ANALYSIS OF TRAFFIC ACCIDENT BY USING MACHINE LEARNING

Julie Marlar[1], Wint Pa Pa Kyaw[2], Soe Mya Mya Aye[3]

Abstract

In the case of road accidents, millions of people are dead each year, and the numbers of accident rates are rising all over the world. As a result, they have a great impact on society in terms of finance and economic. In this paper, Machine Learning (ML) models are used to analyze road accident severity based on road accident dataset. The dataset was collected from Myanmar's Traffic Police Force which shares raw data on annual basis on Yangon- Naypyitaw expressway's traffic accident data in 2022. The dataset is divided into training and testing dataset. The training dataset is used to train our model, and the test dataset is utilized to evaluate the predictions. In the proposed system, the collected data is preprocessed (data cleaning, encoding, transformation) and followed by data training, testing and comparison analysis on the analysis of the ML methods. By this experiment, the road accident contributing factors is vital. By using the continuous variables, this work is predicted road accidents with different accident types and applied with ML technique like Logistic Regression (LR), Adaptive Boosting (AdaBoost), Decision Tree, Adaboost using Decision Tree and Multinomial Naive Bayes. The experimental results showed that Logistic Regression classifier achieves the best accuracy than other classifiers.

Keywords: Logistic Regression, Adaboost, Decision Tree, Multinomial Navie Bayses

Introduction

Road traffic accidents (RTAs) are increasing worldwide and a major of injuries, causing millions of deaths and fatalities, financial and economic expenses on society (S Ahmed, 2023). According to the World Health Organization (WHO), in 2019 Road traffic mortality rate is 20.94 per 100,000 of population ranks. Myanmar stands in the place of 71 in the world. The latest WHO data issued that road traffic accident Deaths in Myanmar reached 11,004 of total deaths in 2020. In recent years, road traffic accidents, especially severe vehicle crashes have increased because of the rapid growth of road traffic (J.Lil, 2023).

Machine learning techniques can be applied in road safety to improve life-threatening problems on the roads. With the advancements of information technology, machine learning becomes increasingly mature, and useful information without preconditions can be found in databases. ML is described as a method that can be used to make provisions for data analysis, decision making, and data preparation for real-life problems. The learning begins with data analysis to identify patterns within the dataset and make future decisions involving societal problems (X.Wang, 2022).

Applications of machine learning techniques in RTAs can help in the modelling for better understanding of RTAs data records and can be used to achieve numerous outcomes such as classification, prediction, and clustering analysis. Classification methods are among the most commonly used techniques in mining traffic accidents, where the goal is building classifiers that can predict the accidents. The main objectives of the proposed system are to explore the factors influencing the severity of traffic accidents on Yangon- Naypyitaw expressway road, to analyze and build the models based on the accident data which is to get a better understanding, and evaluate the causes and effects on the severity of traffic accidents.In this paper, the purpose of the methodology is to set the classification rules for prediction of the best performing five models based on machine learning algorithms are constructed.

---

* Special Award (2023)

[1] Department of Computer Studies, University of Yangon

[2] Department of Computer Studies, University of Yangon

[3] Department of Computer Studies, University of Yangon

Machine Learning Methods

Machine learning involves a group of computational algorithms that can perform classification, pattern recognition and prediction. There are different types of machine learning Algorithms such as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.ML have successfully been implemented in automated stock trading, computer vision, health care, speech recognition, and customer services. Most ML classifiers are influenced by the size of the dataset and capabilities to handle overfitting problems and are being implemented in different environments such as urban and rural settings and on freeways and expressway. ML classifiers employed during the comparative analysis are described below.

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. The idea of the algorithm is to map the results of linear functions to sigmoid functions. The sigmoid function is a mathematical function used to map the predicted values to probabilities. The value must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. Logistic function is shown in Figure 1.

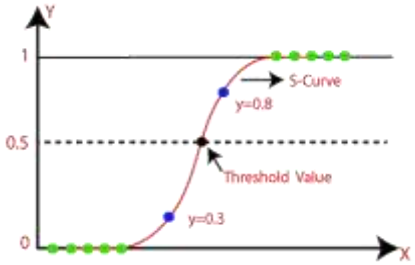


Figure 1: Logistic Regression

Adaptive boosting (AdaBoost)

AdaBoost is the simplest boosting algorithm based on an ensemble decision tree. It uses an iterative adaptive approach in which weights are adjusted at each iteration by assigning higher weights to incorrectly classified instances. Boosting is an ensemble learning method that combines a set of weak learners into strong learners to minimize training errors. This method operates iteratively, identifying misclassified data points and adjusting their weights to minimize the training error. AdaBoost function is shown in Figure 2.

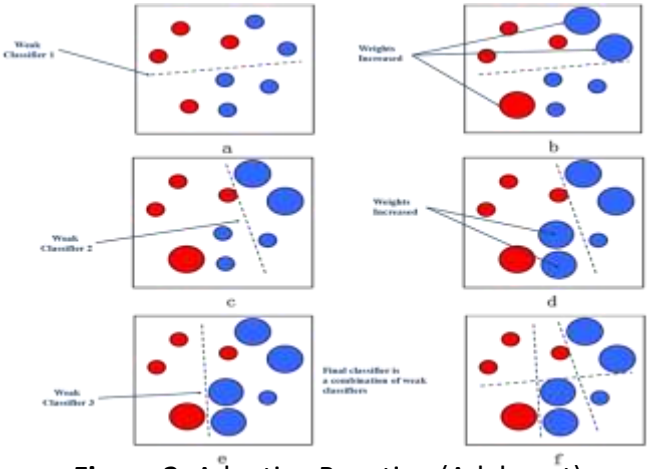


Figure 2: Adaptive Boosting (Adaboost)

Naive Bayes

The NB classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions among variables. Naive Bayes is one of the fast and easy simple ML algorithms to predict a class of datasets. which helps in building fast machine learning models that can make quick predictions. Since the feature set contains continuous variables, the Gaussian NB was chosen. The naive Bayes model is easy to build and particularly useful for very large data sets. Naïve Bayes function is shown in Figure 3.

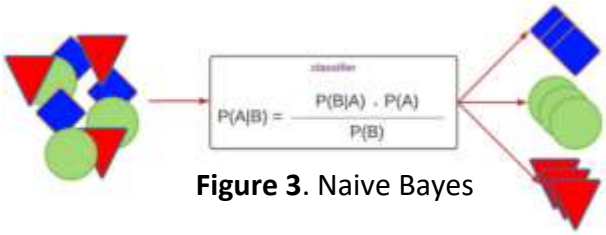


Figure 3. Naive Bayes

Decision Tree

Decision Tree is a Supervised learning technique which is also called 'Classification and Regression Tree algorithm (CART)' used that can be used for both classification and Regression problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision Tree function is shown in Figure 4.

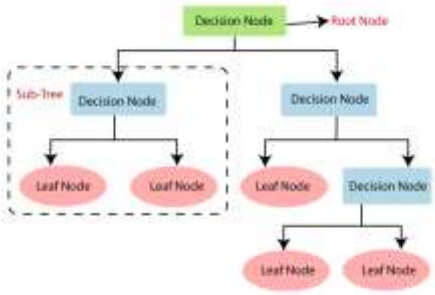


Figure 4. Decision Tree

Implementation of the Road Traffic Accident Data Set

In this study, the dataset is built information about road crashes occurred on Yangon-Naypyitaw expressway during the period from January to December in 2022. Additionally, the dataset contains traffic accident records in this year, having the total number of 165 traffic crashes. The data used is provided from the traffic accident report of Myanmar's Traffic Police Force. This station shares raw data on annual basis on Yangon- Naypyitaw expressway's traffic accident data. The original dataset has 165 records and 14 variables. This research utilized 165 records involving 11 selected variables. The sample dataset was shown in Table 1. Data preparation is performed before each model building.

The process includes various steps such as cleaning, encoding and data transformation. Data cleaning is conducted to correct the missing values and the erroneous values. In this study, dummy coding scheme are used, this categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In this encoding, for each level of a categorical feature, create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category. Likewise, the study chose the factors which are related to the accidents that include, accident type, crash injuries, human factors, reason, season, time, weather condition, environment, road condition, road characteristics and road alignment.

Table 1: Dataset of Selected Features on Road Traffic Accidents

Accident Type	People	Human Factor	Reason	Season
Rollover	Death	No Seat-belt used	Tire puncture	Spring
Collision	Death	No Seat-belt used	Overspeed	Spring
Hit	Minor injury	Seat-belt used	Overspeed	Spring
Hit	Serious injury	No Seat-belt used	Tire puncture	Spring
Rollover	Serious injury	Motorcycle Helmet used	Not Permitted Motorcycle zone	Spring
Hit	Death	Drowsy Driver	Overspeed	Spring
Collision	Death	Seat-belt used	Overspeed	Spring
:	:	:	:	:
:	:	:	:	:
Rollover	Minor injury	Seat-belt used	Tire puncture	Spring
Rollover	Death	No Seat-belt used	Overspeed	Rainy
Hit	Serious injury	No Seat-belt used	Overspeed	Spring
Collision	Serious injury	Seat-belt used	Overspeed	Spring
Hit	Serious injury	No Motorcycle Helmet used	Not Permitted Motorcycle zone	Spring

…

Time	Weather Condition	Environment	Road Condition	Road Characteristics	Road Alignment
Daytime	Fine	Light	Dry	General Road	Flat
Daytime	Fine	Light	Dry	Junction Type	Curve
Nighttime	Mist	Dark	Snow	Barrier Type	Curve
Nighttime	Mist	Dark	Snow	General Road	Flat
Daytime	Fine	Light	Dry	General Road	Flat
Daytime	Mist	Light	Wet	General Road	Flat
Nighttime	Mist	Dark	Snow	General Road	Flat
:	:	:	:	:	:
:	:	:	:	:	:
Nighttime	Fine	Dark	Snow	General Road	Flat
Nighttime	Heavy Rain	Dark	Wet	General Road	Flat
Nighttime	Mist	Dark	Snow	Bridge	Flat
Nighttime	Mist	Dark	Wet	General Road	Flat
Nighttime	Mist	Dark	Snow	Barrier Type	Curve

In developing the system, Python programming language is used and Jupyter notebook which is an open-source IDE that allows us to create and share documents that contain live code, equations, visualizations, and narrative text. The uses include data cleaning, transformation, statistical modeling, data visualization, machine learning, and so on. In preprocessing, the dataset was encoded with dummies method and data transformation by using MinMaxScaler in Scikit-Learn. It can support various languages that are popular in data science such as Python language. In this study, the applied ML classifiers are the Logistic regression, Adaboost, Naive Bayes, Decision Tree and Adaboost with Decision Tree using Python's Scikit-Learn library (T.Bokaba.et.al, 2020).

Workflow Diagram of a Proposed System

During the construction of the RTA model, the stages of the experimental process are illustrated in Figure 5. This process consists of five steps. In the first step, the work flow of the proposed system starts with input of dataset (165 records). In the second step, handling missing data is an essential part of the pre-processing data stage that helps to ensure that absent values are

dealt with sufficiently. After data pre-processing, data training and testing was followed by third step.

In this step, the dataset is divided into training dataset and testing dataset, the system uses the train dataset to train with the model, and the test dataset is used to evaluate the predictions. 80% of the dataset is used for the training and 20% is used for the testing and will be randomly split in the same way. By the fourth step, the models are built and analyzed on two separate datasets. In this way, if we want to train another model, we will be able to accurately compare it with another one, because it will be trained on the same data set.

The comparison analysis of the ML methods and finally, the predicted RTAs model evaluated to test the accuracy of each model in the last step. Evaluation of the models show that the best results are obtained and the goal of classification methods is building models that can predict the accidents. These models are built using training sets of data in which accidents factors are known. To investigate the factors influencing the severity of traffic accidents on Yangon-Naypyitaw expressways and to analyze and model accident data to understand better and assess the causes and effects of the severity of traffic accidents. The scores of different models are measured to evaluate and compare their accuracy.

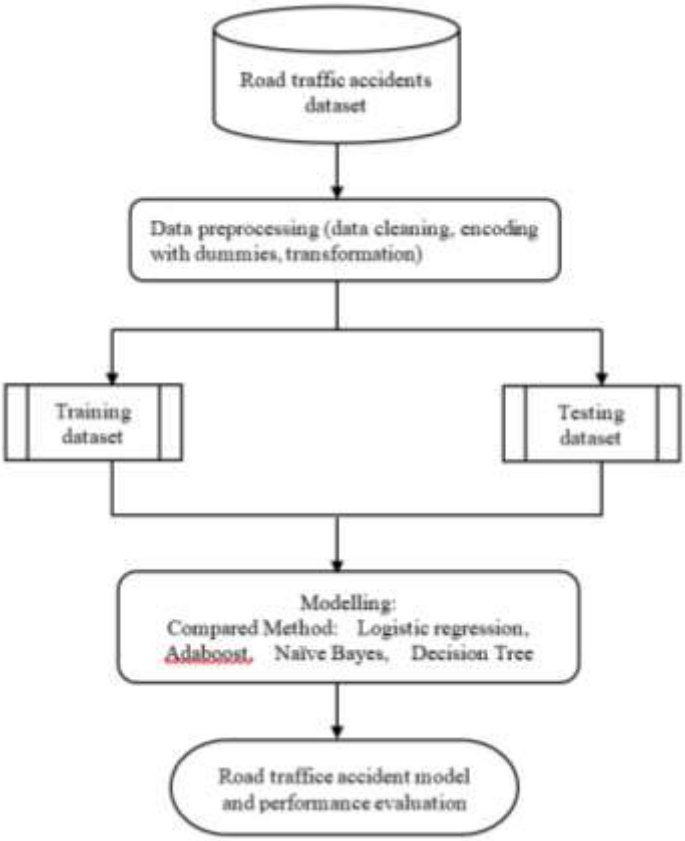


Figure 5. Work Flow Diagram of a Proposed System

Results and Discussion

Visualization of the discovered patterns is important in order to communicate information efficiently using graphs, charts and tables. The paper discusses on data visualization. The data visualization of road accidents is presented with respect to three injuries. The different machine learning algorithms are used to identify the key features of road traffic accident that leads to explore the importance of road accident contributing factors and to evaluate the performance of classification results. The investigation effort establishes the models to select a set of influential factors and to build up a model for classifying the severity of injuries.

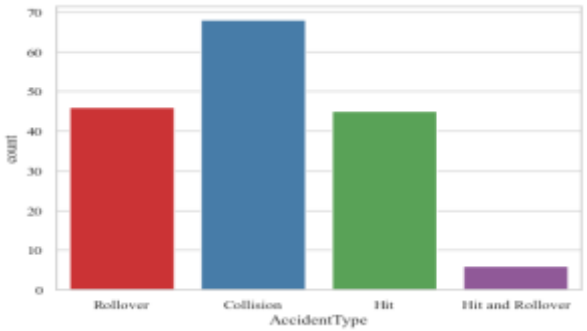


Figure 6: Accident Type

As illustrated in figure 6, it was found that collision is almost 68% - higher rate of accident type. Hit and rollover is 6% - the least accident type. The number of rollover and hit accident type were accounted with 46% and 45% as respectively.

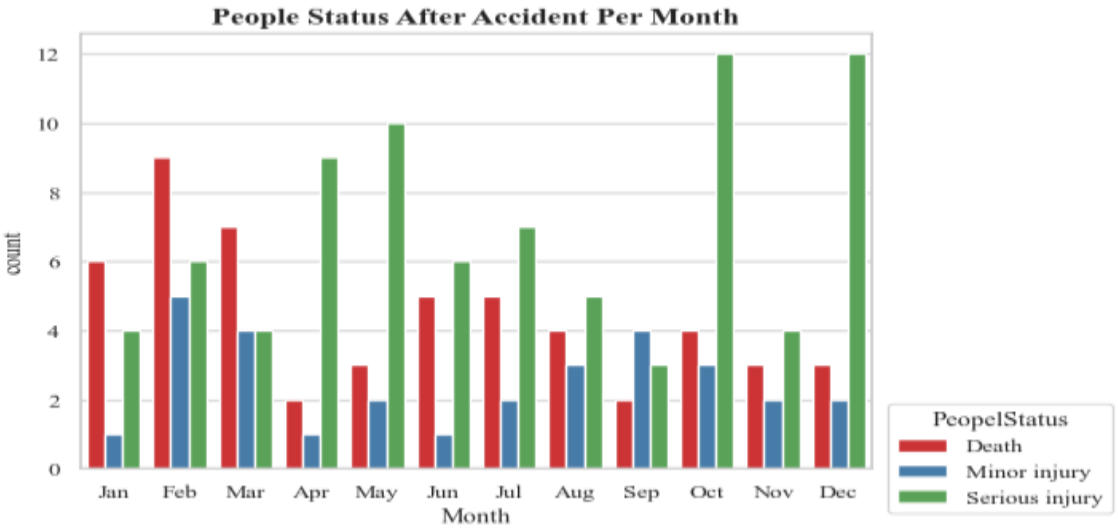


Figure 7: People Status After Accident Per Month

As illustrated in Figure 7, it was found that the serious injuries are caused by 12% in October and December. The higher rate of death case is almost 9% and also minor injury is caused with 5% in February. In Figure 8, when the collision accident occurs, the highest number of people are death by the crashes of vehicles and serious injury that describes 24% and 31%. By the hit and rollover accident type, it is the lowest rate of accident type but the percentage of serious injury and death is equal in 5%.

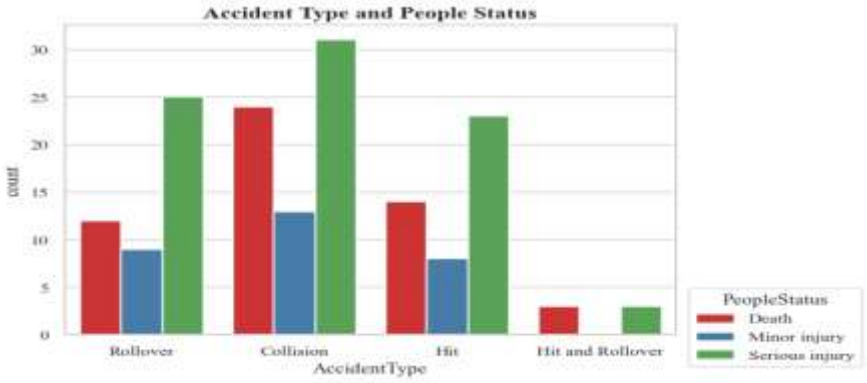


Figure 8: People Status Based on Accident Type

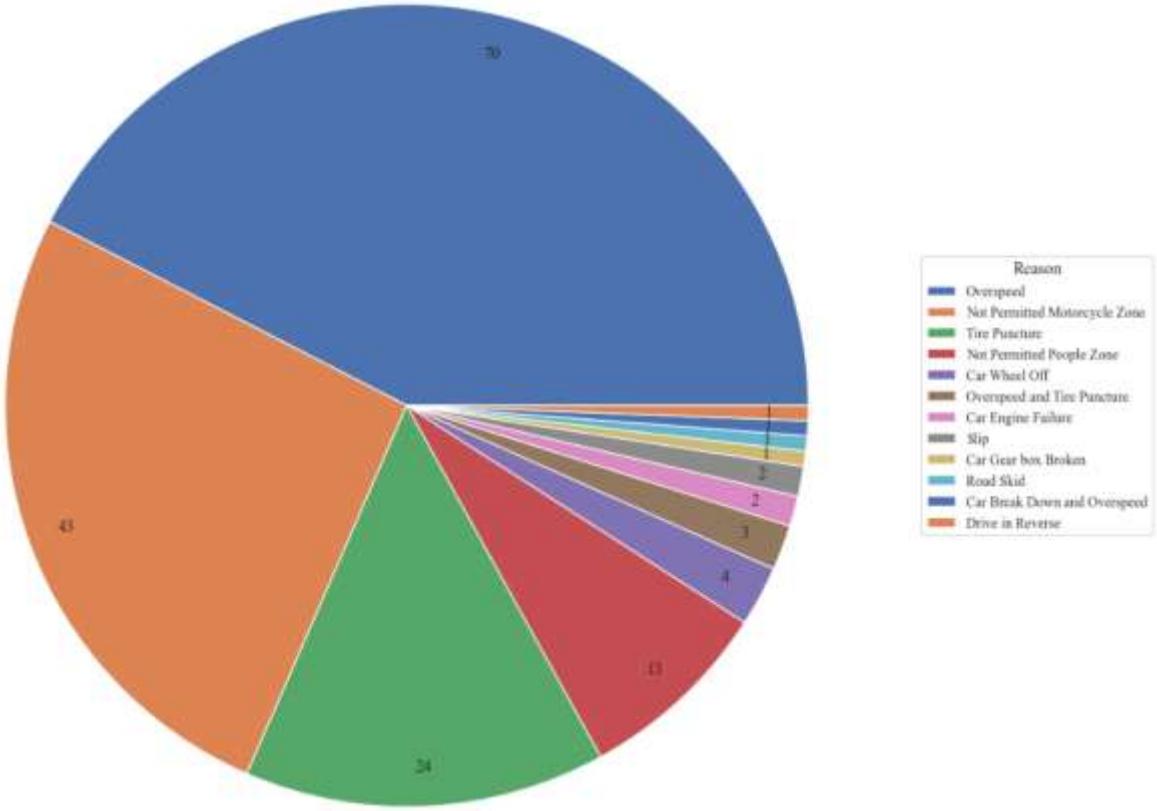
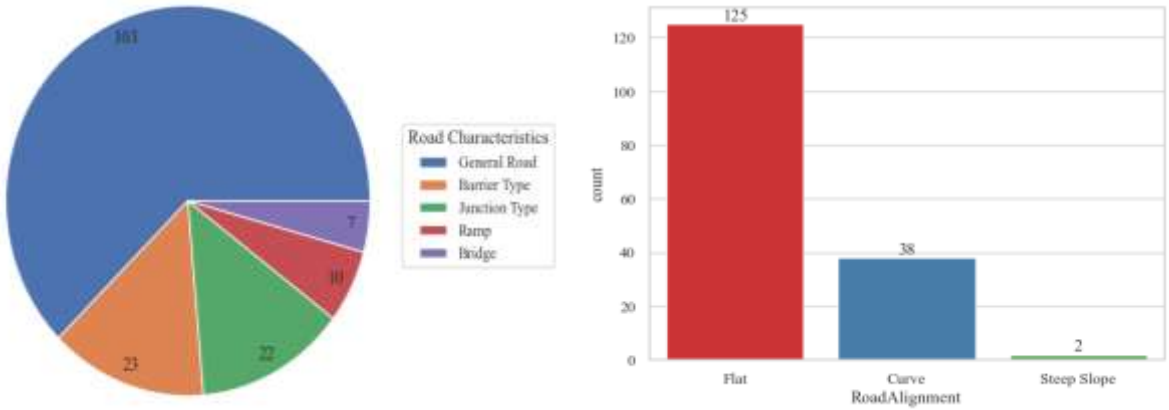


Figure 9: Reason of Accidents Occurred

According to the Figure 9, the reason of accident in which most of accident is happened by the reason of over speed.



(a) (b)

Figure 10 (a) and (b): Traffic Accident Occurred Road Characteristics and Road

According to the road characteristics in Figure 10 (a), since it is based on accidents that occur on the expressway ,most accident is caused on the general road. In Figure 10 (b), road alignment can be seen that the flat alignment is the highest number of accident cases than on the curve and steep slope alignment.

The following the Figures 11, according to the people status based on season, the percentage of serious injuries caused in rainy season significantly increased with 24% in spring and 32% in rainy season, the percentage of death was decreased with the lowest percentage of 13% in summer than in other two seasons.

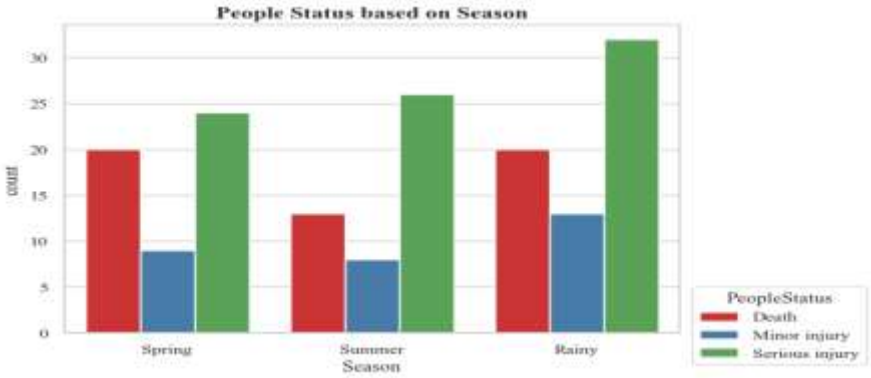


Figure 11: People status based on season

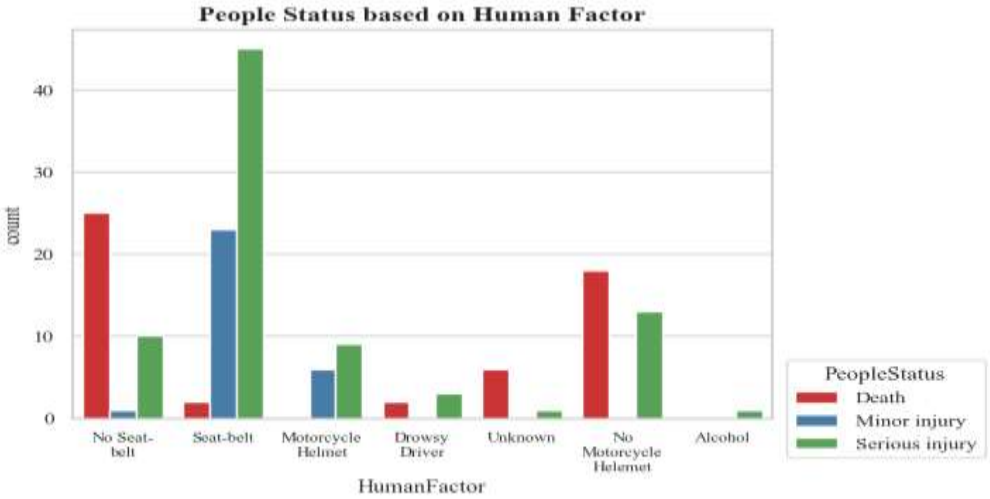
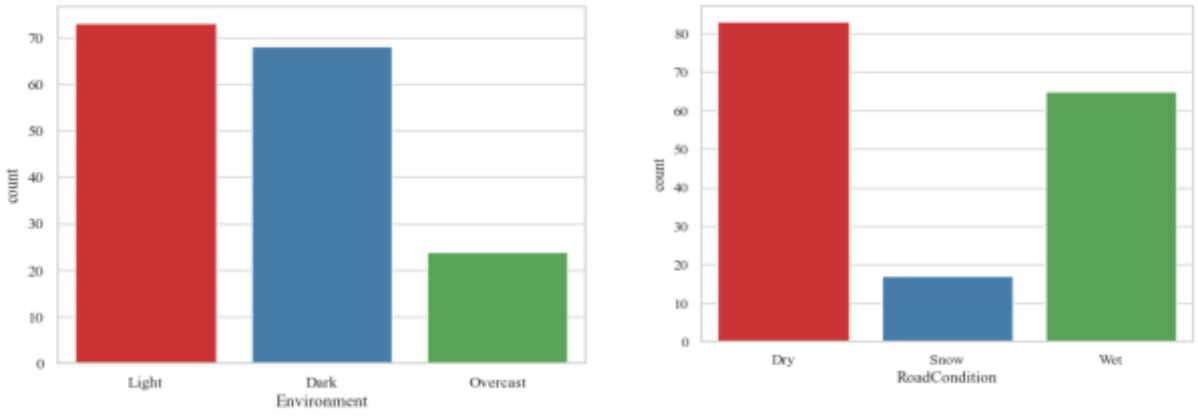


Figure 12: People Status Based on Human Factor

By Figure 12, it is apparent that although most of the drivers wear seat belts, the highest number of serious injuries and minor injuries was achieved with 45% and 23% and the highest number of deaths 25% and 18% were achieved during some drivers are driving without wearing seat belts and without using helmet when driving motorcycle.



(a) (b)

Figure 13 (a) and (b): Traffic Accident Occurred Environment and Road Condition

In the Figure 13 (a), shows that the highest rate of traffic accident causes is light environment with 73% and the lowest rate is overcast with 24 % than dark environment. According to Figure 13 (b), the most accidents cases occur in dry road condition with 83% and the fewest case occurs in snow road condition with 17%. The following Figure 14, shows the highest percentage of death and serious injuries caused in the day-time and night-time which is relatively increased while the percentage is decreased in twilight.

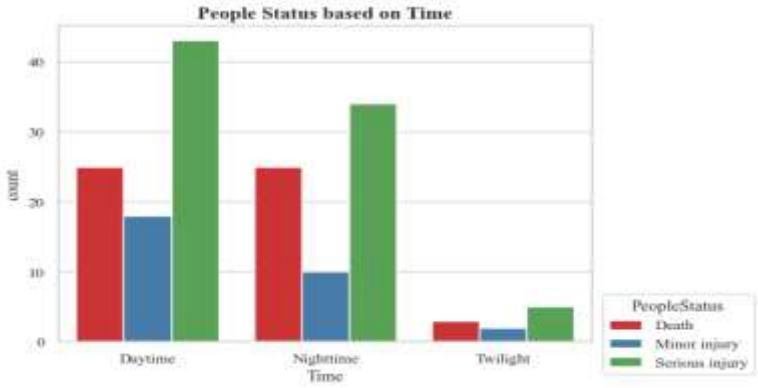


Figure 14: People Status Based on Time

Model Evaluation

The datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. Evaluating a model is the core part of creating an effective model. After building machine learning models by using Logistic Regression, Adaboost, Multinomial Naive Bayes, Decision Tree and Adaboost using Decision Tree, the accuracy of the model is measured to make improvements and continue until achieving a desirable accuracy (J.Wen, 2009).

Table 2: Results of Evaluation Performance

Model	Training Result	Testing Result
Logistic Regression	0.8561	0.8485
AdaBoost	0.7045	0.6061
Decision Tree	0.9470	0.6061
AdaBoost using Decision Tree	0.9697	0.7273
Multinomial Naive Bayes	0.6970	0.7273

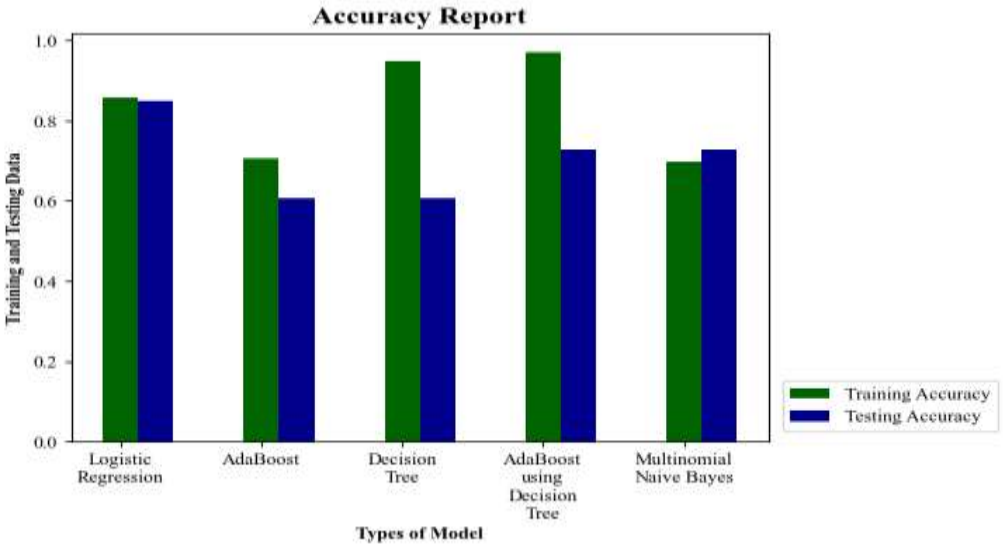


Figure 15: Comparison of Evaluation Performance

According to Table (2) and Figure (15), the experimental result shows that Logistic Regression is superior to the other four models considered. More specifically, the model accuracy for Logistic Regression was 84.85%, while Multinomial Naïve Bayes and AdaBoost

using Decision Tree achieved the second highest accuracy of 72.73%. The test results for both Decision Tree and AdaBoost were identical at 60.61%. As an obvious comparison, Logistic Regression achieved the highest accuracy at 84.85% for testing and 85.61% for training. In contrast, AdaBoost displayed the lowest accuracy, with 70.45% in training and 60.61% in testing, highlighting a notable difference in its performance between the two sets of results. The findings indicate that the most effective machine learning technique for the proposed system is Logistic Regression. This technique exhibits superior performance when compared to other classifiers.

Conclusion

Road accidents are increasing in Myanmar, resulting in 11,004 fatalities, which accounts for 3.05% of total deaths. Reducing fatality and serious injury by 50% are the main targets for most countries. In addition to road accident prevention policy and strategy, it is important to accurately understand and analyze the contributing factors of road accidents and their impacts in order to design safer roads. In this study, 11 factors influencing accident severity are selected and Machine Learning techniques to analyze road accident datasets for continuous variables. It is essential to understand these factors and their influence on a model, which is the primary focus of this research. This research utilized traffic accident data from the year 2022. The five models are used to produce optimal performance and to improve the acceptability of road accident prediction.

This research utilized the dataset of 165 records with 11 selected variables: accident type, people, human factor, reason, season, time, weather condition, environment, road condition, road characteristics, and road alignment. From this analysis, the first result is that collisions are the most common type of accident, and the highest number of deaths in February and serious injuries occurred in December and October. The second results are that most accident happened by the reason of overspeed, general road characteristics and the flat alignment is the highest number of accident cases than on the curve and steep slope alignment. The final conclusion is the highest number of deaths and serious injuries were happened during some drivers' driving without wearing seat belts and helmet and also caused both in the day and at night but the highest rate of accident caused in light environment and dry road condition than others. According to the model analysis, it indicates that Logistic Regression is better than other classifiers, with the highest accuracy in both training and testing outcomes for the proposed system. We intend to continue the analysis as a future scope to add more data to get better results. In further extension, the existing system can be extended with the feature selection process and investigated the performance of this classification problem using different artificial neural networks. For other future extensions, it is suggested to do further analysis in the most prevalent areas of traffic accidents on Naypyitaw -Yangon expressway.

Acknowledgements

I would like to give my warmest thanks to Professor Dr. Wint Pa Pa Kyaw, Department of Computer Studies, University of Yangon for guiding me through all the stages of doing my research. I would like to express my special thanks to Professor Dr. Soe Mya Mya Aye, Head of Department of Computer Studies, University of Yangon for her kind permission to carry out this research. My thanks for my gratitude to U Ye Myint Than, Principal of Co-operative College (Phaunggyi) for giving me a chance to do this research.

References

J. Li1, et.al, (2023), "Predicting the severity of traffic accidents on mountain freeways with dynamic traffic and weather data", Transportation Safety and Environment, ISSN: 2631-4428, DOI:10.1093/ tse/tdad001.

J. Wen, (2009), "Comparison of AdaBoost and logistic regression for detecting colorectal cancer patients with synchronous liver metastasis", International Conference on Biomedical and Pharmaceutical Engineering, ICBPE, DOI:10.1109/ICBPE.2009.5384087.

J.Y. Lee, et.al, (2008), "Analysis of traffic accident size for Korean highway using structural equation models", Accid Anal Prev., 40(6), 1955-1963, DOI: 10.1016/j.aap.2008.08.006.

N. Saxena, (2023), " Analysis of Road Traffic Accident using Causation Theory with Traffic Safety Model and Measures", International Journal for Research in Applied Science & Engineering Technology (IJRASET) , ISSN: 2321-9653; Volume 5 Issue VIII, DOI:10.22214/ijraset.2017.8179.

S Ahmed, et.al, (2023), "A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance", Transportation Research Interdisciplinary Perspectives, Volume 19, 100814, https://doi.org/10.1016/j.trip.2023.100814.

T. Bokaba.et.al, (2020), "Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents", Applied Sciences, DOI: "10.3390/app12020828, https://doi.org/10.3390/ app12020828.

X. Wang, et.al, (2022), "Research on the Prediction of Traffic Accident Severity Based on BP Neural Network", Applied Mathematics, Modeling and Computer Simulation, Vol 30. https://ebooks.iospress.nl/ volume/appliedmathematics-modeling-and-computer-simulation-proceedings-of-ammcs-2022, DOI:10.3233/ATDE221138.