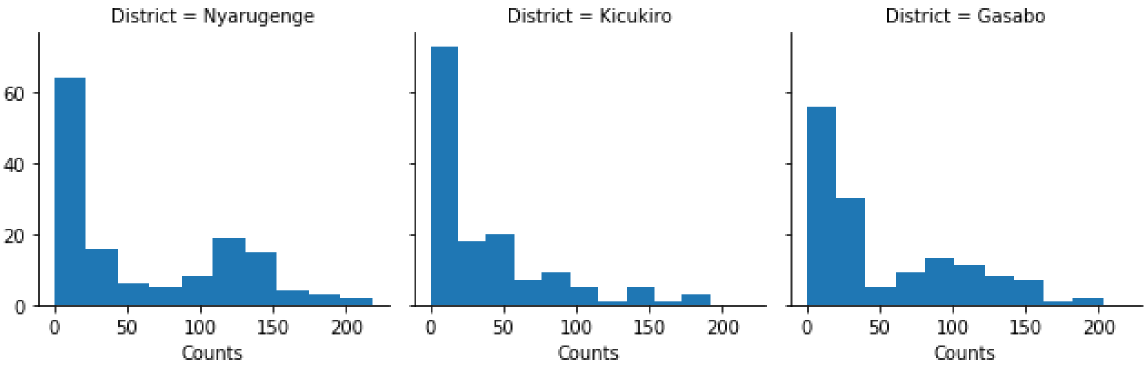


Figure 4.2 The Histogram of City of Kigali accidents distribution

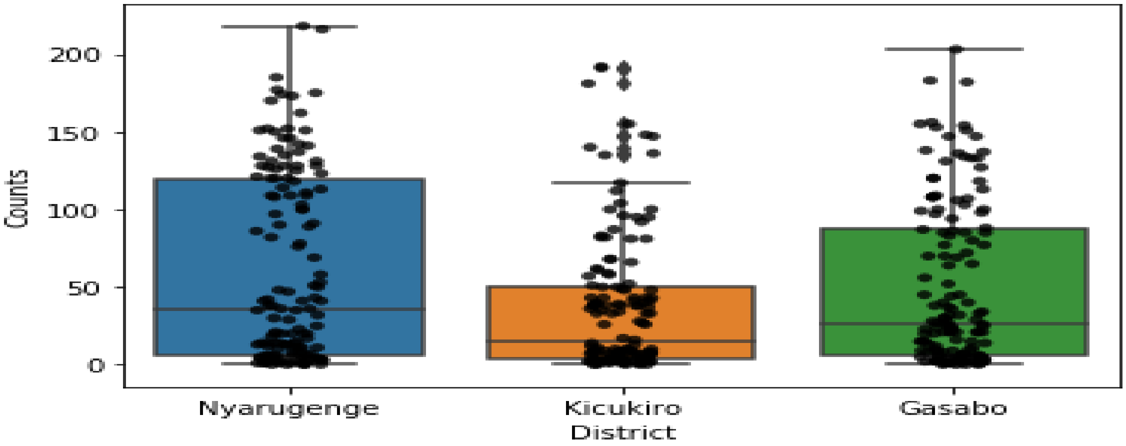


Figure 4.3 Boxplot visualisation of City of Kigali accident distribution

Road accidents database also contain little information about the driver’s behaviour. Some reported characteristics of driver’s behaviour inclined in the road accidents include Drunkenness, over-speeding, negligence and their respective reported numbers of accidents consequences.

4.2 Road data attributes

Road data comes as several accidents caused by the road signs. Other characteristic features or attributes of road network infrastructure was missing. Though the relationship between road signs and the District can be established It will not add value since the road accidents are aggregated on monthly basis. No GPS records were available.

4.3 Vehicle data attributes

The vehicle attributes were reported as an aggregate of several accidents caused by the vehicle mechanical faults. There was no information provided such as make, engine size and type, gearbox, occupational, tyres and vehicle frequency in technical control. Therefore, only mechanical faults accidents number will be included in the training and testing standard linear regression.

4.4 Weather data attributes

The current reporting system of road accident data could not facilitate to easy know at what exact time the road accident took place. Since reported figures are aggregated on monthly basis. Only accidents caused by rain was available. The proper records of the time of the day accidents took place could help in assessing driver's behaviour and vehicle data.

4.5 Selection and data filtering

Subject to the practice in the above sections, the road traffic accidents data management is limited to:

a. Accidents are registered, aggregated and reported monthly across districts.

b. Accidents were reported between January 2010 to December 2015.

c. RNP traffic database only.

The road traffic accidents database contains some incomplete data. The filtering of which resulted in 4230 aggregated records. For consolidation purposes, these records may be reduced in preprocessing for quality checks.

4.6 Data consolidation steps

The dataset contained both driver's behaviour, road data, vehicle data, and weather data to a certain extent. Therefore, in the consolidation process, data points were matched to each other against the selection criterion. Then the consolidated data passed through model training. Visual preprocessing steps followed are as follows:

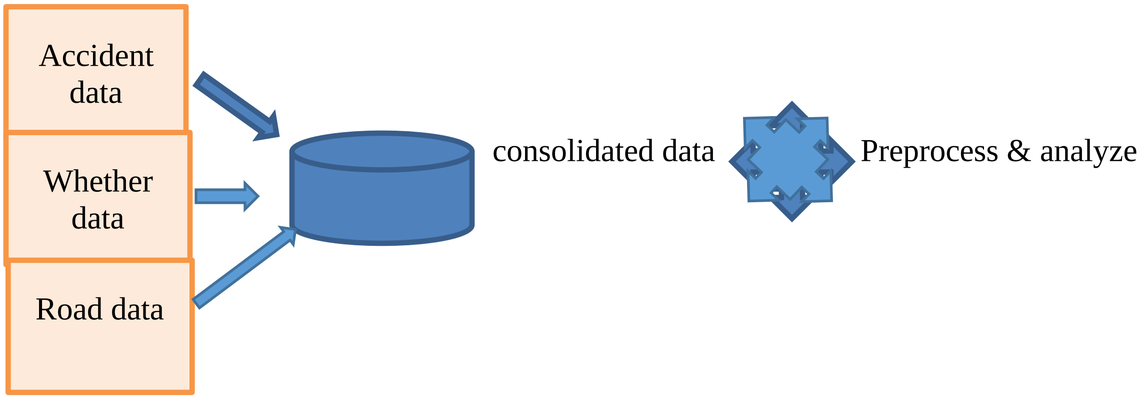


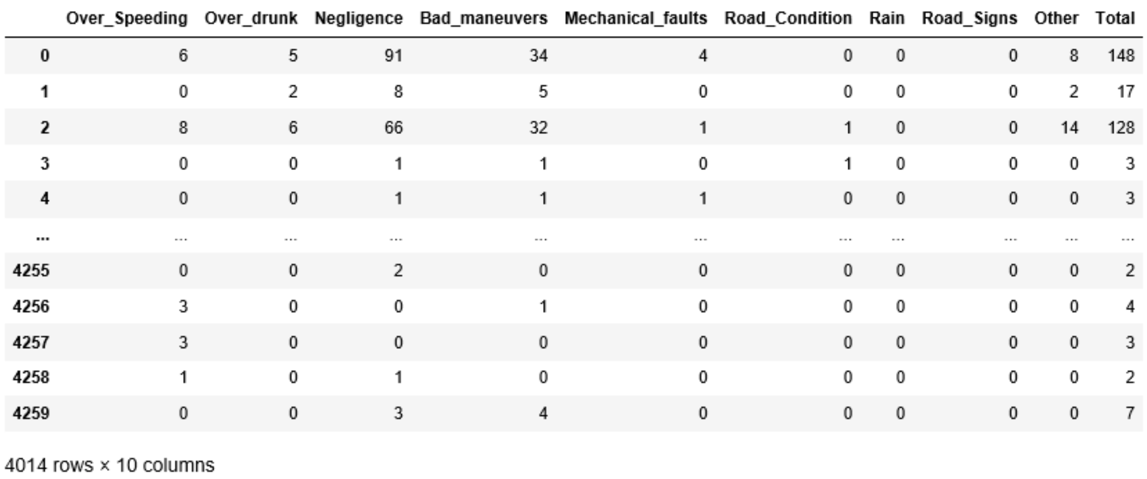
Figure 4.4 Pictorial view of data consolidation processes.

4.7 The final resulting dataset

After cleaning and applying transformation on data sets driver behaviour, weather, routes and vehicle mechanical faults. The 4230 records were stored on local files. These records were matched to the unique Month and District to be able to track back to their original dataset. After preprocessing steps, redundant data were removed and the number of records was reduced to 4014 records. The final training dataset is shown in the following table and figure for illustrations.

Table 4.2 Final sets of features put in the training data

Variable name	**Description**
Over-speeding	# Accident caused by driving out of speed control
Overdrunk	# Accident caused by drink and drive factor
Negligence	# Accident caused by the negligence of drivers
Bad manoeuvres	# Accidents caused by driver's bad decision
Mechanical faults	# Accidents caused by vehicle bad mechanical condition
Road condition	# Accident caused by road infrastructure condition
Rain	# Accidents caused by the weather factor condition
Road sign	# Accidents caused by either violation or inexistent of traffic signs
Other	# Accidents caused by the unspecified reasons
Serious	# Accident classified as Serious
Minor	# Accidents classified as Minor



	Over_Speeding	Over_drunk	Negligence	Bad_maneuvers	Mechanical_faults	Road_Condition	Rain	Road_Signs	Other	Total
0	6	5	91	34	4	0	0	0	8	148
1	0	2	8	5	0	0	0	0	2	17
2	8	6	66	32	1	1	0	0	14	128
3	0	0	1	1	0	1	0	0	0	3
4	0	0	1	1	1	0	0	0	0	3
...	...	...	...	...	...	...	...	...	...	...
4255	0	0	2	0	0	0	0	0	0	2
4256	3	0	0	1	0	0	0	0	0	4
4257	3	0	0	0	0	0	0	0	0	3
4258	1	0	1	0	0	0	0	0	0	2
4259	0	0	3	4	0	0	0	0	0	7

4014 rows × 10 columns

Figure 4.5 Features included in the final training dataset

4.8 Dividing and Normalizing Data for model training and testing

Dividing and normalising involves splitting the dataset into training and test sets. Then the model is trained and test performance based on the splits conditions to prevent under or model overfitting (Van der Aalst, 2010). For this study, to train the model, the 9 inputs variables and 1 output variable (target variable) were used. Of the values across the variables, 80% of them were put into model training for the learning process. The inputs variables have in total 9 variables (put into columns) and 4014 rows. The 80% put forward for model training is equal to 9 variables and 3211. The target variable (Total accidents) has one variable and 4014 rows. The 80% of the target variable put forward for model training is 3211 rows.

After training the model on 80% data of input and target variables, the next step was testing the learned model on the remaining 20% data for both inputs and target variables. The inputs test data was 9 variables with 803 rows. While the target variable test data was 803 rows with one variable. It is common for a model to be trained on large data and tested for small data (Cawley, 2010). The purpose is to see if the model has learned well on the training dataset.

4.9 Data analysis

The analysis of the data was performed to identify correlation patterns between features or variables. Seven features have a strong correlation while three have no low degree of correlation.

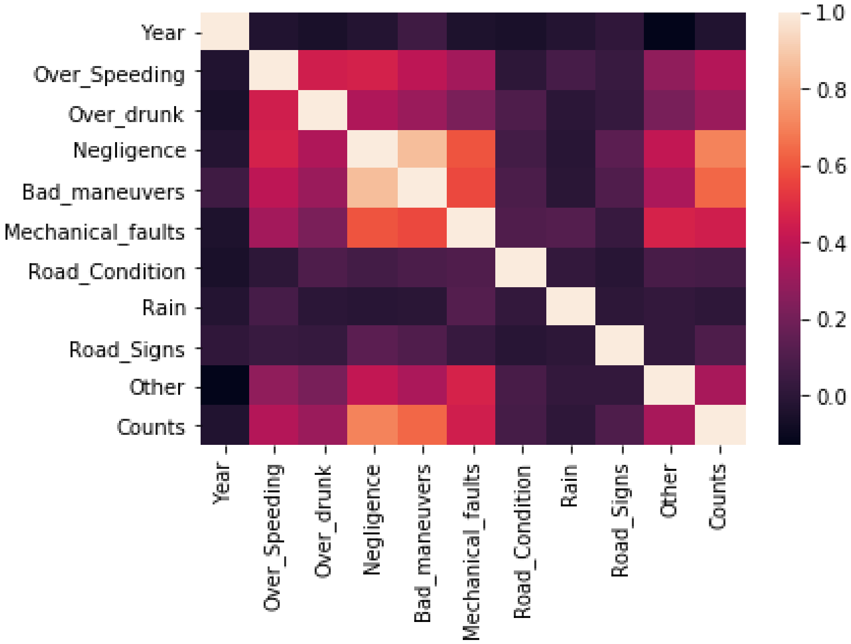


Figure 4.6 Correlation of variables in the data per Month

From the heat map, the diagonal shows that a variable is strongly correlated to itself. And legend from 0.0 to 1 with colour from black to light yellow indicates a direction from weak to strong correlation between variables. For example, negligence appears to be strongly correlated to bad manoeuvres and less correlated to Rain as variables in the dataset.

4.10 Linear regression performance results

The model has been trained and tested, the next step is to show its performance as well as the interpretation of the results. The model has predicted the total accidents on the inputs variable test data. The mean squared error (MSE), Cook's Distance and the R^2 performance metrics were used. The mean squared error for this multiple linear regression model is 0.28. Furthermore, the coefficient of determination R^2 is 1. And the cook's Distance turned to be approximately zero. For better visualization of the model performance results, the following figure in scatterplot style shows the pictorial view.

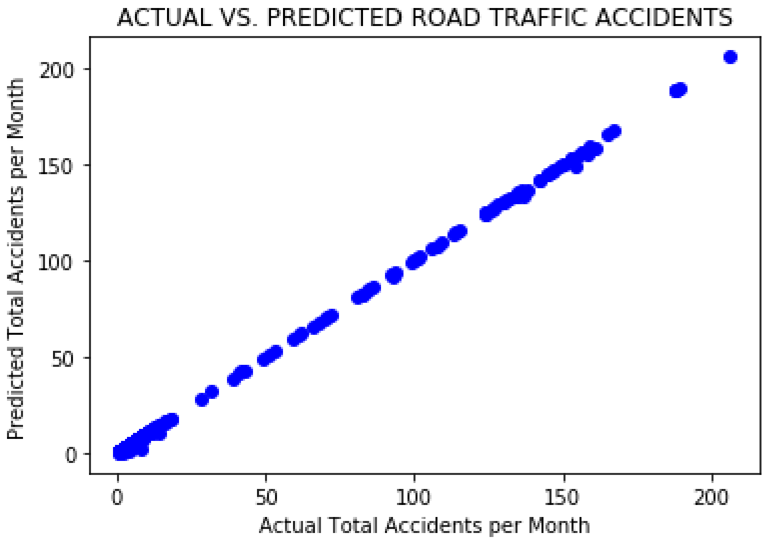


Figure 4.7 Regression model performance

The above model performance; MSE of 0.28 and R^2 of 1 has the following meaning about the model learning and prediction capabilities. The mean squared error of 0.28 means that the model predicted well the total accidents from the learned inputs variables (accidents factors or cause) supplied with. It further means that the actual accidents registered by Rwanda National Police were the same as the results of the model prediction. There is no deviation from the model prediction and the actual accidents records available.

Moreover, the R^2 of 1 means that there is a strong positive correlation between inputs variables (accidents causes or factors reported by RNP in the used dataset for this model) and the prediction results. We perform a linear regression model to rate the variable feature importance (Grömping, 2015). The following illustration shows ratings of variables with the associated weights

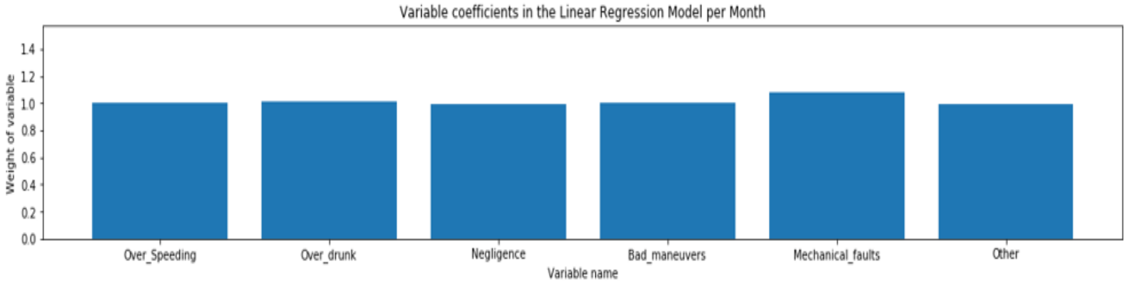


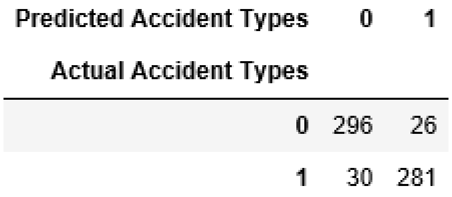
Figure 4.8 Variable coefficients from Linear Regression Model per Month

The y-axis has been arbitrary set to 1.4 because all variable coefficients were below this 1.2. The illustration shows that mechanical faults have a significant influence on the model followed by overdrunk, negligence, bad manoeuvres, overspeeding and finally other unspecified reasons.

4.11 Random forest model performance results

The model has been trained and tested to classify the injury severity in two classes serious and minor. For the model to train and test, the categorical class were converted into numerical classes, 0 stands for Serious injury and 1 stands for Minor injury for computing purposes. The model performance prediction accuracy is 91%.

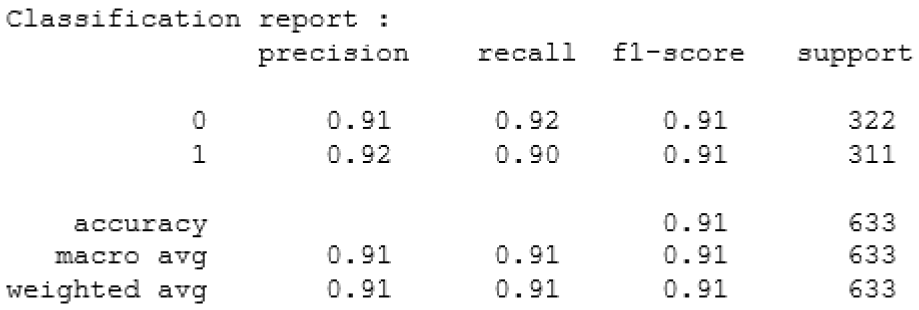
The interpretation of the accuracy score in form of percentage for the classification task is subject to confusion since it does not provide on which class the model has made an error of classification. Here is the same result, in the form of the confusion matrix.



Predicted Accident Types	0	1
Actual Accident Types		
0	296	26
1	30	281

Figure 4.9 Confusion matrix

This confusion matrix is the true measure of this random forest model classification accuracy. The diagonal indicates the correct class classification. This means that model has classified 296 serious injuries correctly and 281 minor injuries. However, the model failed to 26 and 30 severity injuries in their true respective classes. This further indicates that this model has been learning well on the training dataset. The same results have been interpreted in the following classification report for illustration.



```
Classification report :
              precision    recall  f1-score   support

           0       0.91      0.92      0.91       322
           1       0.92      0.90      0.91       311

    accuracy                           0.91       633
   macro avg       0.91      0.91      0.91       633
weighted avg       0.91      0.91      0.91       633
```

Figure 4.10 Random forest performance

From the above F4.10, the model precision is 0.91 on serious injuries and 0.92 on minor injuries, it is high and closer to one on both predictions for each element of both classes. This is a good indicator of how often the model caused the false alarm. The recall error metrics turn to be high, 0.92 on serious and 0.90 on minor injuries, this is also and a good indicator that the model made few prediction errors on the already know element for each class. Since the F-score is the average of precision and recall, it is evident that it must lie between 0 and 0.92. Thus, the result shows 0.91. This model is recommended for use, its prediction accuracy satisfactory since the goal of machine learning is not to find the perfect model that matches the ground reality but to make a good guess enough.

Chapter 5: Conclusion and Recommendation

5.1 Conclusion

This study aimed to construct a predictive model on road traffic accidents data using machine learning. The python language libraries were used to construct a multiple linear regression model for total accidents prediction and the random forest model for classification of injury severities. The multiple linear regression model produced better prediction results at the rate of 100% while random forest got 91% on classification prediction accuracy. These findings are consistent with previous studies in the literature reviewed, in that the machine learning problem is unique and contextual.

Through the preliminary exploration of the dataset used, the main road accident factors in Rwanda are over-speeding and negligence by road users. Furthermore, the City of Kigali accounts for the majority of the road traffic accidents registered by Rwanda National Police for six years in the dataset.

This research study contributes to the existing knowledge, by applying machine learning models to traffic accidents prediction. From the practical point of view, this is the first study of applying machine learning methodical methods to predicting road accidents in Rwanda. This is another step forward apart from the traditional statistical point of view.

Our research experiments showed promising results since multiple linear regression model predictions match the actual road accidents registered. Therefore, I can finally say both models are comparable when used in the prediction of road traffic accidents on the monthly aggregated dataset.

5.2 Recommendations

I have identified that aggregating and reporting traffic accidents data monthly affect the quality of model prediction. Some important attributes have been missing in the aggregated reports. Therefore, it is recommended to register and report road traffic accidents in the database on daily basis.

In addition, I have identified that road features reported in the available data set have little impact on the linear model. It is recommended that reliable information generated by vehicle enabled GPS could improve the model prediction results.

There is a need for research on building an automatic system that could instantly download weather data for providing the observational data in prediction.

References

Alexander, D. L. (2015). 3. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of chemical information and modelling, 55(7)*, pp.1316-1322.

Bangalore, S. R. (June 2000). Evaluation metrics for a generation. *Proceedings of the First International Conference on Natural Language Generation*, (pp. pp. 1-8).

Biau, G. (2012). Analysis of a random forests model. ,. *The Journal of Machine Learning Research, 13(1)*, pp. 1063-1095.

Breiman, L. (2001). Random forests. ,. *Machine learning, 45(1)*, pp.5-32.

Cawley, G. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research, 11*, pp.2079-2107.

Chicco, D. a. (2020). The advantages of the Matthews correlation coefficient (MCC) over the F1 score and accuracy in binary classification evaluation. *BMC genomics, 21(1)*, 6.

Christopher, M. B. (2006). *Pattern recognition and Machine learning.* New York, NY 10013, USA: Springer Science+Business Media, LLC.

Davina, J. F., John, M. W., & James, E. (1993, July). Decision-Making style, Driving style, and self-reported Involvement in Road Traffic Accidents. *Ergonomics*, pp. 2-19.

Deng, X. L. (2016). 1. An improved method to construct basic probability assignments based on the confusion matrix for classification problems. *Information Sciences, 340*, pp. 250-261.

Dogru, N. a. (February 2018). Traffic accident detection using random forest classifier. *In 2018 15th learning and technology conference (L&T)*, (pp. (pp. 40-45)).

Domingos, P. (June 2000). A unified bias-variance decomposition. *In Proceedings of 17th International Conference on Machine Learning* (pp. pp. 231-238). Seattle, WA 98185-2350, U.S.A.: University of Washington.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics, 7(2)*, 137-152.

Harrington, P. (2012). *Machine learning in action. Manning.* Publications Co.

Huang, H. X. (2015). Maximum F1-score discriminative training criterion for automatic mispronunciation detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(4)*, pp.787-797.

IEEE. (2017, March 15). Confusion-Matrix-Based Kernel logistic Regression for imbalanced classification. *IEEE transactions on knowledge and Data Engineering*, pp. 1806-1819.

Jianfeng, X., Zhonghao, Z., Wei, L., & Quan, W. (2016). A Traffic Accident Causation Analysis Method Based on AHP Apriori. *Procedia Engineering 137 (2016) 680-687* (pp. 1-8). Beijing: Elsevier Ltd.

John Winn, C. M. (2013). *Model-based machine learning.* Washington: Phil Trans R Soc.

Kong, E. B. (1995). Error-correcting output coding corrects bias and variance. . *In Machine Learning Proceedings* (pp. pp. 313-321). Morgan Kaufmann.

Mannor, S. S. (2007). Bias and variance approximation in value function estimates. *Management Science, 53(2)*, 308-322.

Martins, M. V. (June 2016). *Mastering Python Data Analysis.* Birmingham: Packt Publishing.

Miao, C., Ajith, A., & Marcin, P. (December 2004). Traffic Accident Analysis Using Machine Learning Paradigms. *Informatica 29 (2005) 89-98*, 1-10.

Mitchell., T. M. (1997). *Machine Learning, 1 edition.* New York, NY: McGraw-Hill, Inc.,

Morstatter, F. W. (August, 2016). A new approach to bot detection: striking the balance between precision and recall. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM)* (pp. (pp. 533-540)). IEEE.

Patel, A., Krebs, E., Andrade, L., Stephen, R., João Ricardo, N. V., & Catherine, A. S. (2016, August 02). The epidemiology of road traffic injury hotspots in Kigali, Rwanda from police data. *BMC Public Health*. doi:https://doi.org/10.1186/s12889-016-3359-4

Quanjun, C., Xuan, S., Harutoshi, Y., & Ryosuke, S. (2016). Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference. *Proceedings of* *the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* (pp. 1-7). Califonia, USA: Association for the Advancement of Artificial Intelligence.

Rwanda National Police. (2019, November 05). *Technology: A Sustainable Solution to Road Safety*. Retrieved from RNP: http://www.police.gov.rw

Sajjadi, M. S. (2018). Assessing generative models via precision and recall. *In Advances in Neural Information Processing Systems*, pp. 5228-5237.

Shanti, S., & Geetha Ramani, R. (October 24-26, 2012). Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques. *Proceedings of the World Congress on Engineering and Computer Science 2012* (pp. 1-6). Vol. I: WCECS, San Francisco, USA.

Stoop, J. A. (1995). Accidents - In-Depth Analysis; towards a method AIDA? *Safety Science 19*, 125-136.

Tesema, T. B., Abraham, A., & Grosan, C. (2005). Rule Mining and Classification of Road Accidents Using Regression Trees. *International Journal of Simulation Systems, Science & Technology, Vol. 6, no. 10-11*, pp. 80-94.

Tibebe, B., & Shawndra, H. (March 2010). Mining Road Traffic Accidents Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. *In Proceedings of AAAI Artificial Intelligence Development (AI-D'10).* Califonia, USA: Stanford.

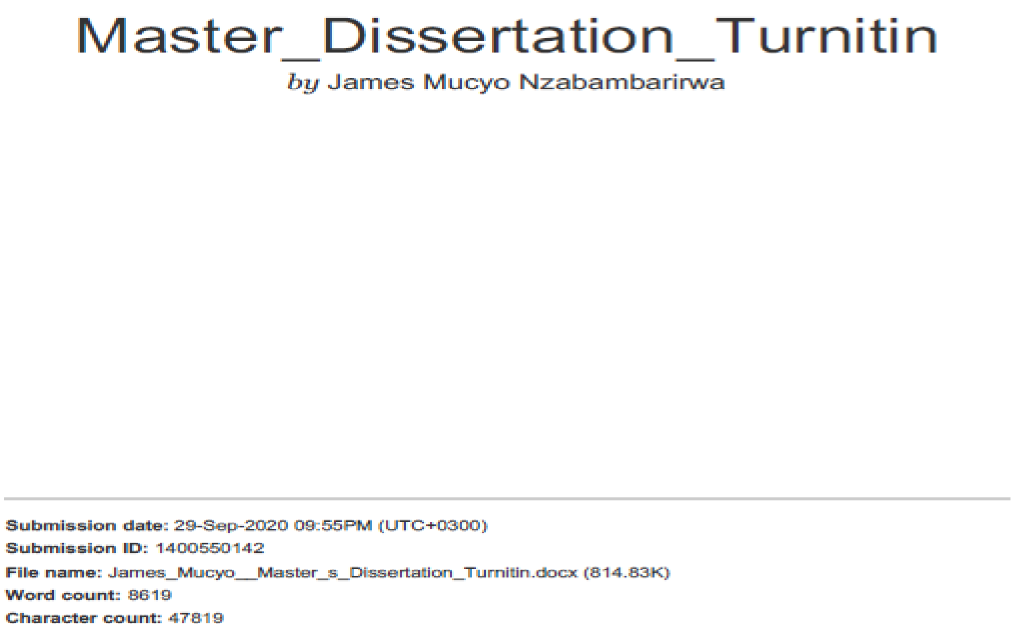
Van der Aalst, W. M. (2010). Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling, 9(1)*, pp.87.

Wahlberg, A. E., Dorn, L., & Kline, T. (2011, January 1). The Manchester Driver Behaviour Questionnaire as a Predictor of Road Traffic Accidents. *Theoretical Issues in Ergonomics Science, Volume 12*, pp. 66-86.

Wang, X. J. (2017). Generalized R-squared for detecting dependence. *Biometrika, 104(1)*, pp.129-139.

World Health Organisation. (2018). *Global Status Report on Road Safety.* Geneva: W.H.O.

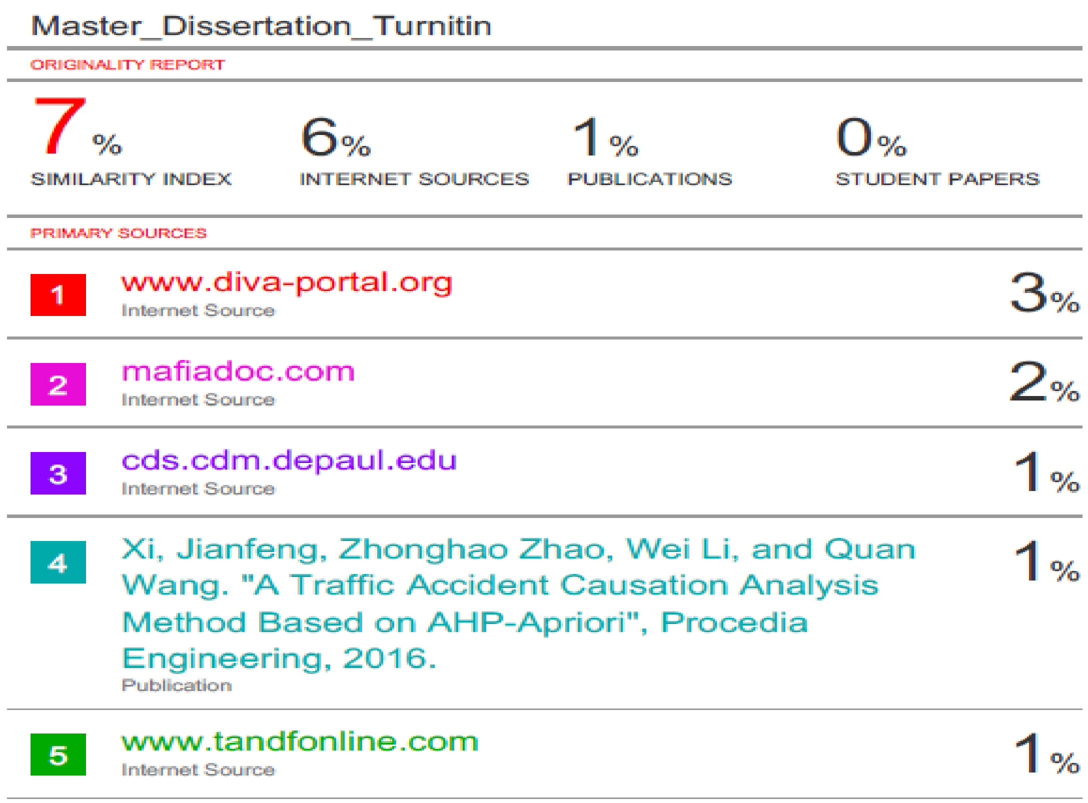
Plagiarism Checking for a level of originality



Master_Dissertation_Turnitin

by James Mucyo Nzabambarirwa

Submission date: 29-Sep-2020 09:55PM (UTC+0300)
Submission ID: 1400550142
File name: James_Mucyo__Master_s_Dissertation_Turnitin.docx (814.83K)
Word count: 8619
Character count: 47819



Master_Dissertation_Turnitin

ORIGINALITY REPORT

7% SIMILARITY INDEX
6% INTERNET SOURCES
1% PUBLICATIONS
0% STUDENT PAPERS

PRIMARY SOURCES

1 www.diva-portal.org
Internet Source
3%

2 mafiadoc.com
Internet Source
2%

3 cds.cdm.depaul.edu
Internet Source
1%

4 Xi, Jianfeng, Zhonghao Zhao, Wei Li, and Quan Wang. "A Traffic Accident Causation Analysis Method Based on AHP-Apriori", Procedia Engineering, 2016.
Publication
1%

5 www.tandfonline.com
Internet Source
1%