Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 1217–1227
DOI : 10.47974/JIOS-1704

Prediction of road crash attributes using machine learning

Ashwini Bagga [†]
Department of Mathematics and Statistics
Manipal University Jaipur
Jaipur
Rajasthan
India

Sumit Srivastava [*]
Department of Information Technology
Manipal University Jaipur
Jaipur
Rajasthan
India

Rajveer Singh Shekhawat [§]
School of Computing and Information Technology
Manipal University Jaipur
Jaipur
Rajasthan
India

Abstract

Road crashes are a significant global problem, resulting in the loss of over 1.19 million lives and causing around 50 million severe injuries each year. However, despite the efforts made by the Government to improve the situation, no significant results have been observed in respect to decline in number of accidents and deaths in India. Road safety is a complex issue and requires a multifaceted approach including adequate research and development in the field. At present the mostly traditional approach is applied in terms of data collection, statistical analysis and decision making. Over time, the Government may start building advanced data analytics solutions for prediction of various road crash attributes. With this intent this paper aims to depict the application of combination of different algorithms including Support Vector Machine, Random Forest, Decision Tree, Artificial Neural Network,

[†] *E-mail:* ashwinibagga@gmail.com
[*] *E-mail:* sumit.310879@gmail.com (Corresponding Author)
[§] *E-mail:* rajveersingh.shekhawat@jaipur.manipal.edu

© Taru Publications

K Nearest Neighbor, Logistic Regression, and Gradient Boosting Algorithm for prediction of road crash attributes.

Subject Classification: *68T07.*

Keywords: *Machine learning, Deep learning, Classification methods, Prediction of attributes, Road safety.*

1. Introduction

Road crashes are often referred to as man-made disasters that result in loss of over 11.9 lakh lives worldwide annually [1]. Shockingly, India ranks on top in the number of road crashes and fatalities globally, according to World Road Statistics [2]. Many road crash cases go unreported as it involves minor or no injury and many are not considered crash-related if the death occurs after a certain number of days. As a result, the estimated deaths due to road crashes in India are nearly double the reported figures [3], which is disturbing. In India, speeding alone is responsible for 72% of road crashes and contributes to 75% of the total road crash fatalities [4]. The World Bank has estimated that in 2016, India incurred a loss of $172.02 billion, equivalent to 7.5% of the nation's GDP [5]. The impact of road crashes extends beyond the loss of life and contributes to a 3 to 4 percent loss of India's Gross Domestic Product [6].

It is imperative to move beyond traditional analysis approaches and introduce scientific techniques for data collection and define certain key indicators or benchmarks. This would enable the use of scientifically proven data-driven solutions which are essential in addressing this issue effectively. Further, measuring the road safety along with finding the relationship between the various roads crashes and the factors contributing to these accidents. The model performance for the high dimensional data is compared for the utilized ML algorithms. The paper aims to demonstrate use of supervised Machine Learning classification algorithms. The utilization of these classification algorithms on extensive road crash data holds immense potential in aiding strategic planning efforts to mitigate road crashes and fatalities. The earlier research done in this field is described in section 2 followed by ML models in section 3, section 4 provides details of the data with some descriptive statistics applied on it to avoid null values. The result inference is studied in section 5 and section 6 concludes the paper.

2. Literature Review

To explore the potential of machine learning in the road safety domain, a comprehensive review of relevant literature has been conducted and is presented in this section.

Cesar Pedraza and Camilo Gutierrez [7] conducted a detailed review of several machine-learning methods for analyzing, visualizing, and forecasting road crashes. They categorized their work into various segments. They explored techniques like Bayesian Networks, Support Vector Machines, Artificial Neural Networks, and Deep Learning. The authors concluded that combining multiple algorithms or strategies yields the best results. They also emphasized the importance of utilizing heterogeneous data sources for forecasting road traffic and crashes.

Taekwan Yoon, et. al [8] used nine years' worth of traffic crash data from the Neebu Expressway (2007-2015) to investigate the relationship between rainfall and accidents. They found that while overall road traffic accidents decreased, fatalities increased specifically during rainy periods in Seoul, according to the findings of the Korean Transport Safety Authority. Their methodologies involved ANN, Random Forest Algorithms, and Decision Trees. The results showed that the Random Forest algorithm produced precise predictions, although the ANN model tended to overfit the level of accident.

Philippe Barbosa Silva, et. al [9] conducted a comprehensive review of research publications discussing machine learning modeling methods for enhancing road safety. The review determined that ANNs were most effective technique for simulating crash frequency. The study also attained fascinating results for crash analysis by using Classification and Regression Trees (CART) analytics in combination with ANN for assessing accident severity.

Ashwini Bagga et. al. [10] presented a study on machine learning methods for preventing traffic accidents. They discussed various mathematical models developed by researchers, including Artificial Neural Networks (ANN), Naive Bayes Classifiers, Association Rule Mining, and hybrid combinations such as Decision Trees. Ashwini Bagga et. al. [11] applied AdaBoost and Decision Tree Classifier to examine the accuracy levels on different sets of features to build prediction models.

Youssef Balouki, et. al [12] recommended a model based on Artificial Neural Networks (ANN) for predicting flight delays. They developed a multi-layer perceptron (MLP) having one input layer, two hidden layers (each having 50 nodes), and one output layer. The model demonstrated

that departure delays and flight distances were the main contributors to flight delays.

Mahmoud Abdel-Sattar, et. al [13] used an ANN classifier to evaluate the color values, polar and cross diameters, and firmness of nine peach cultivars. Their study concluded that the ANN algorithm can help identify peach cultivars dependent upon color, polar and lateral lengths, and firmness values.

Iliyasu Adamu, et. al [14] developed an English character recognition system using Fourier descriptors and backpropagation in an MLPN model. The experiment demonstrated that this approach achieved a decent recognition accuracy of 95%.

Lumacad, Gernel, et. al [15] developed a tool using learner data to classify learners into modified asynchronous, asynchronous, or synchronous learning styles. The MLP NN model achieved a high classification rate, as indicated by a kappa statistic of 0.83 and an f score of 0.91.

Liefeng Bo, et. al [16] discussed the Multi-layer Perceptron and suggested its potential application to other neural network models. They also highlighted the importance of comparing their algorithm with Biomimetic Pattern Recognition, which could be a focus of future exploration.

Friedman, et. al [17] has identified two aspects to understand learning procedure statistically which are error and structural model. The structural model is worthy because it determines the function and the approximator. The error model determines the optimization criterion for the estimation of the structural model.

J.R. Quinlan [18] demonstrated the technology for building decision trees. He suggested that the advances will help in the real-world applications which have incomplete or noisy data. The decision trees generated by the suggested system are accurate and fast to execute.

A. Booth, et. al [19] developed an algorithm that predicts the price impact of order book incidents. They introduced this concept in trading business. They developed a prediction system using ensemble technique e.g., Random Forest. The system was trained using a large volume of books data from BATS and Chi-X exchange.

G. Seni, et. al [20] suggested that bundling competing models into ensembles almost improves generalization and is an effective method to get diverse components. On the other hand, it increases the complexity of the algorithm as they involve many parameters. They further add that by

employing GDF to a modeling process then ensembles can be simpler than their components.

3. Machine Learning Classification Methods

For writing this paper, classification techniques have been demonstrated using Random Forest, Decision Tree, K Nearest Neighbor, Logistic Regression, Support Vector Machine, Artificial Neural Networks and Gradient Boosting algorithms. A brief description of each algorithm is discussed in the subsequent sub-sections.

a. *K Nearest Neighbor (KNN)*

KNN is used for solving regression and classification problems. It finds the numbers of data points for a given value based on a distance metric which is known as Euclidian Distance. Cross-validation method is helpful in selecting best value of k in a dataset [21]. Euclidean distance is the cartesian distance between the two points which are in a hyperplane which can be defined as:

$$dist(y, Y_i) = \sqrt{\sum_{j=1}^{d} (y_j - Yi_j)^2} \tag{1}$$

Steps for K Nearest Neighbor

1. Select Optimum Value for K
2. Calculate Distance
3. Find Nearest Neighbor
4. Selection of Classification

b. *Decision Tree*

Decision Tree is used for solving both regression and classification problems where it builds a tree like structure where leaf node represents a class, internal node represents a test and branches depict the outcome of a condition. In the application of a Decision Tree, the training data is recursively divided into sub-sets based the attribute values [22].

In a Decision Tree the term entropy refers to the level of uncertainty in the data. Entropy for a subset of a dataset can be defined as:

$$Bi = -\sum_{m\in}^{n} m^{p(i,m)\log_2} p(i,m) \tag{2}$$

where,

- S is a sample dataset.
- m is a particular class from M classes
- p(m) is the ratio of the data points that fit to class m to the total instances of data points in sample S

$$p(m) = \frac{1}{n} \Sigma I(y = m) \tag{3}$$

- In the function p(i,m) shouldn't be equal to 0.

Gini Index, a score which evaluates the accuracy of a split. The Gini Index determines a score that lies between 0 and 1, where 0 depicts that all the observations belong to a class and 1 denotes a random distribution of elements within classes. Information Gain is used to measure reduction in Entropy. The information gain for an attribute A, in relation to a dataset is calculated as depicted below in equation 4:

$$IG(A, B) = B - \Sigma \frac{|B_V|}{B} B_v \tag{4}$$

- A is the class label
- |B| is the entropy of dataset
- |BV| is the number of instances in the subset that have the value v for attribute A

c. *Random Forest*

Random forest comprises of a few individual DTs which are combined as an ensemble. The algorithm to create Random Forest is described as below:

1. Randomly choose **"K"** features from total of "**m**" features, where **k << m**
2. Out of the "**K**" features, compute the node "**d**" using the finest split point
3. Break the node into **daughter nodes** by means of **best split**
4. Repeat steps from **1 to 3** until "l" number of nodes has reached
5. Repeat steps **1 to 4** to build forest for "n" ties to create "n" number of trees

With the created random forest classifier, we will make the prediction.

d. *Artificial Neural Networks (ANN)*

A Neural Network is a computational model that is dependent on the functioning of the human brain [23].

Functioning of a Neural Network

There are different layers in a neural network including an input layer, hidden layer and there could be more than one output layers. There are two stages in the process which are known as forward and backward propagation. The forward propagation comprises of input layer, weights, connections, hidden layers, and output layer. Back propagation consists of a loss calculation function, gradient descent, weights adjustment, training and activation functions. In most cases Mean Squared Error (MSE) is used for calculating the loss or as a cost function, which can be described as:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \tag{5}$$

e. *Gradient Boosting*

Gradient Boosting algorithm is used for solving both regression and classification problems. The prediction of each new model is added to the ensemble. This iteration continues till an endpoint is encountered and further iteration is not possible [24].

4. Modeling Observations

For building the prediction model a combination of various classification algorithms were applied on the target variable Accident Type. The classification process was built on the combination of the various ML algorithms. The idea behind trying the hybrid approach is to use the feature importance scores from the base algorithm to identify the most relevant features for further analysis or building a separate model and more diverse ensemble. Averaging or voting on predictions from these diverse learners can improve the overall model's generalization performance. Random Forest is a widely used supervised algorithm. The risk of overfitting is reduced with hyper-parameter namely, number of trees, node size and the sampled feature. Both Random Forests and Gradient Boosting are ensemble methods that build multiple decision trees. With feature reduced at first level will significantly increase the accuracy percentage. The first ML combination works with Random

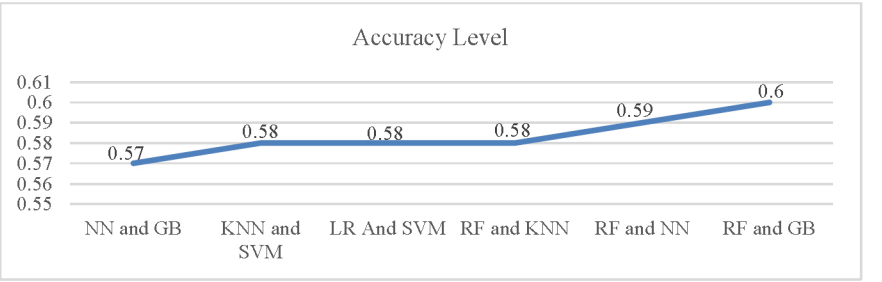


Figure 1

Comparison of Accuracy Level for Combination of Classification Algorithms

Forests as the base classifier followed by KNN, Gradient Boosting, and Artificial Neural Network (ANN) as the next classifier. The results suggest that Random Forest with Gradient classifier shows the highest accuracy values of .60 followed by Random Forest with Artificial Neural Networks as .58.

The next approach tried is Logistic regression acts as an ensemble learner, leveraging the diverse predictions for the other models but provides a single probability output for the final classification. It doesn't directly make predictions but influences the learning process of the Neural Network by penalizing overly complex models that might not generalize well. The data points classified with high confidence by Logistic Regression (e.g., very high probability for a specific class) are considered reliable and passed on. However, Uncertain cases (predictions near the classification threshold) are then fed to a more complex model (e.g., SVM or Neural Network) for a more nuanced decision. Thats why the best result is seen with Logistic Regression combined with SVM for the accuracy score of .58. The interesting facts for both logistic regression and the random forest is that '0' accuracy, recall and precision values for class-4 is seen with most of the ML algorithms. This shows most of the positive prediction were missing in the classification of the data. The KNN and the neural network were also tried as the approach in case any increase in accuracy level achieved. However, the approach chosen initially named Random Forest and Gradient Boosting has achieved the highest accuracy among all the given combination. Figure 1 represents comparison of accuracy level for combination of various classification algorithms.

To predict results based on this data, one might infer that certain types of vehicles are more likely to be involved in specific types of accidents. For example, trucks might be more prone to "Hit From Back"

and "Head On Collision" accidents, while buses might be more prone to "Fall From Vehicle" accidents. However, to make accurate predictions, one would typically need to apply statistical models or machine learning algorithms to the data, considering various factors that could influence the outcomes.

5. Conclusion

The significance of road crashes as a global problem, particularly in India, where high number of crashes and fatalities occur annually. Despite efforts by the government, there has been no significant reduction in the number of accidents and deaths. The paper highlights the need for a multifaceted approach towards road safety, including advanced data analytics solutions and collaboration between government stakeholders and research institutions. The implementation of Integrated Road Accident Database (iRAD) in India is seen as a step towards building advanced data analytics solutions for predicting road crash attributes. The paper also discusses the application of machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree Algorithm for predicting various road crash attributes. It emphasizes the importance of adopting scientific analysis techniques and data-driven methods to provide valuable insights for road safety stakeholders. Aiming for the highest possible accuracy can be achieved through a stacked ensemble approach. Here, you train multiple models (including potentially GBM, ANN, or LR) on the original data, and then use their predictions as features to train a final meta-model. This leverages the strengths of different models but requires more data and computational resources. The inference also suggests that the utilization of machine learning classification algorithms on extensive road crash data holds immense potential in aiding strategic planning efforts to mitigate road crashes and fatalities. Overall, the conclusion underscores the importance of leveraging advanced data analytics and machine learning methods to address road safety challenges effectively.

References

[1] Global Status Report on Road Safety. World Health Organization (2023). https://www.who.int/publications/i/item/9789240086517

[2] IRF World Road Statistics. International Road Federation, Geneva (2020). https://worldroadstatistics.org/

[3] Global Status Report on Road Safety. World Health Organization (2018). https://www.who.int/publications/i/item/9789241565684

[4] Road Accidents in India, New Delhi: Ministry of Road Transport and Highways, Government of India (2022).

[5] Guide for Road Safety Opportunities and Challenges: Low and Middle Income Countries Country Profiles. World Bank, Washington, DC., USA (2019).

[6] What is the road safety crisis? Save Life Foundation, New Delhi. https://savelifefoundation.org/the-problem/

[7] Camilo Gutierrez – Osorio, Cesar Pedraza. Modern Data Sources and Techniques for Analysis and Forecast of Road Accidents: A Review. *Journal of Traffic and Transportation Engineering (English Edition)* (2020).

[8] Jonghak Lee, Taekwan Yoon, Sangil Kwon and Jongtae Lee. Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *MDPI Applied Sciences*.

[9] Bhatnagar, Vaibhav, et al. "Descriptive analysis of COVID-19 patients in the context of India." *Journal of Interdisciplinary Mathematics* 24.3 : 489-504 (2021).

[10] Bagga Ashwini, Srivastava Sumit and Shekhawat Rajveer. Review of the Machine Learning Techniques in Road Crashes. *International Conference on Computation, Automation and Knowledge Management*, 2020, Dubai. 373-376 (2020).

[11] Ashwini Bagga, Sumit Srivastava and Rajveer Singh Shekhawat. Road Crash Data Analysis and Prediction Using Machine Learning Algorithms, *ICCCMLA,* Goa (2022).

[12] Hajar Alla, Lahcen Moumoun, and Youssef Balouki. A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction, *Hindawi Scientific Programming*. Volume 2021 (2021).

[13] Adel M. Al-Saif, Mahmoud Abdel-Sattar, Abdulwahed M. Aboukarima and Dalia H. Eshra. Application of a multilayer perceptron artificial neural network for identification of peach cultivars based on physical Characteristics, PeerJ.

[14] Iliyasu Adamu, Kabir Ismail Umar , Hillary Watsila , Bello Musa Ibrahim. Multi-Layer Perceptron Network for English Character Recog-

nition. *American Journal of Engineering Research*, p-ISSN : 2320-0936 (2017).

[15] Singh, Vijander, et al. "Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine." *Journal of Discrete Mathematical Sciences and Cryptography* 23.8 : 1583-1597 (2020).

[16] Liefeng Bo, Ling Wang and Licheng Jiao. Multi-layer Perceptron with Embedded Feature Selection with Application in Cancer Classification. Institute of Intelligent Information Processing, Xidian University, Shannxi.

[17] J. Friedman, T. Hastie, R. Tibshirani. Additive logistic regression: a statistical view of boosting, *Ann. Stat*. 28 337–374 (2000).

[18] J.R. Quinlan, (1986) Induction of decision trees, *Mach. Learn*. 1 81–106.

[19] A. Booth, E. Gerding, F. McGroarty. Performance-weighted ensembles of random forests for predicting price impact. *Quantitative Financ*. 15 1823–1835 (2015).

[20] G. Seni, J.F. Elder. Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions, Morgan & Claypool, San Rafael (2010).

[21] K Nearest Neighbor Algorithm, https://www.geeksforgeeks.org/k-nearest-neighbours/.

[22] Decision Tree, https://www.geeksforgeeks.org/decision-tree/.

[23] What is a Neural Network, https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/.

[24] Gradient Boosting in Machine Learning, https://www.geeksforgeeks.org/ml-gradient-boosting/.