IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11639

Accident severity detection and prediction

Padmini C[1], Shubhashree K B[2], Pattipati Naga Bhargavi[3], Usha D[4], Vaishnavi K[5]

[1, 2, 3, 4, 5] Atria Institute of Technology, Bangalore, Karnataka, India

Abstract: Despite everything that has been done to improve road safety in India to date, there are always problems lurking around every corner. This circumstance has shown a problem with traffic accidents, affecting public health and the economy of the country. In the past, it was assumed that road accidents and fatalities could not be avoided, but in today's technological age, anything is almost possible. Our research aims to reduce mortality rates by developing a prediction model that considers factors like carriageway/roadway hazards, light conditions, day of the week, special conditions at the accident scene, road class, junction control, junction details, road surface condition, road type, and weather conditions. Machine learning techniques such as random forests (RF), logistic regression (LR), and naive Bayes (NB) were used to create these models. The goal of this research is to analyze data on road accidents in India utilizing the best compatible machine learning classification approaches for estimating road accidents through data mining. Our findings suggest that logistic regression outperformed other machine learning algorithms in terms of accuracy.

I. INTRODUCTION

Globalization has had an impact on many countries. The quantity of economic activity and consumption has skyrocketed, leading in a surge in travel and transportation. A rise in the number of vehicles on the road causes traffic accidents. Given the importance of road safety, the government is striving to identify the causes of traffic collisions in order to reduce the frequency of collisions. Analysis of the limits that cause road accidents is becoming increasingly sophisticated due to the exponential growth of accident data. The act of evaluating data from many sources and translating it into valuable information that can be utilized to make better business decisions is known as data mining. Data classification is a data mining approach for categorizing and analyzing data.

The grim truth of rising accident rates underscores the urgent need for a global increase in road safety. It is not usually the motorist who bears the brunt of the blame. Other elements that may have contributed to the accident include vehicle defects, weather circumstances, and transit conditions.In this study, we look at contributing factors and the educational background of drivers in India's states and union territories in order to draw valuable conclusions on how to enhance road safety in the country. To generate vulnerability groups, we're focusing on utilizing clustering to group similar things in this dataset. The resulting clusters are identified, and a decision tree is used to find the region-wise dominating reason.

II. LITERATURE SURVEY

Machine learning (ML) is utilized in the study (1) to assess and enhance various algorithms through experience. Their research focused on lowering the death rate by creating a predicting model using an unsupervised learning method called k-means clustering, which analyzed road accidents by taking into account various factors such as potholes on the road, sharp turns, and rainfall conditions, as well as providing appropriate and preventative measures to avoid mishaps by representing it on a map and creating a model that anyone could understand. This model eventually attained an accuracy of 81 percent.

Researchers created a predictive machine learning algorithm that used available data to graphically depict important areas on roadways in the paper (2). This model is totally replicable and can be used in any place throughout the world to assist reduce the amount of accidents and deaths caused by vehicle crashes. The Road Lytics model is proposed as a supervised machine learning model that analyzes events between 2017 and 2020 using the Random Forest method.

The authors of the research (3) employed significant characteristics revealed by Random Forest that are substantially connected with the severity of highway accidents. The severity of an accident is influenced by distance, temperature, wind chill, humidity, visibility, and wind direction. This study used Random Forest and Convolutional Neural Network to create an RFCNN ensemble of machine learning and deep learning models for predicting the severity of road accidents.

The study's authors (4) aimed to develop models that could choose a set of influential criteria that could be used to classify the severity of an accident. Furthermore, their research provided a predicted model for future traffic accidents based on existing data. A variety of machine learning approaches are used to generate the models. According to the findings, a rule-based model based on the C5.0 algorithm may accurately recognise the most significant elements

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022
DOI: 10.17148/IJARCCE.2022.11639

indicating the severity of a traffic incident. The RF model's results also suggested that it could be a valuable tool for anticipating accident hotspots, according to the expected model's results.

Researchers discussed highway accidents in the research (5),which have been a major source of concern for both developed and developing countries. Multiple causes, such as atmospheric fluctuations,steep angles, and natural faults, cause numerous road accidents. Their study used a k-means method and machine learning to assess traffic accidents in one of the most populous metropolitan metropolises,Bengaluru, by looking for accident-prone or hotspot locations and their core causes.

According to the authors' study(6) in this publication, teen drivers who failed to manage speed or traveled at a hazardous pace when merging from country roads to highways or approaching intersections were the top causes of teen driver crashes in West Texas. They also failed to yield on their undivided streets with four or more lanes, resulting in serious injuries. The road class, speed limit, and the initial detrimental occurrence are the top three factors that influence the severity of their crash. Their predictive machine learning approach based on Label Encoder and XG Boost appeared to be the best option when accuracy and computational cost were considered. Their findings should aid in improving the safety of rural teen drivers in West Texas and other parts of the country.

The random forest, artificial neural network, and decision tree techniques were employed to identify the strengths and weaknesses of these methods in their article (7). For a period of nine years, researchers looked at three data sets: road geometry, precipitation, and traffic accident data from the Naebu Expressway in Seoul, Korea. To evaluate their model, they employed the out-of-bag estimate of error rate (OOB), mean square error (MSE), and root mean square error (RMSE) (RMSE). The low mean OOB, MSE, and RMSE observed in their data using the suggested random forest model confirm its accuracy.

Researchers in this study (8) used machine learning techniques to dig deeper into traffic occurrences in Bangladesh in order to determine the severity of the accidents. We also highlight the key factors that have a direct impact on road accidents and offer some practical solutions to the problem. Using four supervised learning algorithms, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, and AdaBoost, the severity of accidents was divided into four categories. In the end, AdaBoost performed the best in this essay.

This paper (9), which looks at effort, creates models to choose a set of important factors and build a model for evaluating injury severity. These models were created using a variety of machine learning techniques. Researchers employed supervised machine learning techniques like AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests on traffic accident data (RF). The SMOTE method was used to correct the data imbalance. The RF model could be a useful tool for forecasting the severity of injuries in traffic accidents, according to the findings of this study. The RF algorithm outperformed the LR, NB, and AdaBoost algorithms with 75.5 percent accuracy.

Using data mining and machine learning methodologies, the road accident data analysis in paper (10) aimed to uncover factors that influence the severity of an accident. Accidents can be caused by a variety of things. Some crashes are caused by the driver's actions, while the majority are caused by outside circumstances. Poor weather conditions, such as fog, rain, or snowfall, for example, might result in limited vision, making driving difficult and dangerous. The outcomes of this study were supposed to help authorities take proactive actions in the case of likely crash-prone weather and traffic conditions.

III. PROPOSED MODEL

A. DATA UNDERSTANDING

The data comprehension step in CRISP-DM involves examining the "surface" or "gross" characteristics of the acquired data and reporting on the results. The Ministry of Road Transport and Highways, among other sources, provided information for this study.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11639

Variables	Description
Carriageway/Roadway hazards	Any obstacles on the roadway. It could be animals or pedestrians, potholes, etc. Here we are generally classifying them as hazards present and none.
Condition of light	The lighting during the accident is represented by this variable. Light is present, and darkness is present.
Day of the Week	Days through Sunday to Saturday.
Special Conditions at the accident site	This includes malfunctioning traffic lights, improper signboards, road maintenance issues, road works and fuel-related issues.
Road Class	A- National Highways, B-State Highways, C-District Roads, Village Roads and Unclassified Roads.
Junction Control	This field specifies how the junction is controlled.
Junction Details	Whether the accident occurred at a junction, crossroads, round-about, private driveway, and so on.
Road Surface condition	It depicts the accident scene on the road. Dry, wet, frost, snow, flood are all examples of variables.
Road Type	This talks about the kind of roadway in which the accident occurred. Like: Single Road, Double Road, One-way, Slip Road, U-turn.
Area Type	Whether Urban or Rural areas.
Weather condition	This variable represents the weather conditions at the time of the collision. Rain, Wind, Snow, Fog, Mist, Sunny, and Fine weather are all represented by variables in this variable.
Speed	This variable talks about the speed of the vehicle at the time of the accident. In this dataset, it varies from 10-70 km/hr.
Accident time	This variable indicates the time of day and whether it was day or night when road traffic was severe.
Severity	This is the target variable, and it represents three different injury classes: fatal, serious, and minor.

Table 1: Variable Explanation

B. DATA PREPROCESSING

Exploratory Data Analysis (EDA)

In India, the number of accidents per state or union territory has been researched in relation to a number of parameters, including weather.

	STATE/UT	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	TOTAL
0	A & N Islands	2001	8	23	15	15	14	19	14	19	7	12	13	22	181
1	A & N Islands	2002	12	10	14	16	10	7	16	11	23	21	11	17	168
2	A & N Islands	2003	19	13	15	13	13	12	8	16	17	20	14	15	180
3	A & N Islands	2004	21	14	22	17	13	18	16	19	16	20	15	24	215
4	A & N Islands	2005	19	21	22	17	13	19	21	14	15	19	10	16	206
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
485	West Bengal	2010	1245	1150	1349	1246	1172	1264	1231	1190	1128	1227	1251	1252	14725
486	West Bengal	2011	1350	1179	1314	1148	1220	1241	1185	1074	1112	1214	1161	1270	14468
487	West Bengal	2012	1346	1383	1357	1270	1352	1434	1348	1204	1112	1251	1179	1371	15608
488	West Bengal	2013	1564	1382	1474	1382	1629	1391	1315	1208	1228	1299	1335	1332	16549
489	West Bengal	2014	1516	1396	1473	1385	1627	1439	1416	1356	1391	1395	1415	1394	17105

Fig 1: Month-wise data for Number of Accidents per State/UT

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11639

	STATE/UT	YEAR	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24	Total
0	A & N Islands	2001	2	6	29	40	39	40	18	7	181
1	A & N Islands	2002	2	6	22	41	33	33	23	8	168
2	A & N Islands	2003	2	8	31	35	28	36	25	15	180
3	A & N Islands	2004	2	5	29	42	43	43	37	14	215
4	A & N Islands	2005	0	8	27	28	38	42	50	13	206
...	...	...	...	...	...	...	...	...	...	...	...
485	West Bengal	2010	1241	1397	1721	2508	2272	2296	1831	1459	14725
486	West Bengal	2011	1200	1493	1687	2553	2182	2196	1812	1345	14468
487	West Bengal	2012	1346	1511	1837	2831	2328	2268	1966	1521	15608
488	West Bengal	2013	1442	1911	2136	2900	2246	2366	2137	1411	16549
489	West Bengal	2014	1455	1634	2022	2998	2570	2458	2132	1836	17105

Fig 2: Time wise data for Number of Accidents per State/UT

	STATE/UT	YEAR	TOTAL	SUMMER	AUTUMN	WINTER	SPRING
0	A & N Islands	2001	181	52	32	53	44
1	A & N Islands	2002	168	34	55	39	40
2	A & N Islands	2003	180	36	56	47	41
3	A & N Islands	2004	215	53	51	59	52
4	A & N Islands	2005	206	54	44	56	52
..	...	...	...	...	...	...	...
485	West Bengal	2010	14725	3705	3606	3647	3767
486	West Bengal	2011	14468	3500	3487	3799	3682
487	West Bengal	2012	15608	3987	3542	4100	3979
488	West Bengal	2013	16549	3914	3862	4278	4495
489	West Bengal	2014	17105	4211	4201	4308	4385

Fig 3: Number of Accidents per Season

The bar chart below depicts the amount of road traffic accidents that have occurred in each Indian state, with Tamil Nadu appearing to have the highest number of RTAs. On Google, you can readily find this. Statista, Down To Earth, and other similar websites are good places to start.

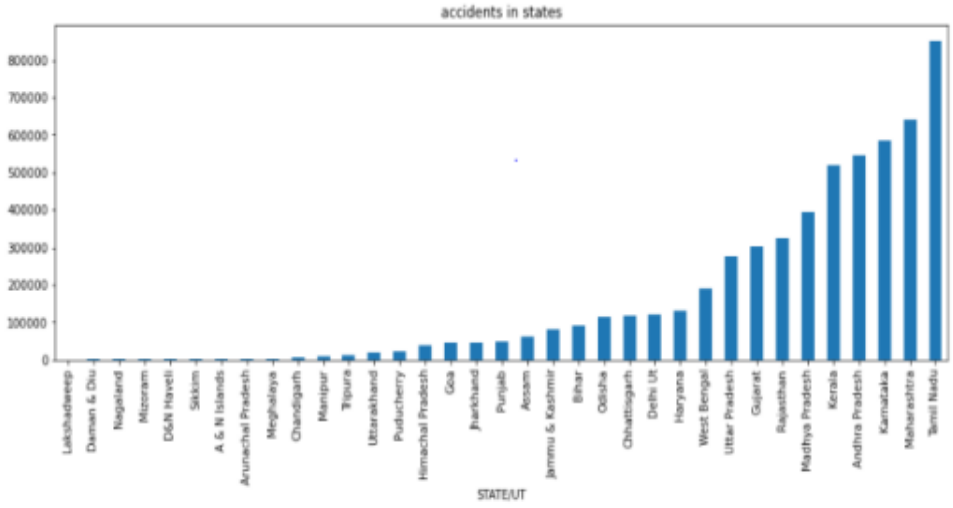

Fig 4: Bar Chart depicting Number of Accidents sorted per State

Data was used to evaluate the impact of weather on India's accident rate. The pie chart below shows that the bulk of our database's accidents occurred during the spring season.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11639

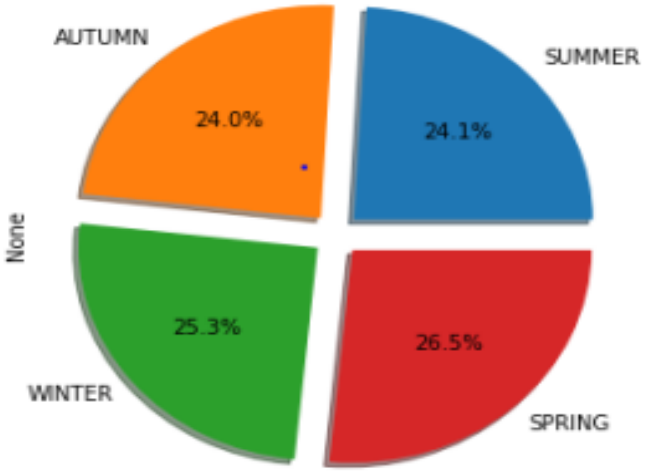

Fig 5: Accidents in Different Seasons

Similarly, data has been evaluated to see how the time of occurrence affects the accident rate in India. The bulk of accidents in our dataset occurred in the afternoon, as indicated in the pie chart below.

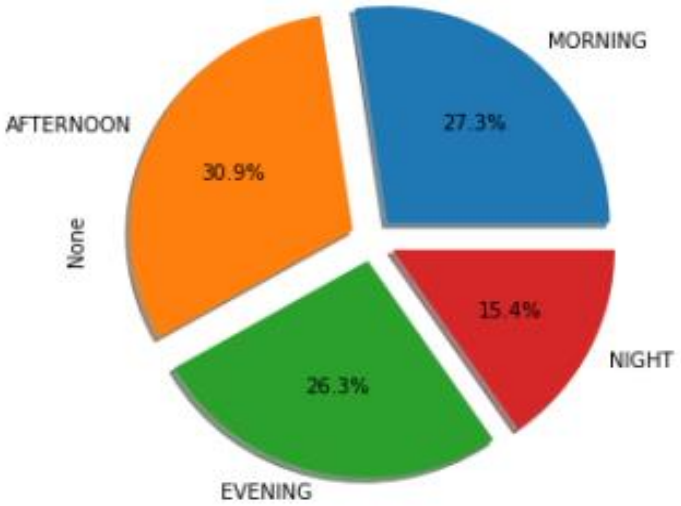

Fig 6: Accidents at Different Times of the Day

The data has been evaluated and investigated for a range of factors, with appropriate data treatment conducted to normalize and use the data for the modeling step.

C. DATA PREPARATION

Raw datasets were not in a proper format that computers could interpret, therefore they only offered a fraction of the information needed to work with. When such datasets were employed, the efficacy of the accident severity prediction model was reduced. As a result, to acquire high-quality data, unnecessary datasets must be deleted. The researchers utilized a costly data preparation technique to extract significant and deciding risk factors before building a model. Data cleaning, missing value handling, outlier treatment, and dealing with absolute values are all thoroughly cleansed before usage.

We got information from a variety of sources. The data has been cleaned and missing values eliminated, as well as the feature values encoded. This information has now been saved in a clean Excel file that will be used as an input data file.

1.Cleaning the data: Removed rows with missing values and dropped columns that were not essential to the analysis.
2.Data Encoding: To generate a simplified version of data for running the model, I encoded the feature variable values.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11639

IV. DATA MODELING

1 .Naïve Bayes Classification

In this case, the Bayes hypothesis is applied. It's easy to set up and works effectively in circumstances with large amounts of data. Applies to Bayes' opinion on the possibility of forecasting the anonymous data set category. This class assumes that the presence of one element in a class has no influence on the existence of other elements in that class. This is easy to build and can be used to evaluate large datasets. A model that can be predicted can be built after training utilizing the risk data provided. Predictability and the intensity of the risk can be forecasted under the correct circumstances.

Y class variables, and feature variables (X1, X2, Xn),

The Bayes hypothesis provides the following relationships:

$P(Y \mid X1\ldots, Xn) = P(Y)\, P(X1 \ldots, Xn \mid Y) / P(X1\ldots, Xn)$

Using the thinking of the Naïve Bayes,

$P(Xi \mid Y, X1 \ldots, Xi\text{-}1\ldots, Xn) = P(Xi \mid Y)$

At all i values, relationships can be provided by,

$P(Y \mid X1 \ldots, Xn) = P(Y)\, P(X1 \ldots, Xn \mid Y) / P(X1 \ldots, Xn) \propto P(Y)\, \Pi P(Xi \mid Y)$

$Y = \text{argmax} = P(Y)\, \Pi\, P(Xi \mid Y).$

2. Logistic Regression

Its is also a collection of rules for supervised classification learning. The miles are used to calculate the likelihood of a binary response based entirely on one or more predictors. They might be either continuous or discrete in nature. When categorizing or distinguishing some recorded gadgets into categories, we utilize logistic regression..

It classifies the statistics in binary shape as the handiest in zero and 1.

One of the main goals of logistic regression is to generate an exceptional match that can describe the relationship between the goal and the predictor variable. The variant of logistic regression that is based on linear regression is called logistic regression. The sigmoid characteristic is used in a logistic regression model to predict the probability of effective and horrible magnificence.

Sigmoid characteristic $P = 1/1+e - (a+bx)$ here P =possibility,

a and b = parameter of version.

3. RANDOM FOREST

It is for classifying and forecasting data that is supervised. The algorithm produces good results in a variety of analyses and is implemented in a wide range of artificial intelligence libraries. The data used in this article is freely available on the internet. However, the government's data, which includes information on budgets, transportation, culture, science, economics, and climate, is public. One of the aspects mentioned by Veljkovic in relation to open data is legal access to the data. The institution cannot erect barriers that prevent anyone with a legitimate interest in the data from accessing it. Furthermore, access must be made easier, and data must be made available in an easy-to-understand and absorb manner, avoiding, for example, photos and PDF lines and favoring data in bulk format.The study's key contribution is the creation of a predictive machine learning model that leverages open data to predict accident severity while taking into account a variety of characteristics. This model is entirely repeatable and may be used in any place throughout the world to assist in reducing the number of accidents and their severity by detecting and predicting them.

V. EXPERIMENTAL RESULTS

The Model was deployed to a local host using the Python Flask framework. The deployment is set up as a web page, making it a simple way to estimate the severity of an accident. The system can be deployed in the user domain, as in the case of a traffic police agency. For parameters such as weather, road conditions, date and time, and so on, the user must fill out a Form Data. When we complete the form and click the Predict Accident Severity button, a pop-up message displays with the severity class projected for the provided data.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022
DOI: 10.17148/IJARCCE.2022.11639

VI.CONCLUSION

In this research we'll look at how analyzing data can help us identify states and union territories of India which are prone to accidents. These clusters are identified so that Machine Learning algorithms may classify them and discover the key causes of road traffic accidents. The main objective of this study was to evaluate and observe accident-causing factors, as well as to predict the accident severity that had occurred or that would occur.

This was done with the help of machine learning techniques like Logistic Regression, Gaussian Classifier, and Random Forest Classifier, which were evaluated using the confusion matrix and accuracy. The system can classify accident severity into Serious, Fatal, and Minor injuries based on numerous significant accident-causing components discovered throughout the data mining process.

REFERENCES

[1] Road Accident Analysis and Hotspot Prediction using Clustering Jayesh patil U.G. Student Department of Computer Engineering
[2] Road Lytics : Road Accidents Analytics Using Artificial Intelligence to Support Deaths Prevention on Highways Kelvin Rinaldi da Luz 1 , João Elison da Rosa Tavares 1 , Jorge Luis Victória Barbosa 1 , Daniel Hernández de le
[3] RFCNN: Traffic Accident Severity Prediction based on Decision Level Fusion of Machine and Deep Learning Model mubariz Manzoor1 , Muhammad Umer2,4, Saima Sadiq2 , Abid Ishaq1 , Saleem Ullah
[4] Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction Daniel Santos * , José Saias, Paulo Quaresma and Vítor Beires Nogueira
[5] Road Accident Analysis using Machine Learning, IEEE Pune Section International Conference (PuneCon) Vishwakarma Institute of Technology, Pune India. Dec 16-18, 2020
[6] Article Factor Identification and Prediction for Teen Driver Crash Severity Using Machine Learning: A Case Study Ciyun Lin 1 , Dayong Wu 2
[7] The Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning : Seoul City Study Jonghak Lee 1 , Taekwan Yoon 2,* , Sangil Kwon 1 and Jongtae Lee 1
[8] Road Accident Analysis and Prediction of Accident Severity by Using ML in Bangladesh by Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain and Amit Kumar Das.
[9] Comparison of ML Algorithms for Predicting Traffic Accident Severity A Framework for Analysis of Road Accidents at the International Conference on Emerging Trends and Innovations in Engineering and Technological Research.

