Predictive Analytics of Road Accidents Using Machine Learning

KALIRAJA C
Department of Computer Applications, Hindustan Institute of Technology & Science, Chennai, India
ckraja1607@gmail.com

CHITRADEVI D
Department of Computer Applications, Hindustan Institute of Technology & Science, Chennai, India
dcdevi@hindustanuniv.ac.in

ANJU RAJAN
Department of Computer Applications, Hindustan Institute of Technology & Science, Chennai, India
sanjur@hindustanuniv.ac.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/ICACITE53722.2022.9823773

Abstract: In today's world of transportation, road accidents are one of the most common problems. The world health organization has released a list of the top ten causes of human death, sadly, traffic accidents are in ninth place. Even in the automobile industry, many inventories make and produce safety features however, traffic accidents are inevitable. In this paper, the machine learning concept is applied to predict the severity of the accident and analyze factors like the number of accidents by year, Number of accidents by state, Accidents on the day of the week, road accidents by state day and hours, accidents ratio between rural and urban areas, Age people involved in the accidents, most dangerous time to drive, with the help of current dataset. This will be effective in improving safety measures and reducing traffic accidents.

Keywords: Road accident, Severity prediction, Predictive analytics, Linear regression, Random Forest.

I. INTRODUCTION

Every year, over 1.3 million people are killed in accidents throughout the world. Furthermore, between 20 and 50 million individuals are hurt non-fatally each year, with many of them becoming incapacitated as a result of their injuries. individuals, their families, and countries as a whole lose money as a result of traffic accidents. However, throughout the same period, the number of traffic accidents increased faster than the number of vehicles, resulting in a higher percentage of individuals killed or injured in the overall population. Predicting Road accidents are one of the most important research areas in traffic safety. The incidence of road accidents is mainly affected by the geometric characteristics of the road, traffic flow, driver characteristics, and road environment[1].The procedures utilized in data analysis were followed in this paper, with the most essential phases being data collection, prediction, and visualization. The suggested method employs a variety of visualization approaches to predicting the severity of accidents several methods of data mining techniques[2].These technologies can make use of statistical models, machine learning methodologies, and mathematical algorithms like neural networks and decision trees. As a result, data mining encompasses analysis as well as prediction. In contemporary data mining initiatives, many important data mining techniques such as association, classification, clustering, prediction, sequential patterns, and regression have been created and developed[3].The process of examining, cleaning, manipulating, and modeling data to extract useful information is known as data analysis. Data analysis aids in the development of more scientific findings and the effective operation of businesses[4].This model aims to make roads more secure and accident-free using machine learning. Datasets containing details about previous accidents in various regions are studied and analyzed and a model is developed which can be used to predict[5]. The primary goal of predictive analytics of road accidents is

- increase efficiency.
- Reduce the severity of the accident.

II. LITERATURE SURVEY

According to the literature review, by Sridevi, N., et al. (2020) The major goal of that model is to anticipate accident-prone locations by taking into account a variety of characteristics that cause accidents. To discover the factors that cause accidents, this model employs the data mining technique of apriori and the machine learning concept of K-Means[6]. Singhal, Shruti, et al. (2021) The goal of this research is to investigate, evaluate, and analyze the performance of six important machine learning approaches to gain a better understanding of how traffic accidents occur. Decision trees, Support Vector Machines, Naive Bayes, Random Forest, KNN, and logistic regression are among the methods studied. The study is based on objective and scientific surveys to detect and further avoid accidents, understand the causes, and the severity of injuries, to achieve the most practical and possible accident reduction[7]. Venkat, Arun, Guru Vijey KP, and Irish Susan Thomas. (2020) Due to a complicated combination of elements such as the driver's mental state, road conditions, weather conditions, traffic, and traffic rule infractions, to name a few, the underlying cause of traffic accidents is difficult to define these days. The costs of traffic fatalities and driver injuries have a substantial social impact. Machine learning techniques are rapidly being used in the field of traffic accidents. This paper provides an overview of existing work on the subject of machine learning-based accident prediction[8].Cigdem, A., and Cevher Ozden. (2018) The severity of injuries in traffic accidents in Adana was categorized in this study, and the elements that influenced the accident outcome were explored. Five key machine learning methods (KNN, Nave Bayes, Multilayer perceptron, Decision Tree, Support vector machine) and one statistical approach (Logistic Regression) were utilized to create prediction models, and their performances were compared, as well as the effective parameters. The study's main purpose is to determine how relevant weather and other factors are in traffic accident incidence[9].Najafi Moghaddam Gilani, Vahid, et al. (2021) Modeling the severity of accidents using the most effective factors allows for the development of a high-precision model that shows the likelihood of each category of future accidents occurring, and it may be used to help authorities prioritize measures. By collecting data on urban accidents, the goal of

this study is to determine the characteristics that influence the severity of the injury, mortality, and property damage only accidents in Rasht city[10-15].

III. Methodology

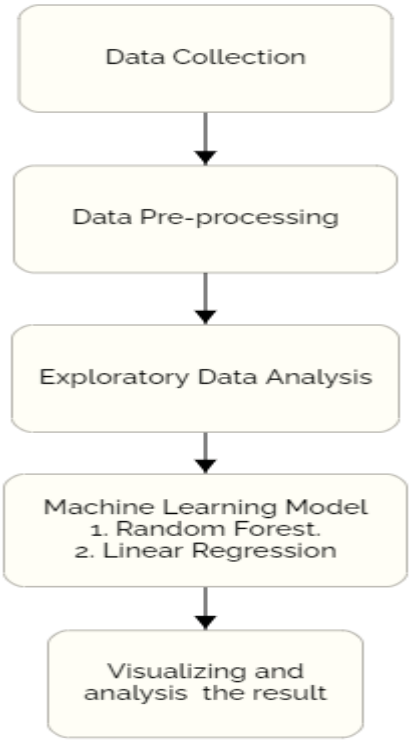


Fig. 1. Represents the workflow of the project.

A. Data collection

Data used in this paper is a set of road accident details that is collected from the website Kaggle (https://www.kaggle.com/). It includes three files to perform an analysis of this data. This data is consisting of three files are acc.csv, accidents.csv, casuality.csv, and has one more file (road _accidents.csv) which is general information about the traffic count for the years 2001 to 2014. The dataset acc has 32 columns and 1048576 rows.

B. Data preprocessing

Data cleaning: The process of recognizing and displaying unnecessary and incorrect data, as well as determining which elements are most relevant, is known as data cleaning.

Null and Missing values: In the acci dataset, there was a 138-missing value in each column Location_Easting_OSGR, Longitude, Latitude, Location_Northing_OSGR, and 151 values in the Time column. Those missing null values are filled by value 0 with the use of the dropna function.

C. Exploratory Data Analysis

Exploratory data analysis is the graphical representation of information and data. Using visual components like charts, graphs, and maps, data visualization tools make it simple to explore and grasp trends, outliers, and patterns in data.

D. Machine Learning Models:

This paper used two machine learning models are:

- Random Forest
- Linear Regression

E. Random Forest:

Several columns from the dataset are utilized for prediction models, including month, hours, years, lon, lat as an X, and severity as a Y. The lamda approach was utilized to predict severity by shifting to 0-1 values instead of 2-3, which contains the highest severity data. The train test ml model is then imported to split the data into test and train with a test size of 0.20, X-train data size of (28992,6), and X-test data size of (7248,6). Fit the train and test data to the random forest classifier after importing it to predict the accuracy[16-18].

F. Linear regression:

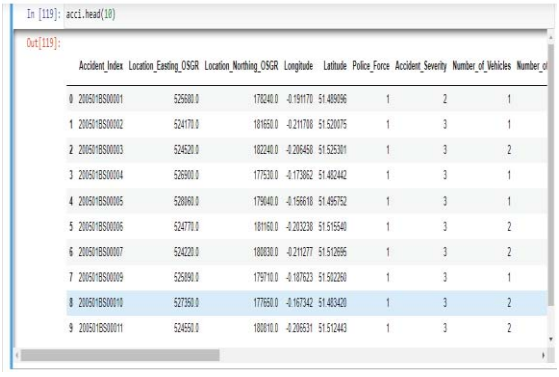
For prediction models, several columns from the dataset are used, including month, hours, years, lon, lat as an X, and severity column as a Y. The lamda technique was used to predict severity by moving to 0-1 values rather than 2-3, which contains the most severe data. The train test ml model is then imported to divide the data into test and train with a test size of 0.20, X-train data size of (28992,6), and X-test data size of (7248,6). once the linear regression model has been imported to anticipate the accuracy, fit the train, and test data.

IV. Experimental Results

This experiment has been done on Jupyter Notebook in Python Language on Windows 10.Important libraries used are pandas, NumPy, seaborn, and matplotlib. pyplot. Datasets have beentaken from the website Kaggle. Analysis of the road accidents dataset has been done by performing the random forest and linear regression technique. The accident data analysishas been done from the year 2001 to 2014[19-23]. The Attributes focused on the analysis is the Number of accidents by year, Number of accidents by state, Accidents on the day of the week, road accidents by state day and hours, accidents ratio between rural and urban areas, Age of people involved in the accidents, most dangerous time to drive and severity of the accident. The analysis is represented through bar graphs, histograms, and heatmap.

A. Analysis and visualization

The datasets were collected from the Kaggle website(https://www.kaggle.com) which is represented in figure 2.



In [119]: acci.head(10)

Out[119]:

	Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of
0	200501BS00001	525680.0	178240.0	-0.191170	51.489096	1	2	1	
1	200501BS00002	524170.0	181650.0	-0.211708	51.520075	1	3	1	
2	200501BS00003	524520.0	182240.0	-0.206458	51.525301	1	3	2	
3	200501BS00004	526900.0	177530.0	-0.173862	51.482442	1	3	1	
4	200501BS00005	528060.0	179040.0	-0.156618	51.495752	1	3	1	
5	200501BS00006	524770.0	181160.0	-0.203238	51.515540	1	3	2	
6	200501BS00007	524220.0	180830.0	-0.211277	51.512695	1	3	2	
7	200501BS00009	525890.0	179710.0	-0.187623	51.502260	1	3	1	
8	200501BS00010	527350.0	177650.0	-0.167342	51.483420	1	3	2	
9	200501BS00011	524650.0	180810.0	-0.206531	51.512443	1	3	2	

Fig. 2. Dataset.

Next applied the data preprocessing method to clean the null values and missing values. Figure 3 represents there are no null values in the data.

Accident_Index	0
Location_Easting_OSGR	0
Location_Northing_OSGR	0
Longitude	0
Latitude	0
Police_Force	0
Accident_Severity	0
Number_of_Vehicles	0
Number_of_Casualties	0
Date	0
Day_of_Week	0
Time	0
Local_Authority_(District)	0
Local_Authority_(Highway)	0
1st_Road_Class	0
1st_Road_Number	0
Road_Type	0
Speed_limit	0
Junction_Detail	0
Junction_Control	0
2nd_Road_Class	0
2nd_Road_Number	0
Pedestrian_Crossing-Human_Control	0
Pedestrian_Crossing-Physical_Facilities	0
Light_Conditions	0
Weather_Conditions	0
Road_Surface_Conditions	0
Special_Conditions_at_Site	0
Carriageway_Hazards	0
Urban_or_Rural_Area	0
Did_Police_Officer_Attend_Scene_of_Accident	0
Hour	0
dtype: int64	

Fig. 3. Data preprocessing.

Next to analyzed the accident details to identify the details to reduce the severity of accidents are the Number of accidents by year, Number of accidents by state, Accidents on the day of the week, road accidents by state day and hours, accidents ratio between rural and urban areas, Age people involved in the accidents and most dangerous time to drive.

The visualizations are done by the matplotlib. Pyplot pre-defined function.

- Here, the visualization of the number of accidents by year between 2001 to 2014 has shown in figure 4.

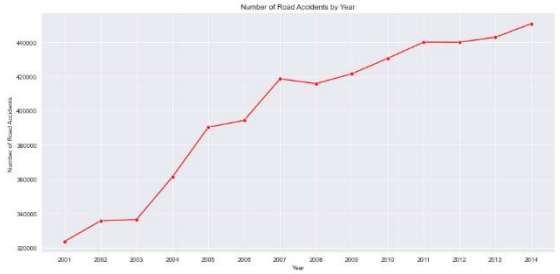


Fig. 4. Accidents by states.

- Here, the analysis of accidents in states in India is displayed in figure 5.
- From the figure, Tamil Nadu, Maharashtra, Karnataka, Andhra Pradesh, Kerala are the top five states that have the most accident records.

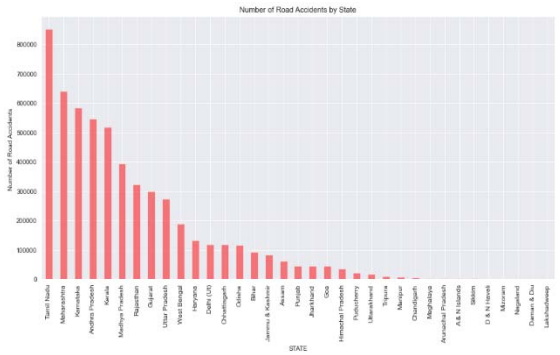


Fig. 5. Accidents by states.

- The analysis of accidents state day and hours are displayed in figure 6, and in Tamandu 15-18 has the top most accident counts.

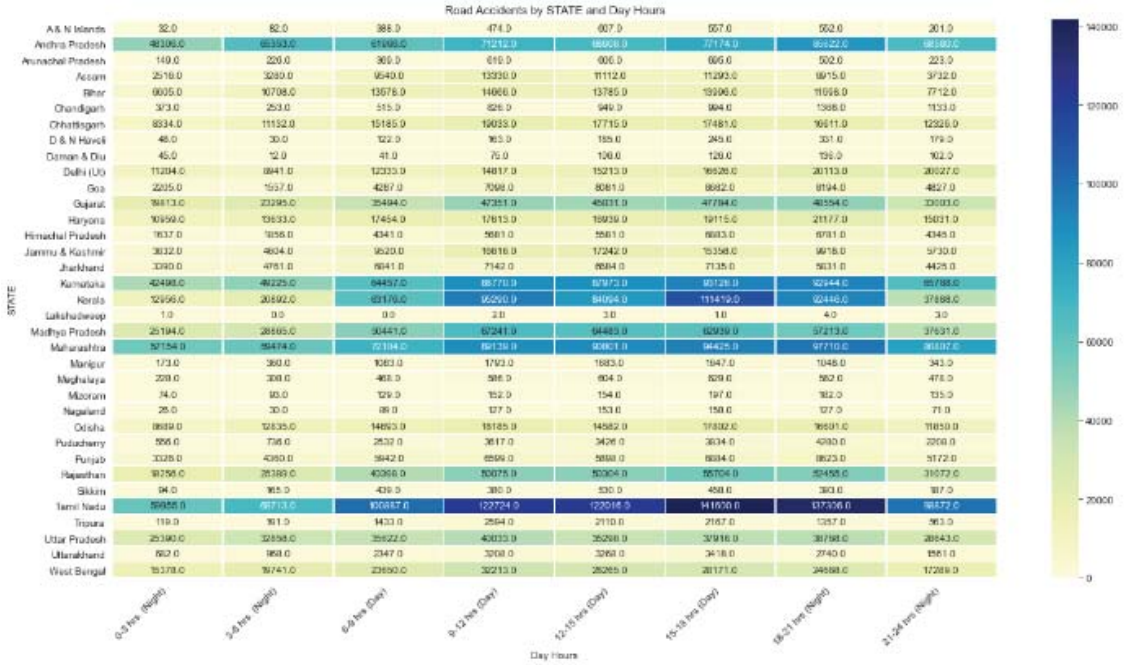


Fig. 6. Road accidents by state day and hours.

- Visualization of accidents on the day of the week in figure 7.
- Compare to all days Friday has the highest number of accident records than other days.

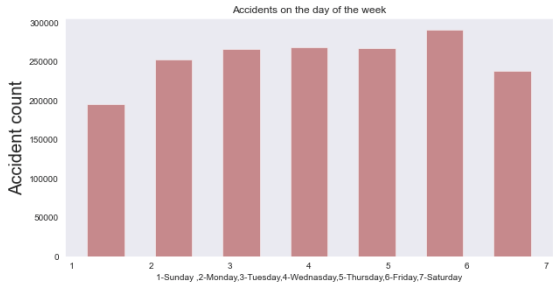


Fig. 7. Accidents on the day of the week.

- Analysis of accident ratio between the rural and urban areas in figure 8.
- Sixty-three of accidents in urban areas, and thirty-seven percent of accidents in rural areas.

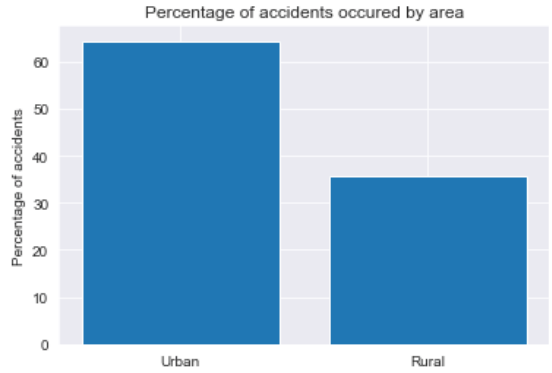


Fig. 8. Accidents in rural and urban areas.

- Visualization of accidents between different age peoples in figure 9.
- Most drivers aged is around 25 to 35 are involved in the accident.

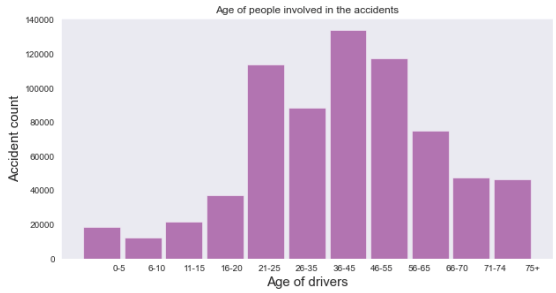


Fig. 9. Age of people involved in the accidents.

- Here, The Highest number of accidents recorded by the hour is in figure 10.
- The highest accident records entered at 4’ o clock as per the data records.

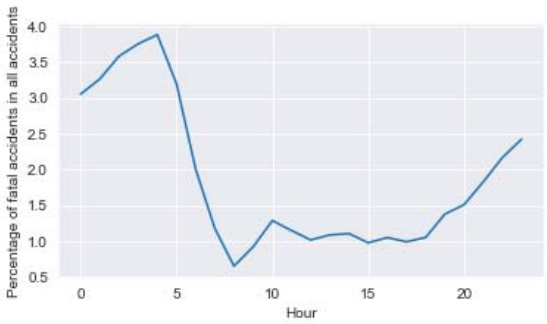


Fig. 10. Accidents by the hour.

Next, the efficiency of machine learning models between Random Forest and Linear Regression is being visualized in figure 11.

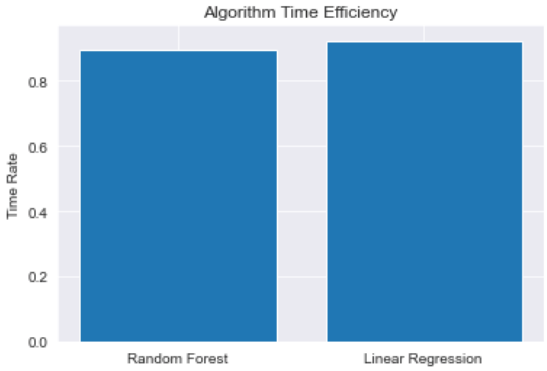


Fig. 11. Algorithm efficiency.

TABLE I. ALGORITHM ACCURACY.

Algorithm	Accuracy
Random Forest	0.89
Linear regression	0.92

Linear regression is the model which has given more accuracy.

V. CONCLUSION

In this article machine learning classification techniques are used to predict the severity of an accident. Machine learning enables us to analyse meaningful data to deliver solutions with greater accuracy than humans. The proposed system for road accident detection works better than the previous hand-designed, the test result is positive, demonstrating the effectiveness of the recommended strategy. Experimental results are encouraging and show the effectiveness of the proposed approach. This model has given more accuracy than the conventional system.

VI. REFERENCES

[1] Road Accident analysis using machine Learning | International journal of research in engineering, science, and management volume-3, issue-5, May-2020 | ISSN:2581-5792 N. Sridevi, M.V. Keerthana.

[2] Singhal, Shruti, et al. "Machine Learning Approach Towards Road Accident Analysis in India." Proceedings of Second International Conference on Smart Energy and Communication. Springer, Singapore, 2021.

[3] Venkat, Arun, Guru Vijey KP, and Irish Susan Thomas. "Machine Learning Based Analysis for Road Accident Prediction." Machine Learning-Based Analysis for Road Accident Prediction (March 7, 2020). IJETIE 6.2 (2020).

[4] Cigdem, A., and Cevher Ozden. "Predicting the severity of motor vehicle accident injuries in Adana-Turkey using machine learning methods and detailed meteorological data." International Journal of Intelligent Systems and Applications in Engineering 6.1 (2018): 72-79.

[5] Data-Driven Urban Traffic Accident Analysis and Prediction Using logit and Machine Learning-Based Pattern Recognition Models | Vahid Najafi Moghaddam Gilani, Syed Mohsen Hosseinian.

[6] H. Nguyen, C. Cai, F. Chen, Automatic classification of traffic incidents severity using machine learning approaches Intel. Trans. Syst. (10), 615-623(2017) CrossRefGoogle Scholar.

[7] R. Mehar, P. K Agarwal, A systematic approach for formulation of a road safety improvement program in India.Procedia-Soc.Behav. Sci. 104,1038-1047(2013) CrossRefGoogle Scholar.

[8] M. Gupta, V. K Solanki. Singh, V. Garcia-Diaz, Data mining approach of accident occurrences identification with effective methodology and implementation. Int. J. Electr. Comput. Eng. 8(5),4033(2018) Google Scholar.

[9] C. Caliendo, M.L. De Guglielmo, I. Russo, Analysis of crash frequency in motorway tunnels based on a correlated random-parameters approach. Tunn. Undergr. Spae technology. 85,243-251 (2019) CrossRefGoogle Scholar.

[10] S. Shanthi, R.G. Ramani, Feature relevance analysis and classification of road traffic accident data through data mining techniques. In Proc. World congress Eng. Comput. Sci.1,24-26 (2012) Google Scholar.

[11] S. Vasavi, extracting hidden patterns within road accident data using machine learning techniques. In.inf.commun. tec. (2018), pp,13-22 Google Scholar.

[12] S. Krishnaveni, M. hemalatha, A prospective analysis of traffic accident using data mining techniques. Int.J. Comput. Appl.23(7),40-48 (2011) Google Scholar.

[13] E. Suganya, & S. Vijayarani, Analysis of road accidents in India using data mining classification algorithm. In 2017 Int. Conf. Inventive comput. Inf.pp.1122-1126(November-2017). IEEE.Google Scholar.

[14] Nitin Kashyap, Hari Raksha K Malali, G Raju, TH Sreenivas, traffic Accident injury and severity prediction using Machine Learning Algorithms, ICCCE 2020,1041-1048,2021.

[15] Shruti Singhal, Bhavani priyamvada, Rachna Jain, Muskan Chawla, Machine Learning Approach Towards Road Accident Analysis in India, Proceedings of Second International Conference on smart energy and Commun,311-322,2021.

[16] Utsav Chaudhary, Army Patel, Arjun Patel, Mukesh Soni, Survey Paper on Automatic Vehicle accident detection and rescue system, Data Science and Intelligent Applications, 319-324,2021.

[17] Arun Venkat, Guru vijey KP, Irish Susan Thomas, Machine Learning Based Analysis for Road Accident Prediction (March 7, 2020), IRJETE 6(2),2020.

[18] Md Farhan Labib, Ahemed Sady Rifat, Md Mosabbir Hossain, Amit Kumar Das, Faria Nawrine, Road accident analysis and prediction of accident severity by using machine learning in Bangladesh, 2019(ICSCC), 1-5,2019.