National College of Ireland

Project Submission Sheet – 2022/2023

Student Name:	Akanksha Tambe, Ayushi Bajaj, Apurva Kumari, Nikhil Salvi, Sanica Kamble		
Student ID:	20225423, 20242638, 21118175, 20179529, 21130701		
Programme:	MS in Data Analytics	**Year:**	2022-2023
Module:	Domain Application of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	12/08/2022		
Project Title:	Predictive Analysis of Traffic Accidents in United States		
Word Count:	3616		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Akanksha Tambe, Ayushi Bajaj, Apurva Kumari, Nikhil Salvi, Sanica Kamble

Date: 12/08/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Traffic Accidents Severity Prediction in United States

Sanica Kamble
Master of Science.in Data Analytics
National College of Ireland
Dublin, Ireland
x21130701@student.ncirl.ie

Ayushi Bajaj
Master of Science.in Data Analytics
National College of Ireland
Dublin, Ireland
x20242638@student.ncirl.ie

Apurva Kumari
Master of Science.in Data Analytics
National College of Ireland
Dublin, Ireland
x21118175@student.ncirl.ie

Nikhil Salvi
Master of Science.in Data Analytics
National College of Ireland
Dublin, Ireland
x20179529@student.ncirl.ie

Akanksha Tambe
Master of Science.in Data Analytics
National College of Ireland
Dublin, Ireland
x20225423@student.ncirl.ie

***Abstract*— Accident analysis has been the focus of a significant amount of study over the last several decades as a direct result of the critical nature of the problem of lowering the number of traffic accidents around the globe. The purpose of the project is to perform a study of accident data from all 50 states in the United States to provide US government agencies and the public with information on patterns and possible causes of traffic accidents as well as what might be done to minimize the number of accidents. The analysis takes into account the number of accidents that occurred each year, the number of accidents that occurred within each state, the number of accidents that occurred each day and hour, severity of the accident, accident-prone areas within each state, and the factors that were responsible for the accidents, such as the weather, wind flow, temperature, location, etc. Python and power BI is used in the development of the analysis platform. The conclusions derived from analysis are that California has the highest number of accidents among all the states in the country and we can see that most of the accidents have occurred when the weather was fair and clear whereas the least number of accidents occurred while it was raining and light snow.**

Keywords*— *Python, PowerBI, temperature, traffic, severity

I. INTRODUCTION

Accidents on the road have grown more frequent in recent years. Every year, around 1.25 million people are killed in motor vehicle accidents, which works out to an average of 3,287 deaths each day. In addition, every year between 20 and 50 million individuals are injured or permanently disabled. Accidents involving motor vehicles are the ninth most common cause of death worldwide and are responsible for 2.2% of all fatalities. Globally, the costs of traffic collisions are USD 518 billion, which amounts to between 1% and 2% of the GDP of nations.[1] Every year, traffic accidents are responsible for the deaths of about 37,000 individuals in the United States, and they leave another 2.35 million people wounded or disabled. Accidents on America's roads cost the country an annual average of $820 per person, or 230.6 billion dollars total. Accidents involving motor vehicles are the single most common way for otherwise healthy United States residents to lose their lives when traveling in other countries. Even though, Road safety analysis has made tremendous progress over the years, particularly in the development of approaches for modelling the link between injury severity and risk variables, acquiring information about the mechanism of accident occurrence, and designing safety regulations and countermeasures. This understanding of the causes of road traffic deaths and injuries has remained a focus, and road safety analysis has made great progress. Over the course of the last few decades, research has been conducted in an attempt to gain a better understanding of the fundamental connection that exists between the frequency of collisions and the various risk factors, including but not limited to road geometry, vehicle type, collision type, seasonal effect, traffic regulation, time of day, driver characteristics, and environmental conditions. Estimating the severity of injuries sustained in car accidents has been done with the use of a variety of statistical methods. In the past, statistical approaches have been the most common approach to both the study and prediction of injury severity.

US-Accidents has a wide range of applications, including the prediction of real-time car accidents, the investigation of accident hotspots, the analysis of casualty data and the extraction of cause-and-effect rules that can be used to predict car accidents, and the research into the influence that precipitation and other environmental factors have on the occurrence of accidents. This page contains the dataset that was taken from Kaggle. It is a database of all traffic accidents that occurred in the United States, including all 50 states. Beginning in February 2016 and continuing through June 2020, a number of different data sources, are being used to continuously collect the information. At the moment, this collection contains the details of around 3.5 million different accidents. The information is supplied in the form of a comma-separated values (CSV) file. The main goal is to determine and study the variables that have the greatest impact on the severity of accidents and offer the prediction model with the highest level of accuracy.

II. LITRATURE REVIEW

XGBoost model has been proposed in [2] to predict urban fire accidents. Box-Cox transformation technique is used for data transformation to transform continual response variables. The proposed methodology claims to have provided a feasible solution in the prediction of urban fire accidents contributing to improvements in the public security.

A statistical study presented in [3] exhibits the effect of various factors other than traffic congestion that leads to traffic fatalities. The study indicated that traffic congestion has the least effect on the accidents and others factors such as weather, vehicular specifications, road infrastructure and human elements played a greater role in the road mishaps.

The analytical study in [4] showed relationship between wind frequency and dust with accidents in the regions of California. The regions that had major dust source due to

high wind frequency reported more accidents than other regions with less wind frequency. This analysis proved that more the wind more the dust in the environment that reduces the road visibility contributing to accidental deaths.

With an increasing number of vehicles, the risks of traffic accidents are increasing. In the research paper proposed in paper [5], a model has been built to predict the risk of traffic accidents based on the data collected by the internet of vehicles. The model is designed in two stages, the first stage is feature extraction and the second stage is feature classifier. The feature extraction process is done using a convolutional neural network, for feature classification is processed using random forest using multiple decision trees. The data about the surrounding environments and driving conditions is collected, pre-processed, and given to the model as an input. With this proposed methodology, the area under the curve (AUC) achieved is 99.22%.

Many predictive models are proposed for traffic accidents. The research paper proposed [6] discusses the multiple facets of road safety and traffic accidents. One of the facets is potholes prediction. The data of 500 imagery inputs is used to train the model based on a convolutional neural network. The model has achieved good accuracy, precision, and recall. Another facet is the detection of objects and objects and vehicles. To train this model imagery and video data are used as inputs. The model can predict objects in the range of 113 km. But, the model failed to detect the objects covered with vegetation, less illuminated objects, and broken or damaged objects. Such a model could be future work for our project. It will expand the road management safety system.

Using real-world traffic and accident data from a Florida highway, this paper[7] develops prediction models for traffic accident severity. Both responders and drivers benefit from accurate severity prediction. The research accurately predicts the number of emergency vehicles needed based on the severity of the accident. The prediction results show that the accuracy and computation costs to support ITS applications are enticing.

The researchers proposed in this paper[8] to use XGBoost to analyze data in order to determine various aspects that influenced the severity of two-wheeled vehicular traffic accidents over the last ten years. A proper legal reform plan to avoid major traffic accidents and enhance two-wheeled vehicle safety control was recommended based on the findings. According to the findings, the active government needs to support the use of advanced technologies to reduce two-wheeler single-wheeler accidents.

In paper [9], there are various factors like human, vehicular, and environmental that are considered to be the consequence of a systemic problem for traffic accidents. On 26 January'2011 , a snowstorm and concurrent accidents occurred which are used as a case study in this paper. This study uses a kernel density smoothing methodology for predicting the different patterns of locations where accidents took place within the area over time. In order to this, a traffic accident data that took place in Virginia is used for augmentation in order to characterize the multiple patterns of accidents.

In paper [10] severity of crash on a freeway is predicted. It discusses the factors affecting freeway crash severity. 11 factors were analysed and visualised for better understanding if their impact on se verity of crash. XGBoost is used for prediction in this study where it concludes that factors like roadside protection facility type, road section type, central isolation facility, lighting condition, and crash occurrence time play a vital role in severity prediction. The accuracy percent of this model is 89.05%.

Paper [11] focuses on severity of crash on old pedestrians in Colorado. The age group above 65 is the primary subject of this paper. It also discusses the factors affecting the severity of c rash using Extreme gradient boosting model. For interpretation of XGBoost model Shapley Additive explanations (SHAP) is implements for result and feature analysis. The study narrows down driver characteristic, older pedestrian characteristics, and vehicle movement as the key features in crash severity prediction among the elderly.

Paper [12] discusses the prediction of accident duration in United States of America. The data is collected emphasizing of five key aspects traffic, location, weather, points of interest and time attribute, then the data quality is improved, further feature extraction is performed for better results. After data preprocessing CatBoost, stacking, XGBoost, LightGBM, and elastic network are used to create a heterogeneous ensemble model. The model shows good accuracy along with leading factor contributing to the prediction.

In the paper [13] researchers created three distinct traffic accident severity prediction models to better understand and predict the impact of road accidents. Second, the random forest method is well-suited to multi-dimensional data and is capable of producing reliable forecasts, which is the case with traffic accident data. By making predictions about the severity of traffic accidents, we can find out whether these elements are influential. The severity of traffic accidents may be affected by a number of variables including the time of day the accident occurs, the weather, the location of the accident, and the illumination. In light of this, there is a practical method for reducing the likelihood of or damage from traffic accidents.

In the following research paper[14] the findings of their studies stated that the that reports of traffic accidents sometimes leave out crucial details, making it difficult to draw accurate conclusions. It has been shown that basic patterns in accident occurrences may be identified to determine the causes of accidents. And also the research highlights the value of a classification system in determining the probable vehicle collision patterns in a road accident data set. The results of this research were intended to be utilized to derive a decision tree that can be used to the prediction of the collision's mode.

III. CHOICE OF METHOD

In this research paper, XGBoost algorithm has been used to predict the severity of the accidents. The predictions are made based on several factors like weather, location,

infrastructure and others conditions along with the number of accidents reported. XGBoost works best on tabular and structured datasets for predictive modelling cases. It performs regularized learning to avoid over-fitting of the model and the model training done in additive manner. It has built-in cross-validation method to specify the boosting iterations. It can also handle missing values and outliers. Other such features of XGBoost are:

- It is computationally fast.
- It is efficient in memory management.
- Supports parallel processing and cache optimization.
- It also performs Auto-tree pruning.
- It takes a multithreaded approach in adding the weak learners.

IV. METHEDOLOGY

The objective of this predictive model is to predict the severity depending on these independent features, and build business applications to improve traffic safety management.

To achieve the objectives of this project, we are going to use Knowledge Discovery in Databases (KDD) methodology. The step-by-step breakdown of understanding the project and building the predictive model is as follows:

1) Data selection:

'US Accidents' is the countrywide traffic accidents data. The dataset includes data from 49 states of the US. It contains independent features. Some of the important independent features are wind speed, start time and end time of the accident, visibility, humidity, temperature, pressure, weather condition, wind direction, etc.

2) Data pre-processing:

The data processor of data cleaning consists of exploring the data thoroughly and taking the necessary steps.

Fig. 1 is the bar plot of missing values in the columns. Hence, the first step of data cleaning consists of treating the missing values. For the features which have missing values more than 40 % of the records are dropped. For the features which have missing values of less than 40%, we have dropped the records containing missing values.

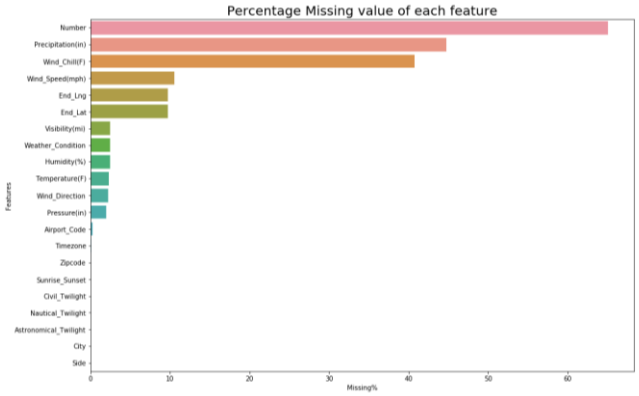

Fig. 1 bar plot of missing values in the columns

3) Data Transformation:

Some features have categorical values. For instance, the feature "Weather_condition" contains records such as fair, clear, overcast, rain, light snow, cloudy, etc. Such other features are mentioned below: Wind direction, weather fair, weather cloudy, weather clear, weather overcast, weather snow, weather haze, weather rain, weather thunderstorm, weather windy, weather hail, weather thunder, weather dust, and weather tornado.

We are going to apply the label encoding technique to convert these categorical values to numeric so that the data in those features become suitable for the machine learning algorithm.

The dataset contains continuous features. Those are as follows: Temperature, visibility, humidity, pressure, wind speed.

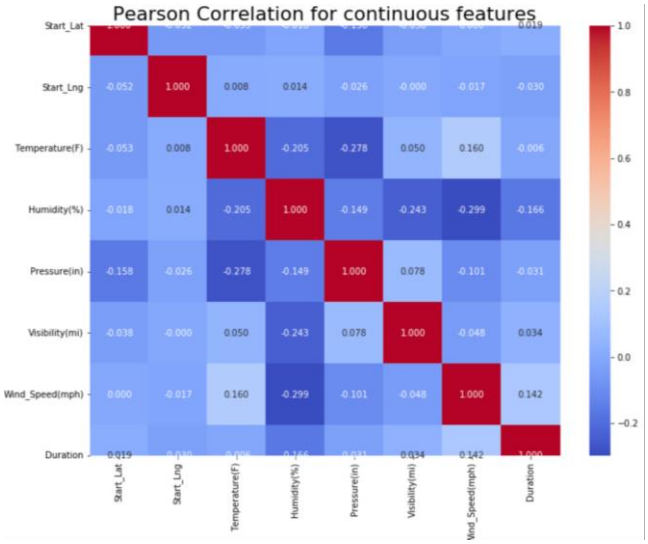

Fig 2 Pearson's correlation matrix.

Fig. 2 is Pearson's correlation matrix has been built to understand the correlation between these features.

It can be observed from the above figure that, there is a fairly negative correlation between pressure and temperature, which means, that if the pressure goes up, the temperature goes down. And the temperature and wind speed have a slightly positive correlation.

Further, we have applied one hot encoding to the categorical features.

4) EDA:

After data collection, pre-processing and transformation, Exploratory data analysis is performed on the cleaned data to analyze different aspects of severity factors responsible for occurrence for traffic accidents based on several factors.

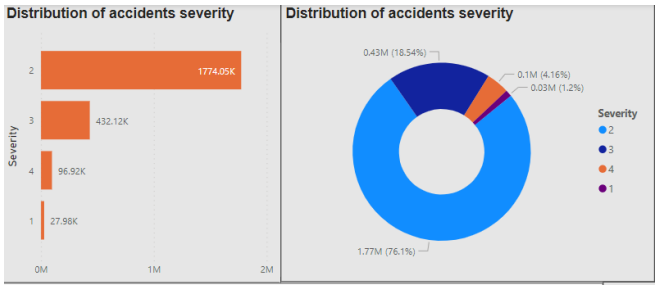

Fig.3 distribution of accident severity

Fig 3 depicts distribution of accident severity. The severity of accidents is categorized into level of 4. The first is a bar graph where number of accidents occurred on the basis of severity is analysed where severity 4 has the highest count of accidents. The other is a donut chart showing the same along with percentage of occurrence of accidents. As shown in the graphs, level 2 is the most common severity, accounting for 76.1 % of all cases. This means that the target variable (label) is unbalanced.

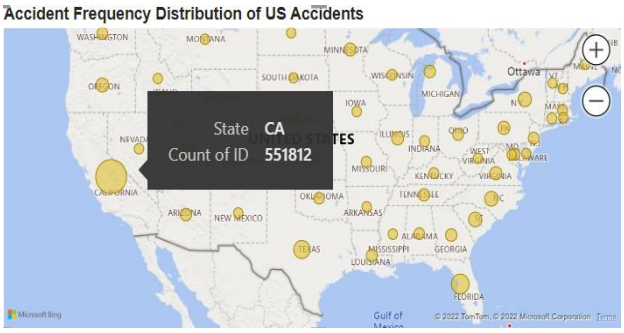

Fig. 4 geographical distributions of accidents

Fig. 4 shows geographical map of United states with different bubbles depicting accidents that took place in every state within US. CA shows highest frequency of accidents.

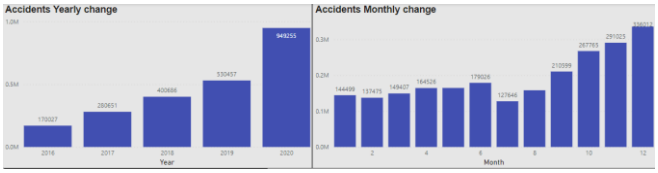

Fig 5 Yearly and monthly change in number of accidents

Fig 5 shows Yearly and monthly change with number of accidents. The first bar graph depicts cases of accidents increasing over years where count has almost reached to a million in 2020.

The second bar graph shows monthly change where it seems that there is increase in the number of the accidents from July(7) to December(12) . Fall and winter is reasonable to be more dangerous. However, the number of accidents in January and February are so much lower than December, which is quite interesting.

Fig. 6 above shows the confusion matrix. Since we have applied label encoding, the values in the target column and predicted column are ranging from 0 to 3. 0 represents the least severity while 3 represents the maximum severity. As we can see, the model can predict the severity of 1 well.

5) Data mining

After data processing, we split the data. The target variable is "severity". This target variable contains 4 ordinal values. Value 1 represents the least severity while value 4 represents the maximum severity. This feature is separated from the independent features. The data is further split into train and test with the ratio of 70% to 30%.

XGBoost algorithm has built on training data. For the evaluation, we have used test data. The model has achieved 90% accuracy.

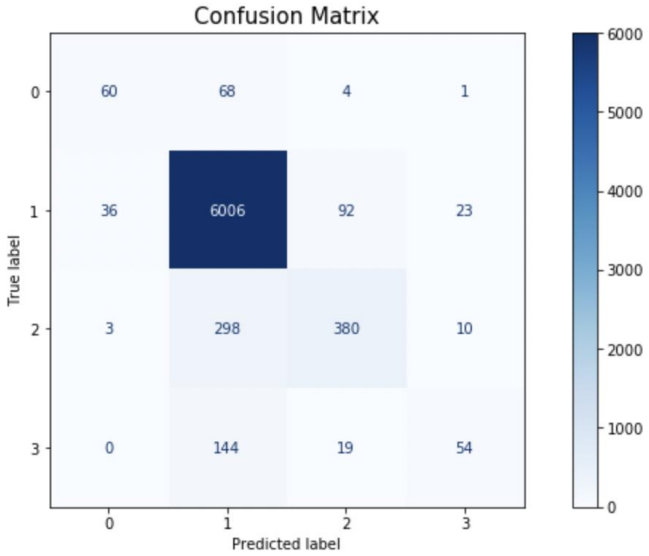

Fig 6 confusion matrix

6) Evaluation:

Fig 7 above shows the classification report. The precision and recall for class 1 are the highest, 92% and 98%. The overall accuracy of the model is 90%.

	precision	recall	f1-score	support
0	0.61	0.45	0.52	133
1	0.92	0.98	0.95	6157
2	0.77	0.55	0.64	691
3	0.61	0.25	0.35	217
accuracy			0.90	7198
macro avg	0.73	0.56	0.61	7198
weighted avg	0.89	0.90	0.89	7198

balanced_accuracy: 0.5563446139857 83
roc_ovo_macro: 0.9092460613043337
roc_ovr_weighted: 0.9230708895432526
<Figure size 432x288 with 0 Axes>

Fig 7 classification report

V. RESULT INTERPRETATION AND BUSINESS VALUE

In this paper the severity traffic accident in US is predicted. The methodology used is XGBoost. The factor 'severity' is predicted from 1 to 4 with 1 being least severe accident and 4 being the most severe. The model created shows 90% accuracy.

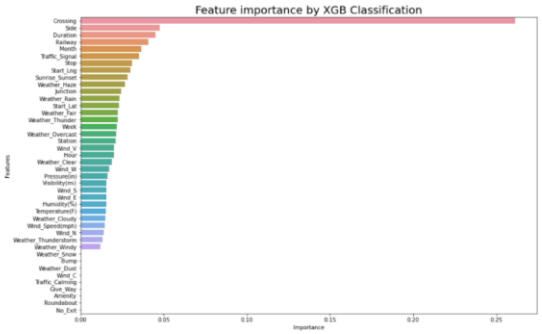

Fig 8 feature importance of the model.

Fig 8 represents the feature importance of the model. In this graph x axis represents features affecting the severity and on y axis it represents importance for XGB model from the scale of 0 to 0.25. Crossing was the most important feature in XGBoost model for severity detection.

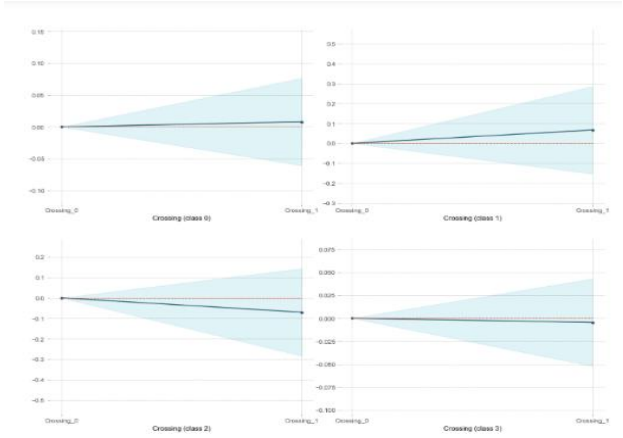

Fig 9 PDP for the feature crossing

Fig 9 represents partial dependence plot (PDP) for the feature crossing. The four graphs are plotted for the four classes. Class 1 and 2 show the higher amount of deviation.

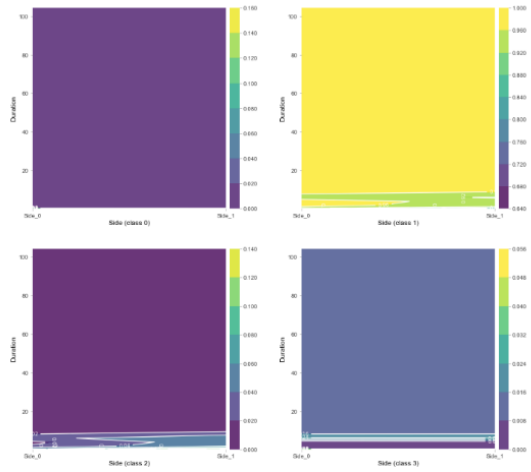

Fig 10 PDP for side and duration.

Fig 10 shows the partial dependence plot (PDP) for side and duration. Here x axis represents the duration and y axis represents the side. The values are color coordinated from ranging from yellow to blue.

Any life related study has great responsibility as well as business values. These feature plots provide better understanding of factors and their impact on any traffic accident. As the plot shows crossing to be the highest contributors it gives authorities the direction to work on for safer travel. It also provides other features like side and duration to be considered hence the responsible bodies could raise awareness regarding the same.

VI. Conclusion

Using predictive analysis, traffic accidents across US were analyzed to predict severity of road accidents. XGBoost model was created for the same which provide the accuracy of 90%. Along with those predictions the factors affecting the severity were analyzed where crossing turned to be the most impactful feature while calculating the severity.

References

[1] S. Rawat, “USA Accidents Data Analysis,” Medium, Jun. 28, 2021. https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02 (accessed Aug. 11, 2022).

[2] X. Shi, Q. Li, Y. Qi, T. Huang, and J. Li, “An accident prediction approach based on XGBoost,” in 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nov. 2017, pp. 1–7. doi: 10.1109/ISKE.2017.8258806.

[3] A. E. Retallack and B. Ostendorf, “Current Understanding of the Effects of Congestion on Traffic Accidents,” Int. J. Environ. Res. Public. Health, vol. 16, no. 18, Art. no. 18, Jan. 2019, doi: 10.3390/ijerph16183400.

[4] A. Bhattachan, G. S. Okin, J. Zhang, S. Vimal, and D. P. Lettenmaier, “Characterizing the Role of Wind and Dust in Traffic Accidents in California,” GeoHealth, vol. 3, no. 10, pp. 328–336, 2019, doi: 10.1029/2019GH000212.

[5] H. Zhao, X. Li, H. Cheng, J. Zhang, Q. Wang, and H. Zhu, “Deep learning-based prediction of traffic accidents risk for Internet of vehicles,” China Commun., vol. 19, no. 2, pp. 214–224, Feb. 2022, doi: 10.23919/JCC.2022.02.017.

[6] M. Jindal, E. Bajal, P. Singh, M. Diwakar, P. Kumar, and S. Singh, “A review on multiple facets of Indian Road safety-classification, challenges and future scope,” in 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Nov. 2021, pp. 1–6. doi: 10.1109/UPCON52273.2021.9667590.

[7] H. M. Alnami, I. Mahgoub, and H. Al–Najada, “Highway Accident Severity Prediction for Optimal Resource Allocation of Emergency Vehicles and Personnel,” in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), Jan. 2021, pp. 1231–1238. doi: 10.1109/CCWC51732.2021.9376155.

[8] Cheol-Woo Kwon, Hyun-Ho Jang, “Comparative Analysis of Severity of Two-wheeled Vehicle Accidents Using XGBoost,” The Journal of The Korea Institute of Intelligent Transport Systems Vol.20 No.4 pp.1-12 DOI: https://doi.org/10.12815/kits.2021.20.4.1

[9] Y. Yang, K. Wang, Z. Yuan, and D. Liu, “Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction,” J. Adv. Transp., vol. 2022, p. e4257865, Apr. 2022, doi: 10.1155/2022/4257865.

[10] M. Guo, Z. Yuan, B. Janson, Y. Peng, Y. Yang, and W. Wang, “Older Pedestrian Traffic Crashes Severity Analysis Based on an Emerging Machine Learning XGBoost,” Sustainability, vol. 13, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/su13020926.

[11] Y. Zhao and W. Deng, “Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning,” Appl. Artif. Intell., vol. 36, no. 1, p. 2018643, Dec. 2022, doi: 10.1080/08839514.2021.2018643.

[12] S. Rawat, “USA Accidents Data Analysis,” Medium, Jun. 28, 2021. https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02 (accessed Aug. 11, 2022).

[13] S. S. Reddy, Y. L. Chao, L. P. Kotikalapudi, and E. Ceesay, “Accident analysis and severity prediction of road accidents in United States using machine learning algorithms,” in 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Jun. 2022, pp. 1–7. doi: 10.1109/IEMTRONICS55184.2022.9795852.

[14] T. Abdullah and S. Nyalugwe, “A Data Mining Approach for Analysing Road Traffic Accidents,” in 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), May 2019, pp. 1–6. doi: 10.1109/CAIS.2019.876958