US Accidents Analysis

Final Report

Group – 15

Aarushi Sharma

Sanil Rodrigues

857-869-7912 (Tel of Aarushi Sharma)

617-602-3507 (Tel of Sanil Rodrigues)

sharma.aaru@northeastern.edu

rodrigues.san@northeastern.edu

Percentage of Effort Contributed by Aarushi Sharma – 50 %

Percentage of Effort Contributed by Sanil Rodrigues – 50 %

Signature of Aarushi Sharma –

Signature of Sanil Rodrigues –

Submission Date – 06/24/2022

TABLE OF CONTENTS

PROBLEM DEFINITION

Most traffic accident studies have relied on small-scale datasets with limited coverage, reducing their effectiveness. Despite the ongoing research, the number of accidents continues to rise, a significant source of concern for everyone. In addition, most accident causes, and investigations are not publicly available to government entities or the public. Without precise information that includes area, cause, contributing factors, and linked activities associated with personnel injuries, identifying injury causative components becomes very theoretical. To address this issue, we're attempting to present a model that can demonstrate the following:

1. Classify the severity of the accidents.

DATA SOURCE

US Car Accidents is a nationwide car accident dataset that spans 49 states in the United States. The accident data was acquired using different APIs that give streaming traffic incident (or event) data **from February 2016 to December 2021**. These APIs transmit traffic data recorded by a range of entities within the road networks, including transportation state departments, law agencies, traffic cameras, and sensors. This dataset currently contains approximately 2.8 million accident records.

Data Source - https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

DATASET DESCRIPTION

The dataset consists of **47 variables/attributes** and approximately **2.8 million records**. The dataset has the following structure:

- 29 String Variables
- 15 Numerical Variables
- 3 Time Stamp Variables

DATASET VARIABLES DESCRIPTION

SNo.	Columns Name	Data Type	Description
1	ID	String	Unique Identifier of Accident Record.
2	Severity	Integer	Severity of Accident 1to 4 – where 1 indicates least impact on traffic (short delay) and 4 indicates significant impact (long delay).
3	Start_Time	Time Stamp	Start Time of Accident (Local Time Zone).
4	End_Time	Time Stamp	End Time of Accident (Local Time Zone)(when traffic flow was dismissed.
5	Start_Lat	Float	Latitude of start point in GPS coordinates.
6	Start_Lng	Float	Longitude of start point in GPS coordinates.
7	End_Lat	Float	Latitude of end point in GPS coordinates.
8	End_Lng	Float	Longitude of end point in GPS coordinates.
9	Distance (mi)	Float	Length of the road extent affected of the accident.
10	Description	String	Natural language description of the accident.
11	Number	Integer	Street number in address field.
12	Street	String	Street name in address field.
13	Side	String	Relative side of the street (Right/Left) in address field.
14	City	String	City in address field.
15	County	String	County in address field.
16	State	String	State in address field.
17	Zipcode	Integer	Zipcode in address field.
18	Country	String	Country in address field.
19	Timezone	String	Timezone based on the location of the accident (eastern, central, etc.)
20	Airport_Code	String	Airport-based weather station (closest one to location of the accident).
21	Weather_Timestamp	Time Stamp	Timestamp of weather observation record (in local time).
22	Temperature (F)	Float	Temperature (in Farenheit).
23	Wind_Chill (F)	Float	Wind chill (in Farenheit).
24	Humidity (%)	Integer	Humidity (in percentage).
25	Pressure (in)	Float	Air pressure (in inches).
26	Visibility (mi)	Float	Visibility (in miles).
27	Wind_Direction	String	Wind direction.
28	Wind_Speed (mph)	Float	Wind speed (in miles per hour).
29	Precipitation (in)	Float	Precipitation amount (inches), if any.
30	Weather_Condition	String	Weather condition (rain, snow, thunderstorm, fog, etc).

31	Amenity	String	Presence of amenity in a nearby location.
32	Bump	String	Presence of speed bump in a nearby location.
33	Crossing	String	Presence of crossing in a nearby location.
34	Give_Way	String	Presence of give_way in a nearby location.
35	Junction	String	Presence of junction in a nearby location.
36	No_Exit	String	Presence of no_exit in a nearby location.
37	Railway	String	Presence of railway in a nearby location.
38	Roundabout	String	Presence of roundabout in a nearby location.
39	Station	String	Presence of station in a nearby location.
40	Stop	String	Presence of stop in a nearby location.
41	Traffic_Calming	String	Presence of traffic_calming in a nearby location.
42	Traffic_Signal	String	Presence of traffic_signal in a nearby location.
43	Turning_Loop	String	Presence of turning_loop in a nearby location.
44	Sunrise_Sunset	String	Period of Day (Day/Night) based on sunrise or sunset.
45	Civil_Twilight	String	Period of Day (Day/Night) based on civil twilight.
46	Nautical_Twilight	String	Period of day (Day/Night) based on nautical twilight.
47	Astronomical_Twilight	String	Period of Day (Day/Night) based on astronomical twilight

DATA COLLECTION

The dataset consists of 2.8 million records and 47 attributes indicating the incidents of US Accidents. Out of all the attributes, we have one target variable **'Severity'** which classifies the severity of the accidents into 4 classes ranging from Severity 1 to Severity 4, 1 being the lowest and 4 being the highest. The dataset includes both numerical as well as categorical variables.

DATA PRE-PROCESSING

1. Some columns in the dataset had a high number of null values. Because those columns were not needed for model building or analysis, they were removed, and the remaining null values were removed. After removing most redundant attributes 2.6 million records and 37 attributes were preserved without any null values.
2. We encoded the categorical variables to numerical variables for further pre-processing, where we intend to apply PCA to the dataset. However, we haven't yet applied the

dimension reduction steps because we need this dataset for our next milestone of Exploratory Data Analysis.

3. The statistics of the attributes show us how the variables are varying throughout the dataset.
4. The ‘Severity’ attribute histogram plot below shows us that ‘Severity’ == 2 dominates the other classes with ~ 2.4 million records classified as the same. In the next milestone, we will focus on sampling the ‘Severity’ variable to make it uniform throughout the dataset.

‘Start_Time’ and ‘End_Time’ contains both Date and Time, to separate those entities we have split the columns as ‘Start_Date’, ‘Start_Time’, ‘End_Date’ and ‘End_Time.

DATA EXPLORATION

The initial US Accidents dataset had 48 attributes or variables and 2.8 million records on which data cleaning and preparation steps were performed to retain most data. By dropping the columns that were not needed for analysis, we preserved 2.6 million records and 37 variables/columns.

Furthermore, the dataset columns had the following data types before converting the categorical columns to numerical or introducing any dummy variables:

1. 20 Object Data Type Variables
2. 2 Int 64 Data Type Variables
3. 13 Float 64 Data Type Variables
4. 13 Bool Data Type Variables

To perform any dimension reduction methods, categorical variables were to be removed. Instead, categorical Variables were converted to numerical variables by label encoding them.
Below are the statistics of all the numerical variables after converting the categorical to numerical columns.
The statistics show how the columns are scattered throughout the dataset to give us a better understanding of what could our model yield as a result and what should we expect after doing an analysis.

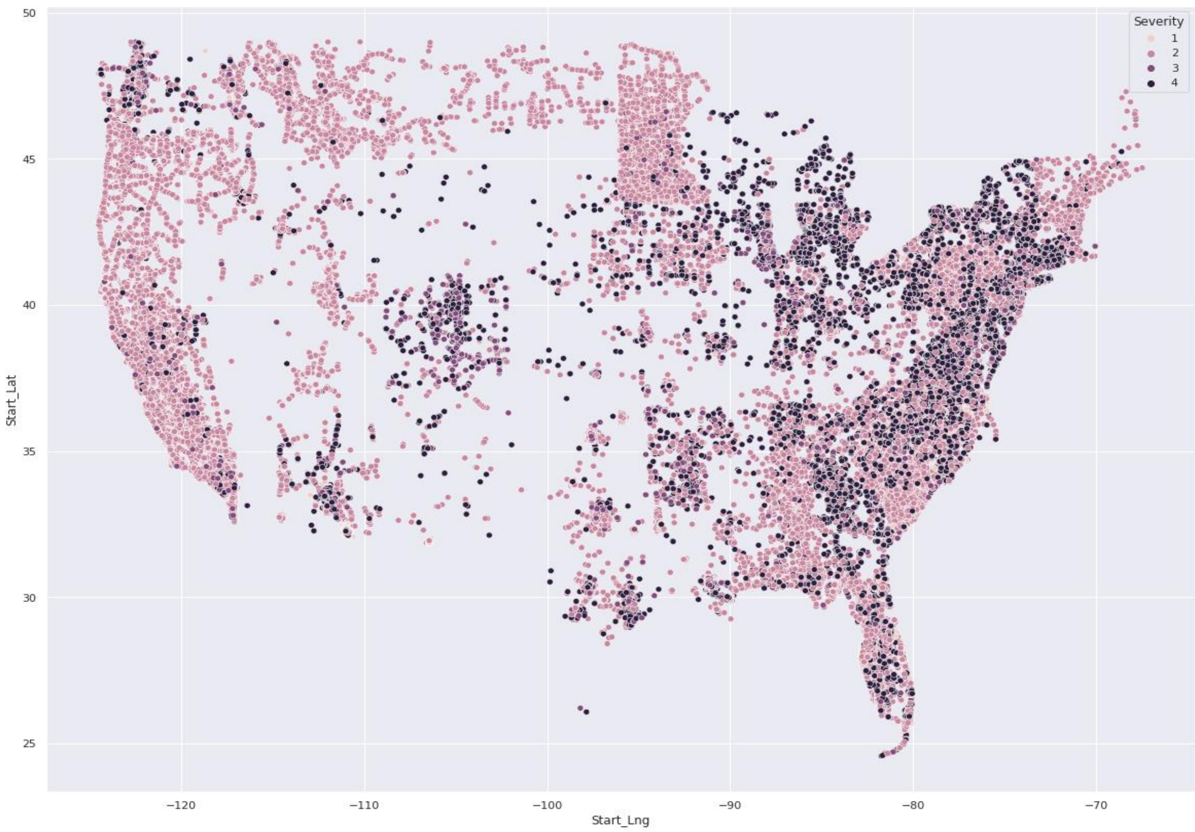
	Severity	Start_Lat	Start_Lng	Distance(mi)	City_Encode	Side_Encode	County_Encode	State_Encode	Timezone_Encode	Temperature(F)	...	Crossing_Encode	Junction_Encode
count	943318.000000	943318.000000	943318.000000	943318.000000	943318.000000	943318.000000	943318.000000	943318.000000	943318.000000	943318.000000	...	943318.000000	943318.000000
mean	2.064917	35.069960	-95.102568	0.274626	4367.522612	0.559909	773.523746	18.620350	1.468370	63.833323	...	0.130028	0.004202
std	0.380617	5.796634	17.794343	0.883254	2297.613923	0.496398	331.079085	15.016261	1.073773	18.162256	...	0.336335	0.064688
min	1.000000	24.566027	-124.517744	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-27.000000	...	0.000000	0.000000
25%	2.000000	30.229957	-117.833342	0.040000	2537.000000	0.000000	553.000000	3.000000	1.000000	51.000000	...	0.000000	0.000000
50%	2.000000	34.976113	-86.136779	0.111000	4796.000000	1.000000	820.000000	8.000000	1.000000	66.000000	...	0.000000	0.000000
75%	2.000000	39.232680	-80.359477	0.255000	6249.000000	1.000000	992.000000	35.000000	3.000000	78.000000	...	0.000000	0.000000
max	4.000000	48.996539	-67.484130	112.968000	8503.000000	1.000000	1410.000000	48.000000	3.000000	196.000000	...	1.000000	1.000000

8 rows × 28 columns

Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Direction_Encode	Weather_Condition_Encode	Wind_Speed(mph)	Amenity_Encode	Bump_Encode	Crossing_Encode	Junction_Encode
943318.0	943318.0	943318.0	943318.0	943318.0	943318.0	943318.0	943318.0	943318.0	943318.0
64.54482793713255	29.421216397862302	9.201041440956255	8.988014646174461	22.6491183248915	7.131648288275984	0.021164654973190377	0.000726160213204879	0.13002826194348036	0.004202188445465898
22.397021186990592	1.0229379555803597	2.4613302827127064	7.042222724050796	20.72621354571581	5.412885254567459	0.14393308971482893	0.026937588492419253	0.33633470376736224	0.06468797796893473
1.0	16.72	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49.0	29.29	10.0	2.0	10.0	3.0	0.0	0.0	0.0	0.0
66.0	29.78	10.0	9.0	10.0	7.0	0.0	0.0	0.0	0.0
83.0	29.98	10.0	15.0	52.0	10.0	0.0	0.0	0.0	0.0
100.0	58.16	100.0	22.0	90.0	1087.0	1.0	1.0	1.0	1.0

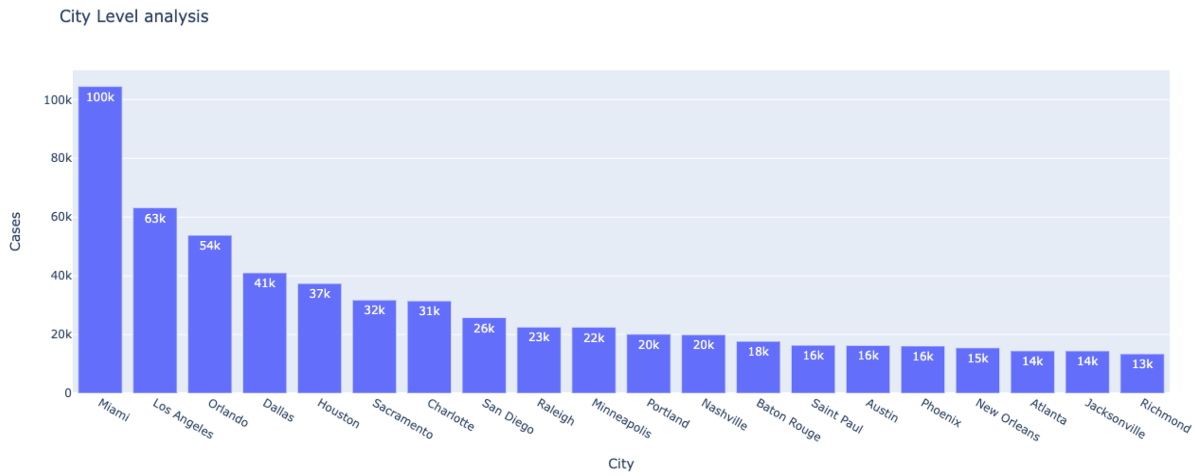
DATA VISUALISATION

1. SEVERITY OF ACCIDENTS



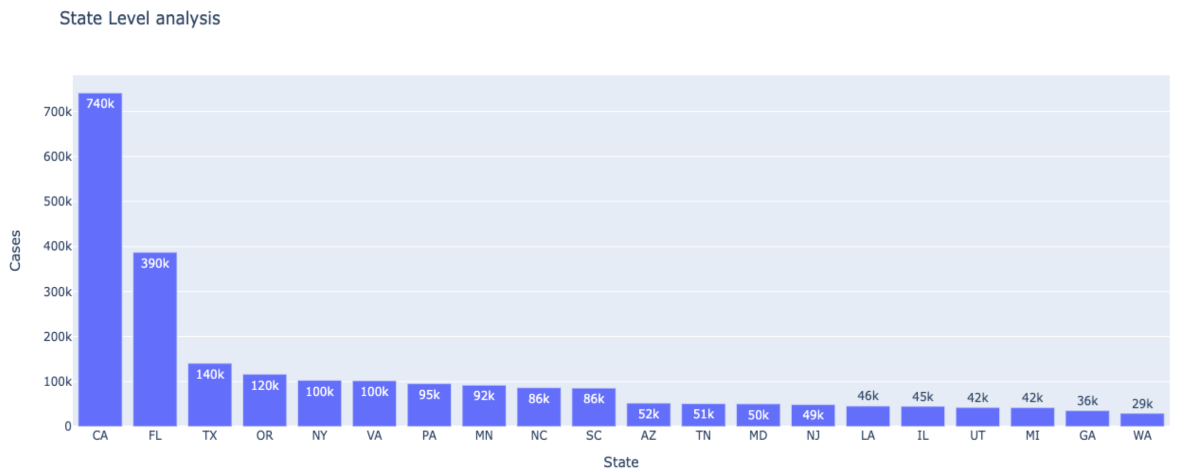
The following visualization tells the severity of the accidents across the United States; we have plotted the latitude and longitude on the x and y-axis in a scatter plot. The points form a map-like structure. The hue in the plot shows the s severity of accidents across the country. We can see that the northeast part of the United States has a high number of fatal accidents compared to the rest of the country. Also, the crowded points show the metropolitan cities, and the points forming a line show national highways.

2. CITY-LEVEL ANALYSIS



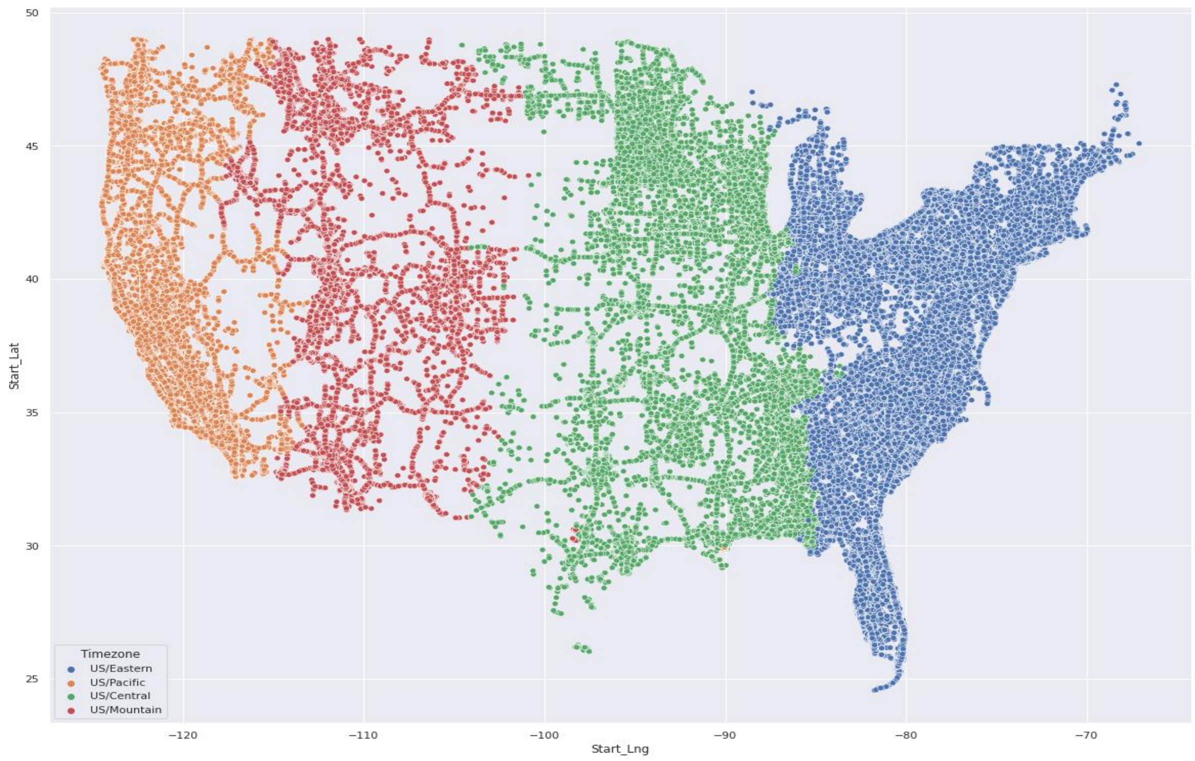
City-level analysis shows us the top 20 cities with the number of accidents. From the following bar chart, we can see that Miami, LA, and Orlando are the top 3 cities in which most accidents take place.

3. STATE-LEVEL ANALYSIS



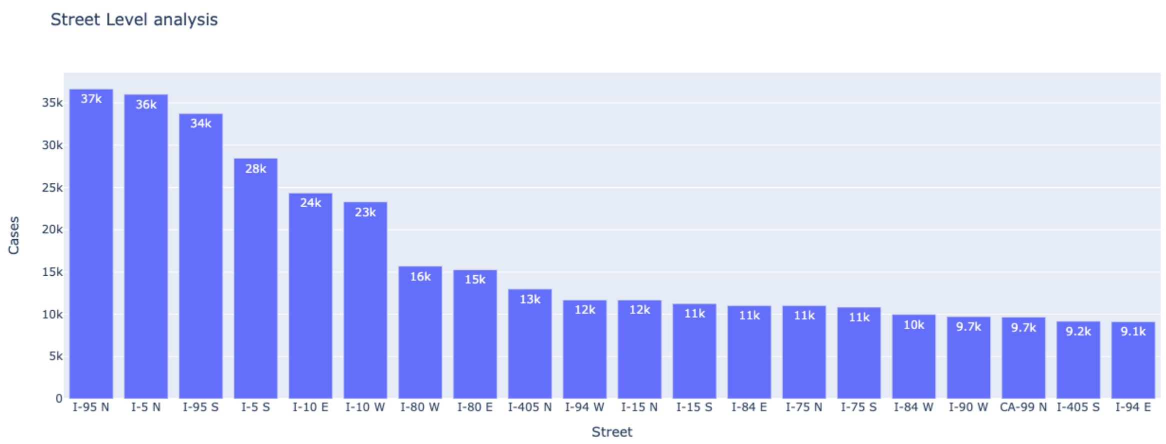
State-level analysis shows the number of accidents according to various states, and we can observe that California, Florida, and Texas have large accident rates.

4. ACCIDENTS IN DIFFERENT TIME-ZONES



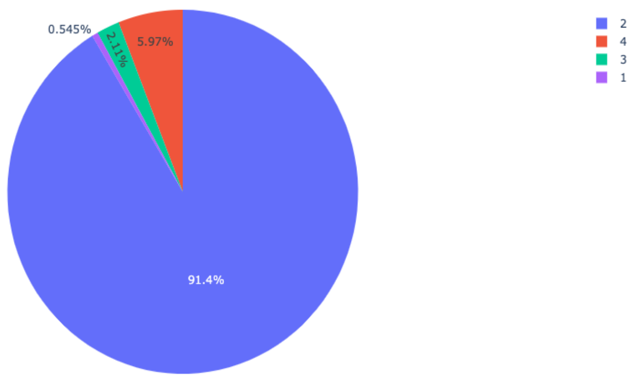
This visualization shows accidents in different time zones. We can see that the Eastern time zone has the highest number of accidents as compared to other time zones.

5. STREET-LEVEL ANALYSIS



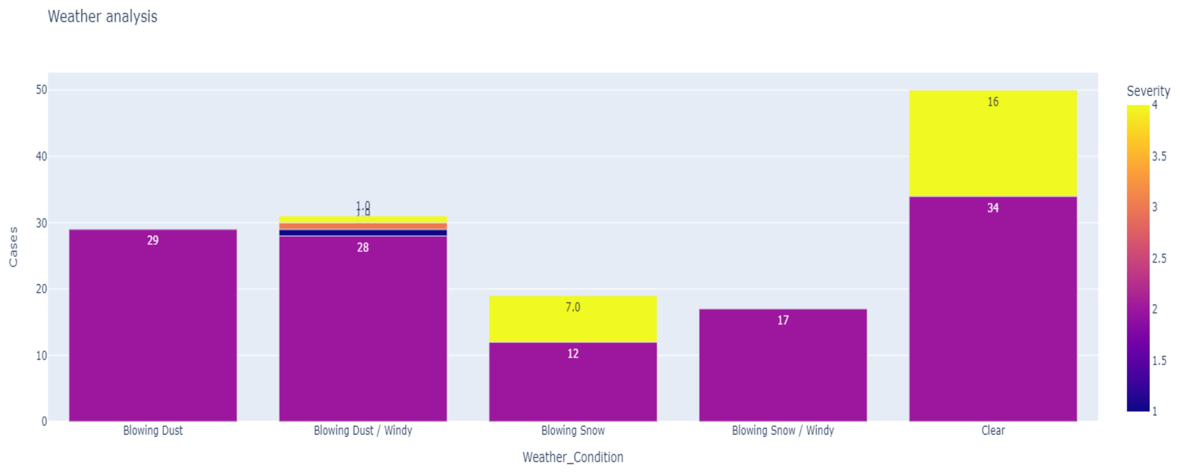
Street-level analysis shows the top 20 accident-prone streets in the United States. This visualization can be used to take necessary precautions in the following street.

6. SEVERITY ANALYSIS



In the severity analysis, we can see that most of the accidents had severity two, which was not that serious; we also have many fatal accidents with severity.

7. WEATHER ANALYSIS



The weather analysis shows us the proportion of accidents that occurred while the weather condition at that time. This tells us the severity level concerning how bad the weather was and can give us an analysis that a lot of high severity accidents occur when the weather condition is harsh.

To perform dimension reduction on the dataset, Principal Components Analysis will be applied to the dataset in the project's next milestone. From this, we will try to reduce the dimensions to a low-dimensional space to capture maximum variation from the variables contributing the most to our model and analysis.

DATA MINING TASKS

The next step after Data Exploration and Visualization was to analyze and implement different models for our Project and then pick one based on the analysis as well as the performance of the models. From the previous milestone, we gathered some insights into how the data was and what should be used to have a good model. Finally, we represent the dataset by creating some visualization to gaze at different patterns from the features. In this stage, we have used two Classification Machine Learning Models to classify the severity of the accidents. For feature reduction, we implemented PCA (Principal Component Analysis) on the dataset, with 28 variables left after data cleaning and preparation (including label encoding and removing features with high linear dependence). After applying PCA, the variance captured by the first five features came up to approximately 39%, whereas the variance captured by 20+ variables equated~90%. Since we could not reduce the number of features significantly with the implementation of PCA, we decided not to go ahead with PCA for our model. Furthermore, the dataset had a total of 0.9 million records; applying machine learning algorithms takes a considerable amount of time; we randomly sampled our dataset taking only 100K records for executing Machine Learning Algorithms.

In total, we have applied five models for our classification project listed as follows:

1. KNN Classification
2. Decision Tree Classification
3. Naïve Bayes
4. Support Vector Machine
5. Logistic Regression

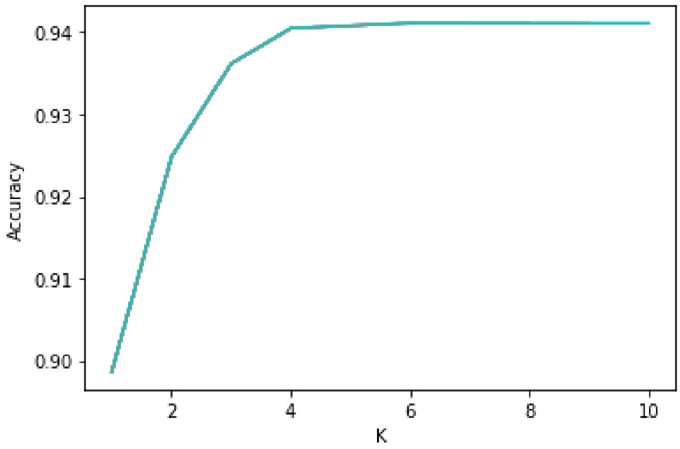
We computed the correlation between the predictors and removed the redundant variables having Pearson Correlation Coefficient greater than 0.7. This reduced the number of columns to 20. We label encoded the categorical variables to change it to numerical columns.

Our dataset was left with a total of 0.9 million records, and applying algorithms takes a considerable time to get the output, so we decided to take a sample, i.e., 100 K records, to use this ML algorithm.

DATA MINING MODELS/METHODS

a. KNN-CLASSIFICATION

The KNN Classification algorithm's main idea is to use identical records from the training dataset to classify new records from the testing dataset. First, the K-Nearest Neighbors method identifies similar records (behaving like neighbors). The new record is then classified as a member of the majority class of the K-Neighbors using a simple majority rule.



As we can see from the graph above and the table below stating the KNN Model measures with different values of K, K = [6,10] provides the highest accuracy and lowest mean squared error. So, K could be any number between 6 and 10.

K	Accuracy	MSE
1	0.89864	0.27592
2	0.92480	0.17832
3	0.93616	0.16896

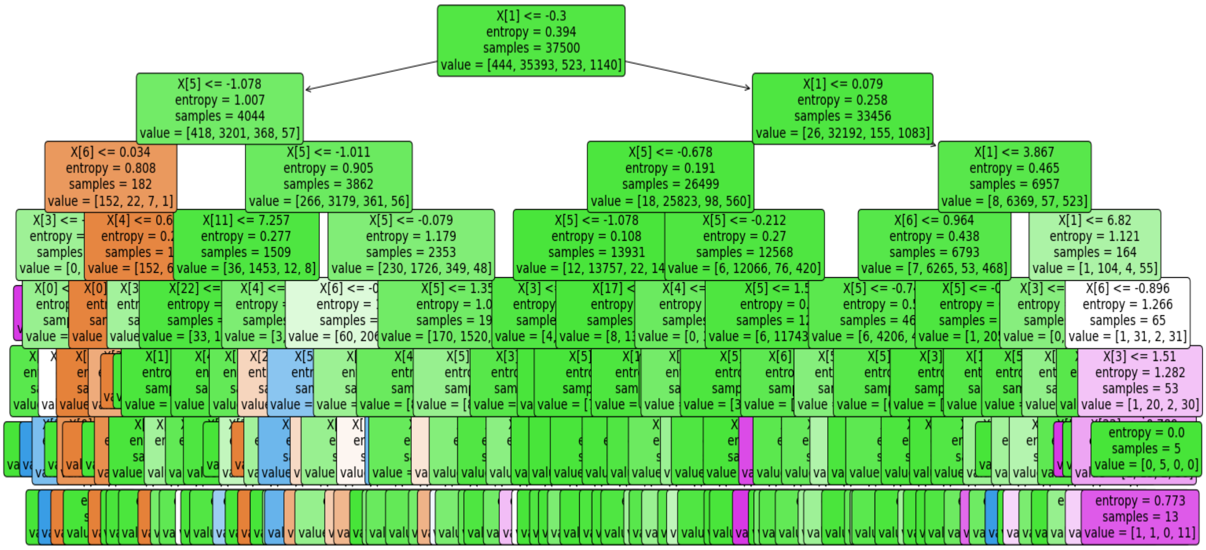
4	0.94048	0.15768
5	0.94080	0.15736
6	0.94112	0.15656
7	0.94112	0.15656
8	0.94112	0.15656
9	0.94112	0.15656
10	0.94112	0.15656

b. DECISION TREE CLASSIFICATION

A Decision Tree is a straightforward representation for categorizing examples. It is a Supervised Machine Learning method in which data is continuously split based on a specific parameter. By splitting predictors into sub-groups, trees generate easily interpretable logical rules.

Depth = 7 was chosen as the ideal case because it produces the highest accuracy and the lowest mean squared error.

Depth	Accuracy	MSE
1.0	0.94112	0.15656
2.0	0.94336	0.15520
3.0	0.94368	0.1544
4.0	0.9436	0.15448
5.0	0.946	0.1528
6.0	0.94616	0.15256
7.0	0.9468	0.15048
8.0	0.94488	0.15416
9.0	0.94448	0.1572
10.0	0.91512	0.25304



c. NAÏVE BAYES CLASSIFICATION

The Bayes Theorem is used to build the Naive Bayes statistical categorization approach. It's one of the most straightforward supervised learning algorithms on the market. The Naive Bayes classifier is a reliable, fast, and accurate algorithm. Naive Bayes classifiers have good accuracy and speed on massive datasets.

Naive The Bayes classifier assumes that the impact of one feature on a class is unaffected by the impact of other features. A loan applicant's value is established, for example, by his or her income, prior loan and transaction history, age, and geographic area. Even though these characteristics are interrelated, they are nonetheless assessed independently. This assumption is considered naive since it makes calculation easier. This assumption is known as class conditional independence.

d. SUPPORT VECTOR MACHINE MODEL

Support vector machines (SVM) is a supervised machine learning approach. Although it is mainly used for classification, it can also be used to handle regression problems.

SVMs divide nearly all points into two groups by defining a decision boundary and a maximum margin. While allowing for specific classification errors.

Support vector machines have taken the place of maximum margin algorithms. Its primary advantage is that it can establish a linear and non-linear decision boundary using kernel functions. This makes it more suitable for real-world applications where data isn't always separable by a straight line.

e. LOGISTIC REGRESSION MODEL

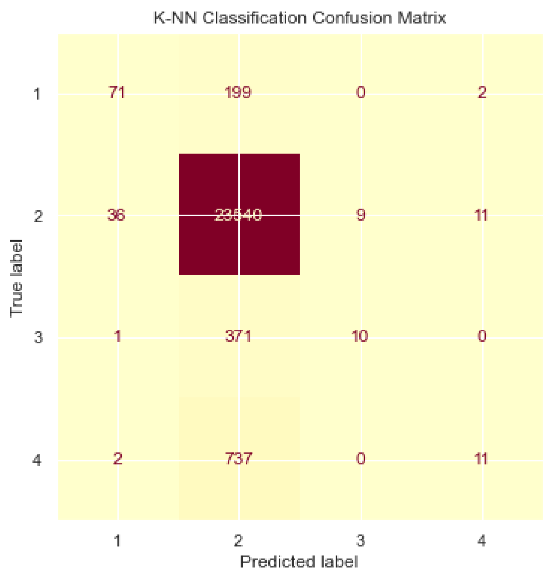
Logistic regression is a statistical strategy for creating machine learning models with a dichotomous (binary) dependent variable. Logistic regression describes data and the connection between one dependent variable and one or more independent variables. Independent variables might be nominal, ordinal, or interval variables. Utilizing a logistic function inspired the term "logistic regression." The logistic function is also known as the sigmoid function. The value of this logistic function is between zero and one.

PERFORMANCE EVALUATION

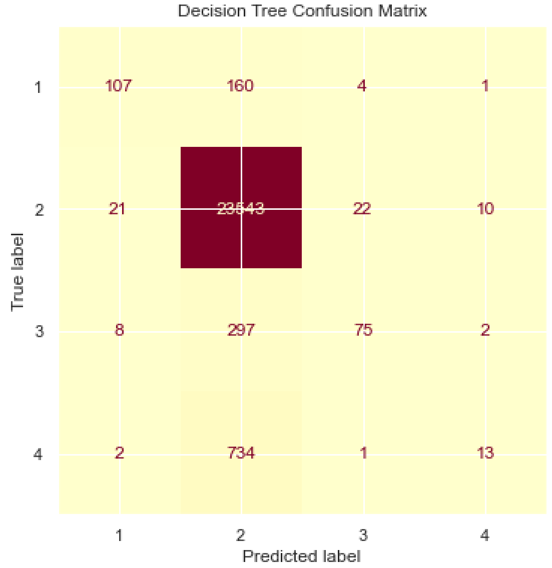
The above model implementations have been used to classify the Severity of Accidents in our dataset. In addition, we have used the confusion matrix and accuracy score for this milestone to conduct a performance evaluation of all five implemented machine learning models. The below confusion/classification matrices indicate the number of data points the model is classifying the data correctly or incorrectly.

We can visualize the performance of classification machine learning models using the Confusion Matrix. This gives us a better idea of how our machine learning models perform when it comesto classifying new data.

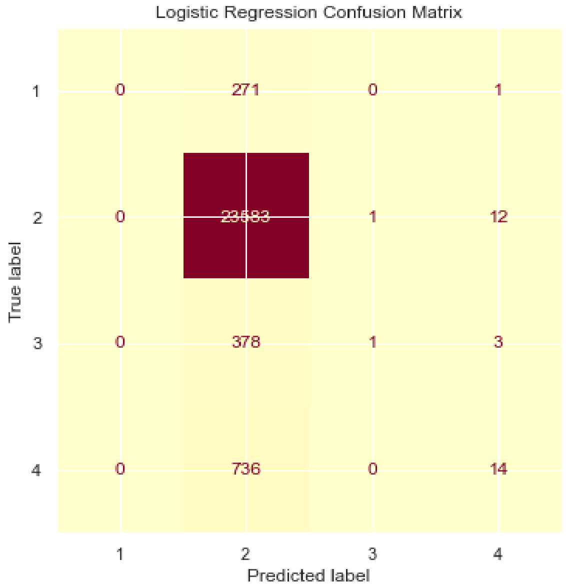
a. Confusion Matrix for K-NN Classification Model



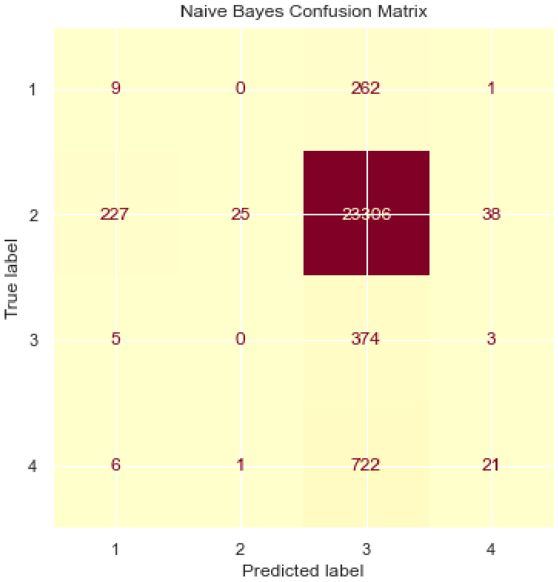
b. Confusion Matrix for Decision Tree Classification Model



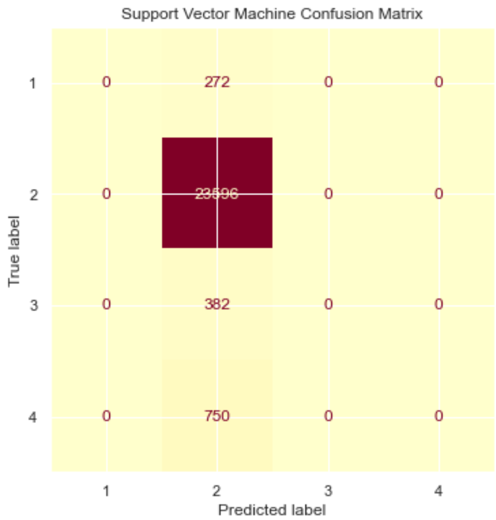
c. Confusion Matrix for Logistic Regression Classification Model



d. Confusion Matrix for Naïve Bayes Classification Model

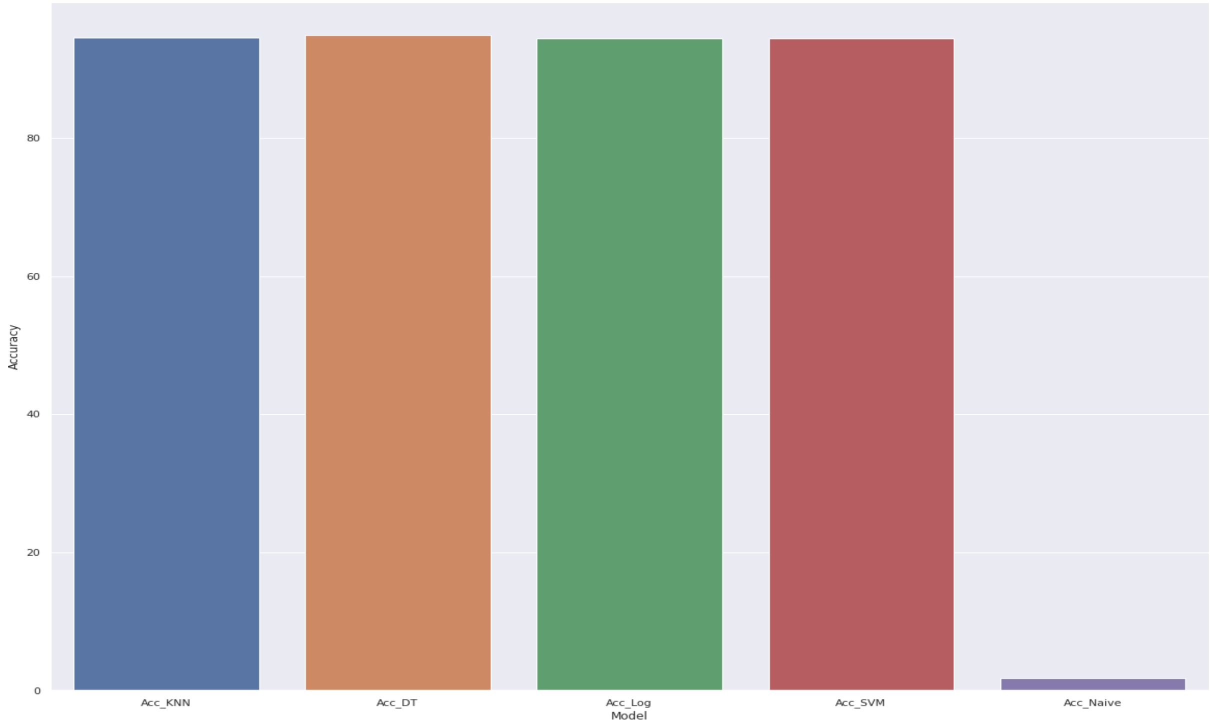


e. Confusion Matrix for Support Vector Machine Classification Model

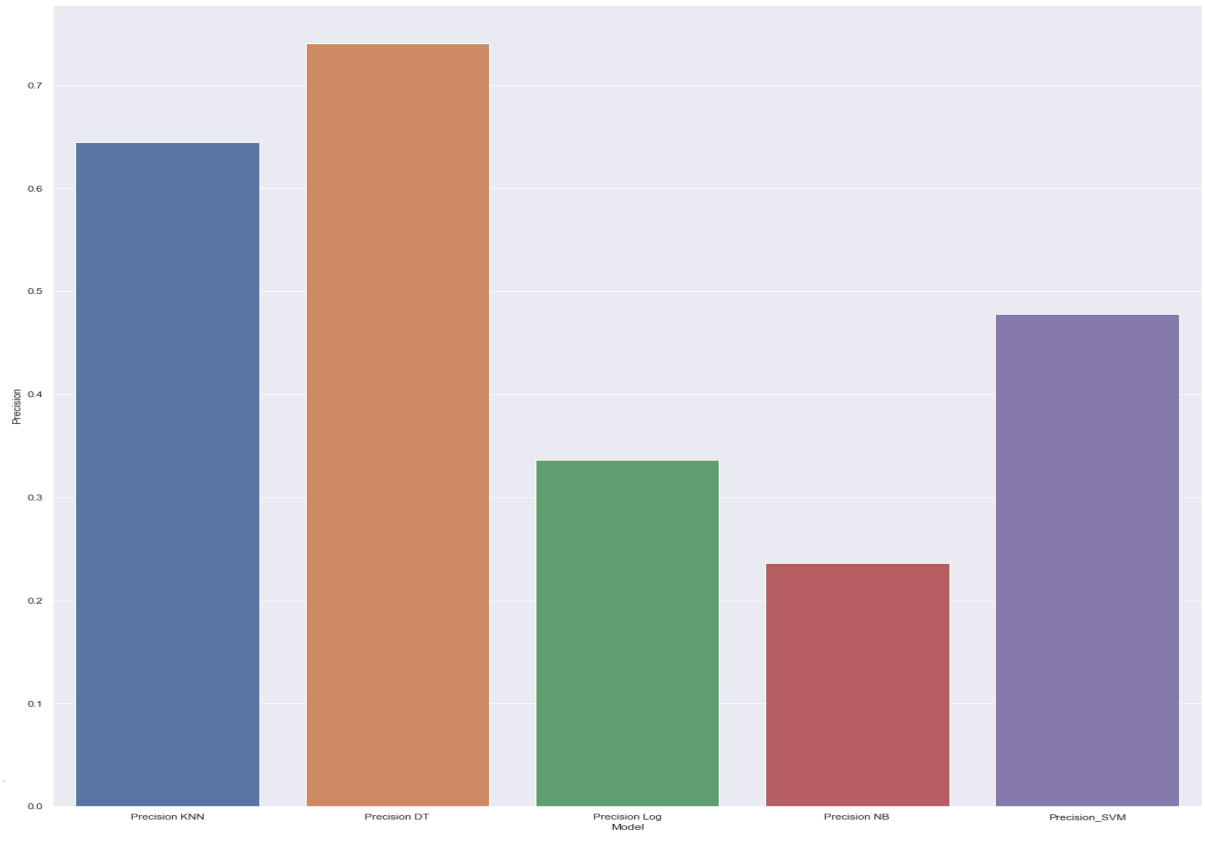


PROJECT RESULTS

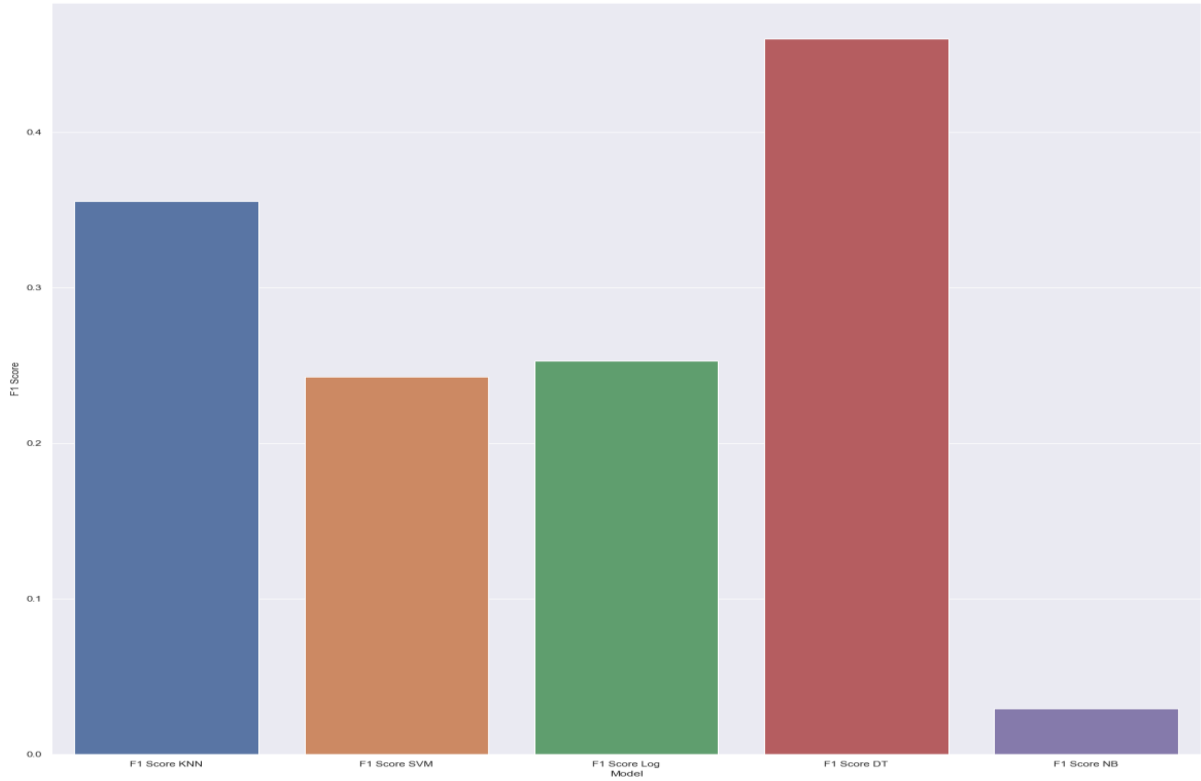
IMPLEMENTED CLASSIFICATION MODELS ACCURACY



IMPLEMENTED CLASSIFICATION MODELS PRECISION



IMPLEMENTED CLASSIFICATION MODELS F1-SCORE



The above graph shows the accuracy of each model when implemented. We can implement any of the first four models from the above measure. Still, since Decision Tree is generating the highest accuracy amongst the five models applied, we would go ahead with Decision Tree Classifier for now.

The Decision Tree model has the highest accuracy of 94.952 %. It also has the highest precision of 0.7406 and f1-score of 0.459.

CONCLUSION

We can finally conclude that the decision tree is the best algorithm in all the measures, which gave us accuracy, precision, and f1 score. Overall, most of the models performed well, except naïve Bayes. The naïve Bayes model performed the worst among all the models giving us the least accuracy, precession, and f1 score.

Overall, the selected topic made us learn all the data mining concepts. However, in some areas, we faced some difficulty, especially regarding data cleaning variable selection and model implementation. Due to the massive size of our dataset, we faced problems during the implementation of the model, such as runtime crashes and ram usage outages. We worked on optimizing by taking sample data and making some changes in the code.