International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper | Volume-9, Issue-9, 2022 Regular Edition | E-ISSN: 2349-7084

Road Accident Severity Prediction Using Machine Learning Algorithms

[1] Anukali Pramod Kumar, [2] D. Teja Santosh

[1] *M.Tech.-A.I. student, Department of CSE, CVR College of Engineering, Vastunagar, Mangalpally, Ibrahimpatnam, T.S., India – 501510.*

[2] *Associate Professor, Department of CSE, CVR College of Engineering, Vastunagar, Mangalpally, Ibrahimpatnam, T.S., India – 501510*

Email ID: [1]769pramodkumar@gmail.com, [2]tejasantoshd@cvr.ac.in

Corresponding author : Anukali Pramod Kumar

Available online at: http://www.ijcert.org

Received: 25/08/2022, Revised: 16/09/2022, Accepted: 26/09/2022, Published: 06/10/2022

Abstract: The majority of fatalities and serious injuries occur as a result of incidents involving motor vehicles. If the traffic management system is going to do its job of reducing the frequency and severity of traffic accidents, it needs a model for doing so. In this paper, we combine the results of three machine learning algorithms—logistic regression, decision tree, and random forest classifier—to build a predictive model. In order to forecast the severity of accidents in different regions, we used ML algorithms on a dataset of accidents from the United States. In addition, we examine vast quantities of traffic data, extracting helpful accident patterns in order to pinpoint the factors that have a direct bearing on road accidents and make actionable suggestions for improvement. When compared to two other ML algorithms, random forest performed best on accuracy. The severity rating in this paper is not meant to reflect the severity of injuries sustained, but rather how the accident affects traffic flow. Accident severity, decision trees, random forests, and logistic regression are all terms that are often used to describe this area of study.

Keywords: — accident severity, random forest, decision tree, logistic regression.

---

1. Introduction

One of the most determined issues today is that of road accidents [1]. It overwhelmingly impacts people's lives financially, physically, and emotionally. Hundreds of billions of dollars are lost annually due to the economic and social effects of traffic accidents and some major accidents that cause the majority of the damage. The WHO estimates that 1.35 million people are killed and over 50 million are injured annually due to traffic accidents worldwide [2]. Also, data shows that automobile accidents are the leading killer of kids and young adults aged 5 to 29 [3]. However, reducing traffic accidents is a huge challenge, especially fatal ones. His preventive method, which is one of the two main ways to make roads safer, works to get rid of dangers before they happen. For this plan to work, it must be possible to predict when accidents will happen and how bad they will be. When we know the causes and common factors that lead to these terrible events, we might be able to take better steps and make better use of our resources.

Machine learning algorithms will be used in this study to come up with a way to predict how bad traffic accidents will be. The main goal of this study is to find out which factors have the biggest effect on how bad an accident is. Second, we need to make models that can accurately predict how bad an accident will be. To be more specific, this model is meant to predict the likelihood of an accident is a severe one without any specific information about the accident itself, such as driver attributes or vehicle type. It could be a recent accident about which we don't know much, or it could be something made up by other models. So, the people who made this project's dataset also made a sophisticated way to predict major traffic accidents in real-time. This solution could help this model predict major accidents better in real-time.

A total of six parts make up this paper. In Part II, we will discuss some related studies. In Section III, we outline the plan's overall structure and methodology. In Section IV, the results of the experiments and how well the accident severity prediction framework works are analyzed and talked about. Within Section V, we cover blatant accident trends. After all that, the study's findings are presented in Section VI.

2. Related Work

This section presents a comprehensive review of current research on estimating accident severity. In a study by Najada et al. [4], accident causes were predicted using data from Hong Kong's transportation system Najada et al. [4], accident causes were predicted using data from Hong Kong's transportation system. They put several different classification algorithms into WEKA and compared how well they performed. Based on their testing, Random Forest proved to be superior to both Naive Bayes and PART. Similarly, Chong et al. [5] developed a model employing ML algorithms to categorize the severity of accident injuries into five groups. The model was created using a combination of artificial neural networks (ANN), support vector machines (SVM), and decision trees. According to their findings, exceeding the speed limit is one of the leading causes of serious and fatal accidents. In addition, the authors conducted another fascinating study [6] that shed light on the process of accident prediction by mining and analyzing large amounts of data. The authors also emphasized the importance of data sampling in the reconstruction of the dataset, as well as the use of prepossessing techniques to ensure the data is complete and accurate.

The research study conducted by Elfar et al. [7] used three machine learning algorithms: logistic regression, random forest, and neural networks to predict traffic accidents and congestion. Two predictive models were also proposed for use in data training. The accuracy of the proposed models was higher than that of the other three classification models tested in this study. Further, the authors' contributions demonstrated how their proposed models can be implemented in a variety of vehicle applications to enhance road safety by providing advanced notice of upcoming traffic slowdowns. Iranitalab [8] showed that linear regression, Naive Bayes, and Random Forest are just a few examples of the ML algorithms that have been shown to be effective in analyzing large datasets for predicting traffic accidents [9]. They developed and deployed a system that speeds up and improves the accuracy of their proposed framework by using the aforementioned ML algorithms in conjunction with a flexible architecture, the Lambda Architecture.

Most of the studies analyzed in the literature reviews were conducted in artificial settings or on relatively small data sets. Some other researchers [9] have found that neural networks can be used to great effect to recognize shifts in driver behavior, allowing for the prevention of potentially disastrous collisions. The authors employed two different types of deep learning models: the Long Short-Term Memory (LSTM) and the Recurrent Neural Network (RNN). Individual drivers' acceleration, deceleration, and speed patterns were used to train the model. According to the findings, the model successfully distinguished between safe and unsafe driving. Recent work [10] investigated problems in the transportation system. On top of that, they developed the framework for mining data in real time about transportation systems. Good analysis of traffic accidents was provided by this study, but unfortunately, due to data limitations, conclusive findings could not be drawn. By conducting a comprehensive literature review, we were able to identify numerous knowledge gaps in the field of accident severity prediction. Most prior research studies only predicted the accident associated with one or two factors, which is insufficient for a real-world situation [11]; ii) many studies do not address the class imbalance problem; iii) unobserved heterogeneity; and iv) most studies only rely on a single accuracy measure to evaluate the performance of the algorithm. Therefore, the overarching goal of this study is to close the aforementioned knowledge gaps.

3. Methodology

The goal of this study is to establish a system for forecasting the severity of accidents. Actions included in the proposed structure, as shown in Figure 1, are: Data on US traffic accidents is downloaded, cleaned, and preprocessed before being split into a training and test set, where predictive models are built using three different machine learning algorithms and then tested on real-world data to determine how accurately they predict accident severity. Finally, we quantify and contrast the efficacy of various algorithms.

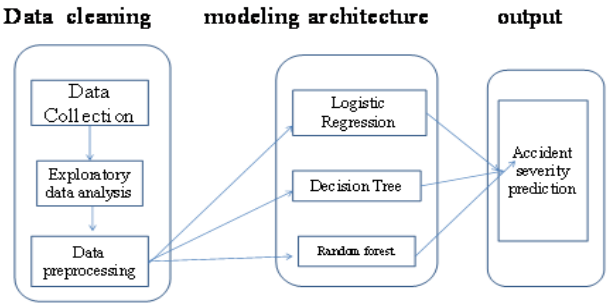


Figure 1: Accident severity Prediction Framework

A. Dataset summary

This database includes accidents from all 49 of the United States' states. Incident data from February 2016 through December 2020 is compiled using multiple APIs that provide data on traffic events in real time. In the United States, transportation agencies, police departments, and road networks all use application programming interfaces (APIs) to share traffic data collected from a variety of sources, including B. traffic cameras and traffic sensors. This dataset currently contains approximately 10,485,67 cases and 49 accident records. The attributes are as follows:

Brief Description Road accident dataset

Traffic Attributes (12):

- **ID**: This is a unique identifier for the accident record.
- **Source**: Indicates the source of the accident report (that is, the API that reported the accident).
- **TMC**: Traffic accidents may have a TMC (Traffic Message Channel) code that contains a more detailed description of the event.
- **Severity**: A number from 1 to 4 that indicates the severity of the incident. 1 indicates the least traffic impact (that is, a

short delay due to an accident) and a 4 indicates a severe traffic impact (that is, a long delay).
• **Start Time**: When an accident occurred, this is displayed here in local time.
• **End Time**: The time the accident concluded is displayed here in your own time zone.
• **Start_Lat**: This variable shows the GPS coordinates of the starting point's latitude.
• **Start_Lng**: The value of Start Lng represents the longitude component of the starting GPS coordinates.
• **End_Lat**: The GPS latitude of the final destination is shown.
• **End_Lng**: The GPS longitude of the destination is shown in the End Lng variable.
• **Distance (km):** The number of kilometers spanning the stretch of road closed due to the accident.
• **Description**: The accident is described in plain English.

Address Attributes (9):
• **House Number**: Shows the house number in the address box.
• **Street**: This option will show the street name in the address box.
• **Side**: Shows the relative side (right/left) of the road in the address field.
• **City**: Displays the city in the address field.
• **Country**: Displays the country in the address field.
• **State**: Displays the state in the address field.
• **Postal Code**: Shows the postal code in the address field.
• **Country**: Show the country in the address bar by selecting "Country."
• **Time Zone**: This displays the time zone that the accident occurred in (Eastern, Central, etc.).

Weather attributes (11):
• **Airport Code**: Tell us which airport weather station is closest to the accident.
• **Weather Time Stamp**: The time stamp of the weather observation record is displayed here (local time).
• **Temperature (F):** The temperature is shown here in Fahrenheit.
• **Wind-chill (F):** Indicator of wind chill temperature in degrees Fahrenheit (Fahrenheit).
• **Humidity (%):** Listed here is the current humidity in percent.
• **Pressure (inches):** Displays the air pressure in inches of mercury.
• **Visibility (mi):** This shows the visibility in miles per hour.
• **Wind Direction**: This indicates the direction in which the wind is blowing.
• **Wind Speed (mph):** How fast the wind is blowing is shown here in terms of miles per hour.
• **Precipitation (inches):** Precipitation, if any, in inches.
• **Weather Conditions**: Shows the current weather conditions (rain, snow, thunderstorms, fog, etc.).
POI attributes (13):
Amenity: An Annotation for a Point of Interest (POI) that identifies the presence of an amenity in the immediate area.
Bump: A POI annotation that indicates there is a speed bump or bump nearby.
Crossing: A POI annotation that indicates there is an intersection at a nearby location.
Give_Way: A POI annotation that indicates there is a Give_Way sign at a nearby location.
Junction: A POI annotation indicating that there is an intersection nearby.
No_exit: A POI annotation indicating that a nearby location has a No_exit sign.
Railway: A POI annotation indicating that there is a railroad in a nearby location.
Roundabout: A POI annotation indicating that there is a roundabout nearby.
Station: A point of interest annotation indicating the presence of a train, bus, or other station.
Stop: A POI annotation indicating that there is a stop sign at a nearby location.
Traffic Calming: Annotations added to POIs can reveal the presence of traffic calming in the immediate area.
Traffic lights: A traffic light is an object of interest that has been annotated on the map.
Turning Loop: When a turn loop is nearby, a POI annotation will show it.
Period of day (4):
Sunrise Sunset: Displays the current time (day or night) based on the position of the sun at sunrise or sunset.
Civil Twilight: Civil twilight is used to indicate the time of day.
Nautical Twilight: Shows the time of day (day or night) based on the time of twilight over the ocean (nautical twilight).
Astronomical twilight: Time of day (day or night) based on astronomical twilight is displayed.

B. Preprocessing

In order to detect and deal with corrupt or missing records, machine learning models must first undergo extensive preprocessing and cleaning of the incoming data. After that, most features underwent EDA (exploratory data analysis) and feature engineering. Data preprocessing is essential because it allows the raw data to be converted into a format that is more conducive to developing a machine learning model. The traffic fatality data set can be downloaded in CSV format. After obtaining the dataset, we filter out any extraneous information by removing duplicate attributes. The next step is to convert the attribute values from strings to numbers. Finally, have the data converted into ARFF [13] format after it has been normalized. Data normalization is a preprocessing technique that ensures comparable results across all attributes by standardizing their ranges (minimum, maximum, and average) without compromising on accuracy.

Exploratory Data Analysis (EDA)

The term "Exploratory Data Analysis" is used to describe the vital process of conducting preliminary investigations on data in order to discover patterns, to spot anomalies, to test hypotheses, and to check assumptions

with the aid of summary statistics and graphical representations.

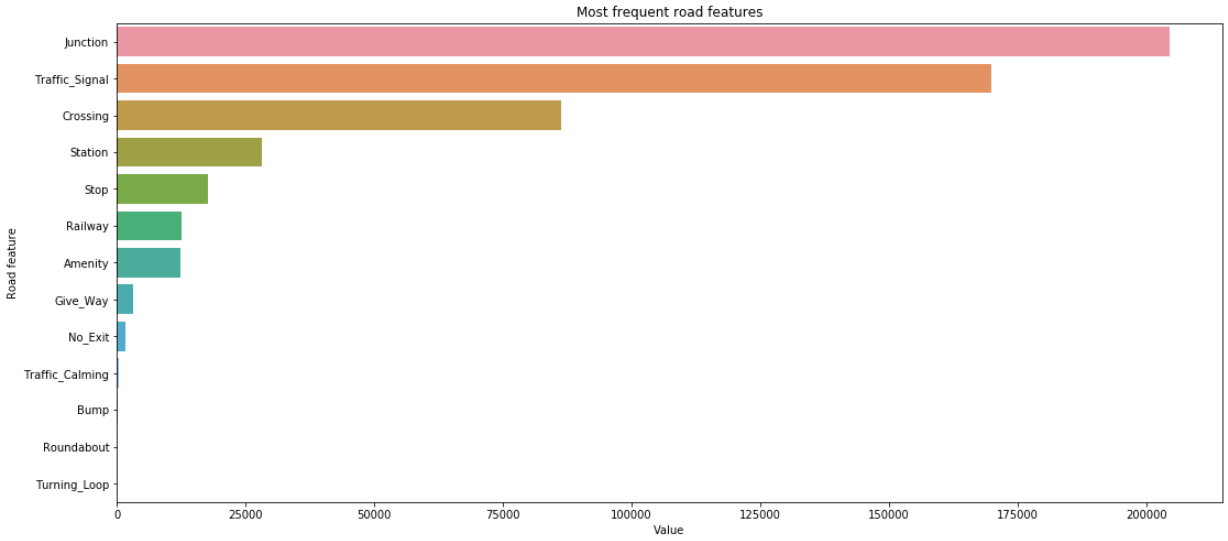


Figure 2: Most frequent Road features

As we can see, most of the accidents occurred near a traffic signal, especially where a junction or a crossing was present. The fourth most common road feature, instead, was the presence of a nearby station, probably because of the high presence of vehicles as shown in figure 2.

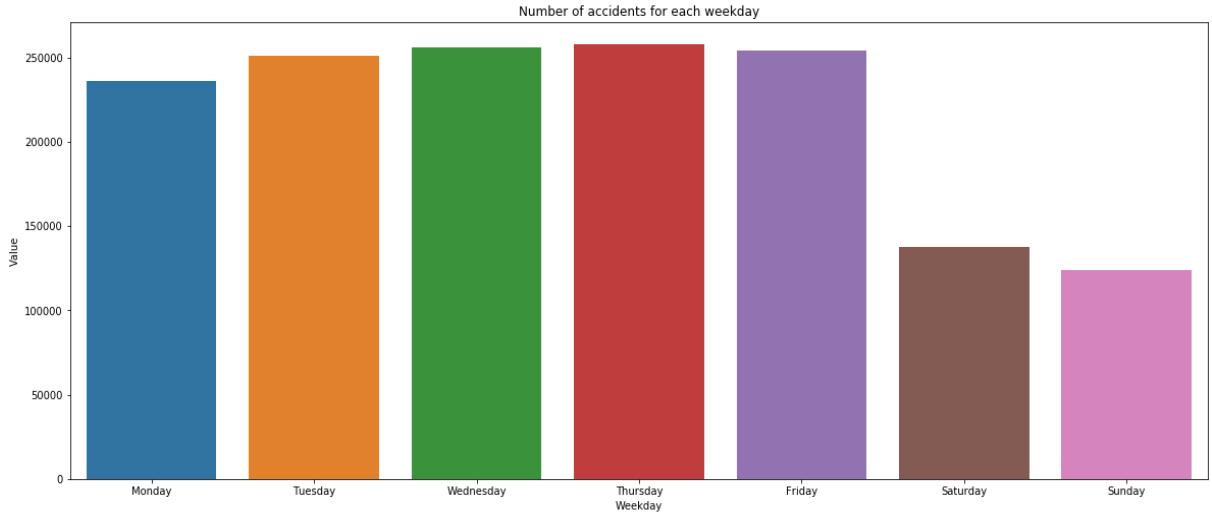
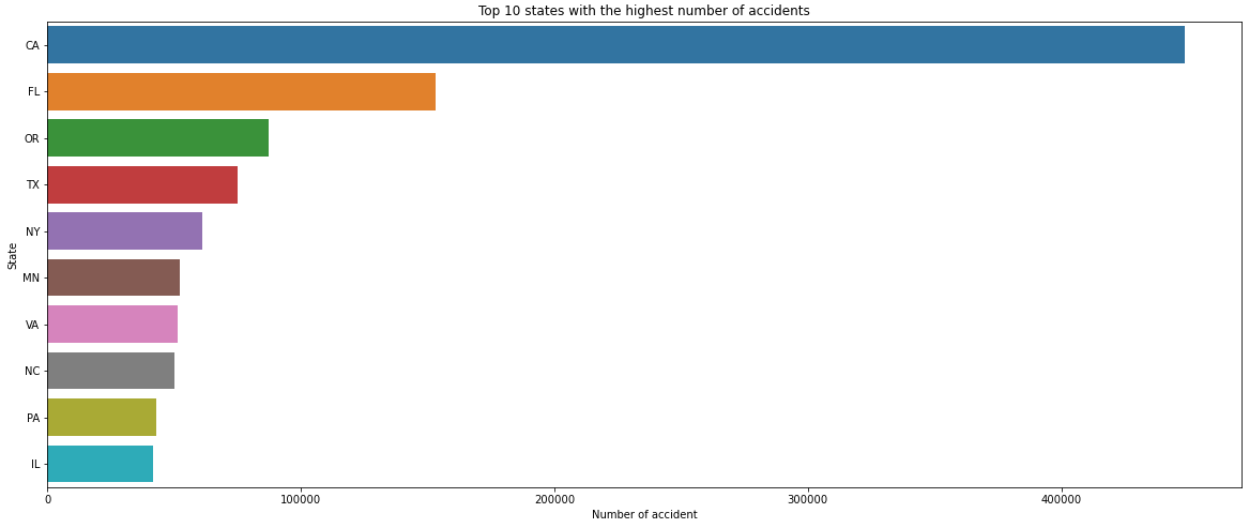


Figure 3: Number of accidents for each weekday

As we can see from the plot above, the days with the most accidents are working days; while in the weekend we have a frequency of at least 2/3 less. This may be due to the fact that during the weekend there are fewer vehicles on the road as shown in figure 3.



As we can see from the map and the plot above California is the state with the highest number of accidents, then we have Texas and Florida as shown in figure 4.

Figure 4: Top 10 states with highest number of accidents

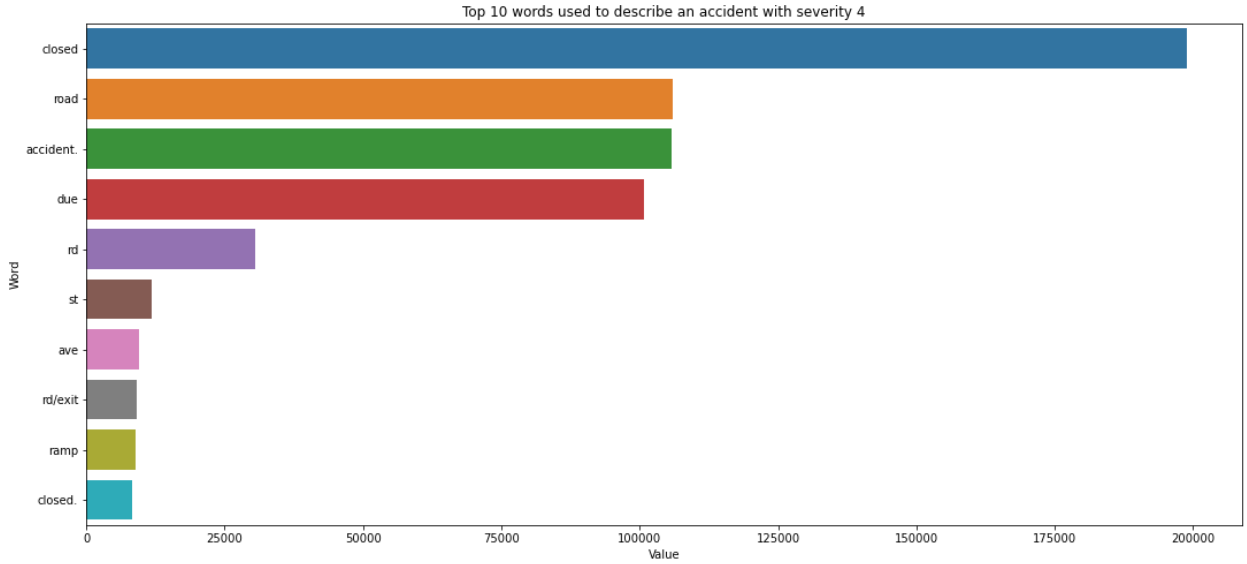


Figure 5: Top 10 words used to describe an accident with severity 4

We can see that the most used word in the description is closed. Subsequent words are accident, due and road as shown in figure 5.

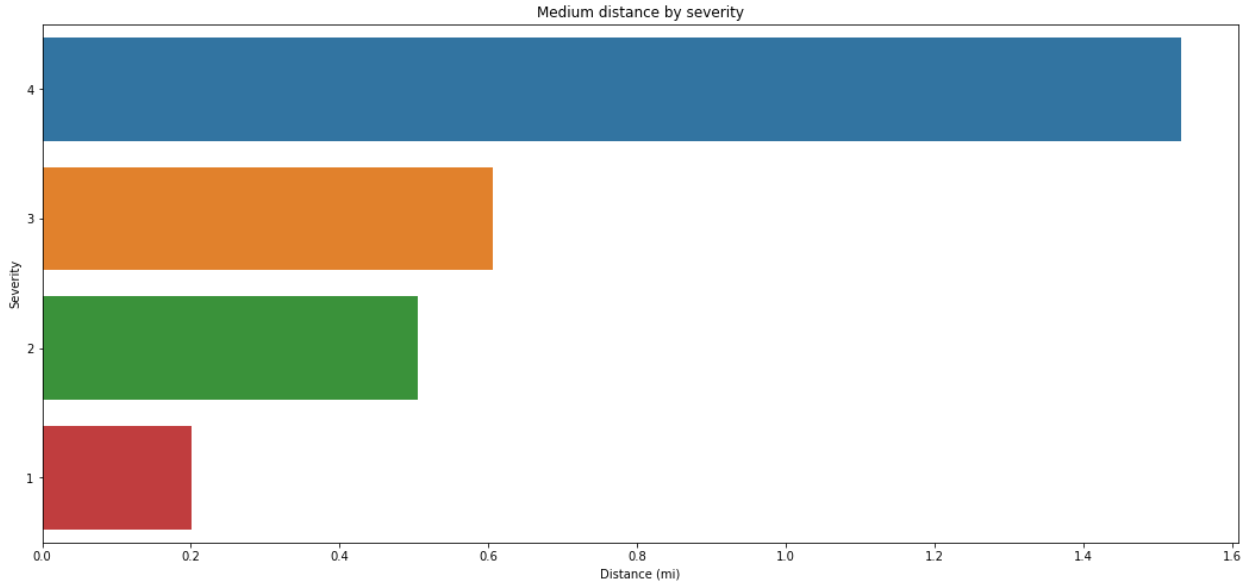


Figure 6: Medium distance by severity

In this graph we can see that the distance of the accident is more or less proportional to the severity, and in fact accidents with severity 4 have the longest distance as shown in figure 6.

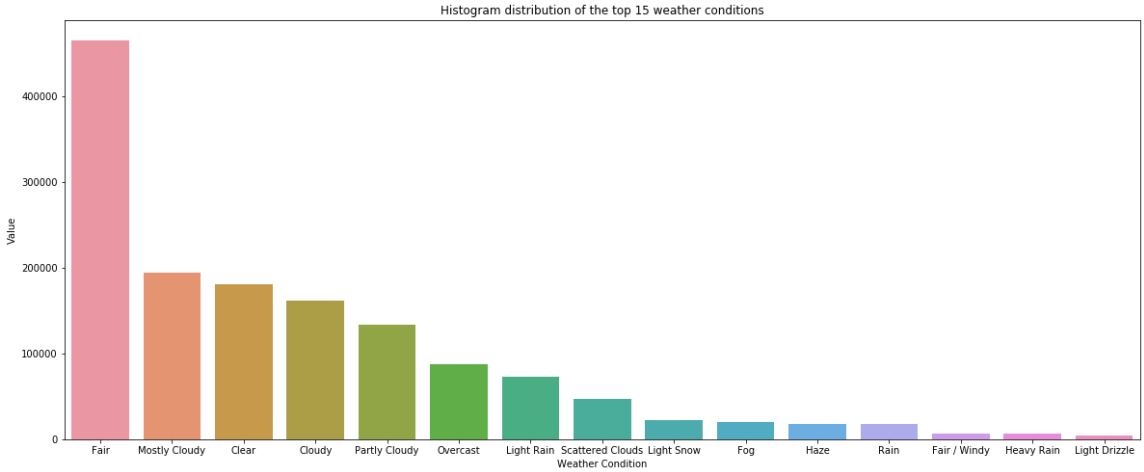


Figure 7: Histogram distribution of the top 15 weather conditions

In most frequent cases the weather condition is clear as shown in figure 7.

C. Machine Learning Algorithms for Prediction

After data preprocessing was finished, we split the data into a training set and a test set. Machine learning algorithms require training data before they can develop a model. Our prediction models include a dependent variable for the attribute class of accident severity. We do this by teaching three machine learning (ML) algorithms to predict the severity of accidents: Random Forest [16], Decision Tree [17], and Logistic Regression [19]. After the classifiers have been trained, the model is given the testing data in order to make predictions about the severity of accidents and compare the results of the various algorithms.

4. Results and Analysis

This section presents and discusses the experimental setup, methodology, and results for three distinct algorithms: random forest, decision tree, and logistic regression. Multiple analyses were performed to compare the three methods and determine which one is most accurate at predicting the severity of traffic accidents.

Jupiter notebook was used to conduct the experiments with pandas and seaborn. Intel i5 7th generation processor, 4GB or 8GB RAM, and 64-bit Windows 10 were used to power the machines that ran the experiments.

Table1: Parameters for the Algorithms

Algorithm	**Parameters**
Logistic regression	Random state=0, solver='lbfgs', multiclass='multinomial'

Decision tree	Random state = 120, criterion="entropy", max depth=4
Random forest	n_estimators=80

Classification accuracy

Recall is a measure of how well you remembered everything, while precision is a measure of how accurate you were. A high recall indicates that an algorithm successfully returned the majority of expected results. When an algorithm has "high precision," it produces more useful results than it does useless ones. Compared to the other algorithms used, RF was found to have a performance of 0.74, making it the clear winner.

The figure demonstrates that among the various algorithms, the random forest yields the best results. In light of this, it's clear that Random Forest outperforms the other two representative algorithms here.

In this case, the outcome proves that random forest is the superior machine learning method. Because it uses variables chosen at random, random forest is more robust to noise than many other algorithms. It can deal with both discrete and continuous information.

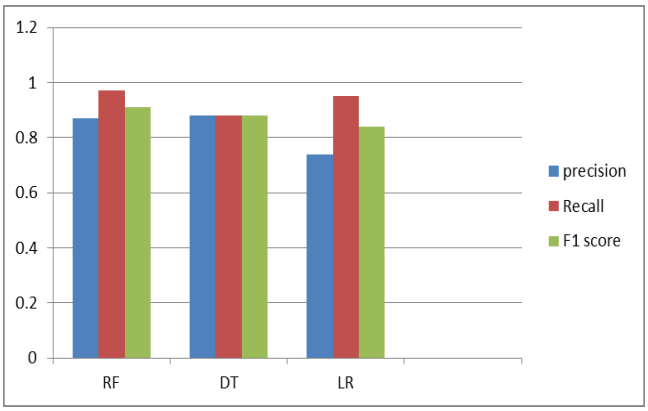


Figure 8: Comparison of algorithms with their precision, recall, F1 score

Table 2: Comparison of Algorithms with their precision, recall, F1 score

S. No	Algorithm	precision	Recall	F1 score
1.	Random forest	0.87	0.97	0.91
2.	Decision tree	0.88	0.88	0.88
3.	Logistic regression	0.74	0.95	0.85

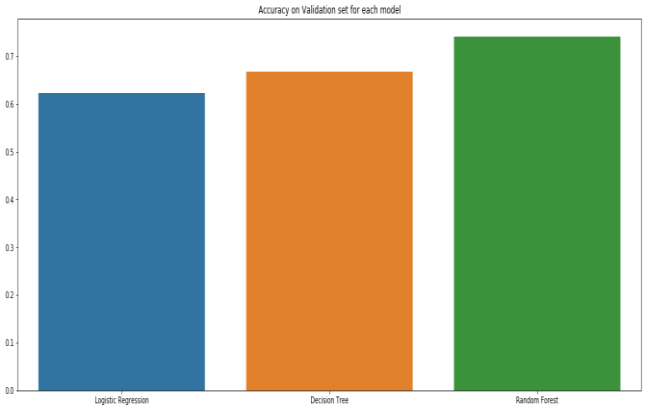


Figure 9: Accuracy on validation set for each model

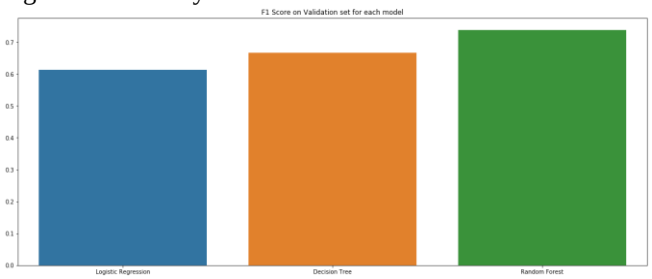


Figure 10: F1 score on validation set for each model

Table 3: The comparison of different algorithms with their accuracy

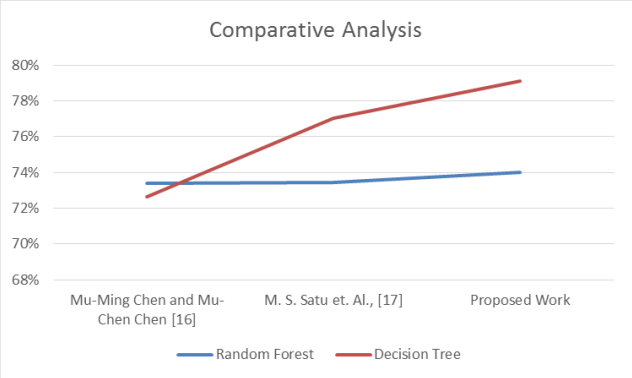
S.No	Algorithm	Accuracy (%)
1	Random forest	74
2	Decision Tree	67
3	Logistic Regression	62

Table 4: Comparison of experimental results with previous study

Author	Technique used	Algorithm	Accuracy
[16] Mu-Ming Chen and Mu-Chen Chen	Machine learning	Random forest, decision tree, logistic regression	73.38%
[17] M. S. Satu, S.Ahamed, F. Hossain, T. Akter, and D. M. Farid	Machine learning	Random forest ,decision tree	73.43%

Proposed work	Machine learning	Random forest, decision tree, logistic regression	74%

The experimental results are compared with previous study, as a previous study does not include stop words: stop words are used to eliminate unimportant words, allowing applications to focus on the important words instead. The proposed one work on large dataset considering the attributes such as traffic attributes, weather attributes, POI attributes, period of day attributes which are not included by previous studies. The visualization of the comparative analysis among the state-of-art works from the literature and the proposed work is shown below.



It is observed from the above figure that there is a good increase of 3.55% accuracy on an average of the considered machine learning algorithms in between the proposed work and the work of [16] and there is a slight increase of 1.34% accuracy on an average of the considered machine learning algorithms in between the proposed work and the work of [17]. These results specify that the proposed accident severity prediction framework by using machine learning algorithms has been implemented successfully.

5. Conclusion

The system was developed using the outcomes of an ML model to predict the extent of damage that could occur in the event of an accident. When estimating severity, comparing ML models yields more accurate estimates. The key influences on the identification of related elements on accident severity and duration also help the government in its efforts to mitigate accident effects. This technological system uses alerts to inform drivers of real hazards ahead. Users would get a lot of use out of a future recommendation system that works as a mobile app and can accurately predict how bad accidents will be while drivers are on the road.

References

[1] T. K. Bahiru, D. K. Singh, and E. A. Tessfaw, "Comparative study on data mining classification algorithms for predicting road traffic accident severity," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018, pp. 1655–1660.

[2] W. H. Organization, Global status report on alcohol and health 2018. World Health Organization, 2019.

[3] A. Jamal, M. Zahid, M. Tauhidur Rahman, H. M. Al-Ahmadi, M. Almoshaogeh, D. Farooq, and M. Ahmad, "Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study," International journal of injury control and safety promotion, pp. 1–20, 2021.

[4] H. Al Najada and I. Mahgoub, "Big vehicular traffic data mining: Towards accident and congestion prevention," in 2016 International Wireless Communications and Mobile Computing Conference (IWCMC). IEEE, 2017, pp. 256–261.

[5] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident data mining using machine learning paradigms," in Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary, 2018, pp. 415–420.

[6] H. Al Najada and I. Mahgoub, "Anticipation and alert system of congestion and accidents in vanet using big data analysis for intelligent transportation systems," in 2016 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2016, pp. 1–8.

[7] A. Elfar, A. Talebpour, and H. S. Mahmassani, "Machine learning approach to short-term traffic congestion prediction in a connected environment," Transportation Research Record, vol. 2672, no. 45, pp. 185–195, 2018.

[8] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," Accident Analysis & Prevention, vol. 108, pp. 27–36, 2017.

[9] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," in 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE, 2019, pp. 272–276.

[10] H. Ibrahim and B. H. Far, "Data-oriented intelligent transportation systems," in Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014). IEEE, 2019, pp. 322–329.

[11] J. Paul, Z. Jahan, K. F. Lateef, M. R. Islam, and S. C. Bakchy, "Prediction of road accident and severity of bangladesh applying machine learning techniques," in 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC). IEEE, 2020, pp. 1–6.

[12] "Road safety data - data.gov.uk," https://data.gov.uk/dataset/ cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data, (Accessed on 12/07/2022).

[13] "Attribute-relation file format (arff)," https://www.cs.waikato.ac.nz/ml/weka/arff.html, (Accessed on 09/07/2022).

[14] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagation for classification," International Journal of Computer Theory and Engineering, vol. 3, no. 1, pp. 1793–8201, 2011.

[15] S. A. Arhin and A. Gatiba, "Predicting crash injury severity at unsignalized intersections using support vector machines and naive bayes classifiers," Transportation Safety and Environment, vol. 2, no. 2, pp. 120–132, 2020.

[16] M.-M. Chen and M.-C. Chen, "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," Information, vol. 11, no. 5, p. 270, 2020.

[17] M. S. Satu, S. Ahamed, F. Hossain, T. Akter, and D. M. Farid, "Mining traffic accident data of n5 national highway in bangladesh employing decision trees," in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2017, pp. 722–725.

[18] M. Taamneh, S. Alkheder, and S. Taamneh, "Data-mining techniques for traffic accident modeling and prediction in the united arab emirates," Journal of Transportation Safety & Security, vol. 9, no. 2, pp. 146–166, 2017.

[19] N. Kamboozia, M. Ameri, and S. M. Hosseinian, "Statistical analysis and accident prediction models leading to pedestrian injuries and deaths on rural roads in iran," International journal of injury control and safety promotion, vol. 27, no. 4, pp. 493–509, 2020.

[20] S. Haynes, P. C. Estin, S. Lazarevski, M. Soosay, and A.-L. Kor, "Data analytics: Factors of traffic accidents in the uk," in 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT). IEEE, 2019, pp. 120–126.

[21] J. Gan, L. Li, D. Zhang, Z. Yi, and Q. Xiang, "An alternative method for traffic accident severity prediction: using deep forests algorithm," Journal of advanced transportation, vol. 2020, 2020.

