Analysis of US Accidents(2016-2020)

JP Purushothama Reddy
Bachelor's of Computer Science
PES University
Bangalore, India
purshothamredddy7843@gmail.com

Nikhil Karle
Bachelor's of Computer Science
PES University
Bangalore, India
nikhilkarle17@gmail.com

Sagar S
Bachelor's of Computer Science
PES University
Bangalore, India
itzsagar3014@gmail.com

***Abstract*—Reducing traffic accidents is an essential public safety challenge all over the world; therefore, accident analysis has been a subject of much research in recent decades. The objective of the project is to analyze the US accident data from 50 states to understand, insights trends, and possible causes of traffic accidents and what could be done to reduce them. We've formulated 3 models namely Decision tree, Random Forest, Logistic Regression. Random forest performed better compared to other models with an accuracy of 75%, We have found the hotspot locations of car accidents to be states like Texas, California, and New York**

I. INTRODUCTION

Approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities. More than half of all road traffic deaths occur among vulnerable road users—pedestrians, cyclists, and motorcyclists. More than 90% of all road fatalities occur in low- and middle-income countries

Also, according to the National Highway Traffic Safety Administration, these accidents cost the United States $871 billion annually. On average, road crashes cost countries 3% of their gross domestic product. Road crashes are the single greatest annual cause of death in the US. citizens traveling, Therefore, it is evident that the impact of road accidents in the United States is substantial-high. It is worthwhile to get a better understanding of the causes of road accidents so that they can be prevented.

Several factors contribute to road crashes and resulting deaths and severity of injuries. These include:

- Poor road infrastructure and management
- Non-road worthy vehicles
- Unenforced or non-existent traffic laws
- Unsafe road user behaviors and
- Inadequate post-crash care.

By understanding each of these factors and through planning, effective management, and evidence-based interventions, road crashes can be predicted and prevented officials and the general public are lacking systems which can show

- ✓ What is the accident-prone area in each state?
- ✓ What day and time are safe to travel?
- ✓ What are the factors responsible for accidents?
- ✓ What is the severity of these accidents?
- ✓ How many deaths happening in accidents?
- ✓ What solution can be implemented to reduce accidents by each state?
- ✓ How can this accident be minimized?
- ✓ How can the State Government improve accident-prone infrastructure?

Having access to accurate and updated information about the current road situation enables drivers, pedestrians, passengers, and Government officials to make informed road safety decisions.

II. LITERATURE SURVEY

A. *[1] Traffic Accident Analysis Using Machine Learning Paradigms (reviewed by JP Purushothama Reddy)*

Introduction of paper

This study is to evaluate a set of variables that contribute to the degree of accident severity in traffic crashes. Applying data mining techniques to model traffic accident data records can help to understand the characteristics of drivers' behavior, roadway condition, and weather condition that were causally connected with different injury severity. This can help decision-makers to formulate better traffic safety control policies. This study used data from the National Automotive Sampling System (NASS) General Estimates System (GES)

Assumptions of the paper

They developed machine learning-based intelligent models that could accurately classify the severity of injuries. Which in turn lead them to a greater understanding of the relationship between the factors of driver, vehicle, roadway, and environment and driver injury severity.

Summary and claims from paper

In this paper, the authors analyzed and investigated the performance of a neural network, decision tree, support vector machines, and a hybrid decision tree – the hybrid approach performed better than a neural network, decision trees, and support vector machines. For the no injury and the possible injury classes, the hybrid approach performed better than the neural network. The ability to predict fatal and non-fatal injuries is very important since drivers' fatality has the highest cost to society economically and socially. Unfortunately, their dataset didn't provide enough information on the actual speed since the speed for 67.68% of the data records was unknown. If the speed was available,

it could have helped them to improve the performance of models studied in this paper

How is this relevant to your work?

In this paper, the authors have Built different machine learning models from the data they had and tested the models with Road traffic data predicted whether the accident could be like no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. Different machine learning models had different Accuracy

By this paper, we got an idea that why not to build a machine learning model based on our EDA and predict whether the accidents could occur or not on different conditions features of the data

*B. **[2]** Review on Road Accident and Related Factors(reviewed by Sagar S)*

Assumptions

The main reasons for accidents are considered to be, the extremely dense road traffic and relatively much liberty of movement given to the drivers. In most cases, road accidents occur either due to carelessness, environment, another important cause for the alarming increase in the number of road accidents is driving of a vehicle in a drunken state. Under the influence of alcohol and other intoxicating substances, the drivers lose self-consciousness and control over the vehicle which ultimately forms the reason for accidents or due to lack of knowledge in road safety awareness of the road user.

Claims and take away from the paper

Based on studies on various safety models, coefficients of employed variables like traffic flow, lane width, etc, various maintenance approaches for preventing accidents can be estimated. These prototypes are useful in determining factors causing accidents and traffic accident distributions as preventive measures can be implemented suitably. Model studies on assessing driver's situation and conduct can help in identifying preventive measures to avoid rider-based accidents. Studies on the use of camera surveillance to monitor predicted accident spots showed the efficiency of its usage in preventing accidents.

Baoji Wang (2002) had investigated a sample of evaluations by drivers regarding Typical road environments related to safety. A face to face survey data of a sample of Sydney drivers was used to estimate an ordered probit model, a method often used in Travel behavior studies. In the survey, 27 scenes were developed to maintain the driver's safety.

The research by G A Handle et al., (2011) reported the rates of personal injury collisions (PIC) over the past decade on the roads of English local authority areas. A significant difference in the improvement rate was observed between urban and rural dimensions and was very much dependent on prior PIC risk levels. The study featured the accident scenario of sites under the continual surveillance of camera and its impact on accidents

*C. **[3]**Road traffic accident retrospective study(reviewed by Nikhil Karle)*

Introduction and assumptions:

A descriptive study was done using information from different sources like Private hospitals, District Hospital, Traffic Police record

The conclusion from the paper

The total number of accidents noticed during the study period was 87 with 9 deaths, i.e. mortality rate is 10.3%. It was found that the age group most prone for an accident is 25-34 years with males' predominance with 70:17, (i.e., 80.5% males: 19.5% females). Maximum mortality was noticed on Saturdays and the time most prone for the accident was between 10.00 AM-11.00 AM but deaths were more in accidents that occurred between 9.00 PM to 12.00 midnight. The most common cause of road traffic accidents was driving above the speed limit (47.1%) followed by consumption of alcohol by a driver (32.1%), rash driving at turns (20.7%). It has been observed that Deaths and injuries were mainly due to rash driving (68%) or due to the consumption of liquor (32%) while driving a vehicle.

How is this Relevance to our work:

This paper gives us a clear idea of how accidents occur and how frequently they occur. With the data available here, we can build machine models that will help us find accident hotspots and the time when more accidents happen. This will help us reduce accidents and fatal injuries.

III. DATASET DESCRIPTION

A. Origin of Dataset

The data has been obtained from the data science community Kaggle where the real-time traffic data has been collected using several data providers, including two APIs which provide streaming traffic event data which makes predicting traffic accidents more realistic. It is real-time traffic data that has been collected between the years 2016-2020.

B. Features and Observations

The dataset consists of 49 features and 351361 observations. The features consist of factors like geographical information, accident time, weather conditions, and other accident-related statistics and findings. The important columns are listed below with their description.

- Source: Indicates source of the accident report (i.e., the API which reported the accident)
- TMC: A traffic accident may have a Traffic Message Channel (TMC) code which provides a more detailed description of the event
- Severity: Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay)
- Start_Time: Shows start time of the accident in the local time zone
- End_Time: Shows end time of the accident in the local time zone
- Distance: The length of the road extent affected by the accident (in miles)

- Temperature: Shows the temperature (in Fahrenheit)
- Humidity: Shows the humidity (in percentage)
- Pressure: Shows the air pressure (in inches)
- Visibility: Shows visibility (in miles)
- Wind_Direction: Shows wind direction
- Wind_Speed: Shows wind speed (in miles per hour)
- Precipitation: Shows precipitation amount in inches, if there is any
- Weather_Condition: Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
- Amenity: A POI annotation which indicates the presence of amenity in a nearby location
- Bump: A POI annotation which indicates the presence of a speed bump or hump in a nearby location
- Crossing: A POI annotation which indicates the presence of crossing in a nearby location
- Junction: A POI annotation which indicates the presence of a junction in a nearby location
- No_Exit: A POI annotation which indicates the presence of no exit in a nearby location
- Railway: A POI annotation which indicates the presence of a railway in a nearby location
- Roundabout: A POI annotation which indicates the presence of a roundabout in a nearby location
- Civil_Twilight: Shows the period of day (i.e. day or night) based on civil twilight
- Nautical_Twilight: Shows the period of day (i.e. day or night) based on nautical twilight
- Astronomical_Twilight: Shows the period of day (i.e. day or night) based on astronomical twilight

C. Data Preprocessing

For our analysis and modeling, we considered a subset of nearly 3 million observations. The first step in data cleaning was removing unwanted columns from the data which would not be helpful for the analysis. We decided to remove features that had unique values for all rows or a single unique value for all the entire column. Columns that had lots of missing values were also dropped.

Outlier detection and handling

Weather-related data had extreme outliers which are practically not possible. In such cases, the outlier values were clipped to the minimum or maximum thresholds. We used various exploratory analyses to find outliers and depending on that we handled our outliers.

Dealing with null values

Our dataset contained null values in most of the columns. In the case of categorical and numerical attributes, missing values were replaced by mode and median of the columns respectively. For Precipitation, null values were replaced with zero assuming the fact that Precipitation values were missing for those instances when there was no rainfall. Null values for some columns were removed by dropping the entire row. It would have been misleading to impute these values so we decided to drop it.

Handling duplicate values

Categorical attributes like Wind_Direction had duplicate values like ‘North’ and ‘N’. The duplicates were handled by keeping the single character representation and replacing the other. The same technique was followed in the case of Weather_Condition were ‘Light Rain Shower’ and ‘Light Rain Showers’ meant the same.

IV. EXPLORATORY DATA ANALYSIS

A. Dataset(Summary)

ut[6]:

	ID	Source	TMC	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	...	Roundabout	Station
0	A-1	MapQuest	201.0	3	2016-02-08 05:46:00	2016-02-08 11:00:00	39.865147	-84.058723	NaN	NaN	...	False	False
1	A-2	MapQuest	201.0	2	2016-02-08 06:07:59	2016-02-08 06:37:59	39.928059	-82.831184	NaN	NaN	...	False	False
2	A-3	MapQuest	201.0	2	2016-02-08 06:49:27	2016-02-08 07:19:27	39.063148	-84.032608	NaN	NaN	...	False	False
3	A-4	MapQuest	201.0	3	2016-02-08 07:23:34	2016-02-08 07:53:34	39.747753	-84.205582	NaN	NaN	...	False	False
4	A-5	MapQuest	201.0	2	2016-02-08 07:39:07	2016-02-08 08:09:07	39.627781	-84.188354	NaN	NaN	...	False	False

3513617rows*49columns

```
Rows     : 3513617
Columns  : 49

Features :
     : ['ID', 'Source', 'TMC', 'Severity', 'Start_Time', 'End_Time', 'Start_Lat', 'Start_Lng
', 'End_Lat', 'End_Lng', 'Distance(mi)', 'Description', 'Number', 'Street', 'Side', 'City', '
County', 'State', 'Zipcode', 'Country', 'Timezone', 'Airport_Code', 'Weather_Timestamp', 'Tem
perature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Directi
on', 'Wind_Speed(mph)', 'Precipitation(in)', 'Weather_Condition', 'Amenity', 'Bump', 'Crossin
g', 'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station', 'Stop', 'Traffic_C
alming', 'Traffic_Signal', 'Turning_Loop', 'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twil
ight', 'Astronomical_Twilight']

Missing values    : 13061803
```

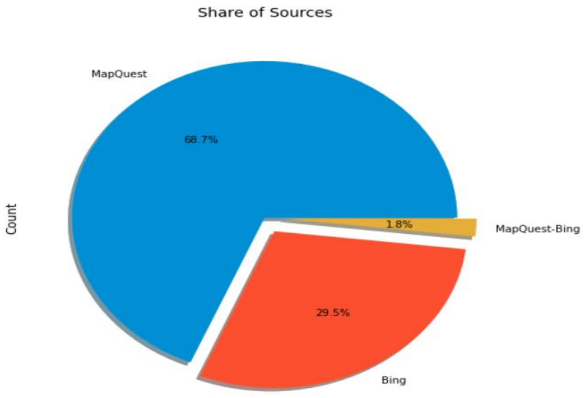
finding out columns with categorical Values, there are

```
Out[8]: Index(['ID', 'Source', 'Severity', 'Start_Time', 'End_Time', 'Description',
       'Street', 'Side', 'City', 'County', 'State', 'Zipcode', 'Country',
       'Timezone', 'Airport_Code', 'Weather_Timestamp', 'Wind_Direction',
       'Weather_Condition', 'Amenity', 'Bump', 'Crossing', 'Give_Way',
       'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station', 'Stop',
       'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop', 'Sunrise_Sunset',
       'Civil_Twilight', 'Nautical_Twilight', 'Astronomical_Twilight'],
      dtype='object')
```

35 categorical columns.

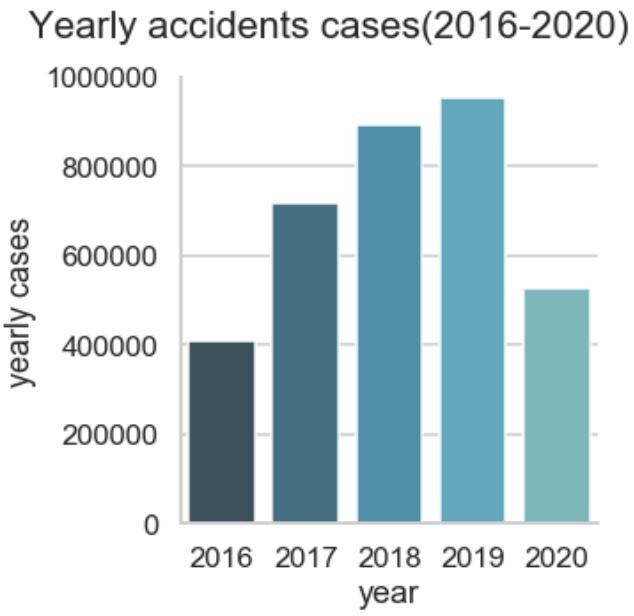
B. Source of the Data

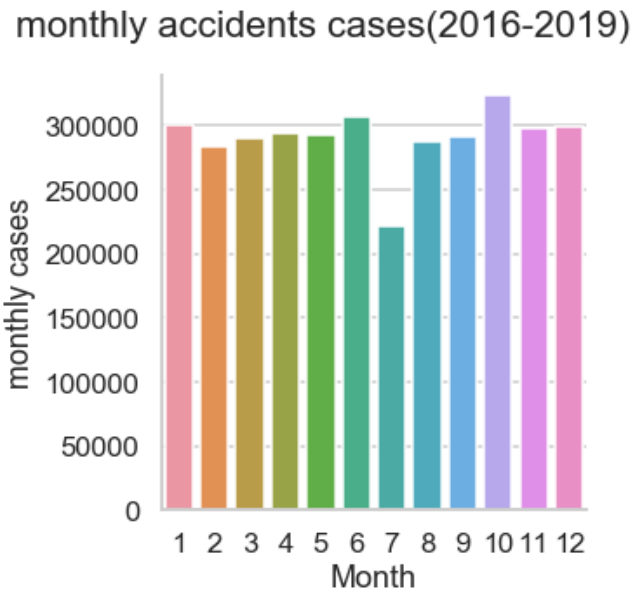
The dataset which we have taken is of streaming data which comes from a different source

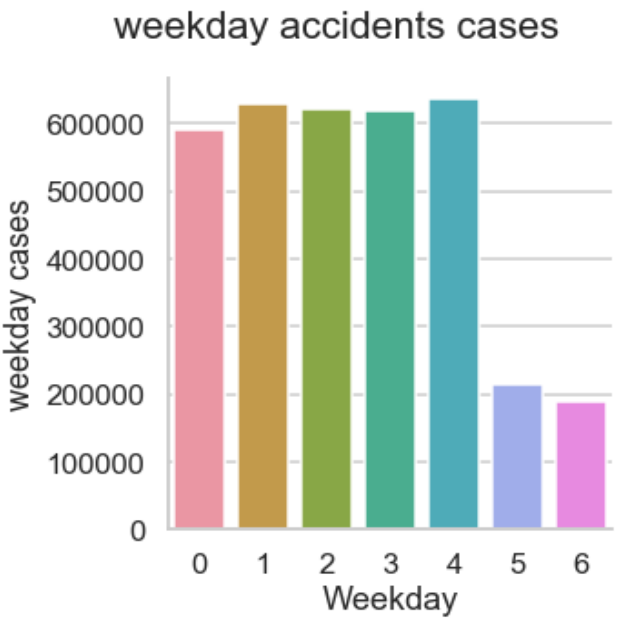


. MapQuest shares 68.7%, Bing 29.5%, MapQuest-Bing (other resources) 1.8%

C. Number of Accidents by year, month, and week

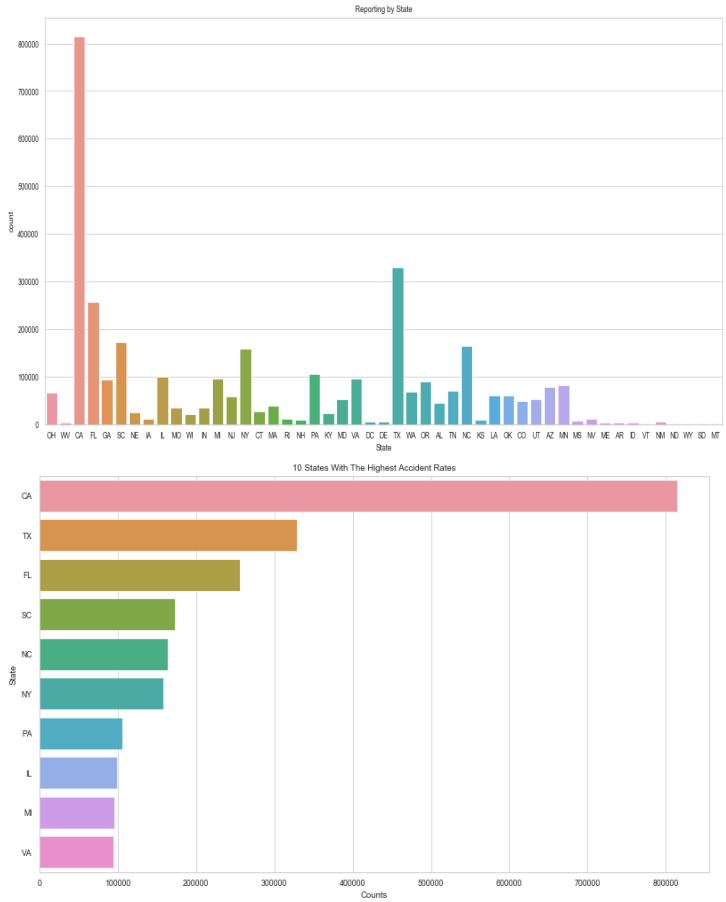






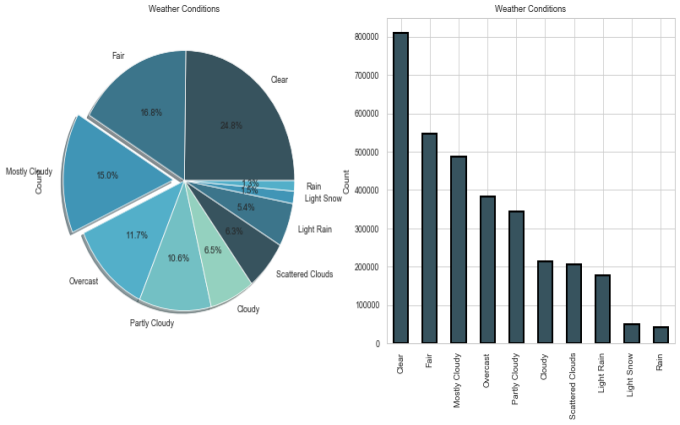
There is a growing trend of year accidents cases
There were more cases during 8-12 months compared to other months, excluding the data from 2020 (because this is the season where the people in the USA travel more)
The number of accidents on a working day is much larger than those on weekend, as fewer people go out to work

D. States and Cities with the highest number of accidents



From the above graphs, we can see that California and Texas have most accident rates

E. Weather conditions causing accidents

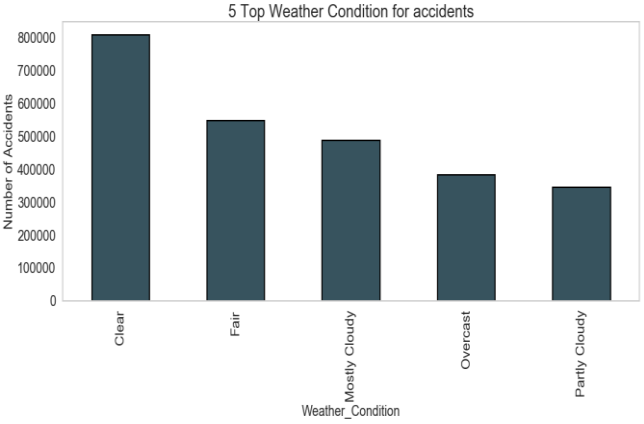


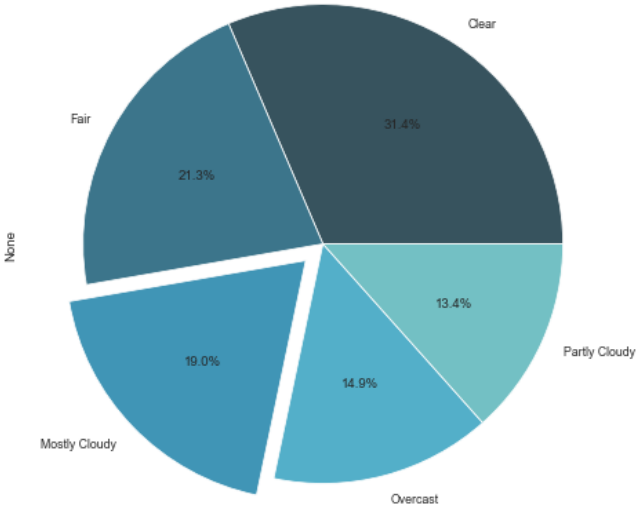
Out[73]:

Weather_Condition	0
Clear	808202
Fair	547721
Mostly Cloudy	488094
Overcast	382485
Partly Cloudy	344815
Cloudy	212878
Scattered Clouds	204660
Light Rain	176942
Light Snow	50435
Rain	42016

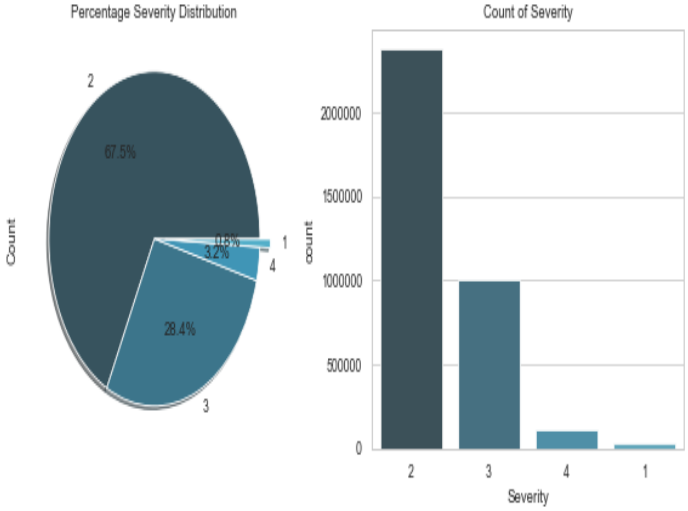
Top 5 weather conditions

We can see the most accidents have occurred when the weather is clear, maybe people drive carefully when the weather is not good





F. The severity of the accidents

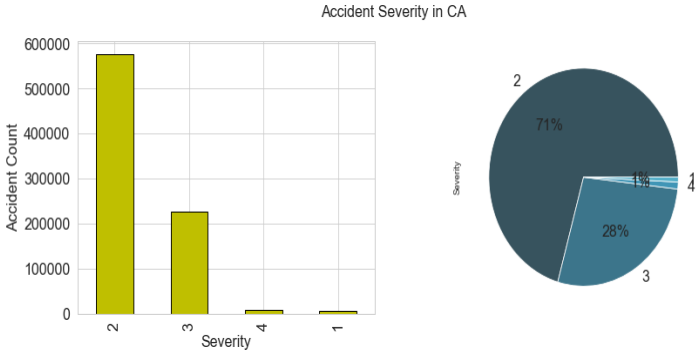


It tells the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on the type of accident 2 indicates medium impact (fracture), similarly, 3 (handicap) and 4 indicates a significant impact (like death)

Top 3 states accidents severity

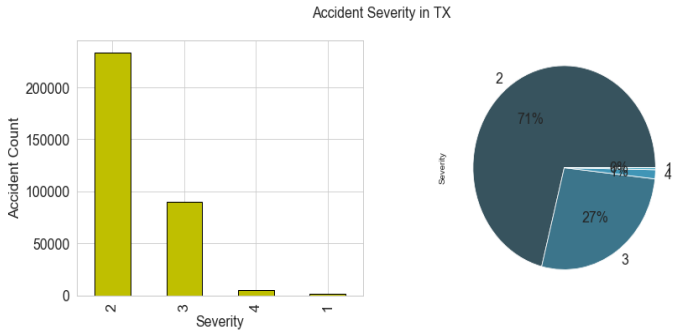
1)Severity of accidents for state California

70% of accidents are of severity level 2

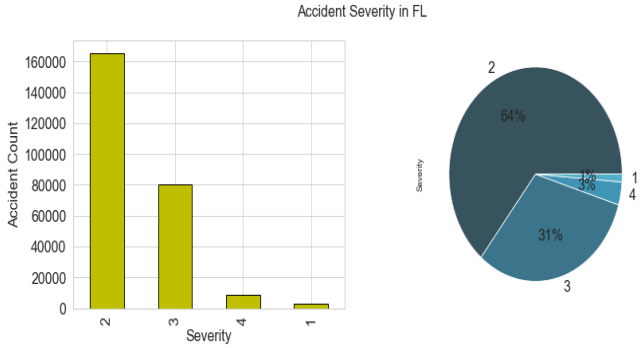


2) Severity of accidents for state Texas

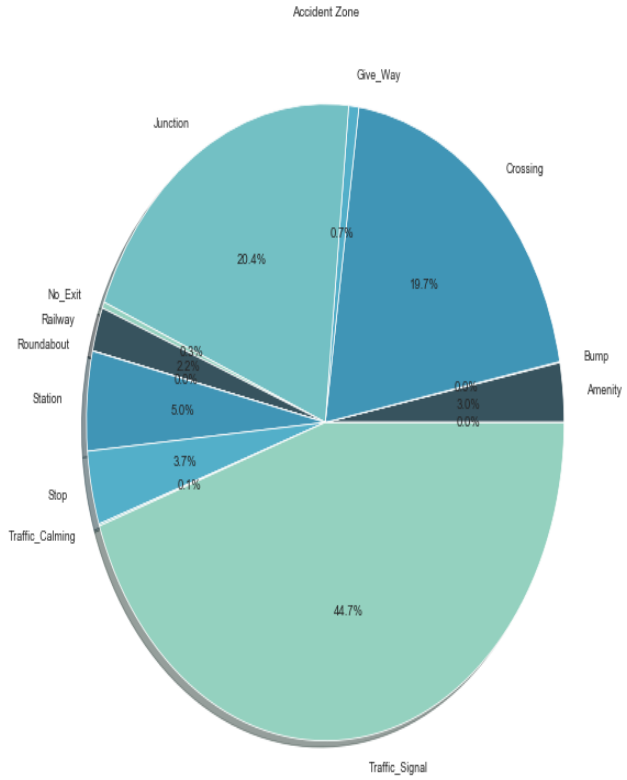
71% of accidents are of severity level 2 and only 1% of severity level 2



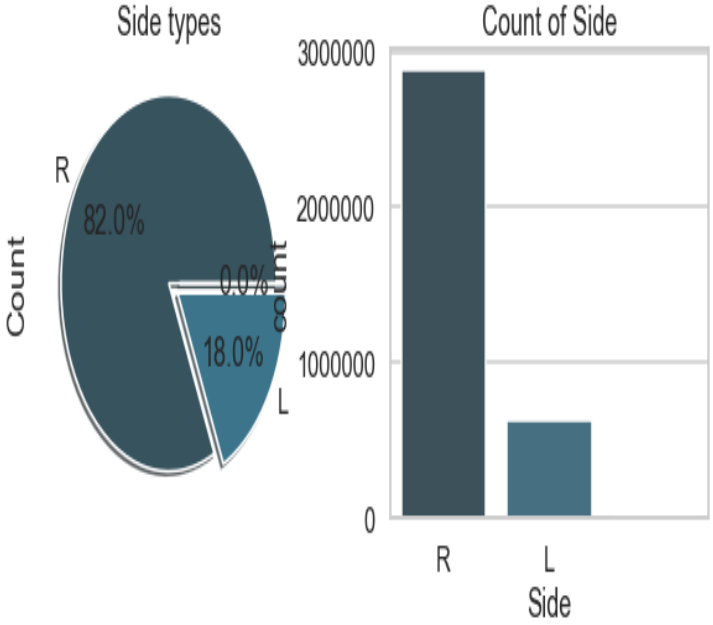
2) Severity of accidents for state Florida
64% of accidents are of severity level 2 and only 3% of severity level 2 and 31% of level 3
from the top 3 accidents prone state Florida has the highest death rate



G. Accident Zones



In this dataset, some Boolean values determine whether the accident happened near a traffic Signal, Bump, Railway, Stop, Crossing, etc.
We can see that most accidents are during Traffic signal (44.7%), Junction (20.4%), Crossing (19.7%), etc. Surprisingly there are almost 0% accidents that have occurred during a bump (because people slow down their vehicle and there is the least chance for accidents)



This graph suggests that more accidents happen on the right side.

V. MODEL BUILDING

We compare three classification models, namely Decision Tree (DT), Logistic Regression (LR), Random forest. For the implementation, we use Python's extensive library for Machine learning

A. Decision Tree

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter. Using the decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain (IG). In an iterative process, we can then repeat this splitting procedure at each child node until the leaves are pure. This means that the samples at each leaf node all belong to the same class. In practice, we may set a limit on the depth of the tree to prevent overfitting. We compromise on purity here somewhat as the final leaves may still have some impurity.
Decision trees (DTs) have been applied as well to analyze the severity of accidents. DTs provides a very useful model to identify the causes of accidents because they can be easily interpreted and, most importantly, decision rules can be

easily extracted from them. These rules can be used by road safety analysts to identify the main causes of accidents. A decision tree (DT) is a predictive model that can be used for both classification and regression tasks. In our case, the class variable has only four states, which means that it is a discrete variable. Hence, in this work, DTs were used to represent classification problems

B. Random forest

Random Forest is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from a randomly selected subset of the training set. It aggregates the votes from different decision trees to decide the final class of the test object.

Random forest is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. It runs efficiently on large databases. It can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. It generates an internal unbiased estimate of the generalization error as the forest building progresses. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

C. Logistic Regression

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems

The linear regression model can work well for regression but fails for classification. In the case of two classes, you could label one of the classes with 0 and the other with 1 and use linear regression. Technically it works and most linear model programs will spit out weights for you. But there are a few problems with this approach. A classification solution is a logistic regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as:

$$\text{logistic}(\eta)=1/(1+\exp(-\eta))$$

The step from linear regression to logistic regression is kind of straightforward. In the linear regression model, we have modeled the relationship between the outcome and features with a linear equation. For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.

Performance Analysis

MODEL	ACCURACY
Decision Tree	0.6759230271533821
Random Forest	0.7577474703202585
Logistic Regression	0.6755714809042862

VI. CONCLUSION

In this report we investigated the problem of real-time prediction of severity of car accidents on traffic using real-time traffic data between 2016 to 2020, addressing an important problem to transportation and public safety. This problem is very challenging due to the imbalanced classes, the spatial heterogeneity, and its non-linear separable nature. We formulated the problem as a multi-class classification problem. In this report, we have also successfully identified the hotspot locations of accidents, various factors accounting for these accidents and the average time is taken to clear the accidents. found that in various variables like a day of the week, an hour of the day, a month of the year, and side of the road for which the number of accidents is very high for certain values. This can help the authorities to prepare in advance during these times and implement stricter traffic rules to decrease the number. Also, we found that during the nighttime the severity of accidents is more which is intuitive. This can indicate that more safety measures need to be taken at night time. We developed several classification machine learning models and evaluated the performance of these models using various evaluation metrics. Results show that our proposed approach (Random Forest) had a significantly improved result accuracy of 75% when compared to our other which had an accuracy of (Decision Tree 67%, Logistic Regression 67%). For future work, we would like to explore advanced machine learning techniques such as ensemble methods, XGBoost, and Neural Networks to predict the severity of traffic on accidents in real-time. We also plan to investigate approaches to predict casualty analysis.