International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 126-133 www.ijaem.net ISSN: 2395-5252

Enhancing Road Safety through Machine Learning: A Case Study on Road Accident Prediction and Prevention

Jacques Ishimwe[1], Nyesheja M. Enan, Phd[2], and James Rwigema, Phd[3]

[1]*Research Scholar, Master of Science in Information Technology, Faculty of Computing and Information Systems, University of Lay Adventists of Kigali, Kigali Campus, Rwanda.*
[2]*Lecturer, Dean of the Faculty of Computing and Information Sciences, University of Lay Adventists of Kigali, Kigali Campus, Rwanda.*
[3]*Lecturer, Head of Masters program, African Center of Excellence in Internet of Things, College of Science and Technology, University of RwandaUniversity*

Date of Submission: 05-09-2023 Date of Acceptance: 15-09-2023

ABSTRACT
This study uses machine learning techniques to predict and prevent traffic accidents, which have a significant impact on individuals and the global economy. The study explores the frequency of accidents on various types of roads, the effect of weather conditions, and the severity of injuries by looking at a wide variety of data, including road type, weather, lighting, accident severity, and vehicle classifications. the study employed Decision Tree Classifier, Logistic Regression, and Support Vector Classifier. To learn a lot about these models' efficacy, the SVC had the highest accuracy, with a range of 89.13% to 89.83%. The Support Vector Classifier (SVC) has continuously demonstrated amazing precision, making it a possible tool for policymakers and safety authorities to pinpoint risk factors and take preventative action to solve concerns with road safety. This study eventually contributes to ongoing efforts using machine learning and data-driven analysis to reduce traffic accidents and save lives.

I. INTRODUCTION

The expansion of nations and populations has resulted in a number of externalities, including an increase in traffic accidents. Every year, traffic accidents cause millions of fatalities in addition to having negative effects on the economy, society, and environment (Philippe, Michelle, & Sara, 2020). One or more of the following factors - people, vehicles, roads, and environments - complexly interact to cause traffic accidents, with people being determined to be both the most significant and most difficult to alter element. A road accident report is always taken, and it includes many accident aspects that may be utilized to further look into the event's potential causes at that specific road stretch. The majority of emerging and underdeveloped nations, however, fall behind the rest of the globe in terms of the accessibility of accurate accident data(Mireille & Jacob, 2023).

The World Health Organization (WHO) estimates that 1.35 million major traffic accidents occur year, gravely injuring 20 to 50 million people globally. If the current trend continues, it will likely move up to the seventh-leading cause of death in the world by 2030 (Shakil, Akbar, Sayan, Mafijul, & Saifur, 2023). Road accidents involving injuries, fatalities, and economic losses are a prominent unintended consequence of transportation systems. 1.35 million people die on the world's roadways each year, according to the World Health Organization. Additionally, road accidents make a considerable contribution to traffic congestion, a critical problem that affects society as a whole(Yetay, Esayas, & Dietrich, 2023).Road geometry, traffic flow, driver characteristics, and the environment around the road all have a major impact on the likelihood of traffic accidents. Numerous researches on hazard location/hotspot identification, accident injury-severity analysis, and accident duration analysis have all been done in order to anticipate accident frequency and study the characteristics of traffic accidents. The mechanism of accidents is the subject of some investigations. Weather and

lighting on the road are other considerations (Swati, Nikita, Pooja, Swapna, & Jayashri, 2013).
Machine learning algorithms provide a cutting-edge method to handle complicated challenges in this era of data-driven judgements. These algorithms offer a tool for calculating and forecasting accident likelihood based on a variety of variables in the context of road safety. Machine learning may be useful in identifying high-risk situations and enabling preventative safety measures by using pattern recognition and predictive modelling (Marwane, El Arbi, Stéphane, & Othmane, 2023).

II. RELATED WORKS

Previously, other researchers tackled the issue of road accidents and here are some papers reviewed to enrich our understanding on this research field. According to Shakil A. , Akbar, Sayan, Mafijul, & Saifur (2023); millions of people worldwide die each year as a result of the escalating problem of traffic accidents. Additionally, they have a substantial financial and economic effect on society. Without considering the many causes that cause them, current research has mostly concentrated on classifying road accidents as a problem and making predictions about them. Their work has taken into account various ensembles of ML models, such as Random Forest (RF), Decision Jungle (DJ), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (L-GBM), and Categorical Boosting (CatBoost), to predict road accidents with varying injury severity. With 81.45% accuracy, 81.68% precision, 81.42% recall, and 81.04% of F1-Score, the comparative findings demonstrate that RF is the best classifier. According to the results, the severity of injuries is greatly influenced by the road type and number of cars involved in an accident. The ML models are retrained using the high-ranked features discovered through SHAP analysis, and their performance is evaluated. The results indicate an increase in the DJ, AdaBoost, and CatBoost models' respective performances of 6%, 5%, and 8%.

As of Saeid, et al., (2023), The study looks at data on road automobile accidents and develops a prediction model by examining key data elements such accident severity, casualty count, and vehicle count. The goal of a preprocessing model is to transform raw data by eliminating useless and missing features, averaging across data characteristics, and removing outliers using the interquartile range. The effectiveness of road accident prediction is investigated using decision trees, random forests, multinomial logistic regression, and naive Bayes, four categorization approaches. Except for naïve Bayes, the results indicate acceptable levels of accuracy for vehicle accident prediction. In order to comprehend the elements impacting road automobile accidents and to highlight important factors for suggesting accident prevention measures, the findings are addressed using a data-driven approach. Finally, methods for building wholesome and sociable towns are offered.Shweta, J, K, & A, (2021), stated that it is a global worry that the number of fatalities from traffic accidents is growing, with a rise of more than 4% per year across all age categories. By 2030, it is predicted that casualty rates would increase alarmingly by 8%. It is unfortunate and irresponsible to allow civilians to die in such instances. A thorough examination must be done in order to address this problem. However, the striking variability of the data on traffic accidents makes it difficult to analyse them. Data segmentation is an essential step in understanding the complexity of this data. As a result, this study suggests the K-means clustering technique as a feasible remedy.The model's secondary goal was to use a Supervised Machine Learning method to extract data, pictures, and hidden patterns. This knowledge may be used to create accident prevention strategies that work. For the purpose of enhancing traffic safety, segmentation and machine learning techniques can be combined. Another work is of David, Jose, & Alexandre (2023); This study has suggested a technique to predict the risk of traffic accidents. The system was created in three stages: data gathering and selection, preprocessing, and algorithmic mining. Using information from the Portuguese National Guard database, accidents that occurred between 2019 and 2021 were covered. The results showed that the most accidents occurred between the hours of 5 and 8 pm, and that rain had the greatest influence on the likelihood of accidents. Additionally, it was shown that when compared to other days, Friday had the highest number of accidents. These findings are essential for those making decisions about how to allocate resources for traffic monitoring in the most efficient way. Finaly we reviewed a work from Francisca, Sanjay, Toochukwu, & Luis (2014), the work stated Road traffic accidents (RTAs), which cause a major loss of human life, are a grave and sad global problem. While RTA rates have decreased in industrialised nations, the same cannot be said for their equivalents in underdeveloped ones. This concerning pattern in developing countries is caused by a rise in the number of cars, a high incidence of novice drivers, and poor upkeep of the road infrastructure. A tailored Artificial

Neural Network (ANN) model is presented in this study for the analysis and forecasting of accident rates in a developing country. The model takes into account important factors including the number of cars, accident incidences, and demographic statistics and uses current data from 1998 to 2010. Along with the feedforward-backpropagation process, activation functions like sigmoid and linear functions are employed. The study of the model's performance shows that it performs better than the existing standard statistical techniques.

Road accident research is done by looking at specific data and posing relevant queries. The current research is responding to the following questions: What Road type either class or surfaces are the riskiest for driving, what lighting and weather conditions are riskiest to drive? What is the annual trend in the number of accidents? And what types of vehicles do accidents frequently?

III. METHODS AND MATERIALS

This study's main goal is to use analytical machine learning techniques to analyze traffic accident data and identify the direct and indirect factors that have a major impact on traffic accidents. In order to do this, a special model was created using two machine learning approaches, Decision Tree Classifier and Logistic Regression and then accuracy of the model was improved by relying on recent and well-structured datasets. The argument for choosing these algorithms is based on the excellent classification accuracy that was attained, and it should be noted that Decision Tree needs less work to prepare the data during pre-processing.

The original dataset has 15 characteristics and 2604 total data points. Based on its kind, each variable in the dataset was categorized as either category or numerical. Dealing with many features, for instance, might affect how well the model works because training time grows exponentially with the number of features. Additionally, it can make over-fitting more likely. Certain incredibly meaningless or superfluous aspects were eliminated to simplify the issue and improve the model's functioning. To deal with missing values that may impede learning, the data were cleaned and pre-processed. Prediction is a supervised machine learning method that excels at assessing predicted data.

Setting aside a portion of the labelled data to evaluate the model's final efficacy is a critical step in the validation process. While dividing the data, it is essential to preserve its statistical properties. The current study divided the labelled dataset into a training set consisting of 80% of the data and a testing set consisting of 20% of the data. The performance of the model was assessed using aaccuracy metric provided by the confusion matrix.

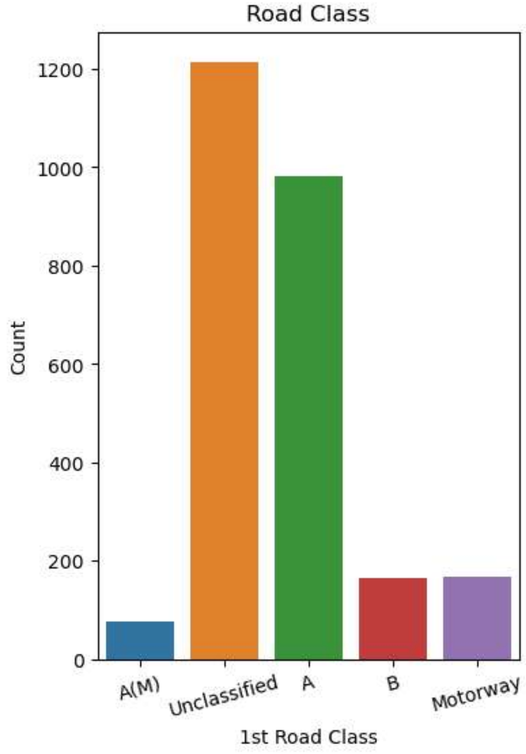
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where used terms in the formula above have to do with binary classification tasks' performance assessment measures. The number of positive cases that are appropriately labelled as positive is referred to as True Positives (TP). The number of cases that are accurately identified as negatives is known as True Negatives (TN). False Positives (FP) are the number of occurrences where something bad is mistakenly categorized as something good. The number of positive cases that are mistakenly labelled as negatives is known as false negatives (FN). Accuracy is just a performance evaluation metric that is calculated using these data.

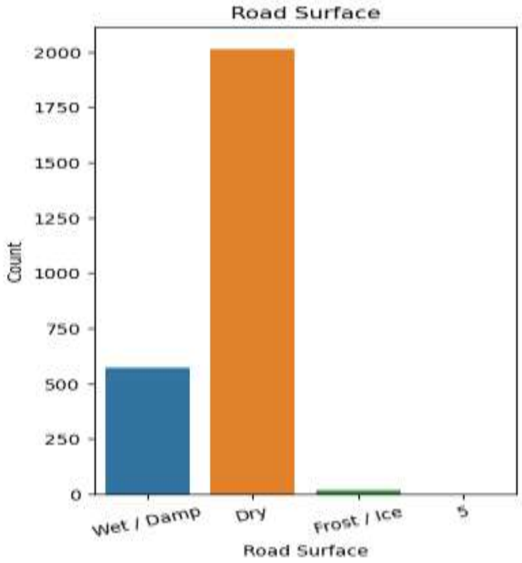
IV. RESULTS AND DISCUSSION

a. Data visualization

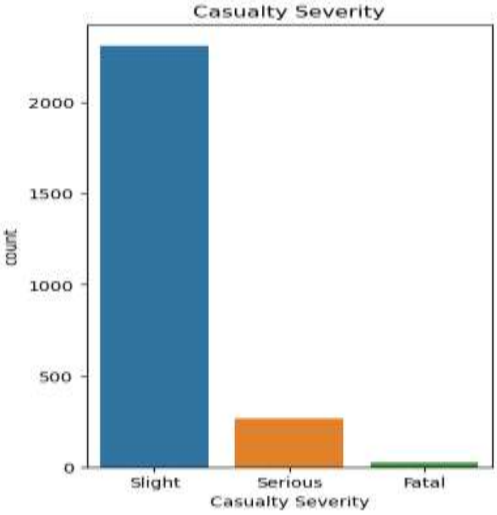
Following the nature of the study, before we answer our research questions, below is the nature of the dataset used; here we are visualizing our dataset to understand the current and past situation of the accidents in the road.



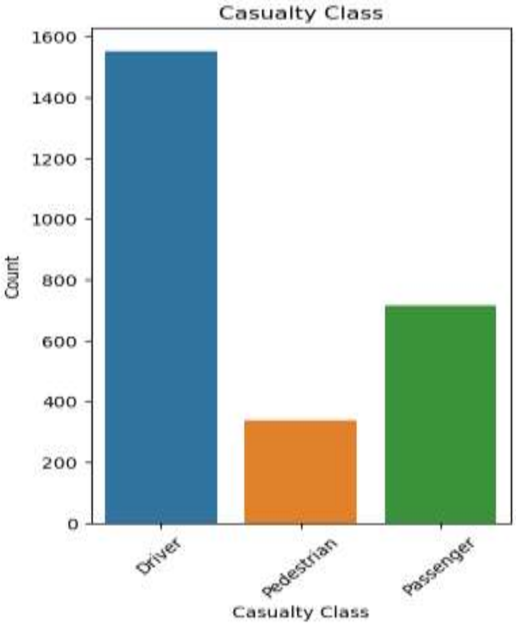
From the above figure, we are visualizing the nature of the roads under where Unclassified roads make up the bulk of roads (1215), followed by Class A roads (981). Less frequently than highways (168), Class B roads are also important (164). Class A(M) roads are the least prevalent type of Class A roads, with only 76 instances of them in the dataset. This implies that the dataset does not contain a significant amount of these big routes, which resemble motorways. The distribution of road categories in the dataset serves as a reminder of the value of having more precise categorization information, particularly for unclassified roads, in order to undertake in-depth network analysis and planning.



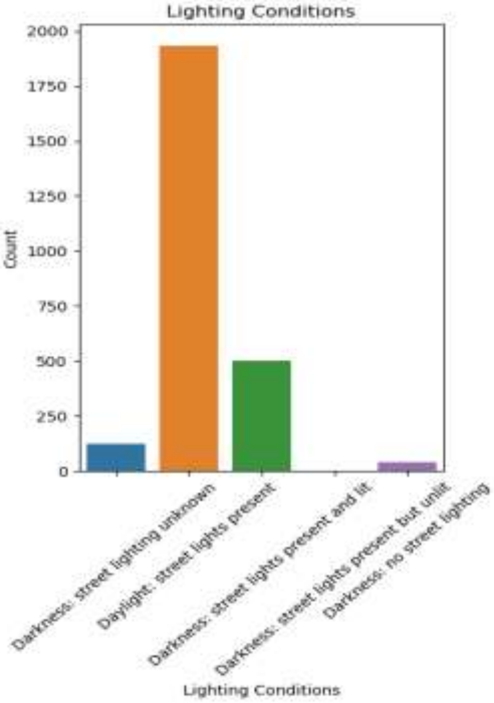
The majority of accidents on the road (77.3%) take place on dry conditions, whereas the risk of accidents is comparatively greater on wet or damp surfaces (21.9%). Additionally, the data shows that accidents on icy or frost-covered roads are uncommon, accounting for only 0.8% of all instances. It is important to highlight that category 5, if applicable, was not a factor in any incidents that were recorded. This study highlights the importance of road surface conditions in accident analysis and emphasizes the need for increased safety measures, particularly during unfavorable weather conditions like wet or icy roads.



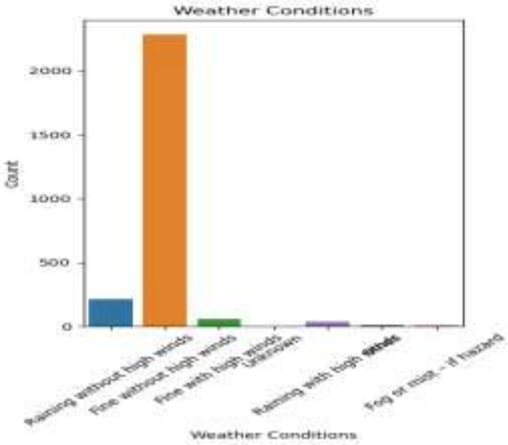
The data reveals that 2,313 instances, or the majority of the casualties, fall into the "Slight" category. This suggests that the occurrences included in the dataset mostly resulted in minimal injury or injuries. But it's vital to remember that there were also 266 incidents that were classified as "Serious," which is a significant amount. This shows that a sizeable but comparatively lesser proportion of events result in serious injuries. There are only 25 examples of "Fatal" casualties in the dataset. Accordingly, fatal accidents were few in the dataset. In conclusion, the data reveals a distribution of injury severity, with the majority of incidences resulting in minor injuries, followed by serious injuries and, mercifully, a tiny number of fatalities.



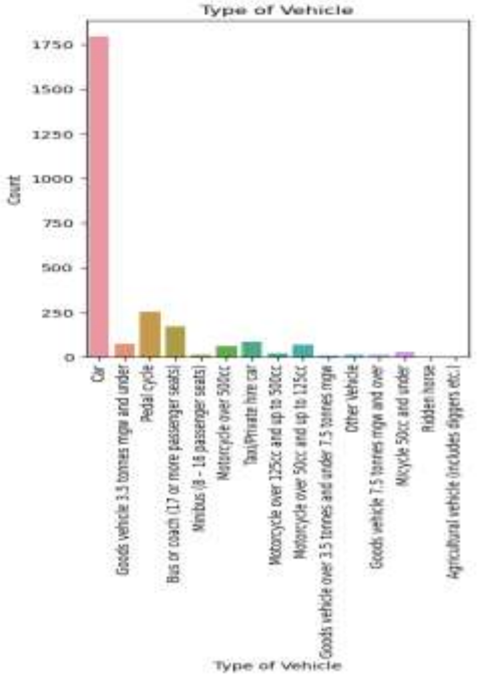
Drivers account for 1552 of the road accident fatalities, followed by passengers (715 instances), and pedestrians (337 cases). This suggests that drivers are the most impacted category, presumably as a result of a higher chance of or involvement in accidents, but passengers are also considerably impacted due to the high number of people who are travelling in cars involved in accidents. A lower percentage of casualties among pedestrians suggests a lesser danger for them. Policymakers and safety authorities must be aware of the distribution of casualties in order to put appropriate road safety measures into place.



The majority of incidents (1,931 instances) happened in daytime with street lights on. The number of incidents that occurred at night while streetlights were on was lower (503 instances). In 126 cases, the illumination was unknown, and in 41 cases, there was no street lighting, therefore the crimes took place in total darkness. Only 3 instances of street lights that weren't illuminated were recorded. It is remarkable that there were just three instances where street lights were there but not working. These numbers provide light on the frequency of different lighting conditions during the incidents that were recorded, with daylight being the most common and various shades of darkness trailing behind.



With 2283 occurrences, the majority of observations throughout the data collecting period show that the weather was "Fine without high winds". With 212 occurrences, "raining without high winds" was the second most frequent occurrence. However, with just 55 and 35 occurrences, respectively, "Fine with high winds" and "Raining with high winds" are less frequent. This shows that there aren't many instances of strong winds in the dataset, whether it's sunny or rainy. In addition, the tiny number of cases classified as "Other," "Fog or mist - if hazard," and "Unknown" circumstances suggests that either these meteorological conditions were not frequently observed during data collection or that it was difficult to describe or standardize how to classify them.



The most prevalent kind of vehicle in road accidents, cars were involved in 1796 events. Pedal bikes were a factor in 255 occurrences, highlighting their weakness. There were 170 events involving buses or coaches with 17 or more passenger seats, showing a substantial number of accidents involving these vehicles. 85 occurrences included taxis or private rental vehicles, while 73 instances involved freight trucks weighing 3.5 tons or less. The dataset keeps track of numerous accidents involving various car kinds. There were 28 occurrences involving bikes under 50cc, 60 involving motorcycles above 500cc, and 71 involving motorcycles between 50cc and 125cc. In addition, there were 11 events involving each other vehicles, goods trucks above 7.5 tons, and goods vehicles between 3.5 and 7.5 tons. There were also 20 motorcycle incidents using engines between 125cc and 500cc, 13 minibus incidents, and 13 incidents involving minibikes. Ridden horses were involved in 2 accidents, while agricultural equipment, such as diggers, was involved in 1. With the use of this distribution, safety measures and policies may be informed about the kinds of cars that are frequently engaged in reported road mishaps.

b. Predictive models results analysis

Accidents, whether mild or serious, can fundamentally alter a person's life. They may cause financial loss, misery, or long-term impairments. Even though car accidents are frequent, you shouldn't ignore the need of becoming informed. That is why we still need to contribute to the field by providing the insight on how accidents can be prevented following the predictions.From the above sections, it was stated that three classifiers were involved in building models in this study, Decision Tree Classifier, Logistic Regression, and Support Vector Classifier. Under python environment, needed libraries such as NumPy, Pandas, Matplotlib, Seaborn, etc. were loaded as well as the dataset. The dataset was split into two sets, training and validation set. Accuracy metric was used to measure the performance of all the models against the dataset. Three trials were conducted to avoid being biased in our conclusions.

\

Table 1 - Models performance, accuracy review

#	**Decision Tree Classifier**	**Logistic Regression**	**SVC**
1	83.69	88.53	89.83
2	84.64	88.53	89.83
3	88.53	83.25	89.13

Three classifiers, Decision Tree Classifier, Logistic Regression, and Support Vector Classifier, were examined in order to increase road safety by using machine learning to predict and prevent traffic accidents. The results showed that the models' degrees of accuracy varied. The Decision Tree Classifier performed best in the third iteration and displayed accuracy ranging from 83.69% to 88.53% throughout the course of the three iterations. Throughout all iterations, Logistic Regression maintained a constant accuracy of 88.53%. The accuracy of SVC ranged from 89.13% to 89.83%, with the third iteration achieving the maximum accuracy. According to the study, the three classifiers perform effectively, particularly the SVC, which constantly displays high accuracy. This demonstrates its potential as a solid tool for forecasting and averting traffic accidents within the parameters of the research.

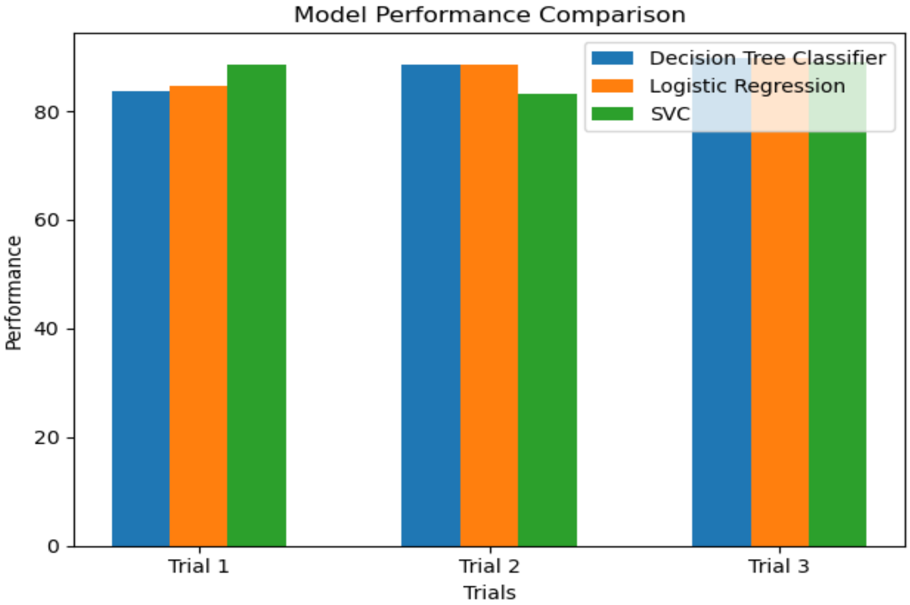


Figure 1 - Comparative analysis of classifiers in research

V. CONCLUSIONS

In order to solve the critical issue of road safety, especially in forecasting and preventing accidents, machine learning techniques have been deployed. These mishaps have a big worldwide influence since they cause a lot of lives, bad economic luck, and social repercussions. Accidents can be caused by a variety of variables, but the research emphasizes the value of exact accident statistics, especially in developing nations. The usefulness of different machine learning models, including Random Forest, Decision Jungle, Adaptive Boosting, Extreme Gradient Boosting Machine, Light Gradient Boosting Machine, and Categorical Boosting, in forecasting traffic accidents was examined in prior research. The findings demonstrated the effectiveness of data-focused techniques in enhancing road safety by demonstrating that these models were able to appropriately identify the severity of injuries.

The current study looked at a dataset with factors including road type, weather, illumination, accident severity, and vehicle kinds using Decision Tree Classifier, Logistic Regression, and Support Vector Classifier. This data's analysis revealed important information on the frequency of incidents on unclassified routes, the effect of weather, and the distribution of injury severity. It also clarified the different kinds of cars that are engaged in collisions. The use of these prediction models is crucial for understanding and reducing the impact of accidents. They have the capacity to inform policymakers and safety authorities, drive preventative safety actions, and detect high-risk scenarios by precisely analysing risk variables. This research contributes to continuing efforts to increase road safety, reduce accidents, and ultimately save lives on our roadways by using machine learning and data analysis.

Three classifiers, Decision Tree Classifier, Logistic Regression, and Support Vector Classifier, were examined in order to increase road safety by using machine learning to predict and prevent traffic accidents. The results showed that the models' degrees of accuracy varied. Throughout all iterations, Logistic Regression maintained a constant accuracy of 88.53%. The accuracy of SVC ranged from 89.13% to 89.83%, with the third iteration achieving the maximum accuracy. SVC is a more reliable alternative for forecasting road safety due to its higher accuracy, which demonstrates its capacity to identify patterns and relationships within the information. These results suggest that SVC is a good candidate for further investigation and implementation in practical machine learning applications focused at enhancing road safety.

REFERENCES

[1]. David, D., Jose, S. S., & Alexandre, B. (2023). The Prediction of Road-Accident Risk through Data Mining: A Case Study from Setubal, Portugal. MDPI - Informatics, 1-10.

[2]. Francisca, O., Sanjay, M., Toochukwu, C. O., & Luis, F.-S. (2014). An Artificial

Neural Network Model for Road Accident Prediction: A Case Study of a Developing Country. Acta Polytechnica Hungarica, 177-197.

[3]. Marwane, B.-l., El Arbi, A. A., Stéphane, C. T., & Othmane, N. N. (2023). A Quantitative Approach to Road Safety in Morocco: Reducing Accidents through Predictive Modeling. E3S Web of Conferences 418, 02004, 1-4.

[4]. Mireille, M.-T., & Jacob, A. A. (2023). Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector. MDPI - Sustainability, 1-6.

[5]. Philippe, S. B., Michelle, A., & Sara, F. (2020). Machine learning applied to road safety modeling: A systematic literature review. Journal of Traffic and Transportation Engineering (English Edition), 775-790.

[6]. Saeid, P. A., Xiangning, L., Kal, T. M., Richard, S. S., Xuhui, W., Bao-Jie, H., & Ali, C. (2023). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. MDPI - Sustainability, 1-15.

[7]. Shakil, A., Akbar, H., Sayan, K. R., Mafijul, I. B., & Saifur, R. S. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. Transportation Research Interdisciplinary Perspectives, 1-8.

[8]. Shakil, A., Akbar, H., Sayan, R. K., Mafijul, B. I., & Saifur, S. R. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. Transportation Research Interdisciplinary Perspectives, 347-361.

[9]. Shweta, J, Y., K, B., & A, G. K. (2021). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms . Journal of Physics: Conference Series, 1-9.

[10]. Swati, M., Nikita, S., Pooja, P., Swapna, B., & Jayashri, S. (2013). Road Accident Prediction Model Using Machine Learning. International Journal For Research in Applied Science and Engineering Technology, 1-7.

[11]. Yetay, B., Esayas, A., & Dietrich, S. (2023). Examining Car Accident Prediction Techniques and Road Traffic Congestion: A Comparative Analysis of Road Safety and Prevention of World Challenges in Low-Income and High-Income Countries. Hindawi, 1-8.