ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020

Estimating the Impact of Traffic Accidents with the Use of Machine Learning Techniques

Bagadi Giridhar, Banda Venkata Ramana, Boddeda Ganesh, Amujuru Venkata Mahesh

Abstract: Accidents involving motor vehicles account for the vast majority of significant injuries and deaths. A model is necessary for the traffic management system to accomplish its goal of decreasing the occurrence and severity of traffic accidents. This work presents a prediction model that is built using the combined results of three machine learning algorithms: logistic regression, decision tree, and random forest classifier. Applying ML techniques to a dataset of US incidents allowed us to predict the severity of accidents in various areas. Furthermore, we analyze massive amounts of traffic data, gleaning useful accident patterns to identify the elements directly impacting road accidents and provide practical recommendations for improvement. In terms of accuracy, random forest outperformed two other ML systems. The impact on traffic flow is the primary metric for severity assessment in this article, rather than the seriousness of injuries. This field is often referred to by a variety of names, including decision trees, logistic regression, random forests, and accident severity.

Keywords: Relevant search terms include: logistic regression, decision tree, accident severity, and random forest.

INTRODUCTION:

The problem of vehicular collisions is one of the most pressing ones right now [1]. To put it simply, it has profound psychological, physiological, and monetary effects on people's daily lives. The societal and economic costs of road accidents, especially the most devastating ones, amount to hundreds of billions of dollars yearly. Every year, road accidents cause the deaths of 1.35 million people and injuries to more than 50 million, according to the World Health Organization [2]. In addition, the top cause of death for individuals between the ages of 5 and 29 is car accidents, according to the statistics [3]. But lowering the number of road accidents, particularly deadly ones, is no easy feat. His approach is to eliminate potential hazards in the future, which is one of the primary means of enhancing road safety. Predicting the frequency and severity of accidents is crucial to the success of this strategy. We may be able to do better with our resources and take greater action if we can identify the reasons and commonalities that contribute to these tragic occurrences. This project aims to develop a method for predicting the severity of traffic accidents by using machine learning algorithms. The problem of vehicular collisions is one of the most pressing ones right now [1]. To put it simply, it has profound psychological, physiological, and monetary effects on people's daily lives. The societal and economic costs of road accidents, especially the most devastating ones, amount to hundreds of billions of dollars yearly. Every year, road accidents cause the deaths of 1.35 million people and injuries to more than 50 million, according to the World Health Organization [2]. In addition, the top cause of death for individuals between the ages of 5 and 29 is car accidents, according to the statistics [3]. But lowering the number of road accidents, particularly deadly ones, is no easy feat. His approach is to eliminate potential hazards in the future, which is one of the primary means of enhancing road safety. Predicting the frequency and severity of accidents is crucial to the success of this strategy.

Associate Professor, Assistant Professor[2,3,4]
Department of CSE

We may be able to do better with our resources and take greater action if we can identify the reasons and commonalities that contribute to these tragic occurrences.

This project aims to develop a method for predicting the severity of traffic accidents by using machine learning algorithms. The severity of an accident is most affected by many variables. Additionally, it is imperative that we develop models capable of reliably estimating the severity of an accident. Without knowing the exact details of the event, such as the driver's characteristics or the kind of vehicle involved, this model can only estimate the probability that an accident will be a serious one. Either other models have made it up or it was a recent accident about which we have little information. The developers of this project's dataset have created an advanced method for real-time prediction of serious traffic accidents. This approach has the potential to improve the model's ability to foretell catastrophic incidents in real-time.

The whole of this paper consists of six sections. We will talk about a few similar research in Part II. We lay out the framework and methods of the plan in Section III. The experimental findings and the efficacy of the accident severity prediction framework are discussed and evaluated in Section IV. Section V discusses patterns of obvious accidents. Section VI presents the study's results after all that.

1. Related Work

A survey of the most recent studies on the topic of accident severity assessment is provided in this section. Using data from Hong Kong's transportation infrastructure, Najada et al. [4] were able to forecast the causes of accidents. By analyzing data from Hong Kong's transportation system, the authors Najada et al. [4] were able to anticipate the causes of accidents. They tested the effectiveness of several categorization algorithms by feeding them into WEKA. Their results showed that Random Forest outperformed Naive Bayes and PART. A model that uses ML algorithms to classify accident injuries into five classes was also created by Chong et al. [5]. Multiple types of artificial neural networks (ANNs), support vector machines (SVMs), and decision trees were used to construct the model. They found that going above the speed limit is a major contributor to accidents that result in significant injuries or death. Furthermore, the authors performed yet another intriguing research [6] that clarified how accident prediction is accomplished by data mining and analysis of massive volumes. Data sampling and prepossessing strategies were also highlighted by the authors as crucial to the dataset reconstruction process, with the former helping to guarantee accurate and full data.

Elfar et al. [7] used logistic regression, random forest, and neural networks as three machine learning techniques to forecast traffic accidents and congestion. The data training process also included the suggestion of two prediction models. Compared to the other three classification models evaluated in this research, the suggested models demonstrated superior accuracy. In addition, the authors' work showed that their suggested models may be used in different kinds of vehicles to improve road safety by warning drivers ahead of time when traffic would be slowing down. Several ML algorithms have shown efficacy in evaluating massive datasets for traffic accident prediction, as demonstrated by Iranitalab [8]. These algorithms include, but are not limited to, linear regression, Naive Bayes, and Random Forest. By combining the aforementioned ML techniques with the Lambda Architecture, a flexible framework, they were able to increase the speed and accuracy of their proposed framework.

Research included in the literature reviews

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020

tended to use either very tiny data sets or artificial environments. Other researchers have shown that neural networks can effectively detect changes in driving behavior, which might lead to the avoidance of catastrophic accidents [9]. In their study, the authors made use of two distinct deep learning models: RNN and Long Short-Term Memory (LSTM). To train the model, we looked at the speed, acceleration, and deceleration tendencies of individual drivers. The results show that the algorithm was able to differentiate between drivers who were safe and those who were not. There has been recent research [10] that looked at transportation system issues. Additionally, they created the infrastructure for real-time transportation data mining. While this research did a good job of analyzing road incidents, the data restrictions, definitive conclusions could not be made. Our extensive literature search revealed some unanswered questions about accident severity prediction. The majority of previous studies failed to address the following issues: (i) the class imbalance problem was not addressed; (iii) unobserved heterogeneity was not taken into account; and (iv) the algorithm's performance was evaluated using only one accuracy measure, which is inadequate for real-world scenarios [11]. The primary objective of our research is, therefore, to fill such gaps in our understanding.

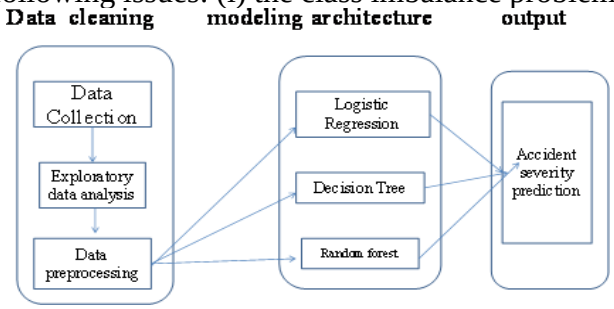


Figure 1: Accident severity Prediction Framework

2. Methodology

Finding a way to predict how bad accidents would be is the driving force behind this research. The steps that make up the suggested framework are as follows, as seen in Figure 1: After downloading, cleaning, and preprocessing data on traffic accidents in the US, it is divided into a training set and a test set. Using three distinct machine learning methods, predictive models are constructed and evaluated on real-world data to ascertain the accuracy with which they predict the severity of accidents. In the end, we compare and contrast the algorithms' effectiveness by quantifying it.

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 datasetTraffic 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

(a brief halt as a result of an accident) and a four denotes a substantial traffic effect (a protracted halt).

Here you may see the start time, in local time, of when the accident happened.

Here you may see the moment the accident ended, expressed in your local time.

The beginning point's latitude may be found in the GPS coordinates stored in the Start_Lat field.

• Start_Lng: The initial GPS coordinates' longitude component is represented by the value of Start Lng.
• End_Lat: Displays the GPS coordinates of the ultimate location.
• End Lng: This variable displays the GPS longitude of the destination.
• Length (kilometers): The length of the accident-closed road in kilometers.
The accident is explained in simple English.

Nine attributes to address:
Use the "House Number" option to display the home number beside the address.
• Street: If you choose this option, the address box will display the name of the street.
• Side: Displays the road's relative side in the address field, whether it's on the right or the left.
• City: Shows the city in the address box when selected.
• Country: Specifies the nation where the address is being entered.
•~State: Shows the current state in the address bar.
The postal code is shown in the address box by default.
• nation: To display the nation in the address bar, just pick "Country."
The time zone in which the accident happened is shown here (Eastern, Central, etc.).

Eleven weather attributes: • Airport Code: Please provide the name of the airport's weather station that is nearest to the scene of the accident.
In this section, you will see the time stamp (local time) of the weather observation record.
Fahrenheit is the unit of measurement used to display the current temperature.
• Wind-chill (F): A numerical indication of the wind chill temperature in Fahrenheit.
• Humidity (%): The present humidity is shown here in percentage form.
Shows the air pressure in inches of mercury (pressure).
You can see the visibility in miles per hour using the visibility (mi) indicator.
The wind's direction is shown by the wind direction indicator.
• Wind Speed (mph): The wind speed is shown here in miles per hour.
• Rainfall (inches): Total amount of precipitation (if any) expressed in inches.
If there is rain, snow, thunderstorms, fog, or any other kind of weather, it will be shown here.
13 properties of POI:
A point of interest (POI) annotation that indicates the existence of an amenity in the nearby region is called an amenity.

There is a point of interest annotation for a speed bump or bump nearby, and it's called a bump.
"Crossing" is a point-of-interest annotation that shows a crossroads in the immediate area.
Points of Interest (POIs) with the annotation "Give Way" show that a Give Way sign is located close by.
Junction: A point-of-interest annotation showing the presence of a nearby junction.
A point-of-interest annotation that indicates the presence of a No_exit sign at a nearby place.
Annotation for points of interest (POIs) that indicates the presence of a railroad in the immediate vicinity.
Roundabout: This point-of-interest annotation shows that a roundabout is in the area.
A station is a kind of point of interest annotation that shows the location of a bus, rail, or other type of station.
Stop: A point-of-interest annotation showing the presence of a stop sign in the immediate vicinity.
The existence of nearby traffic calming measures may be shown by annotations added to points of interest (POIs).
The term "traffic light" refers to a kind of annotated feature on a map.
A point-of-interest annotation will display the location of a turn loop if one is nearby.
Time of day four:
Sunrise At sunset, it shows you the time of day according to where the sun is as it rises or sets.
The term "civil twilight" describes the time of day in the United States.
Displays the current time as a function of nautical twilight, which indicates whether it is day or night on land.
The display of the time of day is dependent on astronomical twilight, whether it is day or night.

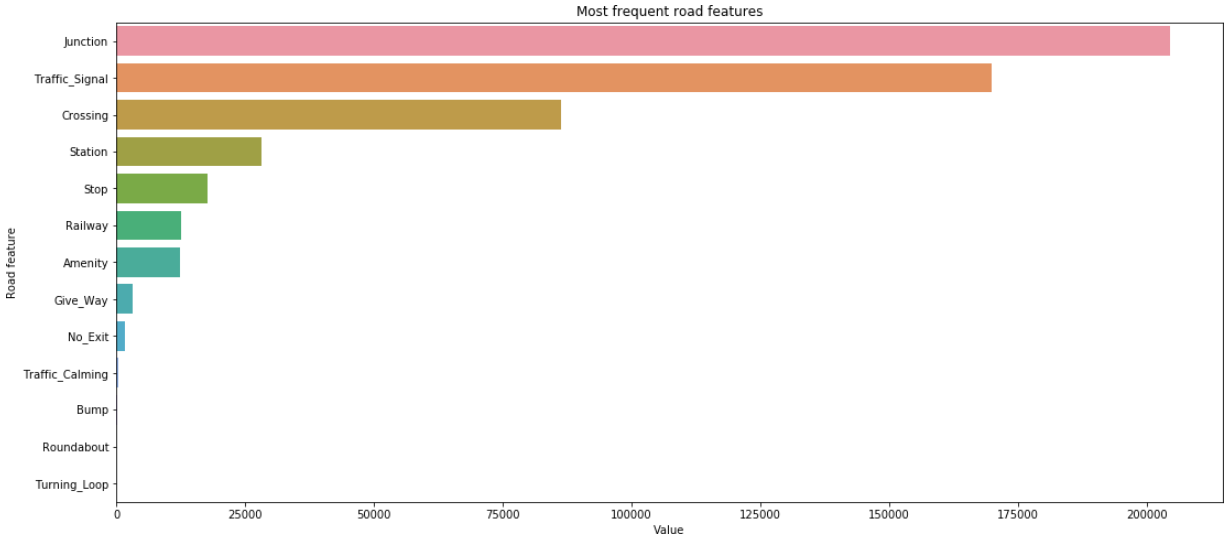
B. Initial Steps

Extensive preprocessing and cleaning of incoming data is necessary for machine learning algorithms to recognize and handle faulty or missing information. Following this, the majority of features were subjected to feature engineering and exploratory data analysis (EDA). An important step in building a machine learning model is data preprocessing, which involves transforming raw data into a more suitable format. The data collection including traffic fatalities is available for download in CSV format. Once we get the dataset, we remove duplicate attributes to remove any unnecessary information. Changing the attribute values from strings to numbers is the next procedure. After the data has been standardized, the last step is to convert it into ARFF [13] format. By normalizing the ranges (minimum, maximum, and average) of all characteristics, data normalization guarantees similar findings across all attributes without sacrificing accuracy.

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020

Analysis of Exploratory Data (EDA)



An essential step in discovering patterns, anomalies, testing ideas, and verifying assumptions, "Exploratory Data Analysis" involves doing first inquiries on data.

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.

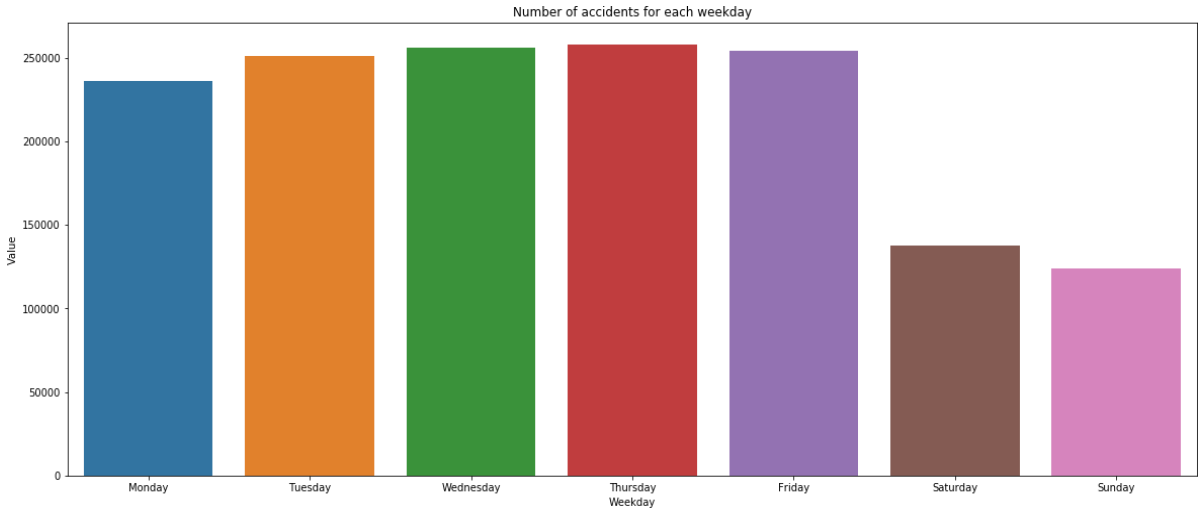
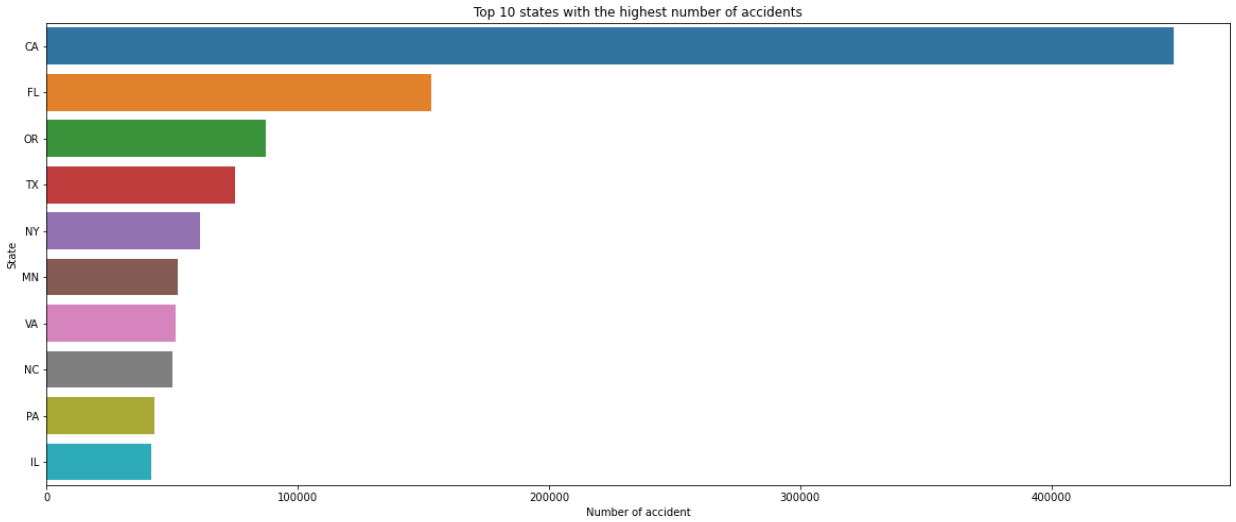


Figure 3: Number of accidents for each weekday

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.

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020



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

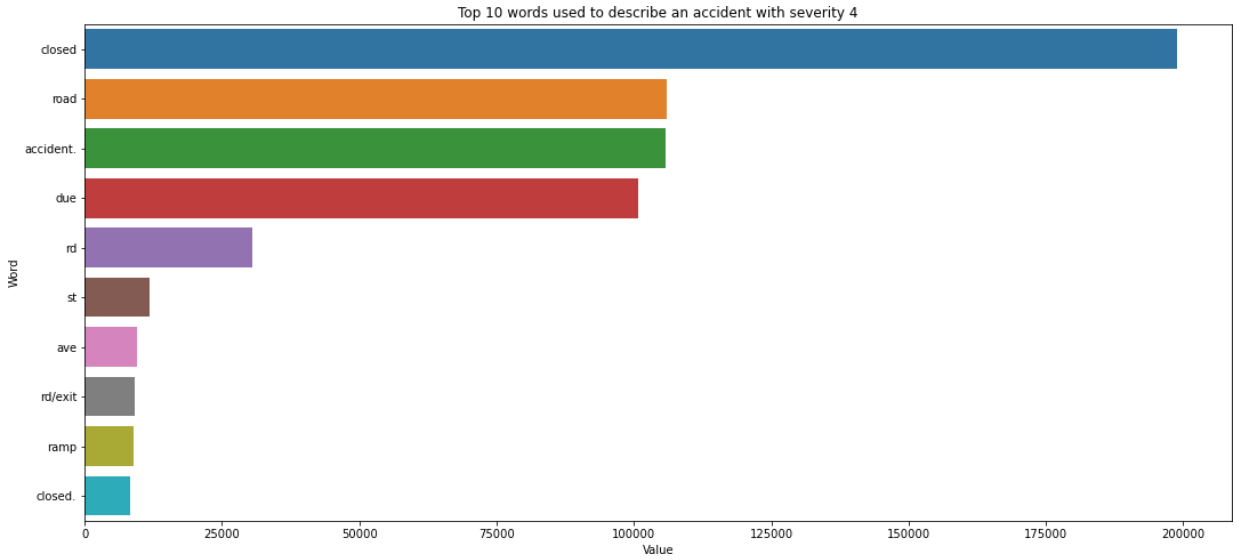


Figure 5: Top 10 words used to describe an accident with severity 4

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.

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020

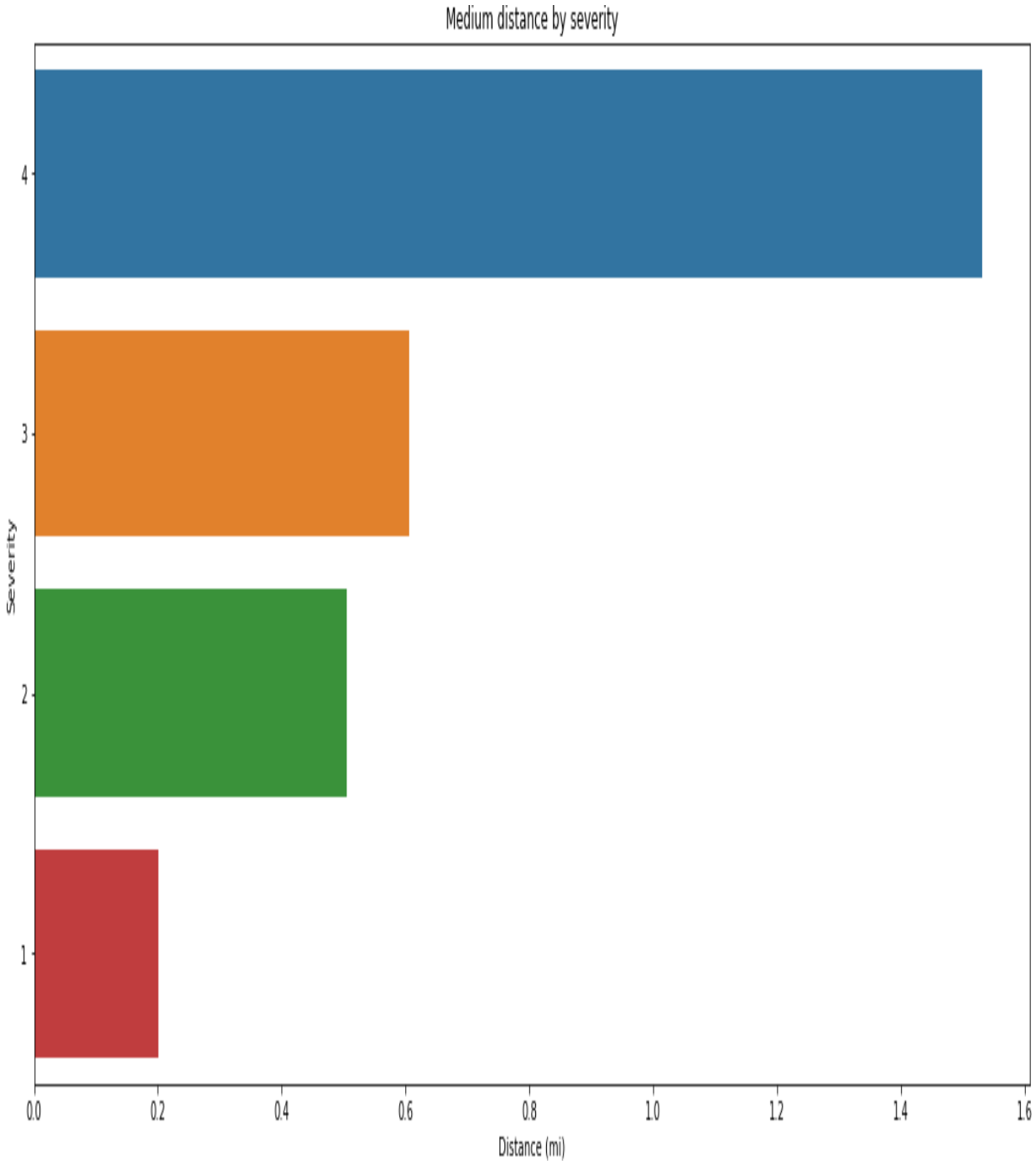


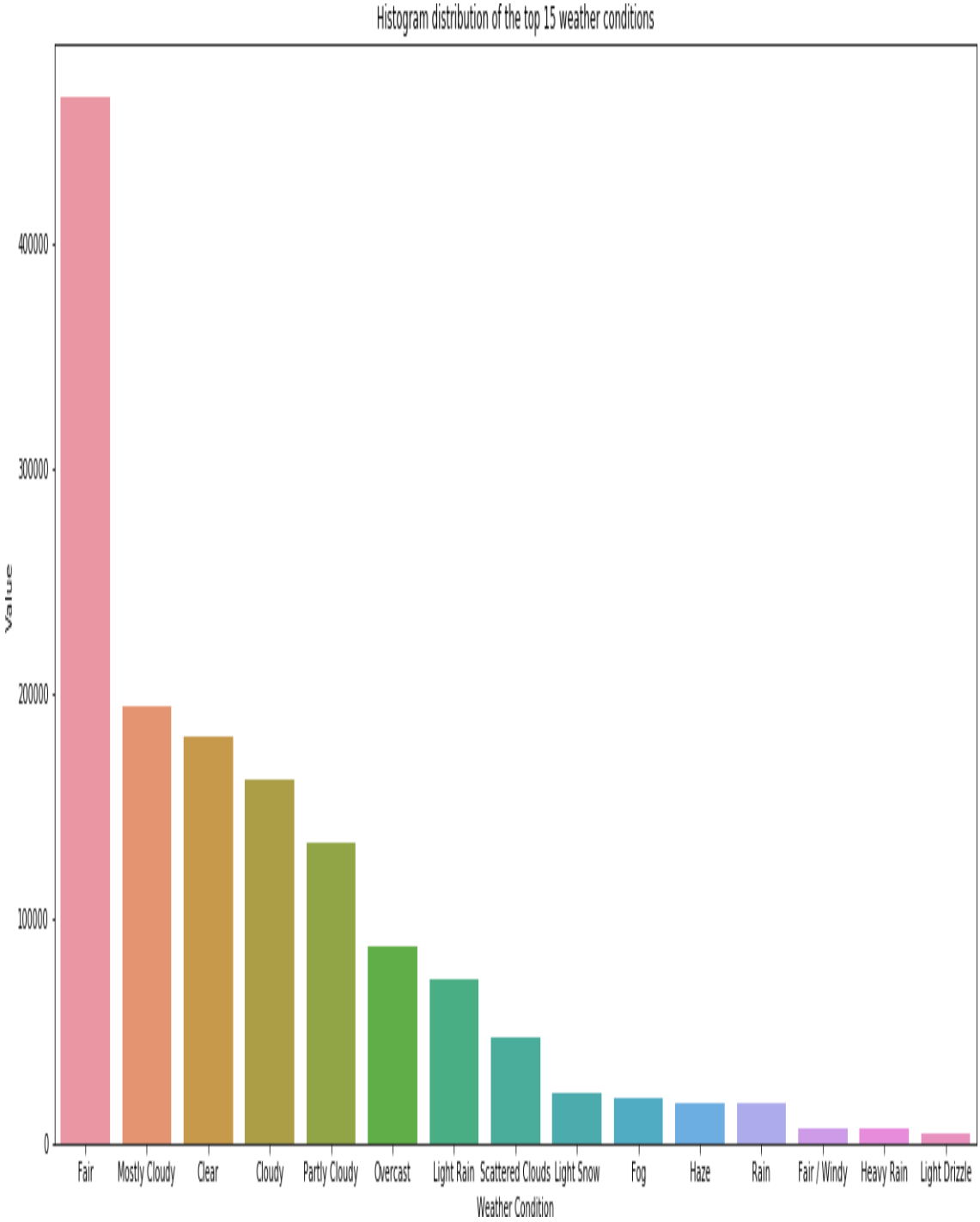
Figure 6: Medium distance by severity

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.

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020

Machine Learning Algorithms for Prediction



After data preprocessing was finished, we split the datainto 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 variablefor the attribute class of accident severity. We do this by teaching three machine learning (ML) algorithms to predictthe 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.

3. Results and Analysis

This section presents and discusses the experimental setup, methodology, and results for three distinct algorithms: random forest, decision tree, and logistic

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

The degree to which you were correct is measured by accuracy, whereas recall indicates how well you recalled all of the information. When an algorithm has a high recall, it means it returned most of the predicted outcomes. In the case of algorithms with "high precision," the number of valuable outcomes outweighs the number of meaningless ones. With a performance score of 0.74, RF easily outperformed the other algorithms tested. You can see that the random forest method outperforms the others in the picture. Given this, it's easy to see that Random Forest is the superior algorithm among the three examples given. Here, the results show that random forest is the best ML technique. Compared to other algorithms, random forest is better at handling noise since it employs variables that are picked at random. Data that is either discrete or continuous may be handled by it.

ISSN2321-2152 www.ijmece .com

Vol 8, Issuse.3July 2020

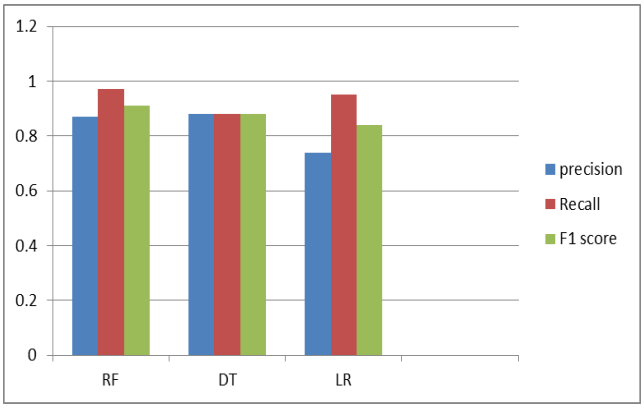


Figure 8: Comparison of algorithms with their precision, recall, F1 s

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

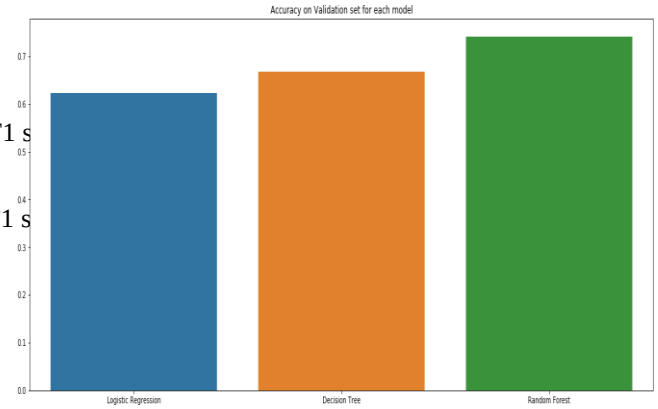


Figure 9: Accuracy on validation set for each model

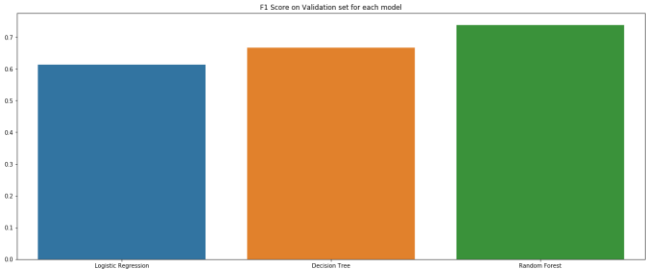


Figure 10: F1 score on validation set for each model

Table 3: The comparison of different algorithms with their accuracy

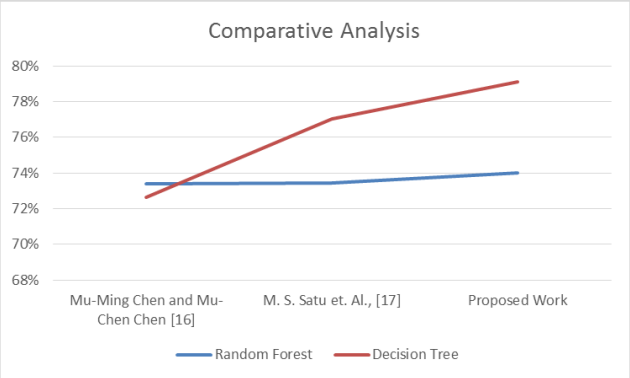
S.No	Algorithm	Accuracy (%)
1	Random forest	74
2	Decision Tree	67
3	Logistic Regression	62

Table 4: Comparison of experimental results with previousstudy

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%

The experimental results are compared with previous study, as a previous study does not include stop words: stopwords 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 workand the work of [16] and there is a slight increase of 1.34%accuracy on an average of the considered machine learningalgorithms 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.

For this system to work, the results of a machine learning model were used to foretell how much damage may happen in the case of an accident. More precise estimations of severity may be obtained by comparing ML models. In addition to assisting the government in its endeavors to reduce the impact of accidents, the identification of relevant factors that affect accident severity and length is a crucial factor. This technical system warns vehicles of impending dangers via alarms. Customers would benefit greatly from an upcoming suggestion system that functions like a mobile app and can reliably foretell the severity of incidents that vehicles may encounter while on the road.

4. REFERENCES

In their 2018 paper presented at the Second International Conference on Inventive Communication and Computational Technologies (ICICCT), T. K. Bahiru, D. K. Singh, and E. A. Tessfaw compare several data mining classification techniques with the goal of forecasting the severity of road traffic accidents. Pages. 1655-1660, 2018 IEEE. "Global Status Report on Alcohol and Health 2018" (WHO). The 2019 World Health Organization Report.

This is a comparative research on injury severity prediction in traffic accidents using ensemble machine learning approaches. The authors are A. Jamal, M. Zahid, M. Tauhidur Rahman, H. M. Al-Ahmadi, M. Almoshaogeh, D. Farooq, and M. Ahmad. The paper was published in the International Journal of Injury Control and Safety Promotion in 2021 and had pages 1–20.

"Big vehicular traffic data mining: Towards accident and congestion prevention," presented at the 2016 International Wireless Communications and Mobile Computing Conference (IWCMC) by H. Al Najada and I. Mahgoub, is referenced as reference 3. Volume 25, Issue 6, Pages 256–261, IEEE, 2017.

(ISDA'04) Proceedings of the Fourth International Conference on Intelligent Systems Design and Applications, Budapest, Hungary, 2018, pp. 415–420, by M. Chong, A. Abraham, and M. Paprzycki.

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

[6] "Machine learning approach to short-term traffic congestion prediction in a connected environment," published in 2018 in the Transportation Research Record, volume 26,

issue 45, pages 185–195, by A. Elfar, A. Talebpour, and H. S. Mahmassani.

"Comparison of four statistical and machine learning methods for crash severity prediction," published in Accident Analysis & Prevention in 2017, was written by A. Iranitalab and A. Khattak.

The 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) included a paper by R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer titled "Comparison of machine learning algorithms for predicting traffic accident severity ". pages 272-276, 2019 IEEE.

In the 2014 IEEE 15th J. Paul, Z. Jahan, K. F. Lateef, M. R. Islam, and S. C. Bakchy's paper "Data-oriented intelligent transportation systems" was published. In the 2020 IEEE 8th R10 Humanitarian Technology Conference (R10- HTC), the authors predict the frequency and severity of road accidents in Bangladesh using machine learning techniques. 2020 IEEE, pages 1–6.

"Data available at data.gov.uk - road safety," The data on road safety may be found at this URL: https://data.gov.uk/dataset/ cb7ae6f0-4be6-4935-9277- 47e5ce24a11f. (Accessed on 12/07/2022).

"Attribute-relation file format (arff)," (accessible at 09/07/2022) https://www.cs.waikato.ac.nz/ml/weka/arff.html .

The 2011 issue of the International Journal of Computer Theory and Engineering, special issue on "Statistical normalization and back propagation for classification" (vol. 3, no. 1, pp. 1793-8201), was written by T. Jayalakshmi and A. Santhakumaran.

"Predicting crash injury severity at unsignalized intersections using support vector machines and naive bayes classifiers," published in 2020 in Transportation Safety and Environment, with an ISBN number of 120-132, is written by S. A. Arhin and A. Gatiba.

In their 2020 article titled "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," M.-M. and M.-C. Chen discuss the use of several statistical methods to predict the severity of road accidents.

Presented at the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), the paper "Mining traffic accident data of n5 national highway in Bangladesh employing decision trees" was written by M. S. Satu, S. Ahamed, F. Hossain, T. Akter, and D. M. Farid. Pages 722–725 in IEEE, 2017.

Article published in 2017 in the Journal of Transportation Safety & Security, by M. Taamneh, S. Alkheder, and S. Taamneh titled "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates," pages 146–166.

In 2020, N. Kamboozia, M. Ameri, and S. M. Hosseinian published an article in the International Journal of Injury Control and Safety Promotion titled "Statistical analysis and accident prediction models leading to pedestrian injuries and deaths on rural roads in Iran." The article can be found on pages 493-509 of the journal's volume 27, issue 4.

Abstract: "Data analytics: Factors of traffic accidents in the UK," presented at the 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), by S. Haynes, P. C. Estin, S. Lazarevski, M. Soosay, and A.-L. Kor. Pages. 120-126, 2019 IEEE.

The article "An alternative method for traffic accident severity prediction: using deep forests algorithm" was published in the 2020 volume of

the Journal of Advanced Transportation by J. Gan, L. Li, D. Zhang, Z. Yi, and Q. Xiang.