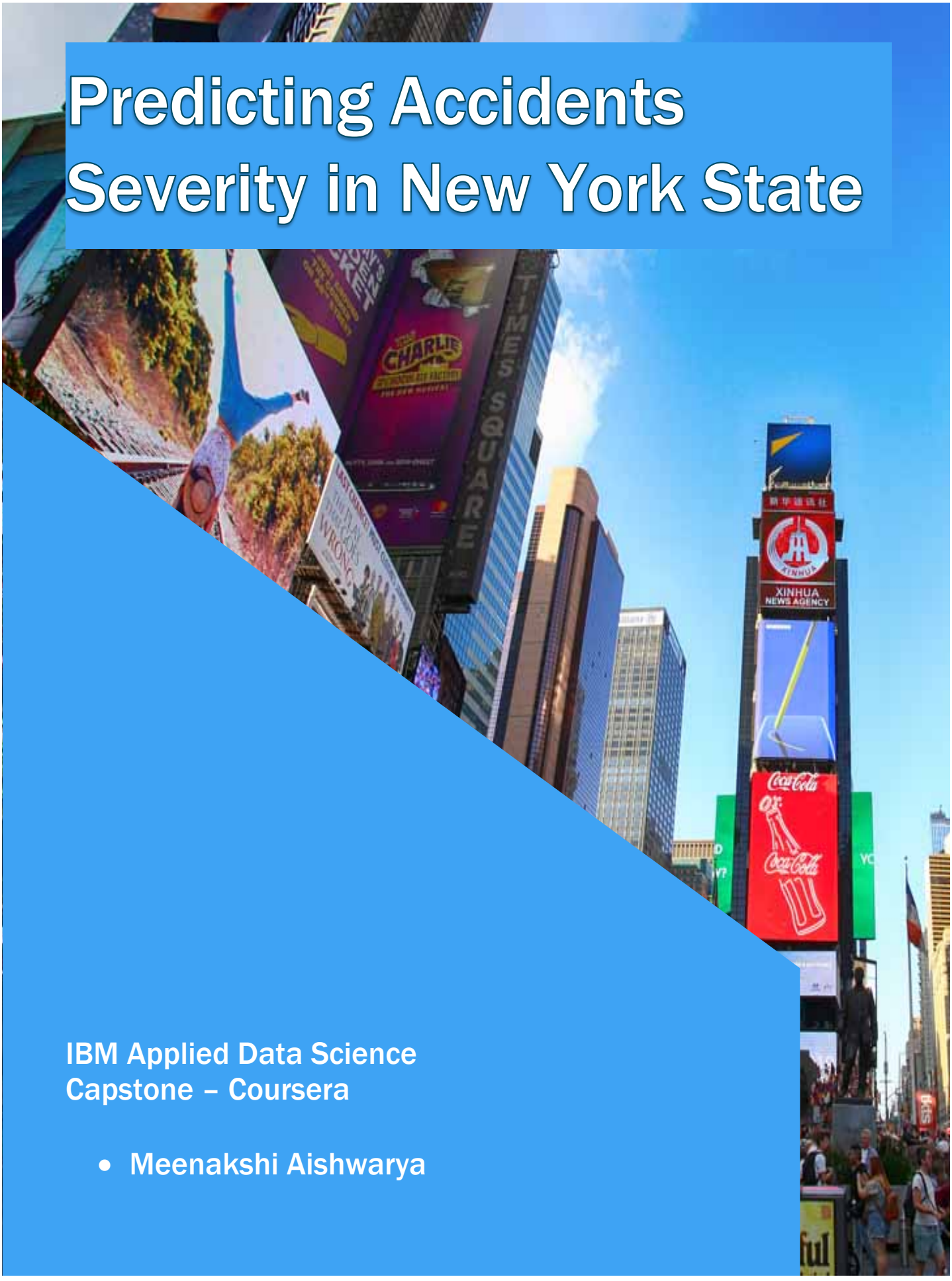
Predicting Accidents Severity in New York State



IBM Applied Data Science Capstone – Coursera

- Meenakshi Aishwarya

Introduction

In today's world the number of vehicles on road are incresing. Be it cars or two-wheelers, taxis, trucks, buses, etc., the number vehicles on the road is increasing by the day. While this shows a good scale of development, it also means that there are higher chances of road accidents. Road accidents can be fatal and very dangerous; thus we must do our best to avoid them.

In this project, the prediction of the severity of the road accidents is done using a Machine Language Model.

Business Problem

The objective of this capstone project is to

1. Analyse accidents severity in New York State
2. To predict the severity of an accident in New York State using machine learning techniques.

The problem question is: What could be the severity of the accident if I go for a road trip today?

Data

To solve the problem, we need the following data:

- List of accidents in New York State
- Date, time, duration and location of each accident
- Cause and severity of each accident

In order to meet these criteria, the source of dataset selected is given below:

https://www.kaggle.com/sobhanmoosavi/us-accidents

the above dataset is a countrywide traffic accident dataset, which covers 49 states of the United States with 49 attributes. It has data from February 2016 to June 2020. The data set is in the form of csv file. So the dataset is should be prepared by removing the unnecessary data. The records from New York State should be selected and the features or attributes of dataset should be identified. The following features or attributes are selected from the dataset:

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Source	Indicates source of the accident report	No
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.	Yes
4	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).	No
5	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
6	Start_Lat	Shows latitude in GPS coordinate of the start point.	No

7	Distance	The length of the road extent affected by the accident.	No
8	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
9	City	Shows the city in address field.	Yes
10	County	Shows the county in address field.	Yes
11	State	Shows the state in address field.	Yes
12	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
13	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
14	Humidity(%)	Shows the humidity (in percentage).	Yes
15	Pressure(in)	Shows the air pressure (in inches).	Yes
16	Visibility(mi)	Shows visibility (in miles).	Yes
17	Wind_Direction	Shows wind direction.	Yes
18	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
19	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
20	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
21	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
22	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No
23	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
24	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
25	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
26	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
27	Station	A POI annotation which indicates presence of Station in a nearby location.	No
28	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
29	Traffic_Calming	A POI annotation which indicates presence of traffic calming in a nearby location.	No
30	Traffic_Signal	A POI annotation which indicates presence of traffic signal in a nearby location.	No
31	Turning_Loop	A POI annotation which indicates presence of turning loop in a nearby location.	No
32	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
33	Hour	Shows the hour at accident	No

34	Weekday	Shows weekday of accident	No
35	Time_Duration(min)	Shows duration of accident	No

Methodology

The methodology of the project had stated with data preparation followed by exploratory data analysis and then building the machine learning model using KNN technique.

Data Preparation

Primarily the year, month, day, hour, weekday and time duration for clearing the accidents are extracted for each data record. The records which produced negative time durations are dropped as the time duration cannot have negative values. Having those records may reflect in wrong prediction of the model. Then the outliers (an observation that lies an abnormal distance from other values in a random sample from a population) are dealt by replacing them with corresponding mean values.

The features which would impact the accident severity such as id, source, TMC, severity, start time, latitude, longitude, distance, side of the road, city, county, state, time zone, temperature, humidity, pressure, visibility, wind direction, amenity, accident location, sunrise or sunset, hour, weekday and time duration to clear the accident are selected from the dataset. From the selected dataset the missing values are dropped and final dataset is selected by selecting New York records from the state feature. Then dummy variables are created to categorial values which help in machine learning algorithm.

Exploratory Data Analysis

For the data analysis, initially the map of accidents is produced based on the latitude and longitude of the accident. It is plotted as a scatter plot using separate color for each county. The map is shown in fig 1.

The time series analysis in fig 2 shows that the maximum number of accidents occurred during October 2017. Even though the trend of time series has a positive slope the number of accidents has increased to maximum during October 2018 and had a decrease. This may be the result of awareness of road safety among the citizens.

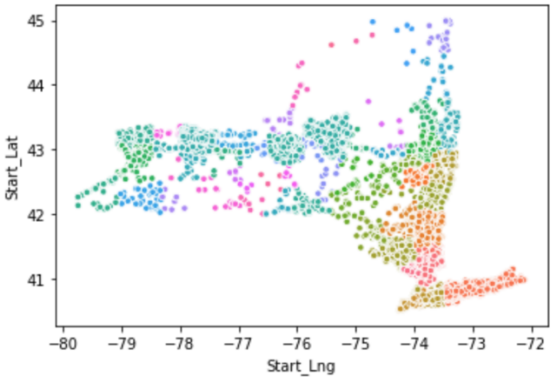


Fig 1: Map of Accidents

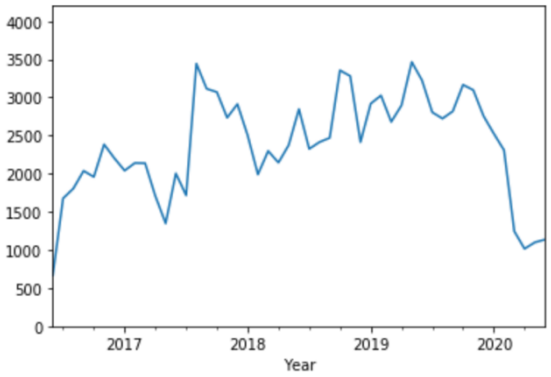


Fig 2: Time Serie Analysis

From the data the groups of severity and number of accidents in each group are found. It is found that there are 4 groups of severity rated from level 1 to 4 where level is least severe and level 4 is most severe. Number of accidents per level are as shown below:

Level of Severity	Number of accidents
2	69142
3	47085
4	162
1	25

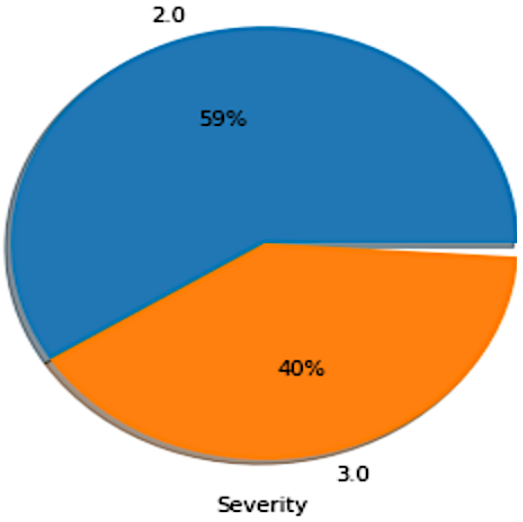


Fig 3: Severity level of accidents

One of the important questions is "When do most accidents takes place?". To answer this question the analysis on daytime vs nighttime and weekday vs weekend is done. The charts of the analysis are given in fig 4 and fig 5. The daytime vs nighttime pie chart shows that 75% of the accidents took place during the day and 25% of the accidents occurred during the night. In the bar chart of weekday vs weekend, it can be seen through comparison that the number of accidents is maximum during the weekdays when compared to weekends. Also, it can be concluded that maximum number of accidents occurred during Tuesday followed by Friday, Wednesday, Thursday, Monday, Sunday and Saturday.

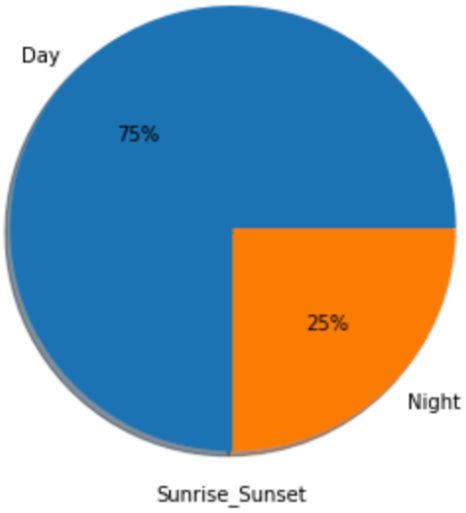


Fig 4: Daytime vs Nighttime

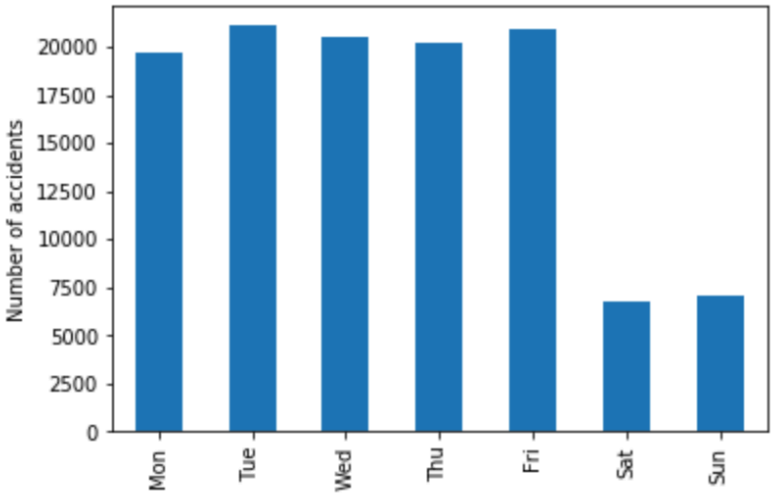


Fig 5: Accidents by Day of Week

The next big question is "Where do most accidents happen?". In order to answer this question, the analysis is done based on county, city and street side. The pie charts of the analysis are shown in fig 6, 7, 8. From the county pie chart, the top 5 counties where the number of accidents is maximum are Westchester (14%), Monroe (13%), Queens (10%), Suffolk (8%) and Bronx (7%). From the city pie chart, the top 5 cities with maximum number of accidents are Rochester (10%), Bronx (7%), New York (4%), Brooklyn (4%) and Albany (3%). From the street side pie chart, it can be said 84% of the accidents occurred on the right side of the street whereas 16% of the accidents occurred on left side of the street.

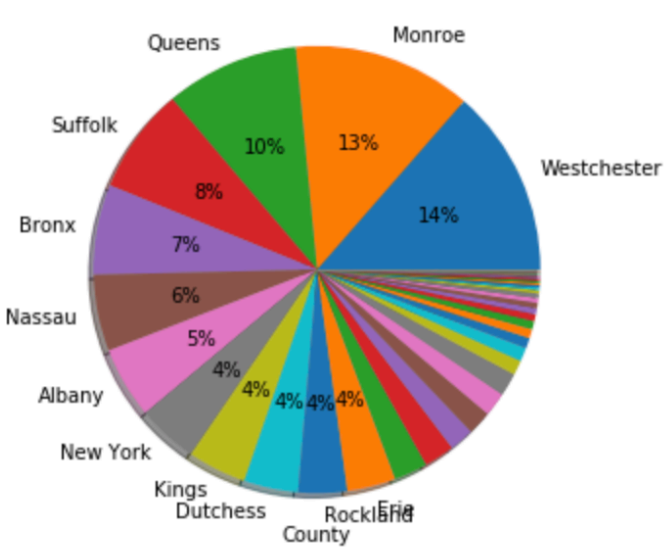


Fig 6: Accidents by County

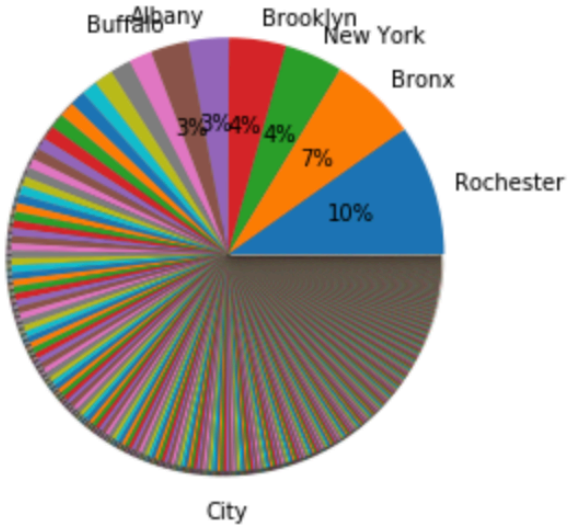


Fig 7: Accidents by City

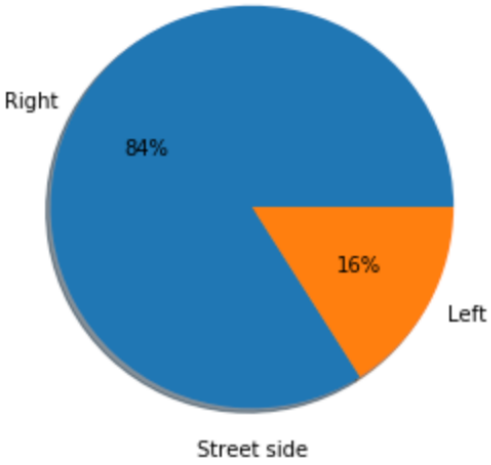


Fig 8: Accidents by Street Side

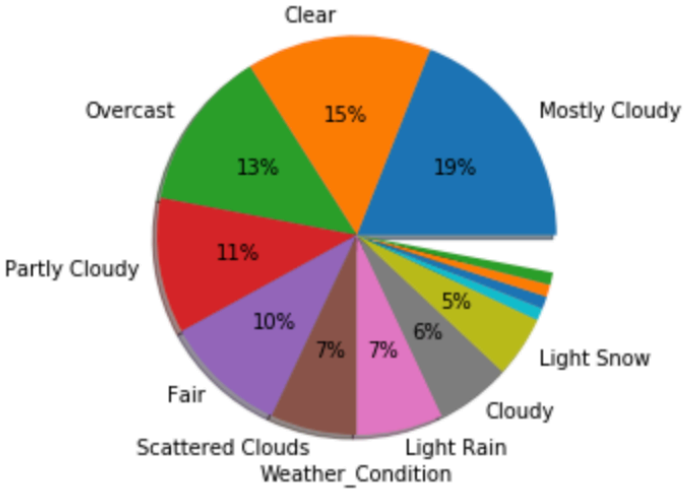


Fig 9: Accidents by Weather Condition

The question "With what weather condition do most accidents happen?" is also analyzed. The pie chart of accidents by weather condition is shown in fig 9. From the pie chart is can be said that 19% of accidents occurred when the weather is mostly cloudy, 15% when the weather is clear, 13% when there is overcast, 11% when its partly cloudy and 10% when the weather is fair.

Machine Learning Technique: K-Nearest Neighbor

The K-Nearest Neighbor (KNN) technique is used to build the supervised machine learning. In KNN technique, assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. This technique as it is robust to the noisy training data and is effective if the training data is large.

In order apply KNN technique, the 80% of dataset is used as training data and remaining 20% is used for test data. Then the algorithm of KNN technique is applied to the data set. The algorithm of KNN technique is as follows:

- **Step-1:** Select the number K of the neighbours
- **Step-2:** Calculate the Euclidean distance of **K number of neighbours**
- **Step-3:** Take the K nearest neighbours as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbours, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbour is maximum.
- **Step-6:** Our model is ready.

Results

The K for the KNN model is chosen through iterative process. The accuracy of the model for different K values is shown is fig 10. The highest accuracy is achieved at K=9 with accuracy of 0.6621.

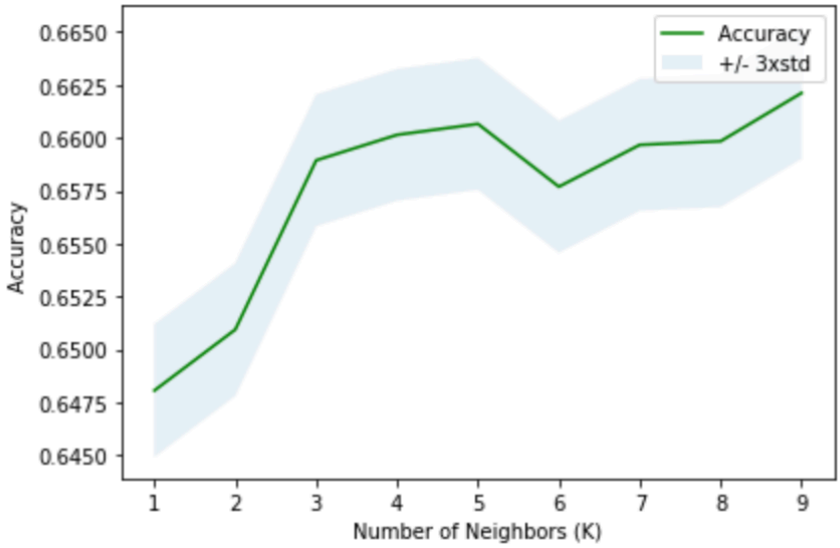


Fig 10: Accuracy of Model at Different K Values

Therefore, the model is built using K=9 and the train set accuracy and test set accuracy of the model is given below:

Dataset	Accuracy
Train Set	0.7362854473805714
Test Set	0.6621139887471545

Conclusion

This project indicates that there are patterns of when, where and under what weather conditions did most accidents occurred. The severity of each accident can be predicted quite accurately with various classification machine learning algorithms. The machine learning algorithm used in this project is K-Nearest Neighbour algorithm. The model gives an accuracy of 66.21% with test data. The further analysis may include the state of driver of the vehicle, road condition, vehicle performance characteristics.