Road Accident Prediction using Machine Learning Approaches

Teres Augustine
Department of Data Science
Christ University
Banglore, India
teres.augustine@science.christuniversity.in

Samiksha Shukla
Department of Data Science
Christ University
Banglore, India
samiksha.shukla@christuniversity.in

2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-6654-3789-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACITE53722.2022.9823499

Abstract— **Road accidents create a significant number of serious injuries reported per year and are a chief concern of the world, mostly in underdeveloped countries. Many people have lost their near and dear ones due to these road accidents. Hence a system that can potentially save lives is required. The system detects essential contributing elements for an accident or creates a link among accidents and various factors for the occurrence of accidents. This research proposes an Accident Prediction system that can help to analyze the potential safety issues and predict whether an accident will occur or not. A comparative study of various Machine Learning Algorithms was conducted to check which model can help predict accidents more accurately. The dataset used for this paper is the government record accidents that occurred in a district in India. Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor, XGBoost, and Support Vector Machine are among the Machine Learning models used in this paper to predict accidents. The Random Forest algorithm gave the highest accuracy of 80.78% when the accuracies of the Machine Learning models were compared.**

Keywords— Machine Learning, Accident Prediction, Vehicle Safety, Traffic

I. Introduction

Road accidents account for a significant share of serious injuries reported every year. However, determining which exact circumstances led to such incidents can be difficult, making it more difficult for local law enforcement to handle the number and severity of road accidents. Yet it is not impossible. Therefore, a system that can predict the occurrence of traffic accidents or accident-prone areas can warn citizens in advance and potentially save lives. A vehicle safety system that detects the possibility of a collision is known as an Accident Prediction System (A.P.S.). It alerts the driver if a crash is impending. Road authorities, designers, and practitioners need prediction tools, also known as Accident Prediction Systems, to improve Road Infrastructure Safety Management. These tools help the concerned authorities to know, learn, understand and analyze the issues concerning the safety of the citizens. They also help identify solutions for improving safety and increasing potential effects in terms of cash reduction. The A.P.S. used in this paper was based on the predictions by various Machine Learning Models.

Traffic accident prediction is not impossible, despite its difficulty. Accidents do not occur randomly; they are impacted by various variables, including drivers' physical status, vehicle kinds, driving speed, traffic situations, road structure, and weather. Studying historical accident records would allow us to understand better the links between these parameters and road accidents, which helps to develop an accident predictor.

Machine Learning algorithms are used for predictions because they allow organizations to make highly accurate estimates about the expected result of a query based on past data, which can be about anything from fraud detection to spam detection to disease likelihood. These give businesses valuable information that they may use to grow their business. In this paper, given an accident dataset, various Machine Learning models are implemented that can give us better accuracies to predict whether an accident will occur or not, thus helping us to save lives.

This paper's primary data is the government accident records of a district in India from 2018-2020. Segment 2 discusses the Literature Review of various articles predicting accidents or accident severity. Methodologies used in this paper are discussed in segment 3. Non-accident records were created, equivalent to the number of accident records to balance the data since only accidents were present. Features like the time of the accident, the driver's age, and the vehicle's age were randomly chosen to make the non-accident data. Then the new data was merged with the original data, and the final dataset was created, which had both accident and non-accident records. Once the dataset was ready, the data was first cleaned, pre-processed and then some exploratory data analysis was done on the cleaned data. Segment 4 discusses the results obtained. On the dataset, various Machine Learning models such as Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbor, and XGBoost were implemented, and then the accuracies were compared. The Random Forest algorithm produced the best results, with an accuracy of 80.78 percent. Finally, Segment 5 discusses the study's conclusion and future work.

II. Literature review

In paper [1], the author discussed predicting road accidents by creating a risk score that calculated the possibility of a fatal or significant accident based purely on information received from person and vehicle data. The situational data was then analyzed to establish the severity of the mishap. Driving Score was developed to identify driver risk factors better by considering each driver's unique features. Every motorist would enter information such as their age and vehicle type to receive a risk score reflecting their likelihood of being involved in a severe accident. The model also guided the driver about the key risk variables. If a low-severity accident was witnessed with no causation, the target variable was set to 0, and if it had severe or deadly repercussions, it was set to 1. Logistic Regression, Random Forest, XGBoost, and Optimal Classification Trees were used for training. The Optimal Classification Trees offered the best results.

The researchers in paper [2] showed a trained predicted website allowing users to enter an origin and a destination. It locates the optimal driving path between the two and will enable users to choose the day and time. The system will find locations along the route that are more accident-prone during that period. The U.K. accident dataset from Kaggle was used, and the grouping was done using the D.B.S.C.A.N. clustering technique. It had high speed, the ability to detect arbitrary-shaped clusters, and resilience to outliers. Some supervised algorithms were also used, including Logistic Regression, Random Forest, and SVM. Random Forest provided the most accurate results.

In paper [3], the author discussed creating negative samples using sampling methods; otherwise, the model will be biased towards No Accident Label. The sampling approach was similar to picking a random record and changing the road section, an hour of the day, or day of the year at random. Add the new sample to the list of negative examples if it wasn't found in the accident data. Repeat until many negative samples are obtained (a few times the number of positive models). The ML approach used was Gradient boosting, particularly with the XGBoost library.

The researchers im paper [4] discussed predicting accident severity using Multiple Logistic Regression and Pattern Recognition Artificial Neural Networks as machine learning solutions. These methods were utilized to determine the most influential variables on accident severity and the best technique for accident prediction. The incidents happened between 18 and 24, and the K.I.A. Pride car had the most significant impact on raising the severity of the accidents. Environmental factors, such as lousy illumination and adverse weather, and the dominant role of dangerous and poor-quality automobiles were contributing factors in raising the severity of accidents. The logit model' performed well in terms of prediction accuracy and performance. Performance and sensitivity analyses validated the ANN model's superior ability to predict and forecast future accidents.

In paper [5], the authors discussed assessing and forecasting the severity of traffic accidents using machine learning algorithms. Kaggle was used to acquire three separate road accident datasets. The algorithms used were SVM and Random Forest. Gender of the driver, colliding area, road condition during the accident, vehicle condition, weather conditions during the Accident, Pedestrian, Passenger, Lack of Lighting were the critical features in the study. SVM provided more accuracy in predicting the road accident severity.

In paper [6], the authors discussed forecasting the severity of traffic incidents using the Deep Forest algorithm and the U.K. road safety dataset. The algorithm was compared with several ML algorithms to prove the superiority of the Deep Forest Algorithm by implementing the ML algorithms on the same dataset. The Random Forests technique was used to extract the main aspects of traffic accidents based on the pre-processed data after it had been pre-processed and cleansed. The accident's severity, the month of the year, the hour of the day, the vehicle reference, the vehicle type, the vehicle manoeuvre, the driver's journey purpose, sex, age band, the engine capacity, the propulsion code, the age of the vehicle, the driver's home area type, the day of the week, the speed limit, the light conditions, this research took into account the meteorological conditions as well as the road surface characteristics. The correlation link between all of the data's attributes was then examined. Finally, the Deep Forests algorithm was used to forecast the severity of a traffic collision. K-fold cross-validation was used to avoid overfitting.

In paper [7], the authors discussed methodologies to dig deeper into traffic accidents to identify the severity of the incidents. A vision-based traffic accident detection system was suggested to automatically identify, record, and report traffic occurrences. This model initially retrieved the cars from a CCTV camera's video picture, followed the moving vehicles, and derived data such as the velocity variation rate, location, area, and direction. The algorithm then used the retrieved characteristics to make choices about the traffic accident. The road collision was anticipated using photographs or videos of the accident individuals posted on social media. These photographs were uploaded to the website, and AdaBoost and Neural Network analyzed them. Deep Learning Neural Networks and AdaBoost were utilized to categorize the severity of incidents into Fatal, Grievous, Simple Injury, and Motor Collision. Accident data records were used to develop models, which helped understand the characteristics of numerous elements such as driver behavior, highway conditions, lighting circumstances, weather conditions, etc. The technique was assessed using YouTube footage of vehicle collisions. In comparison to the dataset in this study, prior video-based accident detection algorithms employed a small number of security cameras. Ada-Boost provides the most excellent overall performance because of its iterative categorization of decision trees.

In paper [8], the authors discussed how predictions of vehicular accidents were determined using a Machine Learning system. The dataset utilized was imbalanced, spanning seven years (2012-2019). The data for the projection was taken from Wolaita Zone's 12 districts and three municipal administrations. The experimental findings, model assessment, and performance measurement revealed that the J48 and Rep tree classifiers' F-measure was equivalent, while the Random Forest tree performed poorly. Pedestrians and passengers were the most common victims in the experiment, which found seven most regularly participating vehicles, 20 locations with a high frequency of accidents, with pedestrians and passengers being the most prevalent victims. The J48 tree was chosen as the best model based on performance, and this experiment resulted in the generation of 23 best rules and the identification of the best features. The most common accident victims, automobiles involved in the most often occurring accidents, and black spot regions for frequent accident occurrences were discovered.

In paper [9], the authors discussed how to classify attributes using Random Forest, predict the rate of accident casualties using Linear Regression, and its visualization of the prediction using graphs. The factors responsible for accidents were analyzed by answering some queries relevant to the study. White-box testing was used to generate the test data. The data from the accident was separated into three stages: pre-collision, collision, and post-collision. Each step gave relevant information.

The researchers in paper [10] showed the risk in percentage present in an area based on the current weather conditions and previously collected dataset by creating an analyzing system. Logistic Regression provided greater accuracy in comparison to Naïve Byes. The system used the logistic regression algorithm to determine the amount of risk in terms of percentage present in the particular area. The predictions were made based on constraints like latitude and longitude, road class, speed limit, weather conditions, etc.

III. Methodology

Machine Learning Models can recognize things and separate them into groups or classes. This paper uses a Supervised learning algorithm, as the labels are specified. Since the problem statement is the classification of accident and non-accident cases, classification algorithms have been chosen. The data received was first cleansed and pre-processed. After this, exploratory data analysis was done on the data by finding the correlation of the variables concerning the target variable. The associated variables with the target variable were plotted. Various Machine Learning Models were used for comparison, like Linear Regression, Random Forest, SVM, K.N.N., XGBoost, Naïve Bayes, as to which of them would provide better accuracy for accident prediction. Random Forest Algorithm gave the highest accuracy.

A. *Logistic Regression*

Logistic Regression is a method for estimating the likelihood of a discrete output when we are given an input variable. The most frequent logistic regression models have a binary result, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be employed when there are more than two discrete outcomes. Logistic Regression is a handy analytical tool for determining if a fresh sample fits best into a category in classification tasks. Because components of cyber security, such as threat detection, are classification issues, Logistic Regression is a valuable analytic tool.

B. *Random Forest*

The Random Forest classifier is an ensemble strategy that trains several decision trees simultaneously using bootstrapping, aggregation, and bagging. Numerous independent decision trees are oriented in parallel on different portions of the training dataset. For the final judgment, the Random Forest classifier aggregates the outputs of the various trees. The Random Forest classifier outperforms most other classification algorithms without the risk of overfitting when it comes to accuracy. The Random Forest classifier does not need feature scaling. Random Forest classifier is more resistant to training sample selection and noise in the training dataset. The Random Forest classifier is challenging to read but easily tweaks the hyperparameter.

C. *SVM*

SVMs are one of the most extensively used machine learning algorithms, and they have been utilized to handle a wide range of medical problems. This method manages data nonlinearity, making them ideal for complex epigenetic data. The model produces a hyperplane by maximizing the margin and minimizing the classification error to partition the data set into various classes. It's best to test the model using multiple kernel functions to acquire the best classification results. SVMs also benefit from being less sensitive to input errors and effective at decreasing outliers in general. They reduce the number of residuals beyond a given estimate's range and are less prone to outliers. They need less memory and are more resistant to overfitting due to their underlying structure.

D. *KNN*

K.N.N. is a non-parametric classification approach. It is one of the most used classification techniques in Machine Learning. The basic idea is that available data is ordered in a space specified by the characteristics that have been chosen. The value of k is provided to the algorithm to determine the new data's class. It will compare the types of the k closest data and classify the new data into the category where it finds similar maximum values. The K.N.N. classification offers several advantages, the most noteworthy of which is its ease of use. It is also a very efficient approach. However, despite its efficiency, calculation times may belong to extensive databases. For finding the value of nearest neighbors, i.e.., the value of k, a trial-and-error method needs to be performed. Outliers can significantly influence the efficiency of this method.

E. *XGBoost*

XGBoost stands for eXtreme Gradient Boosting. It's a distributed gradient boosting library designed to be quick, adaptable, and portable. It creates machine learning algorithms using the Gradient Boosting framework. The same technique may handle issues with billions of instances (Hadoop, S.G.E., M.P.I.).

F. *Naïve Bayes*

The Naive Bayes classifier is a classifier that uses a probabilistic approach based on Bayes' theorem. It states that each feature contributes independently and equally to the target class. Each characteristic contributes similarly and independently to the chance of a sample belonging to a particular category. It's easy to use and fast to compute, and it works well with massive datasets with high dimensionality. It is noise-robust and ideal for real-time applications. Because associated characteristics are voted twice in the model, it performs best when deleted, and the importance of the related qualities is overstated.

IV. Result and discussion

Various parameters were used for building the model. The target variable is "Accident," with values zero and one. Zero is no accident, and one is an accident. After analyzing numerous Machine Learning models, it was discovered that Random Forest provided the highest accuracy.

Figure 1 gives the visualization of essential variables.

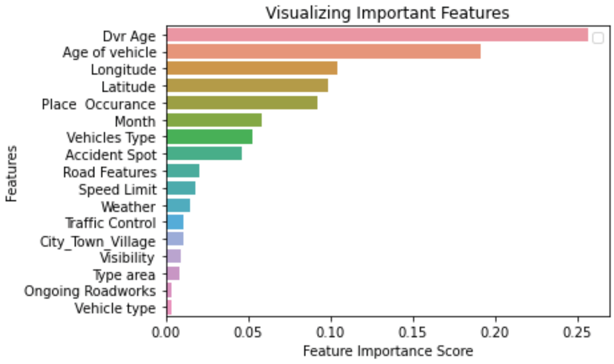

Fig. 1. Visualizing Important Features

Table 1 gives the summary of each algorithm's performance on the accident dataset.

TABLE I. SUMMARY

SNo.	Algorithm	Accuracy	Precision	Recall
1.	Logistic Regression	65.17%	66.71%	66.46%
2.	Random Forest	80.78%	77.75%	86.15%
3.	SVM	67.78%	67.50%	68.35%
4.	KNN	64.17%	63.88%	64.89%
5.	XGBoost	73.19%	71.74%	76.35%
6.	Naïve Bayes	64.71%	63.81%	67.66%

As we can see from the above table, the Random Forest Algorithm gave the highest accuracy, precision, and recall values. Figure 2 shows the confusion matrix of the Random Forest Algorithm.

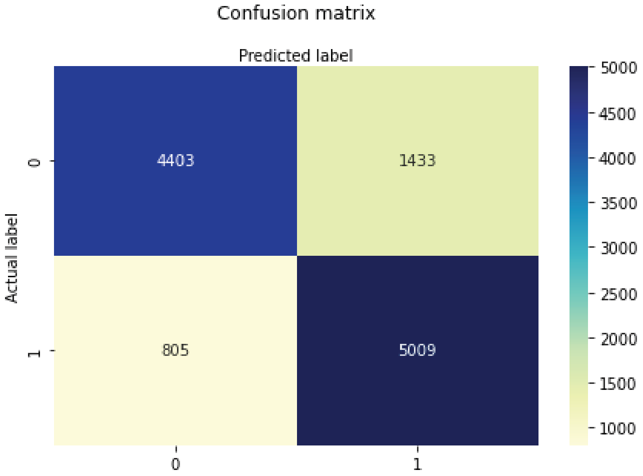

Fig. 2. Confusion Matrix of Random Forest Algorithm

V. CONCLUSION AND FUTURE WORKS

Road accidents cause unfathomable losses to the general people and non-industrial countries like many others. As a result, it has become necessary to control and organize traffic seriously to reduce the number of road accidents. Traffic accidents can be avoided by avoiding possible risks due to the expectations or warnings of a complicated system. Machine learning is a beneficial and outstanding technology for making accurate decisions based on experience to deal with the present scenario. The findings of the investigation phase can be advised to traffic specialists to reduce the number of accidents. The use of machine learning is a practical and effective way to make an accurate judgment based on experience to manage the present situation. The analysis results may be recommended to traffic authorities to reduce the number of accidents. We may apply the presented methodologies to deploy machine learning because of their established and greater accuracy in predicting traffic accident severity. Machine Learning models were used to construct a system that might indicate an accident. People may enter their age, source, and destination into a user interface in the future, and the model will assist in anticipating accidents at that time. This method will help the government reduce accidents, unintended consequences and improve people's safety. This method will also act as a technical means of warning people about and averting traffic accidents.

REFERENCES

[1] Eugenio Zuccarelli, "Using Machine Learning to predict car accidents."

[2] Geraldo Antonio, "Live Prediction of Traffic Accident Risks Using Machine Learning and Google Maps"

[3] Daniel Wilson, "Using ML to predict Car Accident Risk."

[4] Vahid Najafi Moghaddam Gilani, Seyed Mohsen Hosseinian, Meisam Ghasedi, Mohammad Nikookar, "Data-Driven Urban Traffic Accident analysis and prediction using Logit and ML-based pattern recognition method," Hindawi Mathematical Problems in Engineering Volume 2021, Article ID 9974219

[5] Kandasamy Sellamuthu, Akshaya S, Anush J, Arunkumar S, "A Machine Learning Approach to Analyze and Predict the Severity of Road Accidents," Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4241 - 4248 Received 05 March 2021; Accepted 01 April 2021.

[6] Jing Gan, Linheng Li, Dapeng Zhang, Ziwei Yi, and Qiaojun Xiang, "An Alternative method for Traffic Accident Severity prediction, using Deep Forest Algorithm."

[7] Akanksha Jadhav, Shruti Jadhav, Archana Jake, Kirti Suryavanshi, "Road accident analysis and prediction of accident severity using Machine Learning," International Research Journal of Engineering and Technology (I.R.J.E.T.), Volume: 07 Issue: 12 | Dec 2020

[8] Aklilu Elias Kurika, Irfan Ahmad Ganie, Yuliyanti Kadir, Patrick D. Cerna, Price L. Desai, "Predicting Factors of Vehicular Accidents using ML Algorithm," Volume 8. No. 9, September 2020, International Journal of Emerging Trends in Engineering Research

[9] B Mohan Prabhu, Neeraj Chandran, Syed Waseem K, "Predictive Analysis of Road Accidents in traffic violation using ML approach," I.R.J.E.T., VOL: 07 ISSUE: 09 | S.E.P. 2020

[10] Nanditha B, Nidhi Prabhu, Prakruthi M R, Pramatha Nadig H R, Dr. Kavitha K S, "Accident Risk prediction based on Machine Learning," 2020 J.E.T.I.R. August 2020, Volume 7, Issue 8