INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 372-375

Impact Factor: 5.725

ROAD ACCIDENT SEVERITY IN INDIA: A MACHINE LEARNING APPROACH

A. Ramya[1], M. Shanmuga Eswari[2]

[1,2]Dept of Computer Science Assistant Professor, Dept of Computer Science Sri Kaliswari College (Autonomous) Sivakasi, India.

ABSTRACT

One of the world's concerns today is the rate of road traffic accidents. The overwhelming majority of these accidents occur in low and middle-income countries. Road Traffic Accident are one of the leading causes of death in India. This paper adopted the use of four modelling techniques, Randomforest, Decision Tree, Logisitic Regression and XG Boost for short-term road accident forecasting. The comparison between Random Forest, Decision Tree, Logisitic Regression, XG Boost showed that the Logistic Regression 87% is better than the Random forest 84%, Decision Tree 74%, XG Boost 86%.The data used to evaluate the models was obtained from the Kaggle. Prediction positively influences safety enhancements and regulation formulation to prevent future accidents.

Key Words-Traffic Accident, Logistic Regression, Random Forest, XG Boost, Decision Tree.

1. INTRODUCTION

Exploring the influence of accident-causing elements and implementing efficient ways to limit the number of accidents is a pressing issue.In recent years, researchers have been examining the impact of several factors on traffic accidents, with a specific emphasis on people, cars, roads, and the environment. A high and steep roadbed will harm traffic safety, according to research on how road conditions affect collisions. One of the most significant areas of traffic safety research is the prediction of road accidents.The geometry of the road, traffic flow, the traits of the drivers, and the surrounding area of the road all have an impact on the frequency of traffic accidents. Numerous research, such as those on the identification of dangerous locations and hotspots, the analysis of accident injury-severities, and the study of accident length, have been carried out in an effort to forecast accident frequencies and analyze the features of traffic accidents. A few research concentrate on the causes of accidents. The road's lighting and weather are additional variables.

2. LITERATURE SURVEY

Gu et al., (2017) applied Support Vector Machine (SVM) to predict fatal road traffic in China. This research aimed to apply comparative study between SVM, K Nearest Neighbor (KNN) and Bayesian network. The results showed that the prediction model of traffic fatalities based on particle swarm with mutation optimization-SVM obtained higher prediction precision 97% and smaller errors 9% in training and testing data . A.Qasem Al-Radaideh and J. Esraa Daoud (2018) used Decision Tree (Random Forest C4.5/CART/J45), SVM (polynomial Kernel) and ANN back propagation to detect the influential environmental features of RTA in United Kingdom. The experimental results of this study showed that Decision tree (Random Forest) recorded the best accurate result 80.6% in predicting the severity of the accidents in UK . Farhat et al., (2019) applied different data mining techniques tools (Decision Trees and ANN) to predict traffic accidents in Lebanon. The results have shown that ANN using Multi Layers Perceptron (MLP) with 2 hidden layers and 42 neurons in each layer was the best algorithm with accuracy rate of prediction 94.6% and AUC 95.71%. Karthik et al., (2019) applied different data mining techniques methods (J48, Random Forest and Naïve Bayesian) to predict the major causes for fatal accidents in Thanjavur district, India. 10 years accident data containing different RTA factors were collected (Accident Location, Road Bound, Accident Time, Surface Condition etc). J48 registered the highest accurate result 56.96% followed by Naïve Bayesian 54% and the Random forest method 49% .Kumeda et al.,(2019) revealed that lighting conditions, road class and number of vehicles are the key features via fuzzy-FARCHD, random forest, hierarchal LVQ, RBF network (radial basis function network), multilayer perceptron, and naïve Bayes models.

3. METHODOLOGY

3.1. Machine Learning

Logistic Regression: Logistic regression serves as a valuable tool for predicting road accidents by analyzing historical data and relevant features. This process begins with gathering comprehensive historical data encompassing various factors like weather conditions, road type, time of day, vehicle speed, and presence of traffic signals or signage each of which could potentially contribute to accidents. Following data collection, meticulous preprocessing is essential, involving tasks such as handling missing values, encoding categorical variables, and scaling numerical features if necessary. Feature selection is a critical step in this process, aimed at identifying the most influential factors affecting

road accidents. This selection can be informed by domain knowledge, exploratory data analysis, or advanced feature selection techniques. Subsequently, the dataset is divided into training and testing sets to facilitate model training and evaluation. The training set allows the logistic regression model to learn the complex relationships between the chosen features and the likelihood of accidents occurring. Once trained, the model undergoes evaluation using the testing data, where various performance metrics like accuracy computed. This evaluation gauges the model's effectiveness in accurately predicting road accidents based on the selected features. This interpretation aids in understanding the underlying factors contributing most significantly to accidents, thereby informing potential interventions or preventive measures. By employing logistic regression in this manner, stakeholders can gain valuable insights to enhance road safety measures, reduce accident rates, and ultimately improve public safety on roadways.

Random Forest: Random Forest for road accident prediction, a structured approach is crucial. Initially, historical data pertaining to road accidents must be collected, encompassing diverse features like weather conditions, road type, time of day, vehicle speed, presence of traffic signals, and road signage. These features, fundamental to understanding accident dynamics, are pivotal for model development.Data preprocessing is essential to ensure the dataset's readiness for model training. This involves meticulous cleaning, handling missing values, encoding categorical variables, and potentially scaling numerical features for consistency. This step is critical for Random Forest's ability to process and learn from the data effectively.Feature selection follows, aimed at identifying the most influential factors driving road accidents. Techniques such as domain knowledge integration, exploratory data analysis, or advanced feature selection methods aid in selecting pertinent features that contribute significantly to accident prediction.Feature selection, the dataset is partitioned into training and testing sets to facilitate model training and evaluation. The training set is employed to construct multiple decision trees within the Random Forest framework. Each tree is trained on a random subset of features and training data, ensuring diversity and robustness in predictions. During inference, predictions from individual trees are aggregated to yield a final prediction, enhancing predictive accuracy .Performance metrics such as accuracy provide insights into the model's effectiveness in predicting road accidents accurately. Through meticulous application of Random Forest in road accident prediction, stakeholders can gain valuable insights to enhance road safety measures, mitigate risks, and ultimately reduce the occurrence of accidents on roadways.

Decision Tree: Decision trees provide a versatile framework for road accident prediction, enabling stakeholders to analyze historical data and discern patterns indicative of accidents. Beginning with data collection, a comprehensive dataset encompassing various features pertinent to road accidents, such as weather conditions, road type, vehicle speed, and signage, is compiled. Subsequent data preprocessing ensures data cleanliness and compatibility by addressing missing values, encoding categorical variables, and scaling numerical features as needed.

Feature selection plays a pivotal role, guiding the identification of influential factors through domain knowledge, exploratory data analysis, or feature selection techniques. Decision tree models are then trained on the training data..Evaluation of the trained decision tree model using the testing data allows for the assessment of its performance using metrics like accuracy. Leveraging the interpretability inherent to decision trees, stakeholders can visualize the model's decision-making process, discerning the most critical features contributing to road accidents. The trained decision tree model can be applied to predict the probability of road accidents for new instances based on their feature.

XGBoost: XGBoost for this purpose Commence with Data Collection, amassing historical data encompassing a wide array of features such as weather conditions, road type, vehicle speed, and presence of traffic signals. These elements serve as crucial inputs for predicting road accidents. Data Preprocessing, ensure the collected data is clean and formatted suitably for XGBoost model training. Handle missing values, encode categorical variables, and scale numerical features as necessary, ensuring the dataset's readiness.Feature Selection plays a pivotal role in determining the most influential factors in road accidents. Utilize domain knowledge, exploratory data analysis, or sophisticated feature selection techniques to identify pertinent features. Next, proceed with Data Splitting, dividing the dataset into training and testing sets. The training set is employed to train the XGBoost model, while the testing set is reserved for evaluating its performance.Train the XGBoost model by leveraging its ensemble learning technique.Evaluate the trained XGBoost model using the testing data, assessing its performance with metrics such as accuracy. Employ built-in functionalities like early stopping to prevent overfitting and enhance generalization performance.With the model trained and evaluated, proceed to Prediction, leveraging the XGBoost model to estimate the probability of road accidents for new instances based on their feature.By adhering to this structured approach and capitalizing on the strengths of XGBoost, stakeholders can develop robust predictive models for road accident prediction, aiding in risk mitigation and promoting road safety.

3.2 Dataset Used

downloaded the Road Accident in India dataset from Kaggle. This dataset contains 12316 road accident data. There are 32 attributes and 12316 instances in this collection. eighty percent of the training data and twenty percent of the

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 372-375

Impact Factor: 5.725

testing data were separated into two sections of the sample. The proposed technique is implemented using Jupyter notebook and Python. The language Python is open-source. The packages Numpy, Pandas,Scikit-Learn, Matplotlib, seaborn, sklearn. Model _selection, sklearn. metrics, sklearn. linear_model was utilised in this paper. The preferred language for data processing software is Python .

3.3 Preprocessing

Data Cleaning and Data Transformation: Clean up remove duplicates, and remove all rows with null values from the dataset. The dropna()

Method used to remove the rows that contains NULL values.The dropna() method returns a new dataset object unless the inplace parameter is set to True,in that case the dropna() method does the removing in the original dataset instead.

Label Encoding : Label Encoding technique used to convert categorical columns into numerical ones,so that they can be fitted by machine learning models only take numerical data.After applying label encoding ,the Accident_severity column having slight injury element 0 is the label,serious injury element 1 is the label,fatal injury 2 is the label.

Feature selection: In my work in a dataset there are 32 features,and used for most important 25 features.The 25 features : Day_of_week, Age_band_of_driver, Sex_of_driver, Driving_experience, Type_of_vehicle, Defect_ of_ vehicle, Area_ accident_ occuredL anes_or_Medians, Road_allignment,Types_of_Junction, Road_surface_ type,Road_ surface_conditions,Light_conditions, Weather_ conditions, Type_of_collision, Number_ of_ vehicles_ involved, Vehicle_movement, Casualty_class, Sex_of_casualty,Age_band_of_casualty, Casualty_severity, Fitness_of_casuality, Pedestrian_movement,Cause_of_accident, Accident_severity.

4. RESULT

Results shows a comparative study of all the models that were built. These models are evaluated through accuracy.The Table 4.1 below presents the values obtained for Machine Learning models.

Table 4.1. Performance analysis of four Machine Learning Algorithm

Algorithm	Accuracy
Logistic Regression	87%
XGBoost	86%
Decision Tree	74%
Random Forest	84%

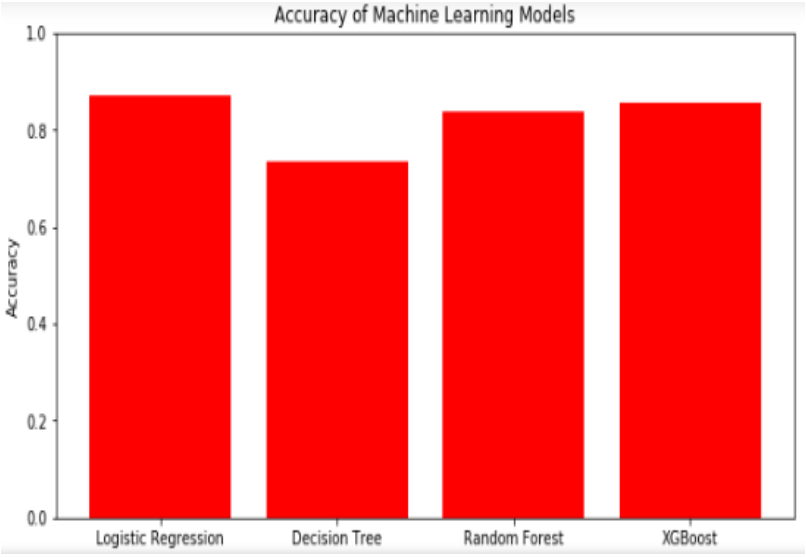

Fig 4.2. Machine Learning Models Accuracy

5. CONCLUSION

This paper aims at using Machine Learning classification techniques to predict the severity of an accident in India. The accuracy of logistic regression is more than other models. The accuracy of RandomForest 84%,DecisionTree 74%, XGBoost 86% are also quite low than Logistic Regression 87%. Therefore, can infer that Logistic Regression do prediction well for dataset.In future, would try to increase our model's accuracy and try to work with a another Machine Learning model for getting more satisfactory results.

6. REFERENCES

[1] Gu. Xiaoning, Li. Ting , W. Yonghui , Z. Liu, W. Yitian , Y. Jinbao , "Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization", Journal of Algorithms & Computational Technology, Volume: 12 issue: 1, pp: 20-29, 2017.

[2] A. Qasem Al-Radaideh and J. Esraa Daoud, "Data Mining Methods for Traffic Accident Severity Prediction", International journal of neural networks and advanced applications, Volume: 5, pp: 1-12, 2018 .

[3] Z. Farhat ,A. Karouni, B.Daya and P.Chauvet, "Comparative Study Between Decision Trees and Neural Networks to Predict fatal Road Accidents in Lebanon", Computer Science & Information Technology, Volume: 9, Number: 11, pp: 01-14, 2019.

[4] D. Karthik, P. Karthikeyan, S.Kalaivani, K.Vijayarekha, "Identifying Efficient Road Safety Prediction Model Using Data Mining Classifiers", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume: 8 Issue-10, 2019, pp: 1472-1474 .

[5] Kumeda, B Zhang, F Zhou, F.Hussain, S Almasri, A Assefa, M. Classification of Road Traffic Accident Data Using Machine Learning Algorithms. In Proceedings of the 2019 IEEE 11th International Conference on Communication Software and Networks , Chongqing, China, 12–15 June 2019; pp: 682–687.