IS597MLC: Final Project

NetID: jyeval2
Student Name: Jayesh Yevale

Identifying Crash Patterns for Injury Severity and Death Occurrences Using Fatality Analysis Reporting System (FARS) for 2022.

Motivation & Objective

We analyzed the crash pattern in car accidents using Fatality Analysis Reporting System (FARS) for latest available data of year 2022. The main motivation for the topic was to understand in depth about the crash pattern observed in accidents of a commuter and make aware for the public regarding the same for highlighting the causes. The data is publicly available on the government website of United States Department of Transportation (US-DOT) in National Highway Traffic Safety Administration (NHTSA). The main objective of this project is to explore and predict the crash patterns found in car accidents in detail, choosing appropriate attributes and comparing it with different machine learning models for better in-depth analysis.

After a detailed brainstorming session, the research questions that could be potential direction for the project can be as follows. What crash patterns based on severity of the injury be formed, so that the drivers can be aware of some situations? What crash patterns based on occurrences of death be formed, so that the drivers can be aware of some situations? Although the government provides a general report which conforms their interest, but there do not exist any reports from the other end (i.e. the drivers). This piqued my interest in applying a ML model for betterment for public awareness. The dataset contains immense details, hence the models such as Random Forest, etc. can be better predictors, and can provide a suggestion to take such precautions to the drivers.

Related Articles

In (Lixin Yan Y. H., 2020), the paper discusses the feature extraction model for traffic injury severity and its application. This paper is applied, and the main objective is to choose appropriate attributes for further analysis. The authors discussed and proposed a novel algorithm known as “Markov Blanket”, which extracts the significant factors leading to the crash injury severity, and it also uses Pearson correlation test for additional correlation measure.

In (Clark, 2003), the paper discusses the trauma system evaluation on FARS dataset. It portrays a general view of the dataset. It mostly is descriptive analysis in a broader aspect for a period to identify the changes throughout the years. It showcases the trends in the Total Mortality Rate over the years for various selected states. The challenge was obtaining the correct and viable data for its correct reporting statistics.

In (Nathaniel C. Briggs, 2005), the paper investigates racial and ethnic determinants of crash fatalities. The authors realized the even though the data of FARS, which is collected from 1975, it did not have the data on race and Hispanic ethnicity until 1999. This was concerning and a need to explore more in the detailed aspect of disparities that existed in crash mortalities among the racial and ethnic subpopulations throughout the United States.

In (Jaime K. Walters, 2024), the paper discusses the alcohol and drug presence in traffic crash fatalities before and after the COVID-19 pandemic. It only explores it among selected certain counties and compares the fatalities using the alcohol or drugs with the local medical examiner and death certificate data. It shows and compares it with blood alcohol and toxicology which is classified in three sections for three years. The main aim was to identify the impaired crash fatalities in detail.

Data

A. Data Collection

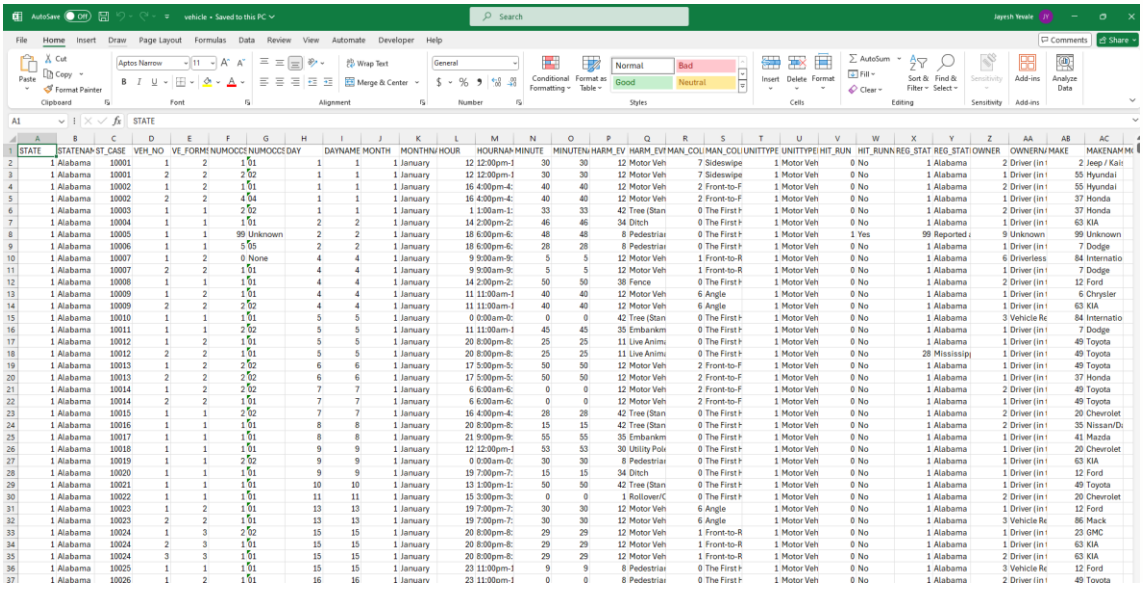
The dataset is of the FARS which is a nationwide census providing public yearly data regarding fatal injuries suffered in motor vehicle crashes which determines the fatality and the injury record census from 1975. It is publicly available on the government website of United States Department of Transportation (US-DOT) in National Highway Traffic Safety Administration (NHTSA) namely on www.nhtsa.gov.

The dataset contains 33 CSV files, and making connections will be tremendous task. In order to focus on major contributing factors, I plan to only choose three most important and directly related to the research problem we intend to focus. The files chosen for the analysis are Accidents, Persons, and Vehicle CSV files. The Accident CSV file contains 39k entries and 80 attributes, it consists of the detailed information on the accident such as location, state, day, weather, etc. The Person CSV files contain approx. 96k entries and 126 attributes, it consists of the detailed persons information such as for a particular accident how many people were involved, their information like sex, injury severity, driver/passenger/pedestrian, history of crashes, etc. The Vehicle CSV contains 60k entries and 201 attributes, it consists detailed information on the vehicles involved in the accident such as crash position, car make, airbags deployment status, speeding, etc.

For the project, I plan to select appropriate attributes for each file to upload separately, and merge using appropriate joins.(Left Outer – Inner Joins). Then, choosing the target variable, for the first problem the target variable is “injury severity”, whereas for the second will be “death occurrence”.

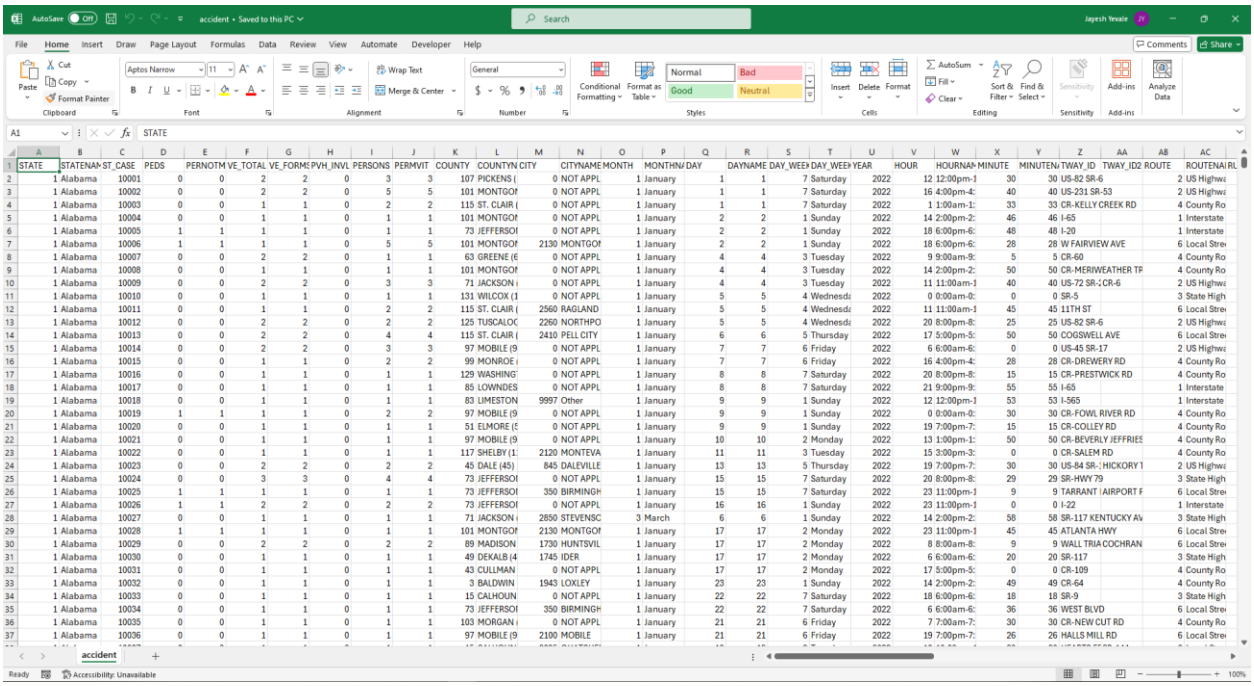
Due to restriction of space, I will only mention selected attributes below:

Vehicle File-



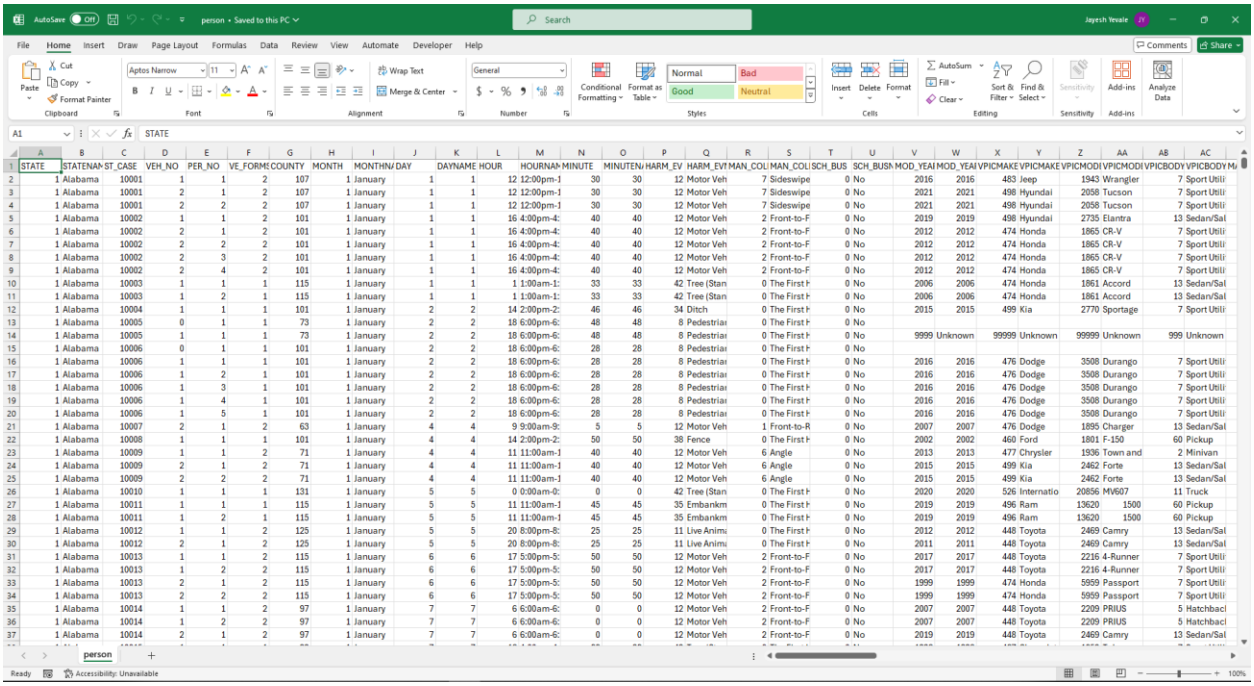
Attribute Name	Description
HARM_EVNAME	The crash happened with what event. (e.g. – Fence, Motor Vehicle, Pedestrian, etc.)
MAN_COLLNAME	Direction of the crash (e.g. – Sideswipe, Front-to-Front, Angle, etc.)
BODY_TYPNAME	The vehicle type. (e.g. – Compact Utility, 4-door sedan, Light Pickup, etc.)

The Accident File-



Attribute Name	Description
ROUTENAME	The accident route information. (e.g. – US Highway, County Road, Interstate, etc.)
RUR_URB	The accident locality. (1 – Rural, 2 – Urban)
LGT_CONDNAME	The light condition when accident happened (e.g. – Daylight, Dawn, Dusk, etc.)

The Person File-



Attribute Name	Description
AIR_BAGNAME	The airbag deployment status after accident for that person. (e.g. – Deployed, Not Deployed, etc)
DRINKINGNAME	The person drinking status. (e.g. – No, Not Reported, Reported as Unknown, etc)
DOANAME	The Death on Arrival status of person (e.g. – Not Applicable, Died at Scene, Died En Route, etc)

B. Data Pre-processing

The dataset is immense and requires joins among each other to extract a final dataset to work with. For the data preprocessing, we may need to eliminate the attributes which are not significant. To identify these, we may need to use our intuition and some test to find the correlation between the target for elimination of irrelevant attributes. We may need to set the variables according to the problem in binary format for feeding in the Machine Learning algorithm.

We selected the target variable as Person Injury, it captures the fatal injuries to the person. The preprocessing is done to separate the deaths vs no deaths effectively. Moreover, most of the variables were one hot coded as the variables are categorical. The variables consist of various tables and are sorted accordingly. The preprocessing is done on almost all the variables for proper consistency and feeding the proper dataset for effective accuracy purposes.

The data which is provided is already preprocessed a bit, as it was effective and easier to navigate using other software. The software used for this purpose was SAS JMP. This helped me to remove all the redundancy and the unknown values present in the target variable column.

Analysis & Methodology

The project will require python libraries such as Scikit-Learn, NumPy, Pandas, and Matplotlib or Seaborn. For data preprocessing, for making appropriate connections, final analysis and displaying the final analysis meaningful results.

In the project, I plan to incorporate machine learning algorithms such as Random Forest, Decision Tree and Logistic Regression. The main objective is to find the most suitable algorithm for predicting. The comparison will be along many factors, but to which I find accuracy to be most important for such cases. The Random Forest algorithm can be helpful for determining the finding and suggestions for the drivers' precautions. The Logistic Regression model can also be helpful for determining the factors which are important in determining the severity or death occurrence.

The evaluation metrics which I wish to consider are Recall, Precision, Sensitivity, F-measure and Accuracy. As the Model trains a lot of variables and attributes, there is a high chance of down sampling for effective runtime and viability. I believe Accuracy can be a good measure for the final impact and choosing the algorithm for both the binary cases.

		Precision	Recall	F1 Score	Accuracy
LR	1	0.9318	0.914	0.9228	0.885
	0	0.7529	0.7967	0.7742	
DT	1	0.9902	0.7902	0.8789	0.8362
	0	0.6048	0.9761	0.7468	
RF	1	0.8122	0.9785	0.8877	0.8136
	0	0.8269	0.3122	0.4533	
GB	1	0.9395	0.9164	0.9278	0.8927
	0	0.7635	0.8205	0.791	

Hence, we can conclude that the various models, which includes Decision Tree, Logisitic regression, Random Forest and Gradient Boosting, are being trained for the dataset after proper preprocessing. We only intent to focus on the target variable as "Person Injury" (i.e. PER_INJ). The column is already populated with either deaths "1" or no deaths "0". The intention is to identify the best model possible according to the f1 score, precision and accuracy.

From the evaluation of the Models, we can conclude that based on Accuracy and other percentages for Deaths (1) cases, the **Gradient Boosting Model ('GB')** is the best model to go ahead with the further analysis, followed by Logistic Regression Model ('LR'). Further analysis can reveal more detailed insights which is not scope of this course.

References

Clark, D. E. (2003). Trauma System Evaluation Using the Fatality Analysis Reporting System. *The Journal of Trauma and Acute Care Surgery*, 1199-1204.

Jaime K. Walters, K. K. (2024). Alcohol and drug presence in traffic crash fatalities before and after the COVID-19 pandemic: Evaluation of the fatality analysis reporting system (FARS) and linked medical examiner-vital records data in Clackamas, Multnomah, and Washington County, Oregon. *Forensic Science International: Synergy, ScienceDirect*.

Lixin Yan, Y. ,. (2019). A novel feature extraction model for traffic injury severity and its application to Fatality Analysis Reporting System data analysis. *Sage Journals*.

Lixin Yan, Y. H. (2020). A novel feature extraction model for traffic injury severity and its application to Fatality Analysis Reporting System data analysis. *Sage Journals*.

Nathaniel C. Briggs, R. S. (2005). The Fatality Analysis Reporting System as a tool for investigating racial and ethnic determinants of motor vehicle crash fatalities. *Accident Analysis & Prevention, ScienceDirect*, 641-649.