**Introduction and Objectives**

Motor vehicle collisions happen very often and can be very damaging financially, mentally, and physically. New York City is very dense and crowded, thus motor vehicle collisions are more likely to happen. The New York City Police Department updates records daily on motor vehicle collisions. As of writing this report, there are about 1.9 million reported incidents. We believe there is a connection between injuries and deaths with the type of vehicles involved in the collisions.

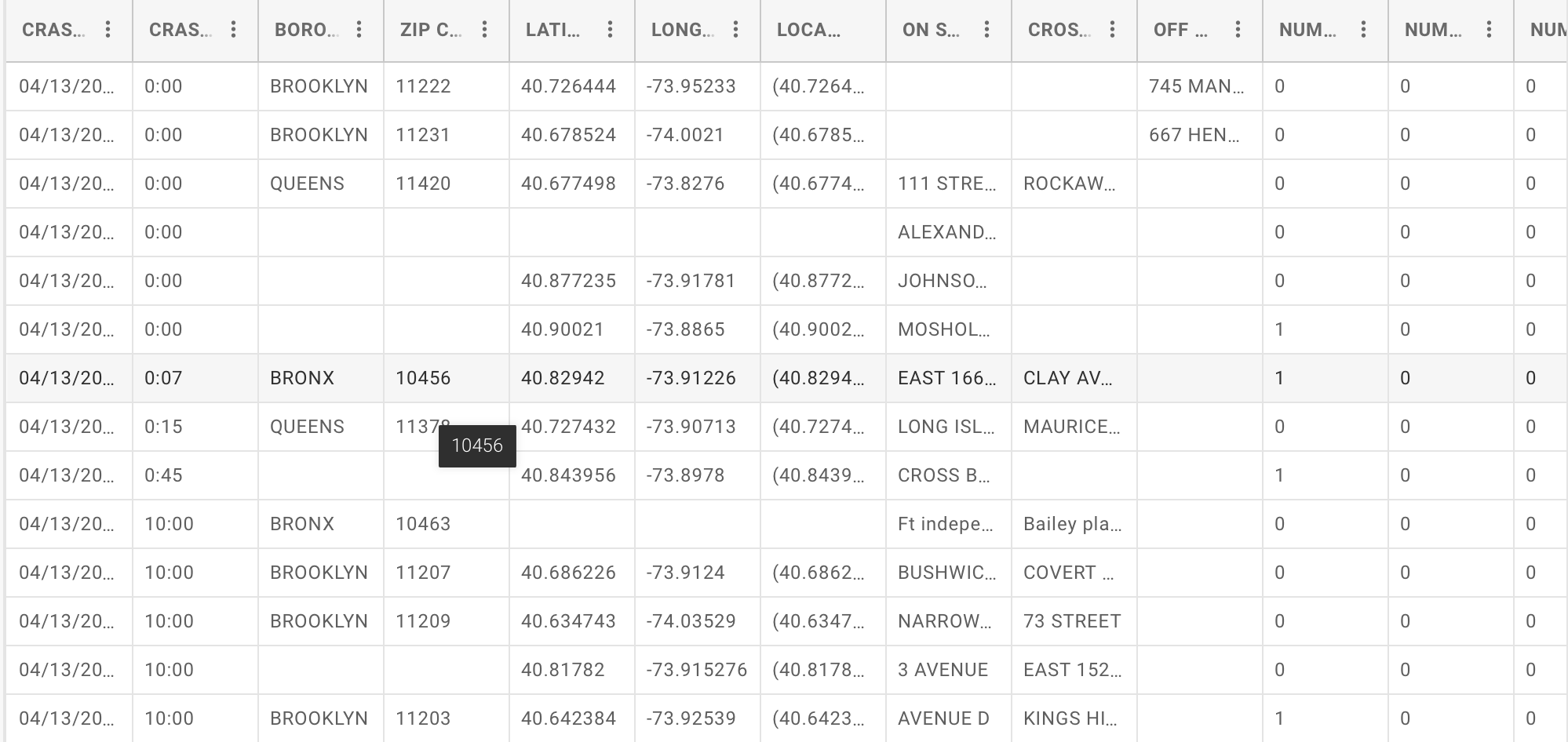
**Description of Data Set**

Dataset Name: NYC Motor Vehicles Collisions - Crashes 2018

Dataset Link: <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>

This dataset covers the information about NYC motor vehicle collisions. This dataset first became available on April 28, 2014. As well the data is frequently updated; the last data upload was October 7,2022. The agency that governs this dataset is the New York Police Department. In addition, the way this information is gathered is through a form (MV-104AN). Whenever there is a crash or an accident in which a party was injured, or damages surpass 1 thousand dollars; this form would be filled out. The information that is collected through the form is exported into this dataset automatically.

Dataset Structure:



Attributes and Objects: There are 29 attributes and 1,949,630 objects.

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type |
| CRASH DATE | Occurrence date of collision | Date |
| CRASH TIME | Occurrence time of collision | Text |
| BOROUGH | Borough where collision occurred | Text |
| ZIP CODE | Postal code of incident occurrence | Text |
| LATITUDE | Latitude coordinate for GCS | Numeric |
| LONGITUDE | Longitude coordinate for GCS | Numeric |
| LOCATION | Latitude, longitude pair | Location |
| ON STREET NAME | Street on which the collision occurred | Text |
| CROSS STREET NAME | Nearest cross street to the collision | Text |
| OFF STREET NAME | Street address | Text |
| NUMBER OF PERSONS INJURED | Number of persons injured | Numeric |
| NUMBER OF PERSONS KILLED | Number of persons killed | Numeric |
| NUMBER OF PEDESTRIANS INJURED | Numbers of pedestrians injured | Numeric |
| NUMBER OF PEDESTRIANS KILLED | Number of pedestrians killed | Numeric |
| NUMBER OF CYCLIST INJURED | Number of cyclist injured | Numeric |
| NUMBER OF CYCLIST KILLED | Number of cyclist killed | Numeric |
| NUMBER OF MOTORIST INJURED | Number of motorist injured | Numeric |
| NUMBER OF MOTORIST KILLED | Number of motorist killed | Numeric |
| CONTRIBUTING FACTOR VEHICLE 1 | Factors contributing to the collision for designated vehicle | Text |
| CONTRIBUTING FACTOR VEHICLE 2 | Factors contributing to the collision for designated vehicle | Text |
| CONTRIBUTING FACTOR VEHICLE 3 | Factors contributing to the collision for designated vehicle | Text |
| CONTRIBUTING FACTOR VEHICLE 4 | Factors contributing to the collision for designated vehicle | Text |
| CONTRIBUTING FACTOR VEHICLE 5 | Factors contributing to the collision for designated vehicle | Text |
| COLLISION\_ID | Unique record code generated by system | Numeric |
| VEHICLE TYPE CODE 1 | Type of vehicle based on the selected vehicle category | Text |
| VEHICLE TYPE CODE 2 | Type of vehicle based on the selected vehicle category | Text |
| VEHICLE TYPE CODE 3 | Type of vehicle based on the selected vehicle category | Text |
| VEHICLE TYPE CODE 4 | Type of vehicle based on the selected vehicle category | Text |
| VEHICLE TYPE CODE 5 | Type of vehicle based on the selected vehicle category | Text |

Preprocessing Steps:

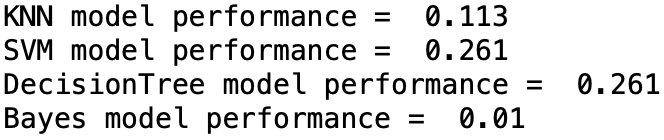
1. Remove all objects with null values.
2. Remove all redundancies.

**Design and Methods**

We will be using classification algorithms such as K-Nearest Neighbor, Support Vector Machines, Decision Trees, and Naive Bayes. Before selecting attributes, the data will be cleaned such as removing objects with null values and removing any redundancies. Certain attributes such as zip code, longitude, latitude, location, off street name, cross street name, on street name, contributing factor 2-5, and vehicle type code 2-5 because all these attributes have a significant amount of NaN values. Using NaN values will result in model errors. Remapping those values will lower the accuracy of the models as well. Also, the attributes crash time, crash date, and collision\_id will be removed as they are irrelavent. The features selection will only consist of the features pertaining to injury and deaths since we are trying to find a correlation between that data and the type of vehicle involved in the collision. The data will be split into training and testing sets. About 60% of the data will be used for training and the rest will be used for testing. To evaluate each model, we will calculate the accuracy score of each model and compare them.

**Results**

As a result of classifying the dataset, the models performed at a low level. The K-Nearest Neighbor model produced an accuracy score of 0.203. The Support Vector Machine model produced an accuracy score of 0.264. The Decision Tree model produced an accuracy score of 0.264. The Naive Bayes model produced an accuracy score of 0.001. It can be seen in the image below that the Support Vector Machine model and the Decision Tree model tied in best performance compared to the K-Nearest Neighbor model and the Naive Bayes model.

****

**Conclusions and Future Works**

In conclusion, the models underperformed with this dataset. The low accuracy scores indicate there may be little to no correlation between the selected features and the desired outcome. We believe the models produced low scores because there are large amounts 0 and NaN values in the dataset, which significantly lowered the accuracy of the models. Low accuracy scores for one dataset does not mean the models are unusable. For future work we would like to apply the same algorithms to a different dataset to see if it underperforms again or find a dataset similar to one used in this project that has more meaningful information.

**Team Member Contributions**

Preprocessing data: Emanuel

K-Nearest Neighbor & SVM Models: Emmanuel

Decision Tree & Naive Bayes Models: Emanuel Talmi

Paper: Emanuel Talmi, Emmanuel