ANALYSING ROAD TRAFFIC ACCIDENTS IN MALAWI: A DATA MINING APPROACH

MASTER OF SCIENCE IN INFORMATICS

CHIKUMBUTSO FELIX PHIRI

UNIVERSITY OF MALAWI

OCTOBER, 2022

ANALYSING ROAD TRAFFIC ACCIDENTS IN MALAWI: A DATA MINING APPROACH

MSc. (INFORMATICS) THESIS

By

Chikumbutso Felix Phiri

BSc (Computer Science) – University of Malawi

Submitted to the Department of Computer Science, Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science (Informatics)

University of Malawi

October, 2022

DECLARATION

I, the undersigned, hereby declare that the thesis is substantially my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

_____

Chikumbutso Felix Phiri

_____

Signature

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with my approval.

Signature: _____ Date: _____

Kondwani Godwin Munthali, PhD (Senior Lecturer)

Supervisor

DEDICATION

I dedicate this work to my family and friends. I owe a special debt of appreciation to my loving parents, Felix and Esther, for the words of encouragement and support. May God continue blessing you both.

ACKNOWLEDGMENT

I greatly thank my supervisor Dr. Kondwani Godwin Munthali for his support and guidance throughout this study. I wish to show my appreciation to Yamikani Phiri as well for taking up the co-supervisor role in this research.

ABSTRACT

One of the significant problems the world faces today is the rate at which road traffic accidents and deaths on the roads are happening. The majority of these accidents occur in developing countries (Ihueze & Onwurah, 2018), and Malawi is no exception. However, to supplement the current safety measures, an analysis of road traffic accidents using data mining techniques was considered. Malawi being a low-income country, it is very crucial to have focus areas when dealing with traffic safety since there are limited budgetary resources. Therefore, this study aimed at digging for patterns in the traffic accident data and modeling the severity of road accidents in Malawi. Using python, three classification algorithms were employed to model the severity of an accident. The algorithms included Decision trees, Logistic regression and Support Vector Machines. These models were evaluated using accuracy, precision, recall, and F1-score. The logistic regression performed better than the other two and it was discovered that the top three attributes that contributed to fatal accidents were accidents involving a moving vehicle and a pedestrian, accidents that occurred at Dawn or Dust, and accidents involving a moving vehicle and a bicycle. Through association rule mining, a series of interesting rules were generated. Road Condition, Weather, posted speed limit and Surface type were the frequent item sets that appeared in all the rules generated.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background to the study

One of the significant problems the world faces today is the rate at which road traffic accidents and deaths on the roads are happening. The majority of these accidents occur in developing countries (Ihueze & Onwurah, 2018) and Malawi is no exception. According to World Health Organization (WHO, 2020), approximately 1.35 million people die every year on the roads, making road accidents the eighth leading cause of death worldwide. Consequently, these fatalities and injuries lead to significant social and economic losses.

Across the world, transportation agencies have taken measures and have made significant investments to improve road safety. In Malawi for example, the Roads Authority introduced rumble strips in areas where accidents frequently occur. These are used to alert inattentive drivers through noise and vibrations. On the other hand, traffic police ensure road users are obeying the laws by issuing fines to those who do not comply. These are just some of the few noticeable measures the country is taking. To supplement the current safety measures, an analysis of road traffic accidents using data mining techniques would help significantly to avoid some of these accidents and reduce their severity.

Data mining is defined as the extraction of hidden predictive information from large databases (Krishnaveni et al., 2011). Aggarwal, (2015) also defines it as the collection, cleaning, processing, analysis, and gaining of valuable insights from data. Krishnaveni et al. (2011) observed that it is a powerful technology with great potential to help companies focus on essential details in their data warehouses. Over the years, data mining has been applied in various industries that are able to collect large amounts of data and has proven to be a success. Throughout the world, road traffic departments have been collecting data about accidents happening in their respective countries.

The data has accumulated so much that it cannot be easily analyzed using traditional methods to extract knowledge, in other words, descriptive and exploratory methods that rely on static dashboards composed of visualizations cannot suffice. This has allowed most researchers to apply data mining techniques in an attempt to reduce the number of traffic accidents or the severity of accidents. Shetty et al., (2017), Taamneh et al., (2017) and many other researchers have used different data mining techniques by analyzing data from previous accidents. Information derived from these studies has been used to come up with measures of improving traffic safety.

1.2 Problem statement

Traffic accidents may appear random, but according to Szalay (2019) they happen when certain circumstances are present in a particular place and time. Having a clear understanding of these circumstances would help with proper planning as authorities work on measures that may reduce the number or severity of these accidents. This can be achieved through analyzing data from past accidents. Road accidents data that is being stored at the Malawi road traffic directorate is usually presented in the form of annual reports where figures of a particular year are compared to that of the previous year. The reports do not articulate the relationships between the contributing factors or predict the severity of accidents. Thus, there is plenty of room to comprehensively understand road accidents, given the huge data sets available with the directorate.

1.3 Research aim and objectives

The main aim of this research is to analyze road traffic accidents in Malawi using data mining techniques. Specifically, the research seeks to:

1 Model the severity of road traffic accidents in Malawi
2 Discover patterns in the traffic accidents data

1.4 Research questions

1 Which model is best suitable for the prediction of accidents severity?
2 What are the patterns in the accidents data in Malawi?

1.5 Research Significance

Data mining techniques, in general, enable organizations to utilize their capital data and promote proper decision-making. Similarly, by analyzing accidents data using data mining techniques, the results of this research will help in decision making. The research will discover some hidden patterns in the data, identify the main attributes that contribute to accidents, and predict the severity of an accident given some attributes. Therefore, this will provide insight into issues to be addressed or policies to be implemented to deal with road safety. On the other hand, the research seeks to contribute to road traffic accidents in general, specifically the approaches used.

1.6 Structure of the research

This research is organized into five chapters. The first chapter is entailing the introduction to the study, and it contains the background, problem statement, research aim and objectives, research questions, and the significance of the research. The second chapter covers the literature review which encompasses a critical analysis of some existing works covering data mining concepts, and a brief discussion on association rule mining and classification models. The third chapter covers the methodology focusing on the approach, the data, study design, and modeling which highlighted the processes involved to get to the results. Chapter four contains the results and discussion. Finally, chapter five, contains a conclusion, recommendations, limitations, and suggestions for future research.

CHAPTER 2

LITERATURE REVIEW

Road traffic accidents contribute to significant health, economic and developmental challenges for many countries (WHO, 2020). It is the eighth leading cause of death for all age groups (World Health Organisation, 2018), to put things into perspective, road traffic accidents kill more people every year than HIV/AIDS, tuberculosis, and diarrhea diseases combined. For each person involved in an accident, numerous others are severely impacted by the cost of extensive medical care, the loss of a family provider, or the additional costs required to care for those with injuries (Gopalakrishnan, 2012). The majority of these accidents occur in developing countries (Ihueze & Onwurah, 2018). Most developing nations still do not have rules in place to safeguard vulnerable road users and to encourage the development of public transportation (Schlottmann et al., 2017). In addition to this, most developing countries post-crash care is insufficient or lacking. As a developing, Malawi is no exception. Schlottmann et al., (2017) , conducted a study on road traffic accidents in Malawi to identify trends and patterns of mortality on scene. Accident records from 2008 to 2012 were collected from the Malawian National Road Safety Council (NRSC). Over the course of these five years, one or more deaths were reported at the scene in almost one-third of RTAs, amounting to 4518 deaths on the road in this period. Pedestrians were more susceptible and had very high death rates when involved in car accidents. Motor vehicle collisions and motor vehicle versus pedestrian collisions were the most common accident types, accounting for 35% and 42% of all RTAs, respectively. In the latest road traffic accident report by the Directorate of Road Traffic and Safety Services (DRTSS), out of the 8194 RTAs that occurred in 2015 in Malawi, 888 were fatal and 706 resulted in serious injuries. It was also discovered that the majority of fatal road accidents involved motor vehicles and pedestrians.

Table 1 highlights the trend over a period of 3 years, from 2013 to 2015. Furthermore, according to the global status report on road safety, 1122 fatal accidents were recorded in Malawi in 2018.

Table 1 Road Accident Trend 2015

ACCIDENT BY SEVERITY	2015	2014	Change 2015-2014 in %	2014	2013	Change 2014-2013 in %
Fatal	888	813	9%	813	818	-0.6%
Serious injury	706	637	10%	637	622	2%
Slight/Minor Injury	2632	2407	9%	2407	2336	3%
Damages only	3944	3470	13%	3470	3580	-3%
Animals	24	28	-14%	28	34	-18%
Grand Total	8194	7355	11.4%	7355	7390	-0.4%

Source: Directorate of road traffic and safety services, road traffic accident situation, annual report 2015.

Across the world researchers have conducted several studies to reduce road traffic accidents frequency and severity (Beshah & Hill, 2010; Feng et al., 2020; Krishnaveni et al., 2011). Traffic safety studies have been conducted by applying GIS methods, and data mining techniques, just to mention a few. GIS methods are applicable where accidents' locations were recorded precisely in terms of geographical coordinates. Over the years, statistical models and data mining techniques have been used in predicting accidents severity. Since datasets are growing at an alarming rate, and traditional statistical techniques prove to have trouble coping with these enormous volumes of data (Sajjad et al., 2017). Data mining techniques handle these situations better. A critical review of the existing literature (Kumar & Toshniwal, 2016; Krishnaveni et al., 2011; Yuan et al., 2017; Feng et al., 2020) shows that data mining techniques like association rules and classification methods are one of the most commonly used methods in the mining of road traffic accidents data. Association rules can identify significant relations between the data stored in large databases and play a substantial role in the frequent itemset mining (Li et

al., 2017). On the other hand, classification methods have the ability to create classifiers/models that can predict the severity of accidents (Taamneh et al., 2017).

2.1 Data-mining Concepts

This subsection, defines and describes some of the data mining concepts used in this research.

2.1.1 Data preprocessing

Usually, when data is collected, it is not in a form fit for processing. To make models efficient and effective, the data must be well prepared and of good quality. Aggarwal (2015) emphasized the need to extract relevant features for the mining process. When the data has a lot of features relative to the number of observations, many data mining algorithms do not work effectively. Data preprocessing is a general term for several processes. It includes feature selection, data cleaning, and data integration. To make the data adequate for processing, it is important to transform it into a format that is suitable for data mining algorithms. It is also crucial to extract relevant features for the mining process, not all features have to be used. The result of this process is a structured dataset, which can be effectively used by computer programs. In some scenarios, data may also be categorical. Most models are not able to handle categorical data. For this reason, different techniques are used to transform the data. One of these methods is one-hot encoding. One-hot encoding generates a sparse matrix or dense array with a binary column for each category. This ensures that categorical variables that are not ordinal are treated as such. For instance, a district coded as 1 is not any less significant than one coded as 2.

Different algorithms can now be applied to clean and transform data. There are several data mining algorithms and the decision of which algorithm to use is solely dependent on one's goal and the data available. This chapter discusses some of the data mining methods and their algorithms.

2.1.2 Dimensionality reduction

Dimensionality reduction is the process of reducing the size of one's feature collection. Reducing the number of features in one's machine learning model might make it simpler, more efficient, and less data hungry.

Imagine planning to build a model that forecasts the amount of rain that will fall in each month. One may have a collection of data from several cities collected over several months. Temperature, humidity, city population, traffic, number of concerts hosted in the city, wind speed, wind direction, air pressure, number of bus tickets purchased, and rainfall totals are among the data points. Not all of this data is pertinent to rainfall forecasting. Some of the characteristics may or may not be related to the desired variable. Evidently, rainfall is unaffected by population, or the number of bus tickets purchased. Other characteristics may be connected to the goal variable, but they are not causal. For example, while the number of outdoor concerts may be related to rainfall, it is not a reliable forecast of rain. There may be a link between the feature and the goal variable in other circumstances, such as carbon emissions, but the influence will be minor. Machine learning models strive to map any attribute in their dataset to the target variable, even if there isn't a causal relationship, because it doesn't grasp causality. This can result in models that are inaccurate and imprecise.

2.1.3 Overfitting and Underfitting

Successful machine learning models can generalize well from training data to any data in the problem domain. This enables us to make future predictions based on data that the model has never seen before. The two most common causes of poor machine learning algorithm performance are overfitting and underfitting.

Overfitting occurs when a model learns the information and noise of the training data and degrades the model's performance on new data. This means that the model will detect noise or random fluctuations in the training data and learns them as ideas. The problem is that these notions do not apply to new data, limiting the generalization ability of the model. On the other hand, underfitting is defined as a model that cannot both model and generalize to new data. The solution to both overfitting and underfitting is usually just trying another

algorithm. However, several other techniques can be used, some of which are cross-validation and regularization. Cross-validation is a technique for assessing the model's efficiency by training it on a subset of input data and testing it on a subset of input data that has never been seen before. Whereas regularization is a type of regression in which the coefficient estimates are constrained or shrunk towards zero. Put differently this method inhibits models from learning a more complicated or flexible model.

2.2 Association rule mining

Association rule mining is a technique that uses machine learning models to analyze data for patterns. It has the capability of producing if-then rules. Association rules use support, confidence lift, and conviction to measure association, and these help us determine whether the occurrence is out of randomness or association. There are scenarios where many rules are generated, but it is crucial to stick to those that matter. For example, If $A \rightarrow B$

Support indicates how frequently the items appear in the data and is calculated as:

$$Support = \frac{frequency(A,B)}{N}$$

Confidence indicates the number of times the if-then rules are found true and is calculated as

$$Confidence = \frac{frequency(A,B)}{frequency(A)}$$

Lift is the strength of a rule. It compares the actual confidence with expected confidence. It is calculated as

$$Lift = \frac{support}{supp(A) * supp(B)}$$

Conviction measures the implication strength of the rule from statistical independence. The chance of A appearing without B if they were dependent is compared to the actual frequency of A appearing without B. it is calculated as:

$$Conviction = \frac{1 - Support(A)}{1 - Confidence(A \rightarrow B)} = \frac{P(A) * P(B)}{P(A \cup B)}$$

Apriori algorithm is a type of association rule mining that uses frequent itemsets to generate association rules. It is based on the concept that a subset of a frequent itemset must also be a frequent item set. A frequent itemset can be defined as an item set whose support value is greater than a threshold value. This ensures we only have a manageable number of rules.

2.3 Classification methods

Classification is one of the methods in data mining for categorizing a particular group of items into targeted groups. The main goal of classification is to predict the nature of items or data based on the available classes of items. Some well-known applications of data mining using classification include Credit or Loan Approval, using data-mining a system is able to detect whether a client should be given a loan or not; Spam detection, if one receives an email that is suspicious automatically it goes into the spam folder and a normal one goes directly into the mail. These are just some of the applications of data mining using classification methods. Classification makes a decision from unseen cases by building on past decisions. The data must be split into training and test. The training data is used to train the model and test data is used to test the effectiveness of the model. There are several types of classification algorithms used in data mining and decision trees, support vector machines and logistic regression are some of them.

2.3.1 Decision tree classifier

Decision trees are a classification methodology wherein the classification process is modeled using a set of hierarchical decisions on the feature variables arranged in a tree-like structure (Aggarwal, 2015). At each node, a decision is made to split the data, and this process is repeated until the data cannot be split anymore. A decision tree classifier has a high predictive performance for a relatively small computational effort. It handles multiple input data, nominal, numeric, and text, and can process data sets that may have errors or missing values. In addition to all this, it does a lot more apart from classification; it can generate if-then rules and even help with feature selection**.**

The core algorithm for generating a decision tree employs Entropy and information gain. Entropy is used to calculate the homogeneity of a sample. Entropy is maximized when there is an equal chance of all values for the target attribute.

Calculation of entropy:

$$(S) = \sum_{(\text{i=1 to l})} -S_i \vee |S| * log_2$$

S = set of examples

S_i= subset of S with value vi under the target attribute

l = size of the range of the target attribute.

Maximum depth(max_depth), criterion, splitter, minimum samples split(mini_samples_split) and minimum samples leaf (min_samples_leaf) are some of the parameters used to help tune a decision tree model. Max_depth helps in determining the height of the tree. This parameter limits the growth of the tree to avoid over-fitting. Criterion measures the quality of a split it takes 'gini' or entropy as inputs. The splitter on the other hand is used as a strategy to choose the split at each node, supported strategies are "best" for selecting the best split and "random" for selecting the best random split. Min_samples_split helps determine the minimum number of samples in a node to be considered for splitting. And lastly, min_samples_leaf helps determine the minimum number of samples needed to be considered a leaf node and this parameter also helps in limiting the growth of the tree.

2.3.2 Support vector machines

Support vector machine (SVM) is one of the well-known and commonly used classification methods because of its strong classification power. SVM creates a discrete hyperplane in the descriptor space of the training data and compounds are classified based on the side of the hyperplane they are located. SVM is a linear classier that works well with high dimension data regardless of the size of the dataset. For a training set of l sample, the learning procedure is represented as solving the following optimization problem:

$$\min_{\propto}: \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \propto_i \propto_j K(X_i, X_j) - \sum_{j=1}^{l} \propto_j$$

$$\text{s.t } 0 \leq \propto_i \leq C, i = 1, \ldots, l$$

$$\sum_{i=1}^{l} \propto_i y_i = 0$$

where y_i is the label of the ith sample x_i, α_i is the Lagrangian multiplier of x_i, C is the upper bound of α_i. K (x_i, x_j) is the kernel, which can map the original data X into a high-dimensional Hilbert space, and can make the samples linear separable in the high-dimensional space. The samples with $\alpha_i > 0$ are called support vectors.

Accordingly, the decision function can be written as:

$$f(x) = sgn\left\{\sum_{i=1}^{n_s} y_i \propto_i \; K(x_i, x) + b \right\}$$

where n_s is the number of support vectors.

Kernel, C, and gamma are some of the parameters used in tuning a Support Vector Machine. Kernel specifies the kernel type to be used in the algorithm. Supported kernels are 'linear', 'poly', 'rbf', 'sigmoid', and 'precomputed'. The parameter C, which is common to all SVM kernels, compensates for the classifications of the training examples compared to the simplicity of the decision surface, this is the regularization parameter. Lastly, gamma is used to determine the effect of a single training example. The larger the gamma, the more other examples must be affected.

2.3.3 Logistic regression

Logistic regression is one of the commonly used machine learning algorithms for classification problems. It is a predictive analysis algorithm based on the concept of probability. It requires that the dependent variable be categorical hence problems with continuous outcomes are not good candidates. Logistic regression is, thus, one of the simplest machine learning algorithms yet provides great efficiency. It has low variance and can be used for feature selection. The best way to think about logistic regression is that it is a linear regression but for classification problems. As opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio

Logistic function $1/(1+e^{-x})$

where x is the input variable.

Parameters C, Penalty, Solver, multiclass, and Max_iteration are used to tune a logistic regression model. C**,** the inverse of regularization strength, helps minimize over-fitting by reducing the variance of the model by constraining the size of the model coefficients. A smaller C indicates stronger regularization just like in SVM. Penalty helps specify the type of regularization. It takes the values 'l1', 'l2', 'elasticnet', or 'none'**.** The solver generally helps figure out what coefficients to be used in the model. It takes the values 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'. The choice of solver depends on the type of dataset. For small datasets, 'liblinear' is a good choice, whereas 'sag' and 'saga' are faster for large ones; For multiclass problems, only 'newton-cg', 'sag', 'saga', and 'lbfgs' handle multinomial loss; 'liblinear' is limited to one-versus-rest schemes. The multiclass *parameter* helps in deciding whether to handle the problem as binary or multinomial. It takes the values 'auto', 'ovr', or 'multinomial'. If the option selected is 'ovr,' then a binary problem is appropriate for each label. Even when the data is binary, the loss minimized for 'multinomial' is the multinomial loss fit across the entire probability distribution.

2.4 Evaluation Metrics

To compare the performance of classification models, different evaluation metrics are used. Some of the commonly used are Confusion Matrix, Accuracy, Precision, Recall, F1 score, Receiver Operating Characteristics (ROC), and Area Under the ROC Curve (AUC). To better understand these metrics, consider having a classification model that predicts whether a patient is COVID positive or negative. There will be four outcomes from this prediction:

True Positive (TP): if the model predicts someone is positive and that person is actually positive.

True Negative (TN): if the model predicts someone is negative and that person is actually negative.

False Positive (FP): if the model predicts someone is positive but the person is actually negative

False Negative (FN): if the model predicts someone is negative but the person is actually positive

Using the terms defined above. Accuracy, precision, and recall can be described as follows:

$$Accuracyscore = (TP + TN)/(TP + TN + FP + FN)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

The F1 score is a statistic that takes precision and recall into account. It's a simple weighted average of precision and recall. If we use the letters P and R to indicate precision and recall respectively, we can represent the F1 score as:

$$F1 = 2\,PR/(P + R)$$

A receiver operating characteristic curve (ROC curve) is a graph that shows how well a classification model performs across all categorization levels. Two parameters are plotted on this curve. True Positive Rate (TPR) and False Positive Rate (FPT).

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(FP + TN)$$

When we calculate the area under this ROC curve, we are calculating another metric that is frequently utilized when dealing with skewed binary targets in a dataset. This metric is referred to as the Area Under the ROC Curve (AUC)

A confusion matrix is an N * N matrix used to compare the actual target values with those predicted by the model. It helps visualize how a classification model is performing and what errors it is making. Figure *1* is an example of a binary confusion matrix:

		Actual Values	
Predicted Values		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

Figure 1 Confusion matrix

2.5 Related Work

Kumar (2016), confirmed that data mining is a reliable technique for analyzing road accidents to get productive results. Different data mining techniques and algorithms have been applied in various research studies regardless of having similar objectives. The type of data collected and the main goals influence the choice of methods used.

Several studies in literature focus on predicting the severity of accidents or identifying factors that affect severity (Beshah & Hill, 2010; Krishnaveni et al., 2011; Taamneh et al., 2017). In these studies, different algorithms were tested on the data, and the best-performing one was chosen for prediction. Chong et al. (2005) and Krishnaveni et al. (2011) agreed that accurately predicting severity can lead to a greater understanding of the relationship between the factors, which could provide crucial information for road accident prevention policy. Chong et al. (2005) used accident data from 1995 to 2000 to investigate

the performance of support vector machines, neural network, decision tree, and a hybrid decision tree. Using accuracy as a performance evaluation metric, the hybrid decision tree and decision tree performed better. Decision trees do not require a predefined relationship between the dependent and independent variables and have also proven to be very useful in handling prediction and classification problems in general. Beshah & Hill, (2010) studied the role of road-related factors on accident severity in Ethiopia. Various classification models were built using a decision tree, naive Bayes, and K-nearest neighbor classifiers. Accuracy was also used as an evaluation metric in this research, all three classifiers produced similar accuracy. The receiver operating characteristics (ROC) curve was then used to evaluate the same models and the results were also similar, it was then concluded that all three classifiers performed similarly well in predicting accident severity.

Taamneh et al. (2017) used accidents data from Abu Dhabi to explore the performance of different data mining techniques in predicting the severity of accidents from 2008 to 2013. The algorithms used were Decision Tree (DT) (J48), Naıve Bayes (NB),Rule Induction (PART), and Multilayer Perceptron (MLP). Using accuracy and area under the curve (AUC) to compare the performance of the algorithms, the results indicated that the Decision Tree (J48) classifier, PART classifier, and MLP classifier performed similarly well. Naive Bayes on the other hand demonstrated low accuracy. Krishnaveni et al. (2011) studied traffic accident records of 2008, which had 34,575 cases. This study classified the type of injury severity of various traffic accidents by applying and comparing Naive Bayes Bayesian classifier, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier, and Random Forest Tree classifier. To reduce the dimensionality of the dataset, a Genetic Algorithm was used for feature selection and the outcome showed that the Random Forest outperformed the other four algorithms based on their accuracy levels. Random Forest must have performed better because it runs well on large databases, it handles thousands of input variables without variable deletion, and also the learning is fast. In addition to this, it has an effective method for estimating missing data while maintaining accuracy. (Yuan et al., 2017) obtained motor vehicle crash data from the Iowa Department of Transportation containing crash records from 2006 to 2013. To predict traffic accidents, four classification models were evaluated, namely, Support Vector Machine (SVM), Random Forest, Decision Tree, and Deep Neural Network (DNN). SVM was used because

it has an efficient library for large-scale data classification. At the same time, classification and regression trees were used because of their ability to handle both numerical and categorical data. Using accuracy, precision, recall, F-Score, and area under the curve (AUC) it was concluded that Random Forest and DNN generally perform better than the other models.

A good number of research studies on Traffic safety also focus on discovering hidden patterns using association rule mining (Tayeb et al., 2015; Das, 2014; Kumar & Toshniwal, 2016). Association rule mining dates back to 1993 when Agrawal, (2016) analyzed the market basket problem to discover exciting collections or related links between items among large amounts of data. For example, by mining market basket data, they observed that 90% of customers who bought bread and butter at the same time would also purchase milk. Continuous expansion in terms of application has led to association rule mining being used in various fields, including analysis of traffic accidents. Kara and Kanga (2016) confirmed that a lot of algorithms have been developed under association rule mining, but from all these algorithms, Apriori is the biggest improvement and easy to implement.

Shetty et al., (2017) described how to mine frequent patterns causing road accidents data. Association rule mining, the Apriori Algorithm specifically was applied to the data, and patterns between road accidents were obtained. In 2014, Das used the association rule mining technique to discover hidden patterns in rainy weather crash data of Louisiana from the year 2004 to 2011. This research showed that the most frequent item in rainy weather was 'single vehicle run-off crashes'. This crash type was mainly associated with a few roadway features like 'on grade curve aligned roadways, curved roadways, and roadways with no streetlights at night. In the same year (Martín et al., 2014) used data mining techniques to improve road safety in Spanish roads. The decision tree was applied to obtain relations like (IF-THEN) that are easily understandable. Feng et al. (2020) analyzed UK traffic accident data from 2005 to 2017 by applying association rules. The results showed that RTAs have a strong correlation with environmental characteristics, speed limits, and location. To give safe driving suggestions and find out variables that are highly associated with fatal accidents, Li et al. (2017) used Association rules by applying the Apriori algorithm and classification models to analyze traffic accidents data. It was discovered that

human factors like being drunk or not, and the collision type, have a stronger effect on the fatal rate. Clustering results showed that some states/regions have higher fatal rates.

For classification, most studies used accuracy as an evaluation metric to determine the performance of a model. A model is considered better than another if it has higher accuracy on test data. However, accuracy alone has proven to be insufficient in other studies, especially where the data is imbalanced. In such cases, other evaluation metrics are used. These include; precision, recall, F-Score, and area under the curve (AUC). One can easily notice how in some studies the same algorithm performed better yet in another it did not. This is mainly to do with the data itself as these models are trained on the data. The quantity, quality of data plus the way it was prepared, and how the parameters were tuned can affect the accuracy of the model. In other words, the methodology matters. In this research, Accuracy, Precision, Recall and F1-score will be used as evaluation metrics to determine the performance of a model.

CHAPTER 3

METHODOLOGY

This chapter highlights the data used and the quantitative methods where we describe the parameters used in the design.

3.1 Data collection

Data for the research was collected from the Malawi Police headquarters in Lilongwe, the central repository for all traffic accidents data in Malawi. To have a significantly large dataset that could allow the extraction of meaningful results, accident records for five years (2016 to 2020) were collected. Secondly, the data set was the most recent data to ensure relevance to the current situation.

The following attributes were captured in the accidents records: accident number, severity, date, time, district, nearest police station, road number, the section of the road, noticeable physical feature close to the accident scene, accident type, road geometry, surroundings, surface type, road condition, weather, other factors, whether an animal was involved in the accident, whether there were any obstructions, whether there was a speed limit sign, speed limit and lighting condition (whether it was during the day or night).

In the data preprocessing phase, some of these attributes were left out as they brought noise to the dataset. Additionally, some input variables containing irrelevant information, redundant information, and attributes with over 80% unknown values were removed.

The data set was carefully reviewed for the issues mentioned above, and the following changes were made: removing the invariant attributes (e.g., police station), removing the descriptive and wordy attributes (e.g., noticeable physical features close to the accident scene and road number), removing irrelevant attributes (e.g., accident ID), removing the attributes with over 80% unknown values (e.g., other factors, speed limit, obstruction, animal and section) and removing the redundant information (e.g., date, time). The final list of the attributes is presented in Table 2.

Table 2 Attributes and value description

Attribute	**Values**
Severity	1: Fatal, 2: Serious injury, 3: Slightly/ minor injury, 4: Damages only, 5: Animal only
District	1:Chitipa, 2:Karonga, 3:Mzuzu, 4:Rumphi, 5:Mzimba, 6:Nkhatabay, 7:Kasungu, 8:Nkhotakota, 9:Ntchisi, 10:Dowa, 11:Mchinji, 12:Lilongwe, 13:Salima, 14:Dedza, 15:Ntcheu, 16:Mangochi, 17:Balaka, 18:Machinga, 19:Zomba, 20:Mwanza, 21:Neno, 22:Blantyre, 23:Chiradzulu, 24:Mulanje, 25:Phalombe, 26:Chikwawa, 27:Thyolo, 28:Nsanje
Accident type	1:Moving+moving head-on, 2:Moving+moving rear end, 3:Moving+moving side, 4:Moving+moving overtake, 5:Moving+moving turn, 6:Single moving rollover, 7:Single moving collision, 8:Moving+pedestrian, 9:Moving+bicycle, 10:Moving+controlled animal, 11:Moving+uncontrolled animal, 12:Moving+other
Road geometry	1:Straight road, 2:Curve, 3:Roundabout, 4:T-junction, 5:Y-junction, 6:+-junction, 7:X-junction, 8:Bridge, 9: Road/Rail crossing
Surroundings	1:Rural area, 2:Urban area, 3:Peri/ urban, 4:Farm/ compound
Surface type	1:Bitumen, 2:Gravel, 3:Earth
Road condition	1:Good/ Fair, 2:Potholes, 3:Corrugated, 4:Slippery
Weather	1:Dry, 2:Rain/Wet, 3:Mist, 4:Windy, 5:Dust
Posted speed limit	1:Speed limit posted, 2:Speed limit not posted
Light condition	1:Day light, 2:Night, 3:Dawn/Dusk

3.2 Methods

In regards to the specific objectives of the research, a decision was made to either use classification methods and Association rule mining. The discovery of hidden patterns in the traffic accidents data was achieved through association rule mining whereas modeling

the severity of road accidents was achieved through classification methods. But before feeding accidents into the algorithms, preprocessing was done. After feeding the processed data into the models, the results were analyzed. For association rule mining, rules were extracted and for classification, the models were compared on how they performed in predicting accident severity. Figure *2* below illustrates the framework described above

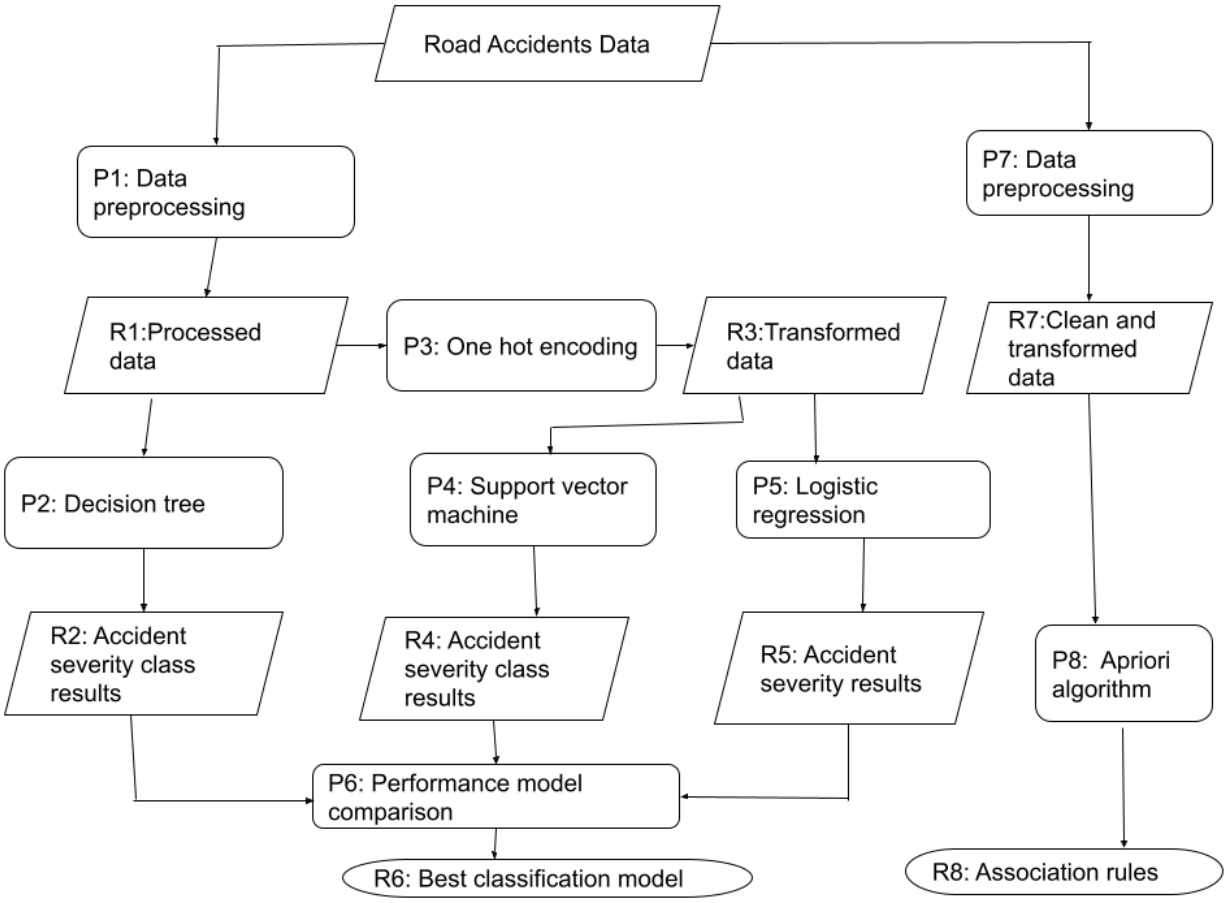

Figure 2 Modelling summary

3.2.1 Association Rule Mining

In this research, association rule mining was used to analyze previous accident data to obtain patterns between road accidents. Apriori algorithm was chosen for its simplicity in implementation and reliability in the extraction of simple rules. The first part of the Apriori

algorithm took two parameters: the data frame from which the rules had to be extracted; and the minimum support (min_support) parameter was set at 0.8. Support as discussed in the literature review indicates how frequently the items appear in the data. While Feng et al., 2020 used a minimum support of 0.4 in the first run and a minimum of 0.6 in the second run, this research used 0.8 to tighten the extraction of association patterns and eliminate occurrence by chance.

The data frame had the following attributes: Severity, District, Accident type, Road geometry, Surroundings, Road condition, Weather, Posted speed limit, and Surface type. All this data had to be transformed before feeding it into the algorithm since it was all categorical. In this research, categorical variables were transformed using one-hot encoding.

The first part produced frequent items as an output, these were used as input data for the last part which was responsibilities for generating the actual rules. While Feng et al., 2020 used a minimum confidence of 0.7, in our research 0.8 was chosen as the minimum confidence to help shortlist the rules. Since confidence indicates the number of times if-then rules are found true, 80% is a good minimum value if we are to consider making some decisions out of the rules generated. Then minimum lift (min_lift) parameter indicated the minimum lift value for the short-listed rules. The output of this last part was a list of rules with their associated attributes like support, confidence, and lift.

3.2.2 Classification

Classification in this research was used to classify accident severity and three algorithms were compared: Decision tree, Logistic regression, and Support Vector Machines. Each algorithm was run twice for modeling with different severity classes. The first run had severity categorized into five: Fatal, Serious injury, Slight/Minor Injury, Damages only, and Animal only. Whereas, the second run had the severity categorized into two categories, Fatal and non-Fatal. In terms of performance, all three algorithms were compared to each other when severity categories were the same. On the other hand, each algorithm's performance was compared to itself upon predicting severity in 5 categories and then 2 categories.

Before feeding the data into the algorithms, it had to go through preprocessing, this involved dimensionality reduction and data cleaning in which case some columns were dropped. For example, the accident number column did not give any information about the actual accident that would be significant in predicting severity as it only identifies the accident uniquely. The animals' column had 35,360 zeros, however, zero was not defined as being a real or missing value. For this research, these zeros were treated as missing values and the whole column was dropped.

The final dataset had the following columns: Severity, District, Accidents Type, Road Geometry, Surroundings, Surface Type, Road Condition, Weather, Posted Speed Limit, and Light Condition. However, some records were deemed erroneous, this is to say, the values captured were not the ones expected for that particular attribute. For example, under the attribute severity, expected values were 1,2,3,4, and 5 yet "severity" was also captured as a value. In most cases erroneous or missing entries are estimated to make the data complete, in the case where "severity" was captured as a value, there was no way of figuring out its real or estimated value between 1 and 5. However, in this research, rows with erroneous data were dropped since they were very few, and it was hard to come up with a method that would justify a value to replace "severity" for example.

3.2.2.1 Decision Tree Algorithm

A decision tree was used as one of the classification algorithms. Since a decision tree model, specifically a classification and regression tree (CART) can be trained directly from categorical data, the output of the preprocessing stage R1 (see Figure 2) was used as input for the decision tree. As a means of doing cross-validation, the data was split into test and training sets. Using the Sci-kit learn library in python to split the data, 80% of records were used to train the model and 20% of records were used to test the performance of the model. A maximum depth of 9 was used after looping through a range of 1 to 25 to choose the best value. Unlike (Yuan et al., 2017) who used a maximum depth of 13, our model began to overfit when the maximum depth went beyond 9. The maximum depth of 9 maximized the accuracy without overfitting. A criterion of entropy was used considering that attributes were in classes and not continuous. For the parameter Splitter, 'best' was used since it considered all features and it chose the best split, furthermore, the random split was also

tested but it produced a slightly lower accuracy. Default values for minimum samples split and minimum samples leaf parameters were used. This was so because any changes to the values had no impact on the performance of the model. All the mentioned parameters were used to help tune the model.

The steps described above were repeated starting from process P1, this time around the severity classes were grouped into 2: Fatal and non-fatal accidents. What was categorized as fatal initially remained fatal whereas serious injury, slightly/ minor injury, damages only, and Animal only were categorized as non-fatal accidents. In process P2 only the maximum depth parameter changed, while the rest remained the same. Maximum depth took the value 6, beyond this the model was overfitting.

3.2.2.2 Logistic regression

Logistic regression was also used for predicting accidents severity. The output from the general preprocessing had to go through one more process before feeding into the algorithm. Since all variables were categorical, and logistic regression only handles categorical variables after they have undergone some transformation, one hot encoding was used to transform the data. The data was split into training and test data as a means of cross-validation. Using the Sci-kit-learn library to split the data, 80% of records were used to train the model and 20% of records were used to test the performance of the model. By using grid search cross-validation for logistic regression the following parameters emerged as the best parameters for the model, C=5.79 and penalty = l2. Since the value that is being predicted (severity) is in 5 categories, this problem had to be solved as a multinomial logistic regression. Hence having multi_class='multinomial' and a solver had to be chosen that supports multinomial classification, which led to the choice of "lbfgs" as a solver. The model was trained using these parameters.

The processes described above were repeated starting from P1. The severity classes were grouped into fatal and non-fatal accidents. The data was split into training and test sets using the same ratio as in the initial run. The parameters were changed and a few more were introduced, multi_class parameter was omitted since the problem became a binary classification upon grouping severity into two classes. After looping through several

options, the following parameter values were chosen to train the model; maximum iteration was increased to 1000, C was changed to 2, and solver was changed to ‘saga’.

3.2.2.3 Support Vector Machine

Just like logistic regression, as some of the input data was categorical, it had to undergo one-hot encoding as well for an SVM. Nonetheless, a few more decisions had to be made before feeding the data into an SVM. Deciding which kernel function to go with to succeed in classification is a difficult task in SVM. Polynomial, linear, and RBF were all tested on the dataset, and RBF emerged to be the best choice for this dataset as it was able to separate the classes with higher accuracy. Just as it was with the case of Chong et al., 2005 where they found RBF to be the best choice. Consequently, the RBF kernel was used to train an SVM in this research, and two extra parameters were used. The parameter C, which is common to all SVM kernels took the value 10. And the parameter gamma took the value of 0.0001. Grid Search cross-validation was used to choose the optimal values for C and gamma after supplying a range of values to the function Chong et al., 2005 and Yuan et al., 2017.

Starting with process P1, the stages were repeated, but this time around the severity classes were divided into two groups. Fatal and non-fatal accidents. The data was split into training and test sets using the same ratio as in the initial run. The parameters for tuning the model remained the same as those in the first run.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Association rules

To discover hidden patterns in the traffic accidents data, the Apriori algorithm was applied. A good number of association rules were generated. To come up with strong rules, that may inform the reality on the ground. Only rules with support and confidence above 0.8 were chosen. Figure *3* shows some of the rules that were created. For example, rule number 1 (road condition_1 => severity_0) reads that non-fatal accidents are a result of good/fair road conditions. Rule number 2 (weather_1 => severity_0) reads that non-fatal accidents are a result of dry weather conditions.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(severity_0)	(road condition_1)	0.854426	0.984177	0.842464	0.986000	1.001853
1	(road condition_1)	(severity_0)	0.984177	0.854426	0.842464	0.856009	1.001853
2	(weather_1)	(severity_0)	0.985919	0.854426	0.842610	0.854644	1.000255
3	(severity_0)	(weather_1)	0.854426	0.985919	0.842610	0.986170	1.000255
4	(surface type_1)	(severity_0)	0.947595	0.854426	0.816102	0.861235	1.007969
5	(severity_0)	(surface type_1)	0.854426	0.947595	0.816102	0.955146	1.007969
6	(weather_1)	(road condition_1)	0.985919	0.984177	0.972854	0.986748	1.002613
7	(road condition_1)	(weather_1)	0.984177	0.985919	0.972854	0.988495	1.002613
8	(posted_speed_limit_2)	(road condition_1)	0.822489	0.984177	0.807276	0.981503	0.997283
9	(road condition_1)	(posted_speed_limit_2)	0.984177	0.822489	0.807276	0.820255	0.997283
10	(surface type_1)	(road condition_1)	0.947595	0.984177	0.944836	0.997089	1.013120
11	(road condition_1)	(surface type_1)	0.984177	0.947595	0.944836	0.960027	1.013120
12	(weather_1)	(posted_speed_limit_2)	0.985919	0.822489	0.811340	0.822928	1.000534
13	(posted_speed_limit_2)	(weather_1)	0.822489	0.985919	0.811340	0.986445	1.000534
14	(weather_1)	(surface type_1)	0.985919	0.947595	0.936562	0.949938	1.002473
15	(surface type_1)	(weather_1)	0.947595	0.985919	0.936562	0.988357	1.002473
16	(weather_1, severity_0)	(road condition_1)	0.842610	0.984177	0.833058	0.988664	1.004559
17	(weather_1, road condition_1)	(severity_0)	0.972854	0.854426	0.833058	0.856303	1.002197
18	(severity_0, road condition_1)	(weather_1)	0.842464	0.985919	0.833058	0.988834	1.002957
19	(weather_1)	(severity_0, road condition_1)	0.985919	0.842464	0.833058	0.844956	1.002957
20	(severity_0)	(weather_1, road condition_1)	0.854426	0.972854	0.833058	0.974991	1.002197

Figure 3 First set of association rules

The support and confidence were both reduced to 0.7 to accommodate some more rules that would be worthy considering as the first set seemed obvious and the rules were few. 91 rules were created out of which none was to do with fatal accidents. This could be because the data collected had more non-fatal accidents, which made up slightly above three-quarters of the dataset. Figure 4 indicates a portion of the 91 rules created upon reducing the support and confidence threshold.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(severity_0)	(road condition_1)	0.854426	0.984177	0.842464	0.986000	1.001853
1	(road condition_1)	(severity_0)	0.984177	0.854426	0.842464	0.856009	1.001853
2	(weather_1)	(severity_0)	0.985919	0.854426	0.842610	0.854644	1.000255
3	(severity_0)	(weather_1)	0.854426	0.985919	0.842610	0.986170	1.000255
4	(surface type_1)	(severity_0)	0.947595	0.854426	0.816102	0.861235	1.007969
...	...	...	...	...	...	...	...
87	(posted_speed_limit_2, road condition_1)	(weather_1, surface type_1)	0.807276	0.936562	0.761664	0.943499	1.007407
88	(weather_1)	(surface type_1, posted_speed_limit_2, road co...	0.985919	0.769213	0.761664	0.772543	1.004329
89	(surface type_1)	(weather_1, posted_speed_limit_2, road conditi...	0.947595	0.798624	0.761664	0.803787	1.006465
90	(posted_speed_limit_2)	(weather_1, surface type_1, road condition_1)	0.822489	0.934646	0.761664	0.926048	0.990801
91	(road condition_1)	(weather_1, surface type_1, posted_speed_limit_2)	0.984177	0.763377	0.761664	0.773910	1.013798

Figure 4 Second set of association rules

The following are some of the rules randomly selected from the 91 generated:

1. Non-fatal accidents are associated with Dry weather, and they occur in urban areas
2. Accidents that occur in Dry weather and the surface type is bitumen are non-fatal
3. Accidents that occur in urban areas are associated with dry weather, good road condition and bitumen surface type.
4. Accidents that occur in daylight are associated with dry weather, the surface type is bitumen and there is no speed limit sign

The following are the top frequent item sets when support and confidence were both set at a minimum of 0.7:

1. Surroundings = urban

2. Road condition = Good/Fair
3. Weather= Dry
4. Posted speed limit = No speed limit posted
5. Light condition = Daylight
6. Surface type = Bitumen

Only fatal accidents were then pulled out from the data and the Apriori algorithm was re-run with support and confidence both equal at 0.8 as a minimum. This was done to also have an insight on the type of associations that lead to fatal accidents without being overshadowed by the huge data set on non-fatal accidents. Support and confidence were both set at 0.8 and 87 rules were created. Figure 5 indicates a portion of these rules. The following are some of the rules randomly selected from the 87 rules generated:

1. Accidents that occur in areas where the speed limit is not posted are usually fatal
2. Fatal accidents that occur in good/fair roads are associated with dry weather
3. Accidents that occur in bitumen surface types and road condition is good are usually fatal

The following are the top frequent item-sets when support and confidence were both set at a minimum of 0.8:

1. Road Condition = Good/Fair
2. Weather = Dry
3. Posted speed limit = No speed limit posted
4. Surface type = Bitumen

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(severity_1)	(road condition_1)	1.000000	0.973474	0.973474	0.973474	1.000000
1	(road condition_1)	(severity_1)	0.973474	1.000000	0.973474	1.000000	1.000000
2	(severity_1)	(weather_1)	1.000000	0.984444	0.984444	0.984444	1.000000
3	(weather_1)	(severity_1)	0.984444	1.000000	0.984444	1.000000	1.000000
4	(severity_1)	(posted_speed_limit_2)	1.000000	0.847028	0.847028	0.847028	1.000000
...	...	...	...	...	...	...	...
83	(surface type_1, weather_1)	(severity_1, road condition_1)	0.891703	0.973474	0.890307	0.998434	1.025640
84	(severity_1)	(road condition_1, surface type_1, weather_1)	1.000000	0.890307	0.890307	0.890307	1.000000
85	(road condition_1)	(severity_1, surface type_1, weather_1)	0.973474	0.891703	0.890307	0.914567	1.025640
86	(surface type_1)	(severity_1, road condition_1, weather_1)	0.903271	0.960311	0.890307	0.985648	1.026384
87	(weather_1)	(severity_1, road condition_1, surface type_1)	0.984444	0.901476	0.890307	0.904376	1.003217

Figure 5 Final set of association rules

4.2 Model comparison

In this research, the decision tree, logistic regression, and support vector machines were compared on how they performed in predicting the severity of accidents. The first run had severity categorized into five: Fatal, Serious injury, Slightly/ minor injury, Damages only, and Animal only. The performance measurements of the three models are listed in table 2. Decision tree and logistic regression performed similarly well in predicting the severity of an accident compared to support vector machines. Table 3 indicates that the Decision tree, Logistic regression, and SVM accurately predicted the severity by 70.37%, 70.35%, and 64.64 respectively.

Table 3 results of classifying accidents severity into 5 categories

	Accuracy	Precision	Recall	F1-score
Decision tree	70.37	67.29	70.37	67.64
Logistic regression	70.35	67.38	70.35	67.66
Support vector machine	64.64	50.33	64.63	56.36

When severity was categorized into two classes, fatal and non-fatal accidents, the models predicted severity with the metrics shown in table 3. This time around all three algorithms produced almost the same accuracy. With all accuracies above 85%, Decision trees, logistic

regression, and support vector machines can all be used in the prediction of severity, and one would expect an outcome with high accuracy. However, for precision and F1-score, the Support vector machine did not perform as good as the Decision tree and logistic regression. Reducing the severity classes significantly helped improve the performance of the models as is evident by the increased accuracy, precision, recall and F1-score. Table 4 indicates that Decision tree, Logistic regression, and SVM accurately predicted the severity by 86.79%, 87.55%, and 85.50% respectively.

Table 4 results of classifying accidents severity into 2 categories

	Accuracy	**Precision**	**Recall**	**F1-score**
Decision tree	86.79	84.49	86.79	84.48
Logistic regression	87.55	86.02	87.55	86.33
Support vector machine	85.50	73.09	85.4	78.81

In this research, both the logistic regression and decision trees performed well in predicting accident severity. However, the logistic regression performed slightly better than the decision tree, because of the binary nature of the problem and the ability of the model to train on categorical data. This, however, differs from what most researchers found in traffic accidents analysis using data mining techniques. Most researchers found decision trees to be the best performing model for accident severity prediction largely because logistic regression was not consider among the models being tested. Beshah & Hill, (2010) in their study of the role of road-related factors on accident severity in Ethiopia, built various classification models among which was a decision tree that performed well. Taamneh et al. (2017) used accidents data from Abu Dhabi to explore the performance of different data mining techniques in predicting the severity of accidents from 2008 to 2013. Using accuracy and area under the curve (AUC) to compare the performance of the algorithms, the results indicated that the Decision Tree was among the best-performing algorithms. On the other hand, Logistic regression is rarely used in literature for accident severity prediction using data mining, but for this research, it is slightly better than the decision tree. Lastly, the support vector machine performed well, but not as good as the other two models, especially in terms of precision. The case is similar to when (Yuan et al., 2017)

obtained motor vehicle crash data from the Iowa Department of Transportation containing crash records from 2006 to 2013. To predict traffic accidents, the Support Vector Machine was not among the best performing models.

4.2.1 Fitting a Logistic Regression

The logistic regression was then chosen as the best model for our data and was then fit to get more insight from the data. All the attributes contributed to determining the severity of an accident. However, some had a bigger impact compared to others. Using the model's coefficients, we were able to determine each attribute's contribution to determining the severity of an accident. The higher the coefficient implied the more likely an accident will be fatal and vice versa. With reference to the Table *5* below. The top 3 attributes that had a higher chance of causing a fatal accident than a non-fatal one were.

1. An accident involving a moving vehicle and a pedestrian
2. An accident that occurred at dawn or dusk
3. An accident involving a moving vehicle and a bicycle.

The bottom 3 attributes had a lower chance of causing fatal accidents and these were

1. An accident involving two vehicles moving side by side
2. An accident that occurred in an urban area
3. An accident involving a moving vehicle and an uncontrolled animal

Below are the coefficients of the attributes that contributed to accident severity.

Table 5 coefficients of the attributes that contributed to accident severity

r	Description	coef
accident type_8	Moving + Pedestrian	5.405831
light condition_3	Dawn/Dusk	3.251403
accident type_9	Moving + Bicycle	3.246903
surroundings_4	Farm/Compound	2.732928
accident type_12	Moving + Other	2.620153
accident type_1	Moving + Moving	2.340571
accident type_6	Single moving rollover	2.274108
road geometry_7	X-Junction	2.022932
weather_5	Dust	1.789397
surroundings_1	Rural area	1.767158

r	Description	coef
road geometry_2	Curve	1.425653
accident type_7	Single moving collision	1.376577
road geometry_8	Bridge	1.37016
weather_1	Dry	1.344641
road condition_1	Good/Fair	1.343218
road condition_2	Potholes	1.228312
surface type_3	Earth	1.170019
road condition_3	Corrugated	1.104156
surface type_1	Bitumen	1.095167
road geometry_5	Y-Junction	1.073332
posted_speed_limit_1	Speed limit posted	1.065488
road geometry_1	Straight road	1.057091
surroundings_3	Peri/Urban	0.982609
weather_2	Rain/Wet	0.963796
accident type_4	Moving+Moving	0.951089
road geometry_9	Road/Rail crossing	0.950392
posted_speed_limit_2	Speed limit not posted	0.910609
road geometry_4	T-Junction	0.813645
road geometry_6	+ Junction	0.799293
weather_3	Mist	0.76422
light condition_2	Night	0.760634
surface type_2	Gravel	0.757195
accident type_10	Moving + Controlled animal	0.68572
accident type_5	Moving + Moving turn	0.559661
weather_4	Windy	0.547475
road condition_4	Slippery	0.532592
light condition_1	Day light	0.392314
road geometry_3	Roundabout	0.350126
accident type_2	Moving + Moving rear end	0.314506
accident type_3	Moving + Moving side	0.306897
surroundings_2	Urban	0.204454
accident type_11	Moving + Uncontrolled animal	0.081729

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

In this research, traffic accidents data was analyzed using two approaches. Association rule mining and classification methods. In both scenarios, the data collected had to go through preprocessing. Decision trees, logistic regression, and support vector machines were applied to the data. Decision trees and logistic regression yielded the most accurate severity predictions compared to support vector machines. These models were evaluated using accuracy, precision, recall, and F1-score. The experiment result revealed that reducing the severity categories from 5 to 2 improves the performance of all the models. However, the logistic regression performed slightly better than the decision tree hence it was chosen to fit the data and get more insight. The coefficients of the attributes were calculated to get a picture of how much an attribute contributes to the severity of an accident. In the end, the type of accident turned out to be an attribute that has a higher chance of determining if an accident is fatal or not. Light conditions and surroundings also showed a higher impact on the severity.

Through association rule mining, using the Apriori algorithm, a series of interesting rules were discovered. Despite having different support and confidence levels, Road Condition, Weather, posted speed limit and Surface type were the frequent item sets that appeared in all the rules generated. Similarly, (Kassu & Anderson, 2018) also discovered that light conditions and weather were among the principal factors affecting the severity of an accident. Tayeb et al., (2015) found that non-fatal accidents were associated with dry road conditions, daylight, and clear weather. The same rule was discovered in this research.

The results from classification and association rule mining are not entirely the same but similar. For example, both classification using logistic regression and association rule mining Apriori algorithm stipulate that non-fatal accidents are common in urban areas and that dry weather plus good/fair road conditions also attributes to fatal accidents. With classification, it was also discovered that accidents that occurred at dusk/dawn had higher chances of being fatal whereas association rule mining found accidents that occurred in daylight were mostly non-fatal, which in a way speaks of the same thing. Here we can conclude light conditions affect the severity of an accident. Poor lighting on the roads, poor sight of the drivers, no road markings, tiredness of the drivers, and high speeds since the roads are usually clear at dawn/dusk contribute to the accidents being fatal as established by Saba Momeni Kho, (2021). As Das, et al. (2019) indicated we conclude that roadway lighting at night would help alleviate crash severity. It was also interesting to note that, contrary to common belief, accidents involving uncontrolled animals are rarely fatal. This research also shows that accidents that occur in rural areas are more likely to be fatal than those that happen in urban areas.

5.2 Recommendations

Regarding the economy of Malawi, decisions to improve traffic safety in the country should be based on data to allow proper allocation of resources. The measures discussed in this section are not meant to be a complete compilation of all possible safety improvements. But provide sufficient information to significantly impact the well-being of road users in Malawi.

In this research, several rules were generated, some of the attributes forming these rules, like weather, we have no control over, but others can be addressed easily. Having functional streetlights, having good roads, and ensuring all road sections have speed limit signs could greatly reduce accident severity. On the other hand, through classification, the top 3 contributing factors to fatal accidents were discovered to be accidents involving moving vehicles and pedestrians, accidents that occur at dawn or dusk, and accidents involving vehicles and bicycles. With this information, a lot can be done to reduce the severity of the accidents.

Numerous factors contribute to pedestrian (and bicycle users) deaths, and it is often necessary to employ a combination of engineering, enforcement, and education measures to be effective (Zegeer & Bushell, 2010). This includes improving highway infrastructure and facilities for pedestrians. Streets should have sidewalks or walkways, as well as better street connectivity for all road users and bicycle facilities. In addition, because pedestrian (and bicycle) safety education has been taught in elementary schools sporadically or not at all, it is also recommended that we develop and implement nationally accepted, well-coordinated pedestrian safety education programs in schools nationwide. And lastly, Police enforcement is key in improving traffic safety, this can deal with both accidents involving pedestrians, bicycles as well as accidents that occur at dawn or dusk. Photo enforcement for speeding and red-light running has been used in some industrialized nations. This is effective because drivers are aware they are being watched every time even when the traffic police officers are not on the roads. Despite this solution being costly, it is a self-sustaining program as it can generate revenue through fines from.

The data collection methods also must be improved. Collecting as much information as possible including human-related data and geographical coordinates of the actual accident area would be necessary. Performing any form of analysis on such rich data will greatly improve traffic safety as it will provide better insight for better decision-making.

Overall, researchers in data mining, when dealing with road accidents should consider combining association rule mining and classification methods. When the two are conducted on the same data and comparing the results, more insight is achieved than just choosing between the two and proceeding with one as is the case with most studies. On the other hand, working with a minimum number of classes possible should make the classification algorithms perform better as was the case in this research.

5.3 Limitation of the research

The main limitation of this research was the data itself. The data had a lot of missing values on attributes that are more likely to contribute to accident severity, for example, the speed limit. On the hand, the data available did not contain human factors that may also contribute to accidents. For example, whether the driver was drunk or not. The data collected on road traffic accidents from road traffic directorate only contained environmental factors, such

as road geometry, surface type, road condition, etc. If the data had both human and environmental factors, the association rules generated would have provided a more realistic picture of what happens on Malawian roads for these accidents to occur.

REFERENCES

Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing Switzerland.

Agrawal, R. (2016). *Mining Association Rules between Sets of Items in Large Databases*. *January 1993*. https://doi.org/10.1145/170035.170072

Amibe, D. (2012). *Final Draft Report on Pilot Global Fuel Economy Initiative Study in Ethiopia*.

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

Chong, M., Abraham, A., & Paprzycki, M. (2005). *Traffic Accident Analysis Using Machine Learning Paradigms*. *29*, 89–98.

Das, S., Anandi, D., Raul A., Karen, D., X. S. & M. J. (2019). Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. *International Journal of Urban Sciences*, *23*(1), 30–48.

Das, S. (2014). *Investigating the Pattern of Traffic Crashes Under Rainy Weather by Association Rules in Data Mining Investigating the Pattern of Traffic Crashes under Rainy Weather by Association Rules in Data Mining*. *January*.

Feng, M., Zheng, J., Ren, J., & Xi, Y. (2020). *Association Rule Mining for Road Traffic Accident Analysis : A Case Study Study from UK*. *February*. https://doi.org/10.1007/978-3-030-39431-8

Gopalakrishnan, S. (2012). A public health perspective of road traffic accidents. *Journal of Family Medicine and Primary Care*, *1*, 144–150.

Ihueze, C. C., & Onwurah, U. O. (2018). *Road tra ffi c accidents prediction modelling : An analysis of Anambra State , Nigeria*. *112*(December 2017), 21–29. https://doi.org/10.1016/j.aap.2017.12.016

Kassu, A., & Anderson, M. (2018). Determinants of Severe Injury and Fatal Traffic Accidents on Urban and Rural Highways. *International Journal for Traffic and Transport Engineering*, *8*(3), 294–308. https://doi.org/10.7708/ijtte.2018.8(3).04

Krishnaveni, S., Hemalatha, M., Professor, A., & Head, &. (2011). A Perspective Analysis of Traffic Accident using Data Mining Techniques. In *International Journal of Computer Applications*, 23, 7.

Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, *24*(1), 62–72. https://doi.org/10.1007/s40534-016-0095-5

Li, L., Shrestha, S., & Hu, G. (2017). Analysis of road traffic fatal accidents using data mining techniques. *Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017*, 363–370. https://doi.org/10.1109/SERA.2017.7965753

Martín, L., Baena, L., Garach, L., López, G., & Oña, J. De. (2014). Using data mining techniques to road safety improvement in Spanish roads . *Procedia - Social and Behavioral Sciences*, *160*(Cit), 607–614. https://doi.org/10.1016/j.sbspro.2014.12.174

Saba Momeni Kho, Parham Pahlavani, B. B. (2021). Classification and association rule mining of road collisions for analyzing the fatal severity, a case study. *Journal of Transport & Health*, *23*.

Sajjad, S., Ehsan, T., & Sara, Z. (2017). The effect of drivers' demographic characteristics on road accidents in different seasons using data mining. *Scientific Journal on Traffic and Transportation Research*, *29*.

Schlottmann, F., Tyson, A. F., Cairns, B. A., Varela, C., & Charles, A. G. (2017). *Road traffic collisions in Malawi : Trends and patterns of mortality on scene*. *29*(December), 301–305.

Shetty, P., C, S. P., Kashyap, S. V, & Madi, V. (2017). Analysis of road accidents using data mining techniques. In *International Research Journal of Engineering and Technology*. www.irjet.net

Szalay, Z. (2019). *Analysis of traffic accident black spots: an application of spatial clustering segmentation method*. https://www.researchgate.net/publication/340062435

Taamneh, M., Alkheder, S., & Taamneh, S. (2017). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety and Security*, *9*(2), 146–166. https://doi.org/10.1080/19439962.2016.1152338

Tayeb, A. A. El, Pareek, V., & Araar, A. (2015). *Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai*. *4*, 1–12.

WHO. (2020). *Road Traffic Imjuries*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

World Health Organisation. (2018). *Global Status Report on Road Safety 2018*.

Yuan, Z., Zhou, X., Yang, T., Tamerius, J., & Mantilla, R. (2017). *Predicting Traffic Accidents Through Het-erogeneous Urban Data: A Case Study* (Vol. 9). https://doi.org/10.475/123_4

Zegeer, C. V., & Bushell, M. (2010). Pedestrian Crash Trends and Potential Countermeasures from Around the World. *Accident Analysis and Prevention*, *44*(1), 3–11.