International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 7 ǁ July 2022 ǁ PP. 283-288

Machine Learning Algorithms for Road Accident Analysis and Forecasting

Vyshnavi K G, Dr. Nalini N
[*1]*Department of Computer Science & Engineering, Nitte Meenakshi Institute of Technology Bangalore, India*
[*2]*Department of Computer Science & Engineering, Nitte Meenakshi Institute of Technology Bangalore, India*

Abstract
Today, one of the biggest challenges is traffic. Because there are more automobiles on the roadway, regulating the traffic has proven complicated. Even when travelling solo, many people want to operate their own vehicles instead public transportation, which culminates in traffic congestion and an increase in traffic density. Road accidents triggered by this challenging traffic circumstance have an impact on both the nation's economy and the public's health. In this study, we analyze and anticipate the number of accidents utilizing machine learning technologies such as decision trees, random forests, and logistic regression. The results indicate that compared to Logistic Regression and Decision Tree, Random Forest performs significantly better in predicting the frequency of accidents.
Keywords: *Machine learning, Random forest, Logistic Regression, Decision Trees,Visualization*

Date of Submission: 01-07-2022 Date of acceptance: 11-07-2022

I. INTRODUCTION

Traffic congestion has grown to be a significant concern in India, which contributes to numerous injuries and serious accidents. According to the National Crime Bureau (NCRB), there were numerous roads and railroads in 2016 that contributed to numerous traffic collisions. Road accidents are a huge worry since they are a current global problem that contributes to many of the major catastrophes that occur. Traffic signals, intersections, and other variables cause movement congestion, and traffic on four lanes is referred to as "uninterrupted flow." A road traffic accident's severity may be reduced by recognizing the primary and contributory reasons. Understanding the main causes and contributing factors may help reduce the severity of traffic accidents.

Numerous global surveys were conducted in 2013, and on the basis of those, the WHO produced a report for those who suffered primarily from traffic accidents, with an average of about 16.6 percent for every 10,000 individuals. All of these contributed to an increase in the population's total death rate. With an average death rate of 10,000 in 2013 and a death rate of 9.2 percent, many countries also had relatively high incomes, which contributed to the 24.1 percent death rate.

Allowing its citizens to be died in traffic crashes is totally unacceptable and dismaying. As a consequence, a thorough investigation is required to regulate this chaotic environment [1].

Everyone will benefit from being able to predict the level of traffic congestion, as this issue is brought on by the population's rapid development and the number of vehicles on the road [2].

II. LITERATURE REVIEW

Traffic safety research has focused on the magnitude of accidents. On the models based on the road crash severity classification, researchers were employing intriguing strategies. Time series data forecasting methods like regression analysis and autoregressive make forecasts about a target utilizing a variety of features. A unique forecasting method known as Autoregressive Integrated Moving Average is employed when past data points are used to make the prediction(ARIMA). Using its own inertia, this method forecasts a series' future values. The time series prediction models for freeway vehicle capacity and occupancy were introduced. Then, a number of enhanced models for forecast time series prediction for traffic based on ARIMA, including seasonal ARIMA and space-time ARIMA, were unveiled. Regression, the average of earlier data points, and other workable adjustments, such as them, were carefully compared to the ARIMA approach. [3].Chinkit Manchanda et al. explains how to use a Convolutional Neural Network (CNN) and a Hybrid Deep Neural Network (HDNN) to anticipate traffic conditions on highways with the help of images, as well as how to predict the amount of accidents at a specific time and place. In terms of traffic volume, it has been noted that HDNN excels [2]

With the aid of deep learning algorithms, Salahadin Seid Yassin et al. also forecasted the quantity of fatalities, which will help with the implementation of traffic control measures [4]. In order to assess the novel

Artificial Neural Network (ANN) simulation's suitability for the prediction of traffic accidents in Jordan, Jadaan et al. developed a model for traffic accident prediction. The findings indicated that the predicted traffic accidents were sufficiently similar to the actual ones [5]. Mussone et alresearched an accident that transpired at a junction in Milan, Italy, employing neural networks. They chose BP learning and feed-forward MLP. Ten input nodes for eight variables made up the model. The ratio between the number of accidents at a specific crossing and the number of accidents at the crossing with the greatest degree of danger was the output node, which was known as an accident index. It was determined that nighttime non-signalized crossings have the highest accident incidence for running over amblers [6]. Dr.G.Parathasarathy et al. focused primarily on the hybrid model, which serves as the fundamental algorithm for a combination of the KNN and SVM algorithms. For a single problem, it is necessary to use both classifiers in order to increase accuracy [7].

Tibebe Beshah Tesema et al, For the Addis Ababa city traffic office, the focus of this investigation is on using adaptive regression trees to build a decision support system for handling analyses of traffic accidents.. Using actual data received from the Addis Ababa the study concentrated on the extent of harm brought on by accidents.. The precision was up to 87.47 percent [8]. Miao Chong et.al, used four machine learning techniques: Hybrid Decision Tree-ANN, Support Vector Machines, Decision Trees, and Artificial Neural Networks Using Hybrid Learning (DTANN). Author examined the performance of decision trees, neural networks, support vector machines, and a hybrid decision tree-neural network in this paper. It was discovered that the hybrid technique outperformed support vector machines, decision trees, and neural networks [9]. Mohammed Balfaqih et.al, used the Classification and Regression Trees (CART), Naive-Bayes Tree (NB), Decision Tree (DT), and Gaussian Mixture Model (GMM) and found that the GMM and CART models performed significantly better on recall and precision [10].

III. DATASET

The dataset was fed into the machine learning model, which was then trained on it. Each new set of information entered into the application form serves as a test data set.The data set included the STATE/UT, a 3-hour time interval, the month, and the aggregate number of accidents.Our aim was to create machine learning-based models that could precisely analyze and forecast the amount of incidents based on month and timeframe.

Table 1 represents details of the datasets

Variable	Description
STATE/UT	Indian States name lists
YEAR	Year in which road accidents occurred
0-3 hrs. (Night)	Road accidents between 12 AM to 3 AM
3-6 hrs. (Night)	Road accidents between 3 AM to 6 AM
9-12 hrs (Day)	Road accidents between 9 AM to 12 PM
12-15 hrs (Day)	Road accidents between 12 PM to 3 PM
15-18 hrs (Day)	Road accidents between 3 PM to 6 PM
18-21 hrs (Night)	Road accidents between 6 PM to 9 PM
21-24 hrs (Night)	Road accidents between 9 PM to 12 AM
Total	Total number of road accidents

3.1 Data Wrangling:

In this phase, we will import the data, examine its integrity, and then trim and clean the provided dataset in preparation for analysis.Data wrangling is the way of converting into other formats, like as merging, grouping, concatenating, etc., in order to examine them or organize them for interaction with other sets of data. To accomplish the analytical objective, Python offers built-in tools that allow users to use these wrangling techniques to a variety of data sources.

3.2 Data collection

In order to forecast the provided data, the obtained data is split into a Training set and a Test set. The ratio of the Training set to the Test set is typically 7:3. The Test set is predicted by utilizing the Data Model, which was constructed using Logistic regression, Random forest, and Decision trees, on the Training set based on the accuracy of the Test set findings.

3.3 Pre-processing

There may be discrepancies in the data because it was obtained with some missing values. To increase the effectiveness of the algorithm and produce better results, data must be preprocessed. A variable conversion must be performed in addition to the removal of outliers.

3.4 Building the classification model

The Random Forest method is successful at estimating the frequency of accidents for the following reasons: It offers improved outcomes for the categorization issue. Good predictions that are simple to comprehend are produced by a random forest. Large datasets can be handled effectively. Compared to the decision tree method, the random forest algorithm is more accurate at forecasting outcomes.

3.5 Construction of a Predictive Model

Data collection is necessary for machine learning, and there is a wealth of historical data. There is enough historical data and unprocessed data for data collection. Raw data cannot be used directly without pre-processing. Then, what kind of algorithm and model is utilized to preprocess. This model has been tested and trained, and it makes accurate predictions with little errors. A tuned model is continually adjusted to increase accuracy.

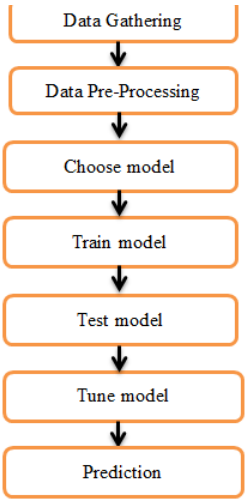

Fig: Process of dataflow diagram

IV. PROBLEM STATEMENT

Today, many urban and metropolitan areas consider traffic monitoring and control to be one of the most important tasks. Road accidents have increased as a result of more vehicles in cities. In order to prevent traffic accidents in many cities, we need to put in place effective traffic monitoring schemes and regulations.

The goal is to develop a machine learning model for real-time road accident prediction that might eventually replace updateable supervised machine learning classification models by predicting outcomes in the form of best accuracy by comparing supervised algorithm.

V. PROPOSED SYSTEM

Current methods for preventing accidents in the communities are plagued by a number of issues. The database that we'll be using is made publicly accessible by a number of institutions and government websites. The most appropriate method will be used to analyze, integrate, and group the collected data based on various constraints. The analysis and identification of the defect and the causes of the accidents will be aided by this estimation. It will also be used as a guide for building roads and bridges to prevent repeating the issues encountered in the past. It will be very helpful to plan the management of such situations based on the predictions provided using the linear regression and multiple linear regression methods.

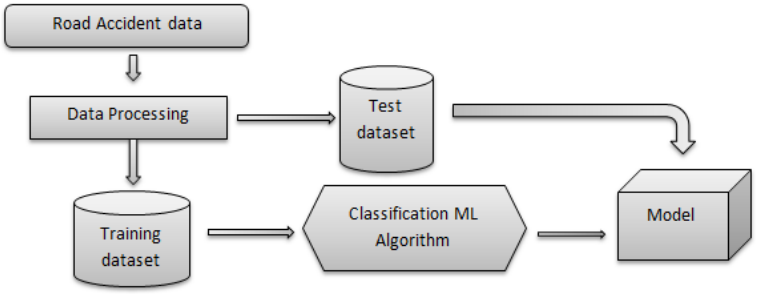

Fig 5.1: Architecture of Proposed model

VI. IMPLEMENTATION

The four key parts of this project are performance, analysis, visualization, and prediction. Python is used to implement this project in Anaconda with Jupyter Lab. Logistic regression, decision trees, and random forest classifications are the algorithms employed in this study.

6.1 Visualization:

In applied statistics and machine learning, data visualization is a crucial ability. In fact, the main focus of statistics is on numerical estimates and descriptions of data. Data visualization provides a vital set of tools for gaining a qualitative insight.This might be helpful when examining and learning about a dataset to find trends, corrupt data, outliers, and much more.

With a little topic knowledge, data visualizations may be used to communicate and demonstrate crucial relationships in plots and charts that are more visceral and compelling to stakeholders than assessments of association or importance. Data visualization and exploratory data analysis are entire topics in themselves, so it will encourage a closer look at some of the books mentioned at the end. Pie charts, bar graphs, histograms, heat maps, and count plots are just a few of the graphs that have been drawn.

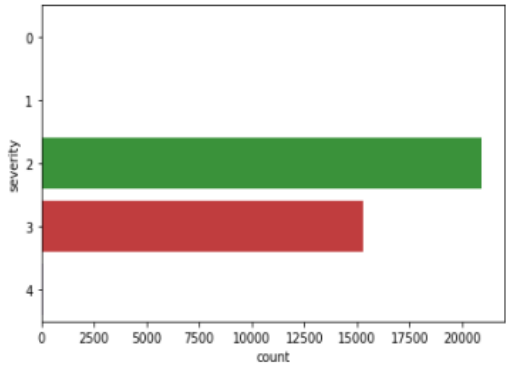

Fig 6.1: Count Plot

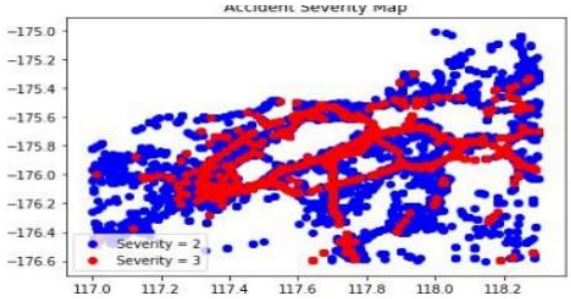

Fig 6.3: Scatter Plot

6.2 Data Analysis and Prediction

Datasets are analyzed in this module based on time period, month, year, and state. Three algorithms—Logistic Regression, Decision Tree, and Random Forest—are used to predict traffic accidents.

The statistical method of logistic regression is used to examine a collection of data where the outcome is influenced by one or more independent variables. The outcome is measured by a dichotomous variable (in which there are only two possible outcomes).The chance of a categorical dependent variable is estimated using

the Machine Learning classification algorithm known as logistic regression. In logistic regression, the dependent variable is a binary variable with data categorized as 1 (yes, success, etc.) or 0 (no) (no, failure, etc.).
Decision Tree is a supervised approach that is referred in machine learning algorithm. The data we provide in this is continuously separated, and a number of predictions are made using the data we take into account. The decision tree's final results are used to calculate the level.
An ensemble method called random forest is capable of both classification and regression. Multiple decision trees comprise the random forest's basic learning concept. When constructing a sample dataset, a random feature and row sampling is used in every design.
Step 1: Get a random sample from the dataset
Step 2: For each sample, the algorithm will build a decision tree.
Step 3: Cast your vote for each expected outcome.
Step 4: The most popular outcome from the anticipated results will be the final prediction
GUI is developed for Analysis and Prediction using Tkinter Package which is included in all python standard distributions. For analysis Time span was divided in interval of 3hrs. Analysis is done base on four parameters: Time span, Month, State,Year.

Fig: GUI for Analysis and Prediction

VII. RESULTANDDISCUSSION

Data set was pre-processed and model was built using Random Forest, Logistic regression and Decision tree algorithms, It was found that Random forest showed outstanding performance with accuracy 99.3\%.Decision Tree and Logistic Regression also performed well with accuracy 97.95\% and 97.2\% respectively. Other performance metrics such as Sensitivity,Specificity,and Confusion Matrix were also used.

Algorithms	Accuracy	Specificity	Sensitivity
Decision Tree	97.95918367	0.99212	0.9
Random Forest	99.31972782	0.992125	1.0
Logistic Regression	97.27891156	0.97637	0.95

The model can be improved in the future by creating a web application that contains more detailed data regardingeach area of the state and forecasts the amount of accidents so that appropriate action can be taken.

REFERENCES

[1]. Md. Farhan Labib, Ahmed Sady Rifat et.al "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh" 2019 7th International Conference on Smart Computing & Communications (ICSCC).
[2]. Chinkit Manchanda, Rajat Rathi, Nikhil Sharma "Traffic Density Investigation & Road Accident Analysis in India using Deep Learning" 2019 International Conference on Computing, Communication, and Intelligent Systems(ICCCIS) ISBN: 978-1-7281-4826-7/19/$31.00 ©2019 IEEE 501.
[3]. R. Silva, S. M. Kang, and E. M. Airoldi, "Predicting traffic volumesand estimating the effects of shocks in massive transportation systems,"Proc. Nat. Acad. Sci. USA, vol. 112, no. 18, pp. 5643–5648, 2015.
[4]. Salahadin Seid Yassin, Pooja "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach" 22 June 2020, Springer

[5]. K. S. Jadaan, M. Al-Fayyad, and H. F. Gammoh, "Prediction of road traffic accidents in Jordan using artificial neural network (ANN)," Journal of Traffic and Logistics Engineering, vol. 2, no. 2, pp. 92-94, 2014.
[6]. Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention, Vol. 31, 1999, pp. 705-718.
[7]. Dr. G. Parathasarathy, T.R Soumya et.al "Using hybrid Data Mining algorithm for Analyzing road accidents Dataset " 3rd International Conference on Computing and Communication Technologies ICCCT 2019.
[8]. Tibebe Beshah Tesema, Ajith Abraham et.al "Rule mining and classification of road traffic accidents using Adaptive Regression trees" I.J of Simulation Vol 6, 10 and 11 ISSN 1473-804
[9]. Miao Chong, Marcin Paprzycki et.al "Traffic Accident Analysis Using Machine Learning Paradigms" 3rd International Conference on Communications and Cyber Physical Engineering (ICCCE 2020)
[10]. Mohammed Balfaqih, Faisal Alqurashi et.al "An Accident Detection and Classification System Using Internet of Things and Machine Learning towards Smart City" 2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE).
[11]. Y. Kamarianakis and P. Prastacos, "Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches," Transp. Res. Rec., vol. 1857, pp. 74–84, Jan. 2003.
[12]. Ajith Abraham, Marcin Paprzycki "Traffic accident analysis using Machine Learning Paradigms" Informatica 29 (2005) 89–98
[13]. Gunwoo Lee, Jae Hun Kim et,al "Machine Learning-Based Models for Accident Prediction" Sustainability 2021, 13, 9137. https://doi.org/10.3390/su13169137
[14]. Bulbula Kumeda, Fengli Zhang, Fan Zhou et.al "Classification of Road Traffic Accident Data Using Machine Learning Algorithms" 2019 IEEE 11th International Conference on Communication Software and Networks