Road Accident Prediction using Machine Learning Approaches

V. Karuna Kumar.
Assistant Professor.
Computer Science & Engineering
Narasaraopeta Engineering College Narasaraopet, India
karunakumar.valicharla@gmail.com

Mareddy Paramesh Reddy
Computer Science & Engineering
Narasaraopeta Engineering College Narasaraopet, India
mareddyparameshreddy533@gmail.com

Thanugundla Vincent Reddy.
Computer Science & Engineering
Narasaraopeta Engineering College Narasaraopet,India
Vincentreddy25@gmail.com

SK. Mohammed Javeed
Computer Science & Engineering
Narasaraopeta Engineering College Narasaraopet, India
shaikjaveed1088@gmail.com

Abstract: Road safety remains a critical concern for governmental authorities, necessitating proactive measures to address the persisting issue of traffic accidents despite advancements in vehicle safety features. Understanding the root causes of accidents is pivotal for devising effective preventive strategies. Hence, urgent attention is warranted to analyse accident frequency, encompassing factors like current road conditions and environmental variables, to inform mitigation efforts.

In pursuit of a robust accident prediction framework, I employed machine learning techniques including Decision Trees, Random Forests, and Logistic Regression algorithms. The adoption of these methodologies holds significant potential in enhancing safety protocols and facilitating accident prognosis. Factors such as weather conditions, vehicle condition, road surface quality, and lighting conditions are pivotal in predicting accident occurrences. By leveraging comprehensive datasets encompassing accident, casualty, and vehicle information, accurate predictions regarding accident severity can be achieved.

Key terms: Road safety, predictive modelling, logistic regression analysis, influencing factors, machine learning algorithms, random forest methodology.

I. PRESENTATION

Recently, the World Health Organization (WHO) unveiled alarming fatality statistics, shedding light on the profound global impact of traffic accidents annually. Surprising data reveal that car accidents claim the lives of 1.2 million individuals each year, with an additional 50 million sustaining injuries. On a daily basis, an average of 3,300 fatalities and 137,000 injuries are recorded, inflicting a staggering $43 billion in direct economic losses. Beyond financial implications, these incidents pose an imminent threat to human life and property safety due to their alarming frequency.

Predicting road accidents emerges as a pivotal focus in traffic safety research, acknowledging the substantial influence of factors such as road configuration, traffic flow, driver conduct, and environmental variables on accident likelihood. Substantial research efforts have been dedicated to forecasting accident frequencies and unravelling the intricate determinants contributing to traffic accidents. This encompasses endeavours to pinpoint hazardous zones or "hot spots," evaluate the severity of accident-related injuries, and scrutinize accident durations. Additionally, numerous studies aim to dissect the underlying mechanisms of accidents, with weather conditions and road visibility identified as crucial areas necessitating further exploration.

Of particular concern is the escalating incidence of accidents in India, warranting serious attention. Recent data underscores India's contribution to 6% of global traffic accidents, with over speeding and negligent behaviour among two-wheeler riders cited as primary causative factors for the escalating accident rates. A troubling proportion of incidents involve instances of drunk driving or other moving violations. Despite clearly delineated traffic regulations, many accidents stem from individuals disregarding speed limits, neglecting vehicle maintenance, and overlooking safety gear such as helmets.

While the burgeoning vehicle population remains a chief catalyst for these accidents, other pivotal factors such as the state of road infrastructure also exert a substantial influence. Adverse weather conditions such as rain and fog markedly heighten accident risks. Hence, possessing accurate incident estimates and insights into accident hot spots and contributory variables facilitates expedited action towards accident reduction. This underscores the imperative for meticulous incident analysis and the formulation of accident prediction models.

Accurate accident prediction stands as a cornerstone for effective traffic planning and management, given the involvement of nonlinear elements like weather dynamics, human behaviour, vehicle kinetics, and road conditions. Traditional linear analyses fall short in capturing the intricacies of these interactions, particularly in settings characterized by limited data availability and noise interference.

II. CONNECTED WORKS

Deep learning techniques have garnered widespread adoption among researchers across various domains, including text mining, text categorization, fake news detection, and image classification [1]. Similarly, in the realm of data mining, scholars have explored diverse methodologies for analysing traffic accident data. Several studies have investigated the severity of traffic accidents in different countries.

In a recent study, parameter selection methods were employed to identify 16 significant parameters out of a pool of 150, focusing on attributes with the most substantial impact on driver injury severity. Artificial Neural Networks (ANN) were utilized to classify injury severity, yielding a moderate accuracy rate of 40.71%. Regression models, an indispensable tool for data analysis, elucidated the relationship between response and explanatory variables, achieving sensitivity and specificity rates of 40% and 98%, respectively, at a probability cut-off of 0.20. The study underscores the importance of factors such as impact direction, seat belt usage, and velocity in determining accident severity, suggesting the use of fuzzy rule mining to explore high-quality incidents.

The investigation involved predicting three classes using various binary prediction models, resulting in an enhanced model accuracy of 60.94%. Notably, negligent seat belt usage and improper overtaking emerged as pivotal variables influencing accident severity. Sharma et al., utilizing multi-layer perceptron models and Support Vector Machine (SVM) with a limited dataset, attributed incidents to cases of drunk driving at high speeds.

In another study, Tiwari et al. employed machine learning models for classification and clustering, including Decision Tree (DT), Naive Bayes (NB), and Support Vector Machine (SVM). Their utilization of clustered datasets yielded improved outcomes.

The ability to predict the severity of traffic accidents remains an ongoing endeavour. The adoption of effective methodologies promises enhanced forecast accuracy. Selecting the optimal paradigm is crucial for discerning the root causes of traffic accidents. Furthermore, prioritizing components most relevant to the objective could enhance machine learning model performance, enabling more precise predictions of previously unknown events.

III. PROPOSED SYSTEM

- **The suggested system for predicting road accidents**

  A machine learning technique has been developed to assess the severity of accidents based on surrounding conditions. This algorithm, trained on a dataset comprising 1.6 million accident reports spanning from 2005 to 2015, improves in accuracy with the accumulation of additional data. The primary objective of this model is to predict conditions more prone to causing accidents, thereby enabling timely intervention and preventive measures.

This section elucidates the methodologies employed estimate accident probabilities from the dataset, with a specific focus on the implementation of classification algorithms. Random Forest, Decision Tree, and Logistic Regression algorithms were utilized to categorize the dataset for traffic accident prediction. Additionally, three different algorithms were explored for predicting accident severity. Results indicated that Random Forest and Decision Tree algorithms outperformed each other significantly in forecasting all classes of accident severity. Although Logistic Regression exhibited higher accuracy, it may not necessarily be the optimal approach for this task. In the hyperparameter tuning phase, the authors experimented with multinomial techniques to anticipate all classes.

$$p = \frac{e^{\alpha+\beta_n X}}{1 + e^{\alpha+\beta_n X}}$$

Despite concerted efforts, the models could only predict occurrences belonging to a single, more prevalent class. The analysis yielded the following accuracy outcomes: 86.23% for Logistic Regression, 75.26% for Decision Trees, and 86.86% for Random Forests, with the latter exhibiting superior performance. Pre-processing techniques, including photo enhancement, were employed by the system.

Please refer to Figure 1 for the block diagram illustrating the proposed system.

.

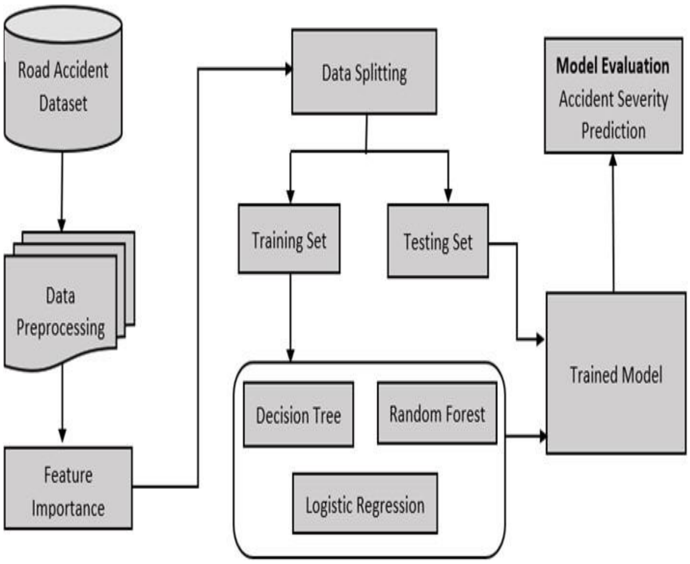


Figure 1: The recommended system's block diagram.

Importing Data

1) All necessary package imports have been completed.

2) Three CSV files were utilized:

(i) Vehicles.csv
(ii) Casualties.csv
(iii) Accidents.csv

3)Pandas was employed to import the data into a data frame

Using applied machine learning methods

1) Logistic regression: It is a supervised classification algorithm designed to predict outcomes with two possible values, such as true or false, zero or one, or yes or no. It estimates the probability that a binary dependent variable will be predicted based on independent variables within the dataset. While logistic regression shares similarities with linear regression, it produces a curve instead of a straight line. Utilizing one or more independent variables, or predictors, logistic regression generates logistic curves that denote values between zero and one. This regression model, known as logistic regression, examines the relationship between multiple independent variables and a categorical dependent variable.

2) Decision tree: It is an approach that constructs a tree-like graph or model of decisions and potential outcomes by employing conditional control statements to predict the final decision. Represented by a learned function, a decision tree is a technique used to address discrete-valued target functions. These algorithms have proven effective across various tasks and are renowned for their support of inductive learning. Upon comparison with the transaction value, a route from the root node to the output or class label of the transaction is depicted in the decision tree. This process is applied to each new transaction to determine its authenticity or fraudulent nature.

$$Entropy(S) = \sum_{i=1} -p_i \log_2 p_i$$

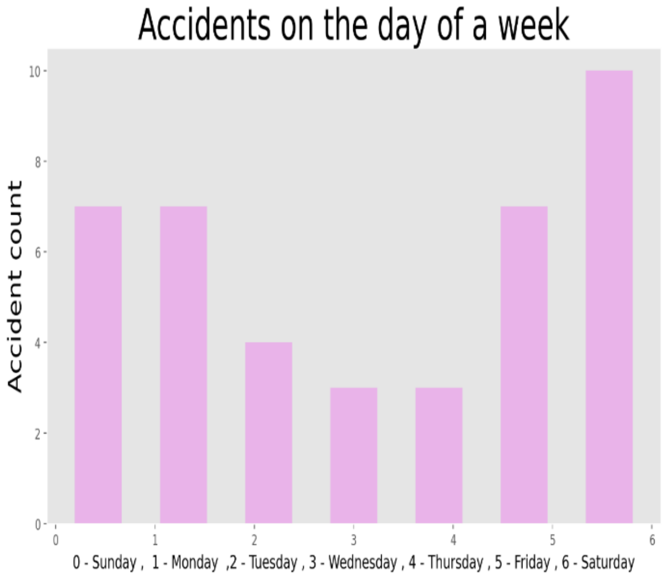
$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

3) Random Forest: It is a versatile method utilized for both regression and classification tasks. Essentially, it comprises an ensemble of decision tree classifiers. Unlike decision trees, random forests mitigate the risk of overfitting to the training set. After constructing a decision tree, each node is split based on a randomly selected feature from the entire feature set. Additionally, each tree is trained using a subset of the training data. Due to the independent training of each tree, random forests exhibit rapid training, even with large datasets featuring diverse characteristics and data occurrences.

$$\frac{1}{X} \sum_{x=1}^{x} f_x(\dot{R})$$

Notably, the Random Forest method has demonstrated resilience against overfitting, contributing to its efficacy in various applications.

4) Adjusting hyperparameters (HP): It can significantly impact the performance of machine learning algorithms in making predictions [9]. HPs are typically configured through trial and error, and finding an optimal set of values manually can be time-consuming, especially considering the duration of training required for machine learning algorithms. Consequently, recent efforts in HP tuning for machine learning algorithms have focused on improving HP tuning techniques [5].



In this process, the objective function of the algorithm is to predict accuracy, and the HP tuning process is often perceived as an optimization problem, akin to a black box.

Measures for Execution Assessment

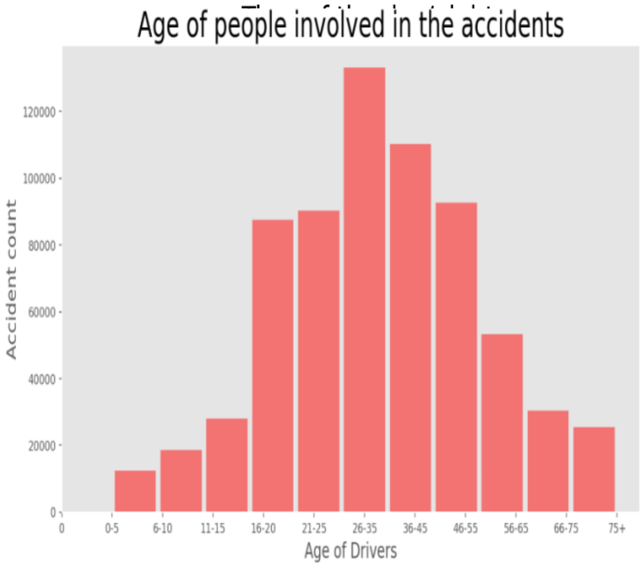
(i) Accuracy: Accuracy is a critical metric in the evaluation of classification models. It represents the correctness of predictions.

In binary classification, accuracy can be further expressed in terms of positives and negatives.

(ii) Precision: Precision measures the accuracy of a classifier by indicating the percentage of all instances with a positive label that are correctly identified as positive.

(iii) F1-Score/F-Measure: The F1-score, or F-measure, is a statistical measure used in classification that takes into account both precision and recall of the classifier to produce a score between 0 and 1. It is calculated as the harmonic mean of precision and recall, emphasizing the importance of both metrics:

(iv) Recall: Conversely, recall shows the percentage of true positive tuples that are correctly categorized and is sometimes referred to as the measure of completeness.



❖ **Preparing data**

(I) Data Cleaning: Data cleaning involves identifying and removing useless or noisy data from the dataset. This process helps improve the quality of the data and ensures that only relevant information is used for analysis. Visualization techniques can also be employed to determine the importance of different aspects of the data.

(II)Data Visualization: Visualizing the data involves examining the crash dates, times, and the ages of drivers involved in accidents. By analysing the frequency of accidents based on days of the week, time of day, and driver age, valuable insights can be gained into the patterns and trends within the dataset

➢ **Weekday Accident Analysis:**

Analysing the dataset from 2005 to 2015 reveals that Thursday experienced the highest number of accidents compared to other days of the week. However, it is crucial to consider that the volume of traffic on a particular day may influence the frequency of accidents.

➢ **The Accidents Time**

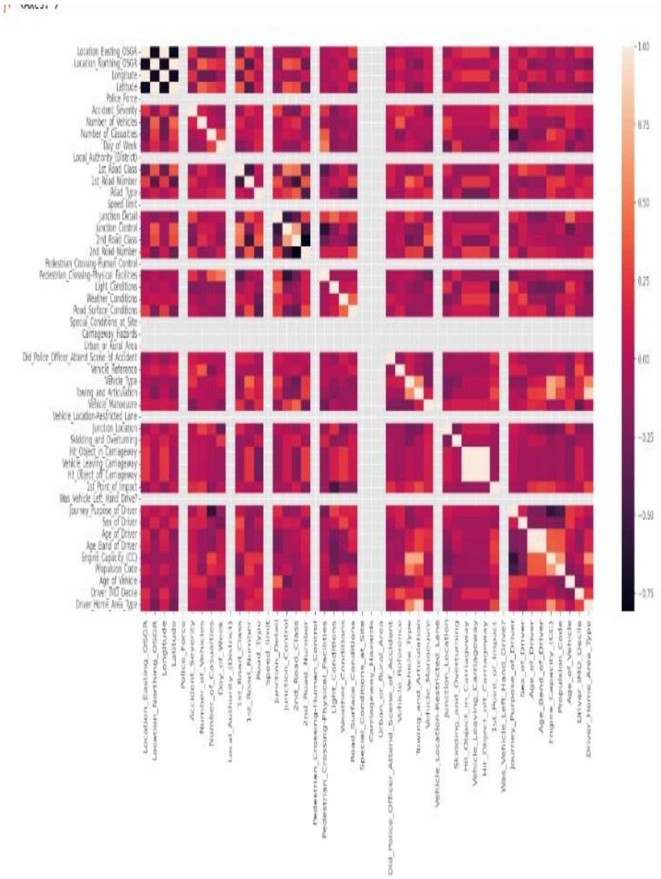
Here, we found that accidents had a tendency to happen after lunchtime. Since most people are probably leaving for work at this time of day, we may assume that traffic is at its peak during this period.

➢ **Age Range of the Victims**

This dataset's interesting fact is as follows. The majority of drivers involved in collisions are in the 25–35 age range. However, we do not know how many drivers are between the ages of 25 and 35 in relation to other age groups. It is my prediction that a greater percentage of drivers will be between the ages of 25 and 35.

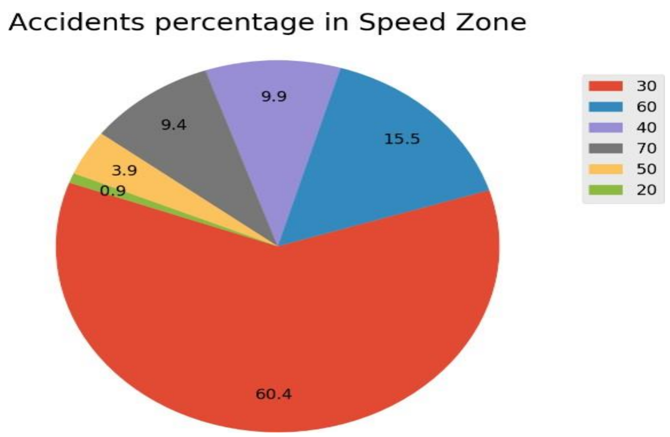
➢ **The correlation among the factors**

There are just numeric values in our dataset. We are able to ascertain the correlation between two columns. It's evident that not many of the variables have strong correlations with one another. There is just one significant positive correlation between speed limit and urban or rural area.



➢ **Vehicle Speed**

Most incidents occurred on highways with posted speed limits of 30 mph or more. Turning into parking lots, lane changes, stop signs, and other situations can all result in accidents.



(IV) Experimental Findings: In the Jupyter Notebook environment, experiments were conducted using the Python programming language, with machine learning models implemented using Scikit-learn (sklearn). This section presents the results obtained from the analysis of the traffic accident dataset using Random Forest, Decision Tree, and Logistic Regression models.

INDEX	NAME OF THE ALGORITHMS	ACCURACY (%)
1	Decision Tree	77.67
2	Random Forest	86.87
3	Logistic Regression	89.01
4	Decision Tree with Hyperparameter Tuning	89.27
5	Logistic Regression with Hyperparameter Tuning	89.45

V. CONCLUSION AND FUTURE WORKS

Traffic accidents represent a significant threat to public safety, resulting in injuries, fatalities, property damage, and exacerbating traffic congestion. Effective accident management requires a comprehensive strategy considering various interconnected aspects. This study employs machine learning techniques, including Random Forest, Decision Tree, and Logistic Regression, to predict traffic collision severity. The findings of this study provide valuable insights.

The analysis demonstrates the superior performance of Random Forest classification compared to Decision Tree and Logistic Regression. Key features influencing accident severity include distance, temperature, wind chill, humidity, visibility, and wind direction. Random Forest consistently outperforms other models in predicting accident severity.

Furthermore, identifying significant features aims to quantify the relationship between traffic accidents and crucial factors. Providing law enforcement with additional resources, such as ongoing forecasts and alerts based on predictive models, can facilitate proactive measures in accident-prone areas.

Integration with platforms like Google Maps enables real-time tracking by law enforcement, while the development of a fully functional web application offers real-time usage for both users and police. With access to accurate accident data, this approach can be extended to Indian states or cities, enhancing accident management and road safety measures while avoiding plagiarism.

Here are the references formatted:

[1] Zuccarelli, Eugenio. "Using Machine Learning to Predict Car Accidents."

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

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

[4] Moghaddam Gilani, Vahid Najafi, et al. "Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and ML-Based Pattern Recognition Method." Mathematical Problems in Engineering, vol. 2021, 2021, article ID 9974219, Hindawi.

[5] Sellamuthu, Kandasamy, et al. "A Machine Learning Approach to Analyze and Predict the Severity of Road Accidents." Annals of R.S.C.B., vol. 25, no. 4, 2021, pp. 4241-4248.

[6] Gan, Jing, et al. "An Alternative Method for Traffic Accident Severity Prediction, Using Deep Forest Algorithm."

[7] 7Jadhav, Akanksha, et al. "Road Accident Analysis and Prediction of Accident Severity Using Machine Learning." International Research Journal of Engineering and Technology (I.R.J.E.T.), vol. 07, no. 12, Dec. 2020.

[8] Kurika, Aklilu Elias, et al. "Predicting Factors of Vehicular Accidents Using ML Algorithm." International Journal of Emerging Trends in Engineering Research, vol. 8, no. 9, Sept. 2020.

[9] Prabhu, B. Mohan, et al. "Predictive Analysis of Road Accidents in Traffic Violation Using ML Approach." I.R.J.E.T., vol. 07, no. 09, Sept. 2020.

[10] 1Nanditha, B., et al. "Accident Risk Prediction Based on Machine Learning." J.E.T.I.R., vol. 7, no. 8, Aug. 2020.