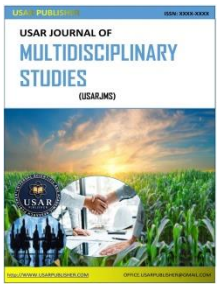
USAR journal of multidisciplinary studies (USARJMS)

OPEN ACCESS

Publish by USAR publisher
Frequency: Monthly
ISSN: XXXX-XXXX (Online)
Volume: 1-Issue-1(March) 2025
Journal home page: https://usarpublisher.com/usarjms/



Modelling and Assessing the Severity of Road Traffic Accidents in Zambia, Using Data Mining Technique

BY

[1]George Tembo,[2]Pethias Siame,

1*Department of ICT, Kwame Nkrumah University, Zambia.
2*Department of Literature and Languages, Kwame Nkrumah University, Zambia.

*Corresponding author: **Pethias Siame**

Abstract

The main aim of this study was to identify and investigate drivers, road, weather, and motor Vehicle-related factors that contribute to the severity of road traffic accidents in Zambia. This research develops a road traffic accident prediction model and compares the performance of Decision Tree (J48), Rule Induction (PART), Naive Bayes, and Random Forest algorithms to select the best-performing algorithm in the prediction of the road traffic accident severity. Raw data was collected from the Zambia Police Service Headquarters Traffic Department to construct the data set required for analysis using WEKA software. The efficiency of the algorithms used in this research was evaluated by comparing the classification accuracy, the Receiver Operating Characteristics curve, and the results shown in the confusion matrix. The results reveal that the J48 algorithm performed better than the other three algorithms. The rules produced by the PART algorithm show that year, province, tire condition, car braking condition, driver's age, driver's license grade, time, and lighting condition are the most important features in the classification of a road traffic accident severity.

Keywords: Road Traffic Accidents Severity, Data Mining, WEKA, Prediction, Modelling,

Introduction

A road accident refers to a collision involving one or more vehicles on the road, or a pedestrian and results in death, injury, or damage to property (Matter et al., 2020). Road traffic injuries place a heavy burden on global and national economies and household finances. With more than 13 million deaths and 20 – 50 million injuries being directly linked to road traffic accidents in the world, the social and economic burden presents a compromising scenario for Zambia as a nation. Road Traffic Accidents follow HIV/AIDS and malaria as the leading causes of death in Zambia (WHO, 2020). This largely affects the economically productive population in the country. In addition, family members are plunged into poverty because of the loss of their usual breadwinner through death or the high costs incurred in medical costs (WHO, 2020). Therefore, relevant authorities in the transportation sector need to make an effort and enact policies or measures that would significantly reduce the impacts of road accidents including fatalities, disabilities, morbidity, and the related costs of medical expenses associated with preventable road accidents.

The Road Transport and Safety Agency (RTSA) was established through an act of parliament under the Road Traffic Act

number 11 of 2002 under the Ministry of Transport and Communications. RTSA is a corporate body responsible for implementing the Policy on road transport and traffic management, Road Safety, and enforcement of road transport and safety laws in Zambia (RTSA, 2019). To ensure safety for all road users, the RTSA has the enforcement, Road Safety Engineering, Education, and Publicity units as well as the Research and Statistics units in place that take care of road user needs. Therefore, this study problematizes modelling and assessing the severity of road traffic accidents in Zambia using data mining techniques.

2. Problem definition

Traffic accidents can be managed and controlled differently for different purposes like traffic congestion, response time estimation, traffic accident duration, and level of accident severity. The number of casualties in a road traffic accident depends on the severity of the accident. A lot of research has been done on road traffic accidents in Zambia but none of the studies used Data Mining Techniques to model the prediction of the severity of a traffic accident. Therefore, there is a need to develop a model that would predict the severity of accidents in Zambia using Data Mining Techniques. The establishment of a set of rules that can be used by the Zambian Traffic Agencies to identify the main factors that contribute to accident severity was another crucial aspect of the study.

3. Related Works

This section attempted to uncover related research in Road Traffic Accident data mining that specifically highlights the prediction of the severity of a Road Traffic Accident. It covered literature related to Data mining and machine learning, Machine learning in road traffic accidents, road traffic accidents, their factors, and road traffic accidents in Zambia.

3.1. Data Mining and Machine Learning Concepts

According to Almahdi et al. (2015), data mining is defined as an approach to discovering useful information from large amounts of data. Data mining techniques apply various methods to discover and extract patterns from stored data. The extracted data patterns are later used to solve numerous problems that are found in various areas such as economics, business, medicine, education, sports, and statistics. The volumes of stored data call for the data mining method because the final analysis results are much more precise and accurate.

3.2. Naïve Bayes

Sameen and Pradhan (2017) define Naive Bayes as a classification approach that adopts the principle of class conditional independence from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result (Beshah and Hill, 2010). There are three types of Naïve Bayes classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes. This technique is primarily used in text classification, spam identification, and recommendation systems (Beshah and Hill, 2010).

3.3. Linear Regression

According to Kale and Baradan (2020), linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes. When there is only one independent variable and one dependent variable, it is known as simple linear regression. As the number of independent variables increases, it is referred to as multiple linear regression Nour et al., (2020). For each type of linear regression, it seeks to plot a line of best fit, which is calculated through the method of least squares. However, unlike other regression models, this line is straight when plotted on a graph (Nour et al, 2020)

3.4. Support Vector Machine (SVM)

According to Arhin, and Gatiba (2020), a Support vector machine (SVM) solves binary classification problems by framing a convex optimization which involves finding the maximum margin separating the hyperplane (Hawkins, n.d.). SVM generalization to SVR is accomplished by presenting an insensitive area around the function, named the ε-tube (Hawkins, n.d.). This tube redevelops the optimization problem to find the tube that best approximates the continuous-valued function. SVR is framed as an optimization issue by first defining a convex ε-insensitive loss function to minimize and discover the flattest tube that comprises most of the training instances (Hawkins, n.d.).

3.5. Random Forest

Random forest is another flexible supervised machine learning algorithm used for both classification and regression purposes. According to De la, et al., (2020), the "forest" references a collection of uncorrelated decision trees, which are then merged to reduce variance and create more accurate data predictions. The final results of this algorithm produce two outcomes, if the dependent variable is numerical, then the final result will be based on the average of all results, however, when categorical it relies on the popular vote of individually developed trees in a particular forest De la, et al., (2020). Random forests have proved to be

more efficient and inherent models when applied in solving regression and classification problems.

3.6. Decision Trees

According to Yannis, *et al.* (2017), decision trees are presented like a tree-like flow chart, whose internal node is represented by the rectangles, and leaf nodes are represented by ovals. The internal nodes may have two or more child nodes and have splits that do the tests for the value of an expression of the attributes. The arcs from an internal node to its children are labelled with specific outcomes of the test and every leaf node has a class label associated with it (Yannis, *et al.* 2017).

3.7. Road Traffic Accident and Its Factors

Many factors contribute to road traffic accidents. These factors can be classified into four categories such as; human error, motor vehicle defects, road defects, and weather conditions (Ramya, *et al.* 2019).

Human error accidents are defined as types of accidents that are caused by humans. These accident factors are further divided into five categories, and these are; driver error, passenger error, pedestrian error, collision with animals, and accidents resulting from obstructions.

Among the driver error factors are, misjudging clearance distance or speed excessive speeding and failure to keep to the near side. According to Arhin and Gatiba (2020), excessive speeding is driving beyond the permissible speed limit on a particular section of the road with a prescribed speed limit. Passenger error factors include; Passengers falling from moving motor vehicles and Negligence on the part of a conductor.

Among the pedestrian error factors are, pedestrians crossing the road, pedestrians walking along the road, standing and playing on or near the road, pedestrians crossing the road while under the influence of alcohol or drugs, and falling ill suddenly.

The other human error factor is Motorists colliding with animals. According to RTSA (2020), Animals wandering about on the road contribute significantly to the number of road traffic crashes. This mostly occurs when the driver is trying to avoid colliding with an animal and ends up losing control of the motor vehicle or colliding head-on with an animal.

According to Kanchele (2016), the conditions of motor vehicles play a critical role in the safety of passengers, drivers, and other road users. These factors include tires, Malfunctioning, failure or binding of brakes, unattended to motor vehicle, vehicle overloaded, steering wheel, springs, defective lights, and smashed windscreen.

The United Nations under its five pillars on road safety proclaims safer roads and mobility for all road users. Road infrastructure has been identified as a key road safety feature. These factors include; road surfaces in need of repair, overgrown vegetation such that road signs are not visible to motorists, unclear road markings or no road markings at all, dust, and obscured view. While Weather conditions include, heavy rain, fog, and wind.

3.8. Machine Learning in Road Traffic Accidents

To examine the association between Road Traffic Accident Severity and driving atmosphere issues, Nour et al., (2020) used several algorithms to develop the correctness of individual classifiers for two Road Traffic Accident Severity groups. Employing a neural network and decision tree separate classifiers, three diverse methods were useful: classifier fusion founded on the Dempster–Shafer procedure, the Bayesian process, and logistic model; information collective fusion based on arcing and bagging; and clustering based on the k-means algorithm. Their experiential consequences designated that a clustering-based classification algorithm works best for road traffic accident classification in Korea.

Lavanya and Divya (2017) used a mixture of cluster analysis, regression analysis, and topographical data organization methods to cluster the same accident information estimate the number of traffic accidents, and evaluate RTA risk in Hong Kong. Their resultant algorithm showed better chance risk estimation equated to estimates built on historical accident records only. The algorithm was more capable, particularly for casualty and pedestrian-related accident scrutinizes. The writers demanded that the planned algorithm might be recycled to help authorities successfully recognize parts with high accident danger, and help as a reference for town planners bearing in mind road protection.

3.9. Road Transport in Zambia

Road Transport plays a vital role in all economic activities in Zambia, contributing to economic growth via quicker mobility of goods, services, and people. Road transport so far accounts for 90 % of all local transportation in Zambia and is without doubt critical to the development of the transport sector and ultimately the general economy (RTSA Annual Report, 2016). Investment in safer vehicles, safer road users, and safer better-conditioned roads is optimally critical for economic development in Zambia.

3.10. Road Traffic Accidents in Zambia

Road traffic crashes, injuries, and fatalities have of late become a global public health and development problem, especially within low- and middle-income countries and Zambia is no exception. Ninety percent of the world's road

traffic deaths occur in low-and middle-income level countries (RTSA Annual Report, 2020). Road traffic crashes and fatalities are disproportionately distributed across population groups. Many of those most affected belong to the most vulnerable populations in society such as pedestrians, cyclists, unsecured passengers, and children below the age of 16 years.

This study is one of the first studies, which are around the analysis of Traffic Accidents in Zambia using data mining methods. According to the reviewed literature, none of the studies used data mining techniques to predict traffic accident severity in Zambia so far. This study will therefore attempt to develop a model to predict the severity of a road traffic accident in Zambia using road traffic accident factors.

4. Research Methods

In this chapter, the different steps taken to develop a road traffic accident prediction model are addressed. The following questions are answered throughout this paper.

1. What are the essential traffic accident features to predict the severity of Road Traffic Accidents?
2. How can data mining algorithms be applied as a predictive tool for determining the severity of a Road Traffic Accident?
3. Which data mining algorithm produces accurate predictions of the severity of a road traffic accident?
4. What are the most interesting patterns or rules generated using the determinant factors of drivers, weather, motor vehicles, and roads that can be used as traffic rules and policies in Zambia?

To answer these questions, this study adopted the CRISP-DM standard data mining methodology. The significant iterative events that take place in this research are business understanding, data understanding, data preprocessing, selection of modelling techniques, model building, and model evaluation. Fig 4 below shows the processes involved in the CRISP-DM methodology.

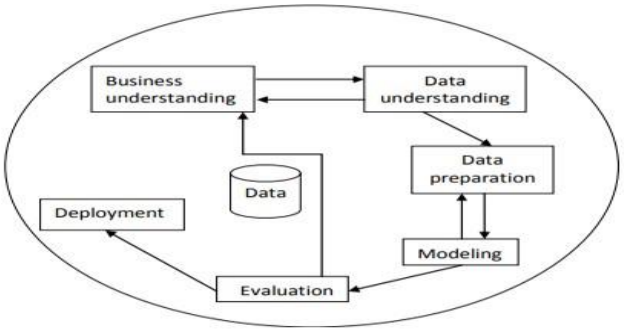


Figure 1: Phases of the CRISP-DM reference model (adapted from the CRISP-DM consortium, 2000)

4.1. Business Understanding

Zambia Police Service Traffic Department and RTSA headquarters are two offices located in Lusaka the capital city of Zambia. They are responsible for the prevention and reduction of traffic accidents in the country. To perform this responsibility the offices have staffs that include traffic police officers and machinery such as Motor Bicycles and different Automobiles. In each provincial head office, road traffic accident data is reported manually. The traffic accident data form contains accident classification, year, time, age of the driver, sex, defects of the vehicle, road surface condition, weather and illumination conditions, driver's license grade, and casualty. Upon occurrence of an accident, the traffic officer during the accident fills in the form and reports to the provincial head office for both RTSA and Zambia Police Service. Every month these accident reports are collected and reported to Zambia Police Service Headquarters and RTSA headquarters respectively. The RTSA statistics and research section generates annual statistics about how many deaths, slight injuries, and serious injuries occurred in road traffic accidents.

4.2. Data Collection and Understanding

This study used secondary data collected from the Zambia Police Traffic Department at Force Headquarters in Lusaka. They are the custodians of all information on reported road transport accidents. The data covered all the ten provinces of Zambia. Data for the period 2016 – 2020 was collected as it was fairly well documented than before. This file contained statistics about 159,698 road accidents that occurred in five years from 2016 to 2020. The data collected is in Excel format. Each accident is described using 23 pieces of information (i.e., attributes) that were recorded at the time. These attributes were classified into four categories: driver-related attributes, road-related attributes, weather-related attributes, and motor vehicle-related attributes. The accidents were classified into three levels based on severity: fatal, seriously injured, and slightly injured.

4.3. Data Cleaning

Real-world databases usually contain incomplete, noisy, and inconsistent data and such impure data may be the source of confusion in the data mining process (Han and Kamber, 2006). Therefore, data cleaning has to turn out to be a must in demand to improve the quality of data to improve the accuracy and efficiency of the data mining techniques. For this research data, a thorough discussion was made with the people in charge of keeping records both at Zambia Police Service headquarters and RTSA and it was discovered that missing attribute values at the time of data entry were recorded as "unknown" and for those records the attribute is inappropriate they simply left it as a blank supposing that it would be obvious. To repair these problems 1130 records with missing or unknown values for a significant number of attributes were removed from the dataset which makes the final number of instances to be 6466. Noisy values for attributes were also deleted and set to blank.

4.4. Data or Attribute Selection

To obtain the most essential features, the researchers selected the most important features manually. Out of 23 attributes, 16 attributes including the class attribute were selected. Table 3.1 below shows selected attributes.

To answer the second research question of the study, the researchers described the most important features in the prediction of the severity of a road traffic accident in Zambia in a tabular form. The researchers selected the attributes after reviewing the literature and making consultations with domain experts. The selected attributes for this study are; year, province, sex of the driver, casualty, cause of the accident, accident severity, month driver's age, driver's license grade, time, weather condition, road surface condition, lighting condition, road defects, car braking condition and tire condition.

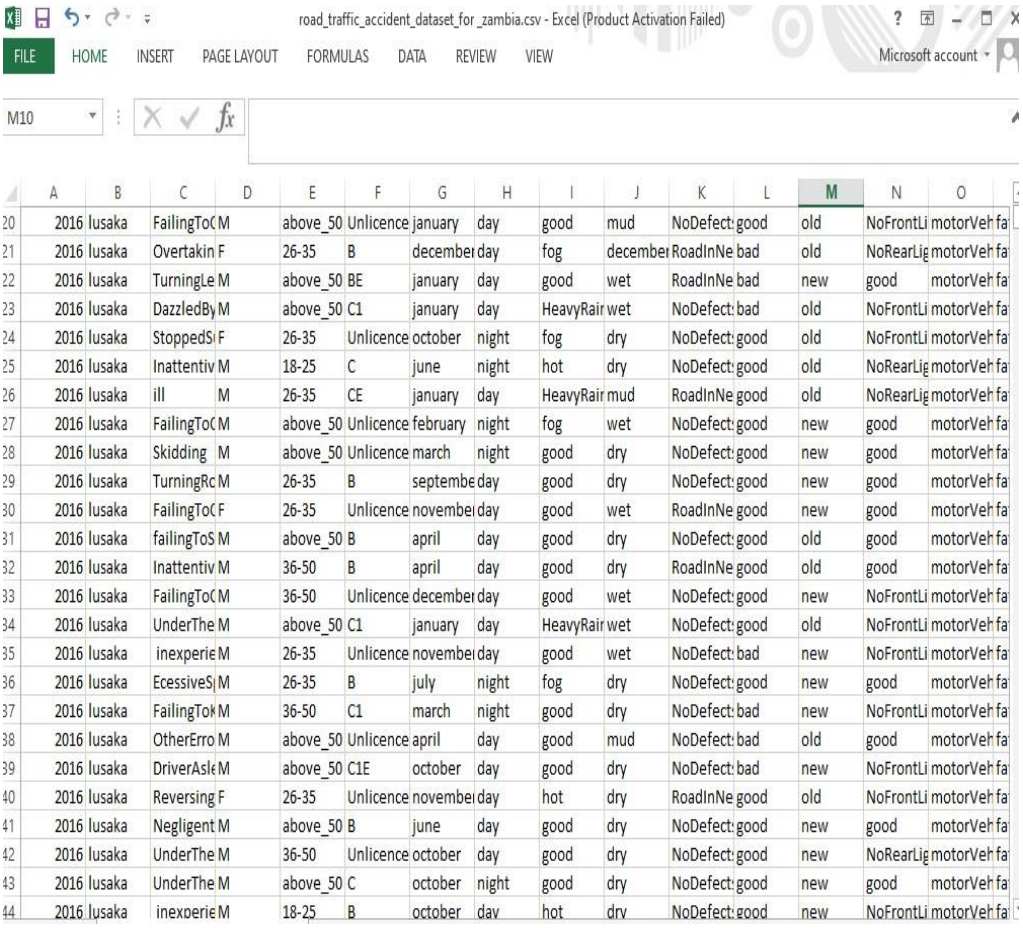
4.5. Data Transformation

Data transformation aims to manipulate the data so that the content and the format are most suitable for the data mining procedure. Therefore, based on the Zambia Police Service classification of drivers' sex, 'DriverSex' resultant from the base attribute drivers' sex to categorize the input values as 'M' for male and 'F' for female. "DriverAge" also results from the driver's age feature to classify the input values as 18-25, 26-35, 36-50, and above 50.

4.6. Data Set Format

The following steps were taken to convert the road traffic accident dataset for Zambia from Excel format to Attribute-Relation File Format (ARFF) which is the standard format for WEKA.

- In a case where a data set is in Excel format, there is a need to change the data into Comma Separated values (.CSV) format. Figure 2 below shows the road traffic accident dataset in CSV format.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
20	2016	lusaka	FailingTo(	M	above_50	Unlicence	january	day	good	mud	NoDefect:	good	old	NoFrontLi	motorVeh	fa
21	2016	lusaka	Overtakin	F	26-35	B	decembe	day	fog	decembe	RoadInNe	bad	old	NoRearLig	motorVeh	fa
22	2016	lusaka	TurningLe	M	above_50	BE	january	day	good	wet	RoadInNe	bad	new	good	motorVeh	fa
23	2016	lusaka	DazzledBy	M	above_50	C1	january	day	HeavyRair	wet	NoDefect:	bad	old	NoFrontLi	motorVeh	fa
24	2016	lusaka	StoppedS	F	26-35	Unlicence	october	night	fog	dry	NoDefect:	good	old	NoFrontLi	motorVeh	fa
25	2016	lusaka	Inattentiv	M	18-25	C	june	night	hot	dry	NoDefect:	good	old	NoRearLig	motorVeh	fa
26	2016	lusaka	ill	M	26-35	CE	january	day	HeavyRair	mud	RoadInNe	good	old	NoRearLig	motorVeh	fa
27	2016	lusaka	FailingTo(	M	above_50	Unlicence	february	night	fog	wet	NoDefect:	good	new	good	motorVeh	fa
28	2016	lusaka	Skidding	M	above_50	Unlicence	march	night	good	dry	NoDefect:	good	new	good	motorVeh	fa
29	2016	lusaka	TurningRc	M	26-35	B	septembe	day	good	dry	NoDefect:	good	new	good	motorVeh	fa
30	2016	lusaka	FailingTo(	F	26-35	Unlicence	novembe	day	good	wet	RoadInNe	good	new	good	motorVeh	fa
31	2016	lusaka	failingToS	M	above_50	B	april	day	good	dry	NoDefect:	good	old	good	motorVeh	fa
32	2016	lusaka	Inattentiv	M	36-50	B	april	day	good	dry	RoadInNe	good	old	good	motorVeh	fa
33	2016	lusaka	FailingTo(	M	36-50	Unlicence	decembe	day	good	wet	NoDefect:	good	new	NoFrontLi	motorVeh	fa
34	2016	lusaka	UnderThe	M	above_50	C1	january	day	HeavyRair	wet	NoDefect:	good	old	NoFrontLi	motorVeh	fa
35	2016	lusaka	inexperie	M	26-35	Unlicence	novembe	day	good	wet	NoDefect:	bad	new	NoFrontLi	motorVeh	fa
36	2016	lusaka	EcessiveS	M	26-35	B	july	night	fog	dry	NoDefect:	good	new	good	motorVeh	fa
37	2016	lusaka	FailingTo	M	36-50	C1	march	night	good	dry	NoDefect:	bad	new	NoFrontLi	motorVeh	fa
38	2016	lusaka	OtherErro	M	above_50	Unlicence	april	day	good	mud	NoDefect:	bad	old	good	motorVeh	fa
39	2016	lusaka	DriverAsle	M	above_50	C1E	october	day	good	dry	NoDefect:	bad	new	NoFrontLi	motorVeh	fa
40	2016	lusaka	Reversing	F	26-35	Unlicence	novembe	day	hot	dry	RoadInNe	good	old	NoFrontLi	motorVeh	fa
41	2016	lusaka	Negligent	M	above_50	B	june	day	good	dry	NoDefect:	good	new	good	motorVeh	fa
42	2016	lusaka	UnderThe	M	36-50	Unlicence	october	day	good	dry	NoDefect:	good	new	NoRearLig	motorVeh	fa
43	2016	lusaka	UnderThe	M	above_50	C	october	night	good	dry	NoDefect:	good	new	good	motorVeh	fa
44	2016	lusaka	inexperie	M	18-25	B	october	day	hot	dry	NoDefect:	good	new	NoFrontLi	motorVeh	fa

Figure 2: Road traffic accident dataset in. CSV format

- The CSV file is then converted to. ARFF format is a format accepted in WEKA data mining software.
- To convert data to. ARFF format, then. CSV file is opened in a text editor, for this research 'notepad' was used.
- The next step is to add header relation e.g. @Relation road Traffic accident.
- After that add the file with headers equal to the number of instances in your Excel file e.g. @attribute drivers Age {18-25, 26-35, 36-50, above-50} @attribute drivers' sex {m, f}. This presents a file with two columns living put the class label.
- Add the class label relation e.g. @attribute CLASS {fatal, slightly injured, seriously injured} this has three classes.
- Lastly appending the header with @data and then saving the file as.ARFF.

Figure 3 below shows an ARFF file for the road traffic accident data set for Zambia. The data set consists of 6466 instances and 16 attributes including the class attribute.

Figure 3: An extraction of the resulting ARFF file for road traffic accidents dataset for Zambia.

5. Model Development and Evaluation

This research used a total number of 6466 road traffic accident records in Zambia that occurred over five years, from 2016 to 2020. The study used fifteen independent variables with accident severity as a class variable to build the models to predict the severity of a road traffic accident in Zambia. The research also aimed to determine the most important variables that contribute to road traffic accident severity. WEKA data mining software was used to build the prediction models (classifiers).

To achieve this four well-known classification algorithms were applied to the road traffic accident dataset. These classification algorithms are; Decision Tree J48, Naïve Bayes, Random Forest, and Rule Induction PART. Several parameter settings were evaluated. To choose the parameter that produced the best-performing classification, model. To evaluate the performance of each algorithm two methods were employed. Initially, the whole dataset was used as a training set. Then the accuracy was evaluated using cross-validation with 10-folds. The rules that were generated by the Decision Tree J48 algorithm and Rule induction PART were used to establish the main contributing factors to road traffic accident severity. Figure 4 shows the output of the preprocessor in WEKA with all the 16 attributes loaded.

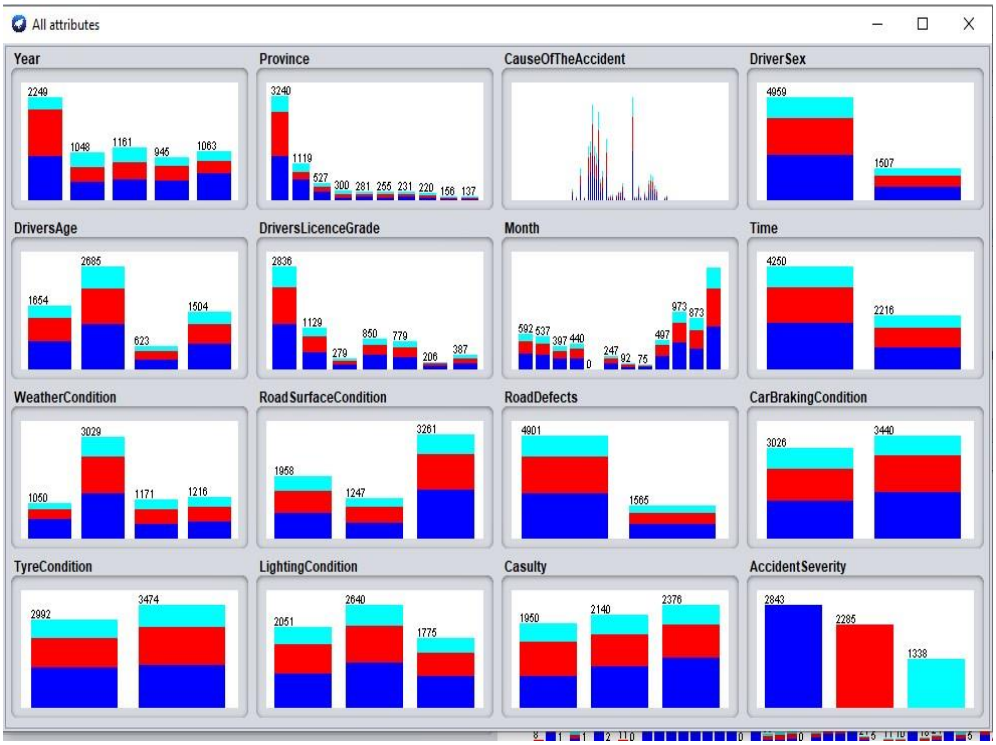


Figure 4: Output of the Preprocessor Visualization of all attributes

5.1. Decision Tree (J48) Prediction Results

The table below shows the performance of the Decision Tree classification algorithm in terms of prediction accuracy for both the training set and 10-fold cross-validation evaluation approaches

algorithm	sample	Accident severity	Correctly classified instances	Incorrectly classified instances	Accuracy (recall)	ROC area	Time (sec)
J48	Using training set	Fatal	2380	463	83.71%	0.875	0.02
		Seriously injured	1671	614	73.13%	0.881	
		Slightly injured	696	642	52.02%	0.885	
		overall	4747	1719	73.4148 %	0.879	
	Cross-validation (10-fold)	Fatal	1913	930	67.3%	0.661	0.07
		Seriously injured	1235	1150	54.0%	0.661	
		Slightly injured	299	1039	22.3%	0.617	
		overall	3447	3019	53.3096 %	0.652	

The table shows that 3447 out of 6466 road traffic accidents were correctly classified giving a classification accuracy percentage of 73.4148 %. The confusion matrix also shows that 2380 out of 2843 fatal accidents were correctly classified, 1671 out of 2285 seriously injured instances were correctly classified and the number of instances that were correctly classified as slightly injured is 696 out of 1338 instances. The building of the model took 0.02 seconds.

The model that was built by using the J48 algorithm based on the 10-fold cross-validation classified 3447 out of 6466 records of the dataset correctly and the model had an accuracy of 53.3096%. The confusion matrix also shows that the total number of 1913 out of 2843 fatal accident instances were correctly classified, 1235 out of 2285 were correctly classified as seriously injured and 299 out of 1338 instances were correctly classified as slightly injured. The time taken to build the model was 0.07 seconds.

5.2. Naïve Bayes Prediction Results

The table below presents the prediction results for the Naïve Bayes classifier for both training set and 10-fold cross-validation.

Table 2: Naïve Bayes prediction results.

Algorithm	Sample	Accident severity	Correctly classified instances	Incorrectly classified instances	Accuracy (recall)	ROC area	Time (s)
Naïve Bayes	Using training set	Fatal	1990	853	70%	0.669	0.26
		Seriously injured	1086	1199	47.5%	0.677	
		Slightly injured	263	1075	19.7%	0.680	
		overall	3339	3127	51.6393 %	0.674	
	Cross-validation (10-fold)	Fatal	1928	915	67.8%	0.639	0
		Seriously injured	1026	1259	44.9%	0.648	
		Slightly injured	219	1119	16.4%	0.651	
		overall	3173	3293	49.1%	0.645	

Using the training set, the Naïve Bayes prediction model classified 3339 instances correctly out of 6466 road traffic accident instances. The confusion matrix also shows that 1990 out of 2843 instances were correctly classified as fatal accidents, 1086 out of 2285 instances as seriously injured and 263 out of 1338 instances were correctly classified as slightly injured. The model had an accuracy of 51.6393 %. The model was built in 0.26 seconds.

When 10-fold cross-validation was applied to the same dataset using the Naïve Bayes classification algorithm, out of the 6466 records of the dataset, 3173 records were correctly classified and the model shows an accuracy of 49.0721%. The confusion matrix also shows that 1928 out of 2843 fatal accidents were correctly classified, 1026 out of 2285 seriously injured were correctly classified and 219 out of 1338 slightly injured were correctly classified. The model was built in 0 seconds.

5.3. Random Forest Prediction Results

Table 3: Random Forest prediction results

Algorithm	Sample	Accident severity	Correctly classified instances	Incorrectly classified instances	Accuracy (recall)	ROC area	Time (s)
Random Forest	Using training set	Fatal	2843	0	100%	1.000	2.38
		Seriously injured	2285	0	100%	1.000	
		Slightly injured	1338	0	100%	1.000	
		overall	6466	0	100%	1.000	
	Cross-validation (10-fold)	Fatal	2053	790	72.2	0.685	4.49
		Seriously injured	1135	1150	49.7	0.678	
		Slightly injured	153	1185	11.4	0.653	
		overall	3341	3125	51.6703 %	0.676	

The prediction model developed using the Random Forest algorithm using a training set produced an accuracy of 100%. This implies that all the 6466 road traffic accident instances were correctly classified as fatal, seriously injured and slightly injured. The time taken to build the model is 2.38 seconds.

Random forest classification using 10-fold cross-validation produced 3341 correctly classified instances out of the 6466

records of the dataset thus giving it 51.6703% accuracy. The confusion matrix also shows that 2053 out of 2843 fatal accidents were correctly classified, 1135 out of 2285 accidents were correctly classified as seriously injured and 153 out of 1338 accidents were correctly classified as slightly injured. The time taken to build the model was 4.49 seconds.

5.4. Rule Induction PART Prediction Results.

The table below shows the prediction results produced by the Rule induction PART algorithm for both the training set and 10-fold cross-validation.

Table 4: Prediction Results for Rule Induction PART

Algorithm	Sample	Accident severity	Correctly classified instances	Incorrectly classified instances	Accuracy (recall)	ROC area	Time (s)
PART	Using training set	Fatal	2381	462	83.7%	0.914	1.56
		Seriously injured	1773	512	77.6%	0.906	
		Slightly injured	835	503	62.4%	0.919	
		overall	4989	1477	77.1574 %	0.9112	
	Cross-validation (10-fold)	Fatal	1684	1159	59.2%	0.636	2.11
		Seriously injured	1089	1196	47.6%	0.621	
		Slightly injured	304	1034	24.9%	0.566	
		overall	3077	3389	47.5874 %	0.616	

The results obtained from the PART prediction algorithm using the training set indicate that; the model classified 4989 out of 6466 road traffic accident instances correctly giving the model an overall accuracy of 77.1574 %. The confusion matrix also shows that 1773 out of 2843 fatal accidents were correctly classified, the model further classified correctly 835 out of 2285 seriously injured accidents and 835 out of 1338 slightly injured accidents were correctly classified. The time taken to build this model is 1.56 seconds.

The results from PART algorithm classification using 10-fold cross-validation show that the model classified 3077 instances correctly out of 6466 total instances, giving it 47.5874% accuracy. The confusion matrix also shows that 1684 out of 2843 fatal accidents were correctly classified, 1089 out of 2285 were correctly classified as seriously injured and 304 out of 1338 slight injured were classified correctly. The time taken to build the model is 2.11 seconds.

5.5. Comparison of Classification Algorithms Result Using 10-Fold Cross Validation

Table 5: classification algorithms comparisons

Classification model (classifier)	Correctly classified instances	Incorrectly classified instances	accuracy	ROC area
J48	3447	3019	53.3096 %	0.652
Naïve Bayes	3173	3293	49.0721 %	0.645
Random Forest	3341	3125	51.6703 %	0.676
PART	3077	3389	47.5874 %	0.616

The table above shows the comparison of four classification algorithms in terms of classification accuracy for the models that were built using 10-fold cross-validation. The results show that the J48 classifier out-performed the other three classifiers, Random Forest algorithm was the second in terms of performance, followed by Naïve Bayes and then the PART classifier with an accuracy of 53.3096 %, 51.6703 %, 49.0721 %, and 47.5874 % respectively.

5.6. Evaluating Classifier Performance Using Receiver Operating Characteristics (ROC)

The study of the prediction of the used prediction methods, the Receiver Operating Characteristics (ROC) curve, also known as the relative operating characteristic curve, was studied. The area under the ROC measures the total

discriminative ability of a test. An entirely random test has an AUC of 0.5, whereas a perfect test has an AUC of 1.00. Since WEKA Explorer does not produce a multiple number of ROC curves for multiple classifiers, therefore, to achieve this we used the knowledge flow application of WEKA. The knowledge flow environment has a lot of tools. In this evaluation experiment, the researcher used the Arff loader, class assigner, class value picker, cross-validation fold maker, classifier performance evaluator, and the model performance chart together with the four algorithms to be compared. The road traffic accident dataset was loaded into the knowledge flow environment by selecting the configure option in the Arff loader tool. The cross-validation fold maker was set to 10 and the ROC curve for each class was generated.

The evaluation was performed by clicking the run button and the ROC curve was generated and visualized by selecting the show chart option in the model performance chart tool. The results of the experiment show that, in the fatal class J48, Naïve Bayes Random Forest and PART produced ROC of 0.661, 0.539, 0.685, and 0.636 respectively. In the seriously injured class J48, Naïve Bayes Random Forest and PART produced ROC of 0.66, 0.648, 0.678, and 0.621 respectively. For the slightly injured class, J48, Naïve Bayes Random Forest, and PART produced ROC of 0.617, 0.651, 0.653, and 0.566 respectively.

Random forest algorithm out-performed with a total ROC value of 0.673 followed by the J48 decision tree with 0.652 then Naïve Bayes and PART with 0.645 and 0.616 respectively.

5.7. Rule Induction PART Rule Extraction

To come up with the rules that can be used by policymakers, RTSA, and the Zambia Police Service to reduce and control traffic accidents, the researcher used the following research question.

What are the most interesting patterns or rules generated using the determinant factors of drivers, weather, motor vehicles, and roads that can be used as traffic rules and policies?

To answer this research question, the researcher interpreted the rules generated by the Rule induction PART algorithm with the help of domain experts and related literature. Table 5.6 below shows the rules generated by the PART algorithm.

Table 6: Rules generated by the PART classifier

Class attribute	No of rules	Generated Rules	Total number of instances / misclassified instances
fatal	450	Year = 2016 AND Province = CopperBelt AND TyreCondition = new AND CauseOfTheAccident = UnderTheInfluenceOfDrink/drug:	(34.0/3.0)
		Year = 2016 AND Province = lusaka AND Casulty = Pedestrian AND CarBrakingCondition = good AND CauseOfTheAccident = FailingToObeyTrafficSign/signal:	(11.0/4.0)
		Year = 2016 AND Province = lusaka AND Casulty = Pedestrian AND CarBrakingCondition = good AND CauseOfTheAccident = FailingToKeepToNearSide AND DriverSex = M AND DriversLicenceGrade = UnlicencedDriver	(10.0/1.0)
		Year = 2016 AND	(40.0/6.0)

The number of generated rules for fatal, seriously injured, and slightly injured severity accidents is 450, 320, and 138 respectively. Due to space constraints, the table above presents only the most significant rules recognized in this study.

5.8. Decision Tree J48 Pruned Tree Results

Figure 6 below is a pruned tree generated from the Decision Tree J48 algorithm.

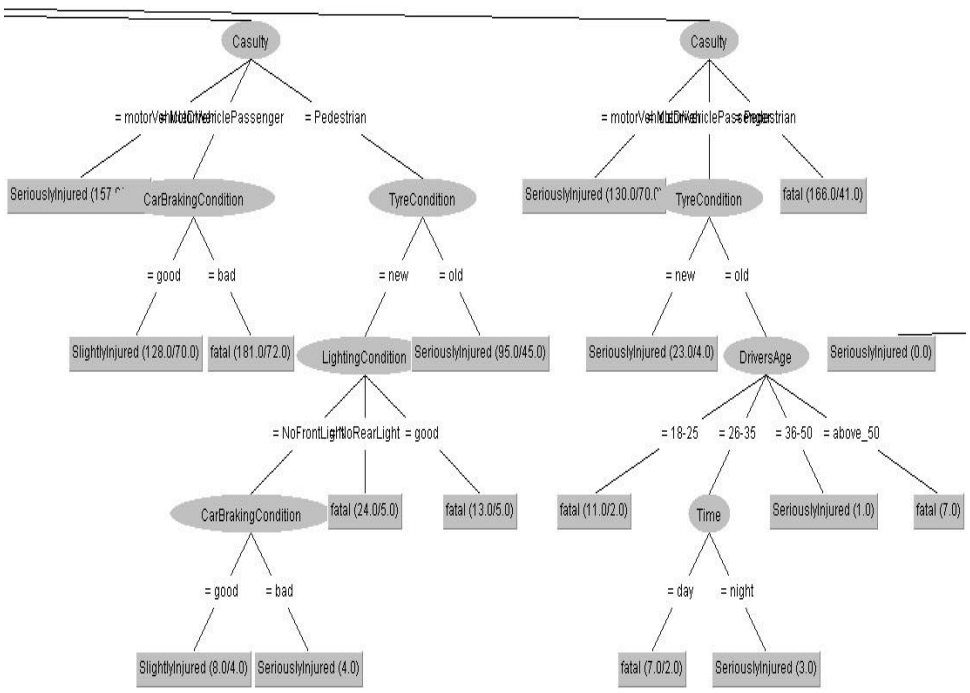


Figure 5: Decision tree J48 pruned tree.

The decision rules show that the most important factors in the prediction of road traffic accident severity are; casualty, tire condition, car braking conditions, lighting conditions, drivers' age, and time when the accident occurs.

6. Discussion of Results

All the objectives of this study were accomplished. This section therefore discusses the research findings and recommendations.

6.1. Data Collection

To achieve this goal, the Zambia Police Service Traffic Department and RTSA were engaged in serious discussion to acquire relevant knowledge on how they operate, how road traffic accidents are classified in Zambia in terms of severity, and the factors that contribute to road traffic accidents. The collected data covered all the ten (10) provinces of Zambia. Data for the period 2016 – 2020 was collected as it was fairly well documented than before. The data collected contained 159,698 road traffic accident records. The data collected is in Excel format each accident is described with 23 pieces of information (i.e., attributes). Road traffic accident attributes were classified into four categories: driver-related attributes, road-related attributes, weather-related attributes, and motor vehicle-related attributes and they categorized accident severity as fatal, seriously injured, and slightly injured.

6.2. Development of the Dataset

To achieve our first objective, which is developing a road traffic accident dataset for data mining in Zambia, the data was extracted from an Excel file and then combined to form one file. This stage involved getting only the fields that are applicable for data mining and all the derivables were also selected and extracted from the file. The data was then cleaned and transformed so that the content and the format were suitable for the data mining procedure. WEKA accepts data in Attribute Relation File Format (.ARFF).

6.3. Undertaking Feature Selection on the Road Traffic Accident Attributes

To obtain the most important features, the researcher selected the most important features manually based on a deep understanding of the learning problem and what the attribute means. Out of 23 attributes, 16 attributes including the class attribute were selected. The selected attributes include year, province, driver's sex, casualties, cause of the accident, month, driver's age, driver's license grade, time, weather condition, road surface condition, lighting condition, road defects, car braking condition, tire condition, and accident severity.

6.4. Model Development

The third objective of this study was to develop a data mining model to determine and predict the severity of a road traffic accident in Zambia. This task involved the specification of the selected machine learning algorithms, hyper-parameter settings, selection of data mining tool, and building of the models. The models were developed using two test options, first using a training set and then using 10-fold cross-

validation to find out the best-performing suitable model to be recommended in the prediction of road traffic accident severity prediction in Zambia.

After the features to be used in constructing a model were selected, classification algorithms were applied to the road traffic accident dataset for Zambia. The researcher used the J48 decision tree, naïve Bayes, Rule Induction (PART), and Random Forest algorithm in this study. Several parameter settings were evaluated. To choose the parameter that produced the best-performing classification, model.

The actual Building of the model was achieved using a data mining software called WEKA. To build the model, the road traffic accident dataset for the Zambia file which is in ARFF format was uploaded into WEKA Explorer and it was trained and tested using specified algorithms and the set parameters.

WEKA data mining tool exhibited some limitations in terms of several operations, the software couldn't handle a large dataset due to limited memory size. Nevertheless, the models were developed successfully.

6.5. Classification Accuracy

The evaluation of the performance accuracy of the classifiers was done in two ways, the first one was by checking the classification accuracy on the classifier output window in WEKA and the second one was by performing a performance comparison test for all classifiers in WEKA's experimenter. Both experiments were based on 10-fold cross-validation. The results obtained from both experiments show that the Decision Tree classifier performed better in terms of accuracy with 53.31% and Rule induction PART came out last with 47.59% classification accuracy. According to the prediction accuracy results obtained, the J48 classifier is more accurate in predicting road traffic accident severity using 10-fold cross-validation. The fact that the J48 algorithm performed better proves that it is a better algorithm to use when predicting road traffic accident severity. This is because cross-validation is highly recommended for testing models rather than using a training set.

6.6. Receiver Operating Characteristic (ROC)

The results of the models when compared in terms of the ROC curve show that random forest out- outperformed three other classifiers by producing a better ROC curve in all the three classes (i.e., fatal seriously injured, and slightly injured). The slightly injured class performed poorly in all scenarios this is because the slightly injured class was underrepresented as they represent a small percentage of total accidents. Since a random test has an ROC of 0.5, the classifiers performed well because they all produced ROC values above 0.5. This still qualifies the J48 decision tree algorithm for better performance in this test and is suitable to be used in building a road traffic accident severity prediction model.

6.7. Rule Induction (PART) Rules.

According to the first rule in the fatal class most of the fatal accidents that happened in the Copperbelt province involving cars with good braking conditions were caused by drunk drivers. Therefore there is a need to sensitize drivers on the Copperbelt on the dangers of drinking and driving.

The rules also show that Lusaka province had the highest number of fatal accidents followed by the Copperbelt province. These results support the finding by Chomba et al (2017). The year 2016 had the largest number of fatal accidents and most of them happened in November. The rules further indicate that most of the fatal accidents were caused by the driver and the casualties were mostly passengers. According to the third rule (Year = 2016 AND Province = Lusaka AND Casualty = Pedestrian AND CarBrakingCondition = good AND CauseOfTheAccident = FailingToKeepToNearSide AND DriverSex = M AND DriversLicenceGrade = UnlicencedDriver:) generated by the PART algorithm in the fatal class, most of the Road Traffic accidents that occurred in Lusaka in the year 2016 with pedestrians being casualties were caused by unlicensed male drivers failing to keep to the nearside. This could be a result of impatient driving habits exhibited by male drivers. The rules produced by the PART algorithm prove that the classifier can be used to predict and assess the severity of an accident based on the road traffic accident data provided.

The rules also show that the age group between 26 and 35 is the most vulnerable age group to road traffic accidents in Zambia. This is logical because this age group lacks experience, tends to drive fast, drinks and drives, lacks respect for other road users, and is less responsible. The year of the accident is a significant input variable because it appears in almost all the rules.

As the research shows, Lusaka and Copperbelt are provinces with a high number of seriously injured accidents and fatal accidents. So the office has to improve the roads and perhaps assign more traffic police officers in these provinces by collaborating with road transport safety agencies in Zambia.

In addition, it is observed that those who don't have driving licenses cause fatal accidents. Therefore, rules and regulations should be strict on those who drive without a driving license.

The results From the PART Rules showed that most of the accidents occurred in good weather and daylight conditions respectively, which indicates that weather and light conditions are not the main factors for traffic accidents.

6.8. Decision Tree (J48) Pruned Tree

To further find answers to our last research question, the pruned tree that was generated by decision tree J48 was explored. The pruned tree produced by the J48 Decision Tree algorithm shows that casualty, car braking condition, lighting condition, tire condition, driver's age, and time are the primary splitters in the classification tree. The tree shows that, when the car has new tires, no front light with good braking conditions, the accident severity is slightly injured. On the other hand, an accident that occurs in a vehicle under the same conditions but with bad breaking conditions has a severity of serious injury. The same results were obtained by Sameen and Pradhan (2017b) in their research, Severity prediction of traffic accidents with recurrent neural networks in Switzerland.

The decision tree rules further state that when the tire condition is old and the driver's age is between 26-35 and the time is day the accident severity is fatal while an accident that happens under the same conditions but happens during night time the severity of is seriously injured this rule is justified by RTSA (2019), that most of the motorists reduce on the speeding at night and there is less traffic during night time. The rules produced by the J48 Decision Tree can be further summarized as; in an accident that involves a pedestrian as a casualty, and the tire condition is new, the car has no front light but with good braking condition, the severity of an accident will be slightly injured. In an accident that involves a pedestrian as a casualty, and the tire condition is new, the car has no front light but with bad braking conditions the severity of an accident will be seriously injured. The tree further proves that when the car's braking condition is good, the accident severity is slightly injured. And when the car's braking condition is bad the accident severity is fatal.

7. Conclusion and Recommendations

7.1 Conclusion

This paper used the J48 Decision Tree, Naïve Bayes, Random Forest, and Rule induction (PART) to predict the severity of an accident based on 6466 records from ten (10) provinces of Zambia over five years from 2016 to 2020. For the decision Tree (J48) model, the overall prediction accuracy for the whole dataset used as a training set and with 10-fold cross-validation was, 73.4148 % and 53.3096 % respectively. For the Naïve Bayes model, the overall prediction accuracy for the whole dataset used as a training set and with 10-fold cross-validation was 51.6393% and 49.1% respectively. For the random Forest model, the overall prediction accuracy for the whole dataset used as a training set and with 10-fold cross-validation was, 100% and 51.6703 % respectively, and for Rule Induction (PART) model, the overall prediction accuracy for the whole dataset used as training set and with 10-fold cross-validation was 77.1574 % and 47.5874 % respectively.

Based on the results obtained from the performance of the four algorithms in terms of classification accuracy in the classification of road traffic accident severity, the decision tree algorithm outperformed the other three algorithms using the training set and came out second in terms of classification accuracy using 10-fold cross-validation. J48 Decision tree algorithm outperformed the other three algorithms in terms of accuracy using 10-fold cross-validation and came out third using the training set. This implies that the Random Forest algorithm is the best classifier to use when building a model for road traffic accident severity prediction using the training set while the J48 algorithm is the best to use when building a road traffic accident prediction model using 10-fold cross-validation. Therefore, the J48 Decision Tree algorithm is the most suitable classification algorithm for predicting accident severity using the traffic accident dataset for Zambia developed in this research.

The most important factors associated with the fatal severity of a road traffic accident are province, tire condition, car braking condition, driver's age, time, and lighting condition. The rules also show that most of the accidents happen in Lusaka followed by the Copperbelt province this is logical as the two provinces have a large population of vehicles. As the research shows, Lusaka and Copperbelt are provinces with a high number of seriously injured accidents and fatal accidents. So the office has to improve the roads and perhaps assign more traffic police officers to these provinces by collaborating with road transport safety agencies in Zambia. In addition, it is observed that those who don't have driving license are likely to be involved in fatal accidents. Therefore, rules and regulations should be strict on those who drive without a driving license.

The results From the PART Rules showed that most of the accidents occurred in good weather and daylight conditions respectively, which indicates that weather and light conditions are not the main factors for traffic accidents. The rules also show that the age group between 26 and 35 is the most vulnerable age group to road traffic accidents in Zambia. This is logical because this age group has a lack of experience, tends to drive fast, drinks and drives, lacks respect for other road users, and is less responsible.

7.2. Recommendations

The researcher makes the following recommendations based on the results of this study. RTSA and Zambia Police Service should take measures to store all its records with all the necessary attributes in an electronic format and to make all decisions based on collected records.

RTSA and Zambia Police Service could optimize their traffic accident prevention and control efforts by employing data mining technology.

To make accurate predictions of road traffic accident severity, RTSA, and Zambia Police Service could use the J48 Decision Tree as it produced the highest classification accuracy rate.

Policymakers should put strict measures on driving without a driver's license and should strictly follow the rules generated by decision trees and PART algorithms to prevent the occurrence of traffic accidents with a certain severity.

There is a need to train as many officers as possible in the use of data mining and machine learning techniques so that they can be able to understand the meaning of the rules and patterns produced in machine learning.

References

Arhin, S. A. and Gatiba, A. (2020b). 'Predicting Crash Injury Severity at Unsignalized Intersections using Support Vector Machines and Naïve Bayes Classifiers', *Transportation Safety and Environment*. Oxford University Press, 2(2), pp. 120–132. doi: 10.1093/tse/tdaa012.

Assi, K., Ketty P., and Ling, H. (2020). 'Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique : A Promising Protocol', (Ml).

Bahiru, T. K. (2018). 'Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity', *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, (Icicct), pp. 1655–1660.

Beshah, T. and Hill, S. (2010). *Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia*.

De la, F. A. et al., (2020). A Comparison of Machine Learning Techniques for LNG Pumps Fault Prediction in Regasification. *IFAC Papers OnLine*, 53(3), pp. 125-130.

Chomba, C., Kunda, D., Chimbola, O., and Kaliki, B. (2017). 'Incidences and Fatalities of Road Traffic Accidents in Zambia for the Period 2008 – 2013.' *A Prelude to Sustainable Road Transport Sector Development for Socioeconomic Development*, 19(1), Pp. 137–162.

ENGINEERS (2019) 'व ा र ा ष ि ा क प ा रत ि ा व ा दन Annual Report Annual Report', *Fresenius.Com*, (December), pp. 2–2.

Hawkins, J. (n.d.). *Support Vector Regression*. *Vc*, 67–80.

Kale, Ö. A. and Baradan, S. (2020) 'Identifying Factors that Contribute to Severity of Construction Injuries using Logistic Regression Model'. *Teknik Dergi/Technical Journal of Turkish Chamber of Civil Engineers*. Turkish Chamber of Civil Engineers, pp. 9919– 9940. doi: 10.18400/TEKDERG.470633.

Kanchele, C., Kanyenda, E., and Mabo, M. (2016). 'Annual Road Traffic Accident Report'. Available at: https://www.rtsa.org.zm/wp-content/uploads/2019/09/RTSAAnnual-Accident-Report-2016.pdf.

Lavanya, B. and Divya, B. (2017a). *Predictive Analytics on Accident Data Using Rule Based and Discriminative Classifiers*. Available at: http://www.ripublication.com.

Lavanya, B. and Divya, B. (2017b) 'Predictive Analytics on Accident Data Using Rule Based and Discriminative Classifiers'. *Advances in computational service and technology*, 10(3), pp. 461–469.

Matter, S. A. *et al.* (2020). 'Predicting Risky and Aggressive Driving Behavior among Taxi Drivers: Do Predicting Risky and Aggressive Driving Behavior among Taxi Drivers: Do Spatio-Temporal Attributes Matter?', (June). doi: 10.3390/ijerph17113937.

Nour, Mohamed K *et al.* (2020). *Road Traffic Accidents Injury Data Analytics*. *International Journal of Advanced Computer Science and Applications*. Available at: www.ijacsa.thesai.org.

Ramya, S. *et al.* (2019). 'Accident Severity Prediction Using Data Mining Methods'. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. Technoscience Academy, pp. 528–536. doi: 10.32628/cseit195293.

Sameen, M. I. and Pradhan, B. (2017a). 'Severity Prediction of Traffic Accidents with Recurrent Neural Networks', *Applied Sciences (Switzerland)*. MDPI AG, 7(6). doi: 10.3390/app7060476.

The CRISP-DM consortium (August, 2000). *Step-by-Step Data Mining Guide*. Available at: URL: http://www.crisp-dm.org/CRISPWP-0800.pdf. Accessed on April 28, 2011.

UNDP (2017). 'Annual Report 2017'. *NTT Docomo*, 21(5), p. 12. Available at: file:///C:/Users/green/Downloads/Yemen HC Annual report 2017.pdf.

WHO (2020). World report on road traffic injury prevention, Switzerland, Geneva.

Yannis, G. *et al.* (2017). 'Road Traffic Accident Prediction Modelling: A Literature Review'. *Proceedings of the Institution of Civil Engineers: Transport*. Thomas Telford Services Ltd, 170(5), pp. 245–254. doi: 10.1680/jtran.16.00067.

Zheng, M. *et al.* (2019). 'Traffic Accident's Severity Prediction: A deep-learning approach-based CNN network', *IEEE Access*. *Institute of Electrical and Electronics Engineers Inc*., 7, pp. 39897–39910. doi: 10.1109/ACCESS.2019.2903319.