COVID-19 Impact on US Traffic Accidents

(Team 036-Bavithra Radhakrishnan, Chandra Sekhar Nookarapu, Lohitha Rajasekar, Sravya Karamched, Swaminathan Murugesan, Uday Bag)

Introduction

Automobiles have been in the preeminent position as the primary means of transportation over several decades. It enhances the lives of individuals and the society, but benefits come with a price. Every year on average, 6 million accidents occur and around 3 million people are injured, and more than 90 people die every day in the US [1]. Because of this frequency, traffic accidents are a major cause of death, cutting short millions of lives per year.

Problem Statement

Traffic accidents are a main public safety issue, and it is important to have a system that can predict the occurrence of traffic accidents or accident-prone areas that can potentially save millions of lives. Regardless of the progress of research in US traffic accidents, the number of accidents is not decreasing, and this is a concerning issue. And to add on, the COVID pandemic is new and to our knowledge we think there is limited research conducted so far that explains the impact of COVID on US -traffic accidents. To address these challenges, we have proposed a solution in this project that gives better insights about COVID on US traffic accidents. We believe this will benefit the government in making the right decisions and policies that reduce the number of accidents and fatality rates in the United States.

Survey

A great deal of research has been conducted and addressed regarding the traffic accidents. Most of the research conducted is related to traffic analysis, prediction of accident location, severity of accidents etc. [2,4,10] A survey was conducted to study factors that can be used to avoid accidents happening in various cities and countries [3]. The author concluded that Association rule is an important method to analyze road traffic accidents. Currently, most of the research uses Linear regression [7,8,9,11,12,13,14] and Random Forest [15,16,17,18] algorithms to study and predict the trends in traffic accidents.

It is also important to understand the factors that influence the road accident. In [5], the author used the Apriori algorithm, Naive Bayes, IBK and K-means clustering to determine the factors affecting road accidents. The IBK algorithm was found to be superior to the other algorithms. In [6], the author used supervised learning algorithms to classify the severity of accidents into Fatal, Grievous, Simple Injury and Motor Collision as four categories and found Adaptive Boosting to be better than the others in terms of performance. Despite all this analysis conducted in the past, these data are not available for future references. Most of the research conducted is not available to the government and public. Therefore, there is a gap between the tremendous amount of research conducted and the actual implementation in real time.

Proposed Method

In this project, we aim to build an Interactive User Interface that provides an in-depth analysis and visualization of US Traffic Accidents. The analytical UI shows the discover patterns, and extract cause and effect rules that measure the impact of precipitation or other environmental stimuli on accident occurrence. We used the recent dataset to study the positive/negative impact of COVID-19 pandemic on traffic accidents. While analyzing traffic accidents and behavior, we explored if the past trends and predictions are valid in the COVID-19 era and find the correlation (if any) between them.

Innovations

Today, many UIs and dashboards provide the statistics of US Traffic Accidents based on various factors. We are working to develop our project with the following innovations:

1. Leveraged recent datasets to study the impact of COVID-19 on traffic accidents and find correlation (if any) between COVID cases and traffic accidents.
2. Unlike the existing solutions, we used different features (like aggregations and dummy variables) for our analysis using machine learning algorithms.
3. We analysed and located the most concentrated traffic location, having numerous accident criteria and applied ML model to find a comparative analysis on COVID data. Similarly, we identified less populated, rural locations with less historical accident count and plan to study accident impacts due to COVID cases and find similar/outlier trends.
4. We plan to develop dashboards that show and compare both traffic accident and COVID case counts at State, County and Month level across the US.

Data Source

The accident dataset was taken from kaggle and COVID dataset was collected from the US government website. AWS S3 bucket was used to store both the datasets.

Data Collection

Our project used two large-scale datasets containing about 4.2 million instances of traffic accidents across the United States over the last five years (2016-2020) and the most recent COVID-19 data (year 2020). We imported the dataset into AWS S3 bucket for further data manipulations.

Data Cleaning

1. Traffic Accident Dataset - This dataset contains accident data collected from February 2016 to December 2020. After analyzing the initial dataset, we have identified the relevant features and excluded irrelevant columns. All the null values were imputed with common measures like mean and mode. To maintain the same granularity level with the COVID dataset we have flattened the whole dataset to County and Date level.
2. COVID Dataset - This dataset contains county wise COVID counts from Jan 2020 to Feb 2021. Each column represented the date for which count of COVID cases were captured. We converted the dataset using Python melt to match the granularity of the accident dataset. In addition, we have captured the lockdown dates for different states collected from different sources to enrich the covid dataset.
3. Combined Dataset - Accident dataset was merged with COVID dataset using State, County and Date. Since County names were mismatched between the two datasets, we leveraged fuzzy matches to fix the data. Missing COVID counts in the beginning of January 2020 were defaulted to zero.

Data Manipulation

Data manipulation is a necessary step to rationally predict the accident count for our comparative analysis. In order to accomplish the task, we modified the following columns:

1. Temperature: In the original accident dataset, temperature was recorded based on the timestamp of the accident, however, we integrated the data to day level. Due to this change, we need a single temperature value for a day. Hence, we calculated the mean of the temperatures per day and replaced the existing values.
2. Weather condition: The initial dataset consisted of more than 100 unique weather conditions. To simplify the further computations, we obtained 7 unique values based on the original values and replaced the existing values.

3. Region: To create an effective visualization, we added a new column 'Region' to the dataset. We used a K-means clustering algorithm to segment the states into 4 clusters. This column holds four values (West, Northeast, South, Midwest).
4. Bump/Junction/Crossing/Stop/Traffic_Signal: These columns were casted from boolean values to integer values so that they can be used as features in the prediction model.
5. Dummy/Indicator Variables: For easy interpretation of the regression results, the updated weather condition values were expressed as dummy variables.
6. Accident count: We computed traffic accident count for each state, county and day and captured the values in the 'Acc_count' column.

Final columns

Severity	Start_Lat	Start_Lng	County	State	Bump	Junction	Crossing	Stop	Traffic_Signal	Region	Date
Temperature	Covid_Count	Acc_Count	Weather_Condition_Cloudy	Weather_Condition_Fair	Weather_Condition_Fog	Weather_Condition_Rain					
Weather_Condition_Snow	Weather_Condition_Storm	Weather_Condition_Windy									

Experiment/Evaluation:

For this project, we split the tasks into the following two parts:

1. **Model Building:**

   We have built multiple models for our analysis.

   Model-1: Prediction of Accident counts (2020) using 2020 data (training-0.7/testing-0.3)
   For this model, we filtered the data for the year 2020. We split the data into 70/30 training/test sets. Linear regression is used to predict the accidents count for the year 2020. The independent variables include Bump, Junction, Crossing, Stop, Traffic_Signal, Temperature and Weather_Conditions. We applied Principal Component Analysis (PCA) algorithm for dimensionality reduction and ran the model on different sets of states and counties. Metrics like SSE, SSR, SSTO, F-statistics were used for model evaluation. The results show that densely populated locations had better metric values compared to sparse populated locations.

   Model-2: Prediction of Accident counts (2020) with COVID-19 as an additional feature
   In this scenario, we added COVID count as an additional feature to the above model. The purpose of this model is to evaluate if COVID case count has any effect on the traffic accident counts. The results show that there is no significant correlation between the counts.

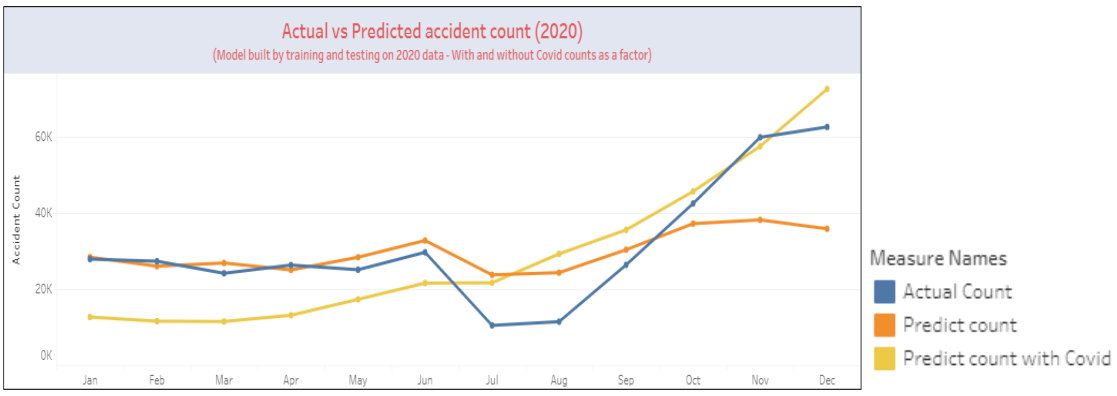


Figure 1.1: Actual vs Predicted accident counts for 2020

Model-3: Prediction/Trend analysis of accident counts for 2020 using 2016-2019 data.
For this case, we used 2016-2019 data as a training dataset and linear regression model is built to predict the accident counts for 2020. Additionally, we also developed a time-series model to analyze for trends in the data. The purpose of these models is to explore if the past trends and predictions are still valid in the COVID era. We observed that the R-squared values were better for densely populated counties compared to sparse populated counties. We also found that the past trend of increase in the number of accidents continued during the COVID-19 era.

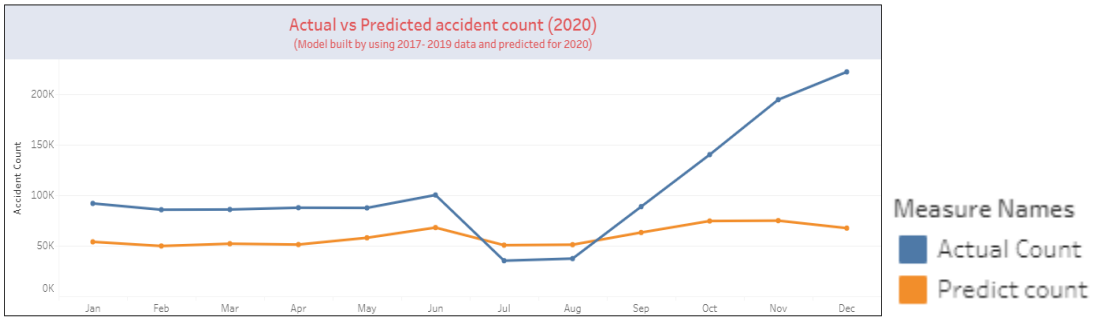
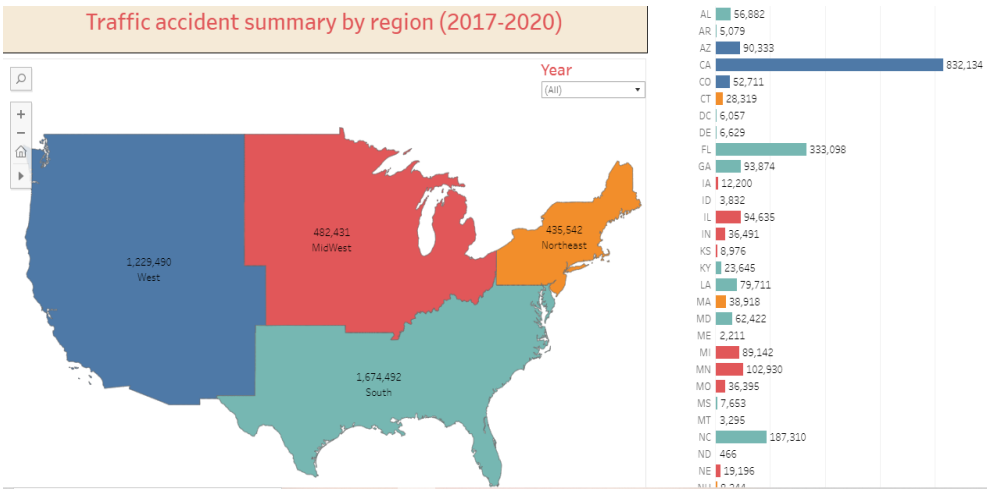


Figure 1.2: Predicted accident counts for 2020 based on 2016-2019 data

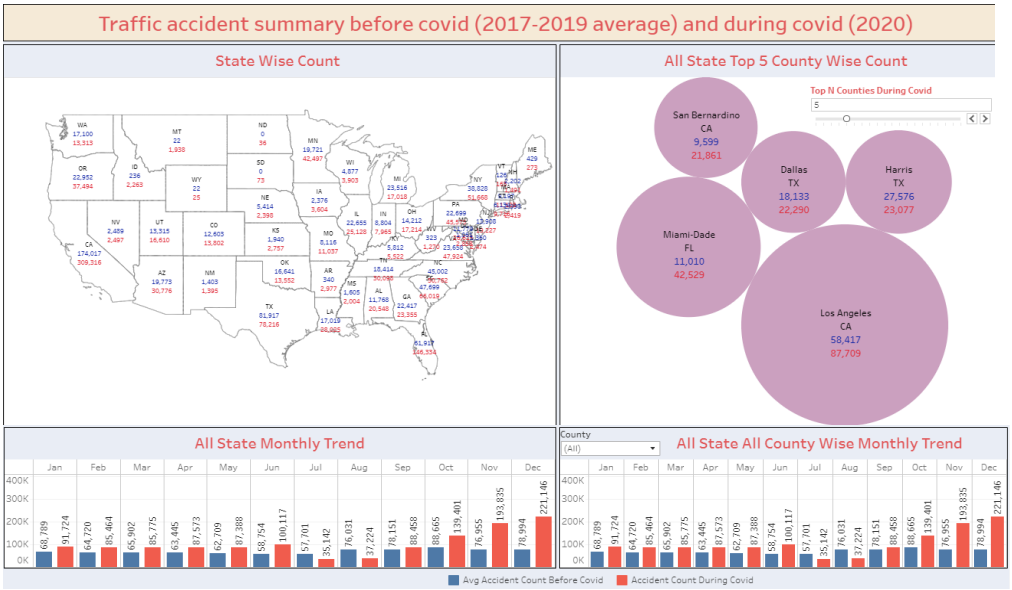
2. **Visualization:**
We built an Interactive UI using Tableau data visualization software. We developed the following dashboards which show the analysis of US traffic accidents: https://public.tableau.com/profile/bavithra.r1051#!/vizhome/USAccident_Covid_data_Analysis/Introduction?publish=yes

- **Accident Count by Region**
  - Displays the accident summary by 'Region'
  - USA is divided into four regions namely West, Midwest, Northeast and South
  - The map shows accident counts for all the four regions and also the counts for the respective states
  - There is a filter for 'Year' to enable customized visualization

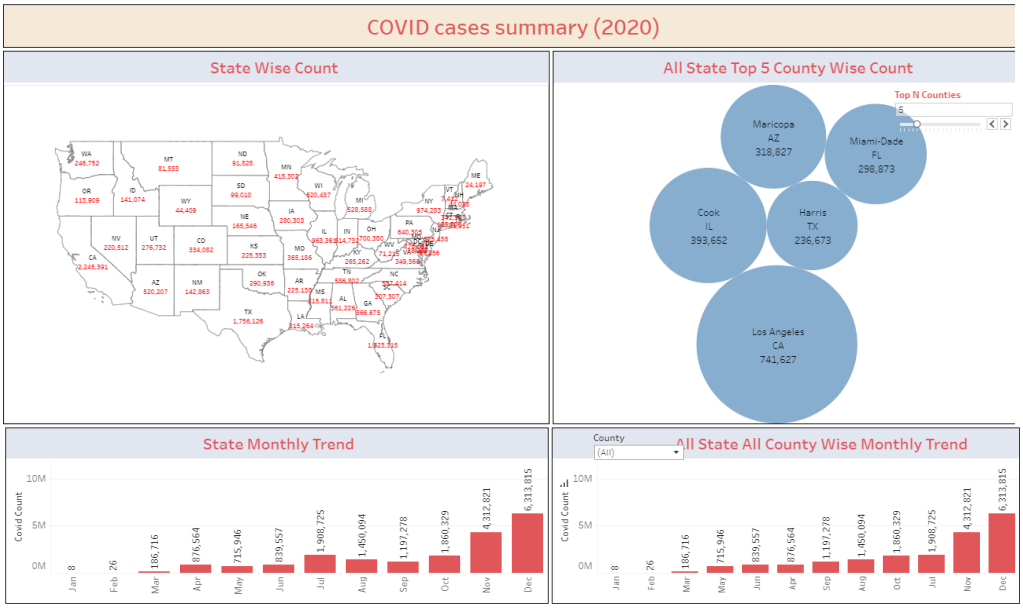


- **Accident Count by State/County**
  - Shows a comparison between the accident counts before (average of 2016-2019) and during COVID-19

- By default, the dashboard shows summary for every state, top 5-20 (*configurable*) accident prone counties in USA
- On filtering a state, top 5-20 (*configurable*) accident prone counties, month wise trend for state/county is displayed



- **COVID Count by State/County**
  - Shows the COVID-19 spread across different states in USA and the top 5-20 (*configurable*) counties with highest COVID counts
  - By default, the dashboard shows summary of COVID counts for every state and top 5 counties in USA
  - On filtering a state, top 5-20 (*configurable*) counties with highest COVID counts, month wise COVID cases trend for state/county is shown



Conclusion/Future Work:

Our qualitative results show that there is necessarily no significant correlation between the traffic accidents and COVID-19. However, they provide a considerable starting point for future guidance in implementing new traffic rules and regulations. This shows that there was a limited impact of COVID-19 on traffic accidents. We observed that the past trend of increase in the number of accidents continued during the COVID-19 era. We also noticed that there was a dip in the accident count during the mid of 2020(July, August) in majority of the states and this might be due to the lockdowns and stay-at-home orders enforced by the government and then there was a sudden spike during late 2020. The fluctuations in accident counts have been a challenge in our dataset due to the different APIs used to gather the data. Although, the level of detail needed to interpret this fact is not always easily available to the public.

While analyzing the predictions of our model, the features available in the datasets is ranked as the 5th factor of cause of accidents. To improve the accuracy of the model, identifying the right data sources and considering the additional causes as features for model building is suggested in the future.

We hope our analysis will contribute to providing better quality of public care, traffic management and implementing corrective actions to reduce the traffic accidents in the future.

.

Teamwork Distribution:

All team members contributed similar number of hours in all the phases- project idea, literature survey, proposal and execution

Task	Owners	List of Activities
Data (Accident & COVID) Collection, Parsing & Cleaning, Processing, Storage	ckala3, skaramched3, ubag3(Accident) bradhakr3, lrajasekar3, smurugesan3(COVID)	✓ Data Cleaning ✓ Data Combining ✓ Data Manipulations
ML Algorithms: K-means clustering, Linear Regression, PCA, Time-series	ckala3, skaramched3, ubag3, lrajasekar3, bradhakr3	✓ K-means for additional columns ✓ PCA for dimensionality reduction ✓ Linear Regression for Accident count prediction ✓ Time-series for trend analysis ✓ Model evaluation
Visualization: UI design, plots, graphs	bradhakr3, smurugesan3, lrajasekar3	✓ Experiment with different visualization options ✓ Build dashboards to represent data at various levels ✓ Publish the final dashboard
Cloud Infrastructure	ckala3, skaramched3, ubag3, smurugesan3	✓ Configuration of S3 ✓ EMR cluster configuration ✓ SageMaker Notebook Setup
Project Objective	ALL	✓ Define Experiments and evaluation methods ✓ Validate the project objectives by performing the experiments ✓ Collect metrics to verify the hypotheses
Project Reports	ALL	✓ Proposal Report/Slides/Video ✓ Progress Report ✓ Final Poster/Report/Presentation

References:

[1]National Center for Statistics and Analysis http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/NASS.html

[2] Johnson, Emi, et al. "Study on Road Accidents Using Data Mining Technology." *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*. IEEE, 2018.

[3]Moosavi, Sobhan, et al. "A countrywide traffic accident dataset." *arXiv preprint arXiv:1906.05409* (2019).

[4]Moosavi, Sobhan, et al. "Accident risk prediction based on heterogeneous sparse data: New dataset and insights." *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2019.

[5]Sugetha, C., et al. "Performance evaluation of classifiers for analysis of road accidents." *2017 Ninth International Conference on Advanced Computing (ICoAC)*. IEEE, 2017.

[6]Sakhare, Apeksha V., and Prajakta S. Kasbe. "A review on road accident data analysis using data mining techniques." *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE, 2017.

[7]Labib, Md Farhan, et al. "Road accident analysis and prediction of accident severity by using machine learning in Bangladesh." *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. IEEE, 2019.

[8]Cuenca, Laura Garcia, et al. "Traffic accidents classification and injury severity prediction." *2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*. IEEE, 2018.

[9]Lee, Yongbeom, et al. "A Machine Learning Approach to Prediction of Passenger Injuries on Real Road Situation." *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*. IEEE, 2018.

[10]Nikam, Swapnil Kisan. "ANALYSIS OF US ACCIDENTS AND SOLUTIONS." (2020).

[11]Al Mamlook, Rabia Emhamed, et al. "Machine learning to predict the freeway traffic accidents-based driving simulation." *2019 IEEE National Aerospace and Electronics Conference (NAECON)*. IEEE, 2019.

[12]Luo, Yong. "The fuzzy regression prediction of the city road traffic accident." 2009 International Conference on Industrial Mechatronics and Automation. IEEE, 2009.

[13]Silva, Charith, and Mo Saraee. "Predicting road traffic accident severity using decision trees and time-series calendar heatmaps." *2019 IEEE Conference on Sustainable Utilization and Development in Engineering and Technologies (CSUDET)*. IEEE, 2019.

[14]Hadjidimitriou, Natalia Selini, et al. "Machine Learning for Severity Classification of Accidents Involving Powered Two Wheelers." *IEEE Transactions on Intelligent Transportation Systems* 21.10 (2019): 4308-4317.

[15]Dogru, Nejdet, and Abdulhamit Subasi. "Traffic accident detection using random forest classifier." *2018 15th learning and technology conference (L&T)*. IEEE, 2018.

[16]Aburas, Abdurazzag, Scott Eyono, and Omesan Naidoo. "Vehicle accident foresight system using bigdata intelligent random forest algorithm." *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2018.

[17]Zhang, Jian, et al. "Comparing prediction performance for crash injury severity among various machine learning and statistical methods." *IEEE Access* 6 (2018): 60079-60087.

[18]Aburas, Abdurazzag, Scott Eyono, and Omesan Naidoo. "Vehicle accident foresight system using bigdata intelligent random forest algorithm." *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2018.

[19]National Center for Statistics and Analysis. (2020, December).Early estimate of motor vehicle traffic fatalities for the first 9 months (Jan–Sep) of 2020 (Crash•Stats Brief Statistical Summary. Report No. DOT HS 813 053). National Highway Traffic Safety Administration

[20] Mohammed, Ali. (2018). Classification of Traffic Accident Prediction Models: A Review Paper

[21] L. G. Cuenca, E. Puertas, N. Aliane and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE), Singapore, 2018, pp. 52-57, doi: 10.1109/ICITE.2018.8492545.

[22] Retallack, A. E., & Ostendorf, B. (2019). Current Understanding of the Effects of Congestion on Traffic Accidents. International journal of environmental research and public health, 16(18), 3400. https://doi.org/10.3390/ijerph16183400 7.

[23] G. Meena, D. Sharma and M. Mahrishi, "Traffic Prediction for Intelligent Transportation System using Machine Learning," 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), Jaipur, India, 2020, pp. 145-148, doi: 10.1109/ICETCE48199.2020.9091758.

[24] Y. Luo and S. Zhang, "The fuzzy regression prediction of the city road traffic accident," 2009 International Conference on Industrial Mechatronics and Automation, Chengdu, China, 2009, pp. 121-124, doi: 10.1109/ICIMA.2009.5156575.

[25] C. Silva and M. Saraee, "Predicting Road Traffic Accident Severity using Decision Trees and TimeSeries Calendar Heatmaps," 2019 IEEE Conference on Sustainable Utilization and Development in Engineering and Technologies (CSUDET), Penang, Malaysia, 2019, pp. 99-104, doi: 10.1109/CSUDET47057.2019.9214709.