The purpose of this project is to find out an analysis of the probability of death based on the perpetrator's demographics, the location, and the time of the shooting in New York city.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | New York City On The Cheap | | NYPD SHOOTING INCIDENT REPORT | |  |

NOVEMBER 29, 2022

Bayat Vahabodin

***Contents***

[Introduction 1](#_Toc120709649)

[Dataset Understanding 1](#_Toc120709650)

[Attributes types and descriptions 2](#_Toc120709651)

[Data Preparation 3](#_Toc120709652)

[Attribute values 6](#_Toc120709653)

[Modeling & Evaluation 9](#_Toc120709654)

[kNN Clustering Method 9](#_Toc120709655)

[Decision Tree Method 10](#_Toc120709656)

[Random Forest Method 11](#_Toc120709657)

[Discussion of Results 15](#_Toc120709658)

[Conclusion 15](#_Toc120709659)

***Figures***

[Figure 1: Dataset 4](#_Toc120709660)

[Figure 2: distribution of the class attribute 5](#_Toc120709661)

[Figure 3: confusion matrix with 10- fold cross validation 10](#_Toc120709662)

[Figure 4: Visualize decision tree 11](#_Toc120709663)

[Figure 5: Random Forest Method for STATISTICAL\_MURDER\_FLAG as the class 11](#_Toc120709664)

[Figure 6: Number of shootings per month 12](#_Toc120709665)

[Figure 7: Number of shootings per year 12](#_Toc120709666)

[Figure 8: Number of shootings per borough 13](#_Toc120709667)

[Figure 9: Number of shootings per borough and time category 13](#_Toc120709668)

[Figure 10: Number of shooting per borough and perpetrator age category 14](#_Toc120709669)

[Figure 11: Number of shootings per borough, perpetrator age group, victims age group and time category 14](#_Toc120709670)

***Tables***

[Table 1: Value of Month\_Category attribute 6](#_Toc120709687)

[Table 2: Value of Year attribute 6](#_Toc120709688)

[Table 3: Value of Time\_Category attribute 7](#_Toc120709689)

[Table 4: Value of Boro attribute 7](#_Toc120709690)

[Table 5: Value of STATISTICAL\_MURDER\_FLAG attribute 7](#_Toc120709691)

[Table 6: Value of PERP\_AGE\_GROUP attribute 7](#_Toc120709692)

[Table 7: Value of PERP\_SEX attribute 8](#_Toc120709693)

[Table 8: Value of PERP\_RACE attribute 8](#_Toc120709694)

[Table 9: Value of VIC\_AGE\_GROUP attribute 8](#_Toc120709695)

[Table 10: Value of VIC\_SEX attribute 8](#_Toc120709696)

[Table 11: Value of VIC\_RACE attribute 9](#_Toc120709697)

[Table 12: kNN clustering with 10-fold cross validation, Seed 1 9](#_Toc120709698)

[Table 13: kNN clustering with 10-fold cross validation, Seed 2 9](#_Toc120709699)

[Table 14: kNN clustering percentage split of 70%, Seed 1 10](#_Toc120709700)

[Table 15: kNN clustering percentage split of 70%, Seed 2 10](#_Toc120709701)

# Introduction

Since at least the 1800s, crime statistics have been kept track of in New York City. Since the post-World War II era, they have increased.

The NYPD keeps statistics that are utilized as a management tool to cut crime, enhance practices and training, and increase openness with the general public and governmental oversight bodies.

As the crack epidemic grew, crime rates peaked in the late 1980s and early 1990s. After that, they steadily decreased into the 2000s.

When the agency implemented CompStat in 1994, crime successfully decreased to historic lows not seen since the 1950s thanks to management, statistics, and accountability.

On the citywide, borough, and precinct levels, the department publishes current crime-related statistics in the seven primary crime categories, as well as historical crime data.

The New York City Police Department (NYPD) made a significant effort to lower crime throughout the 1990s by implementing CompStat, broken windows policing, and other tactics. Following, there was a decline in crime that has been variously attributed to the end of the crack epidemic, the rise in incarceration rates across the country, gentrification, an ageing population, and a decrease in lead poisoning in youngsters.

# Dataset Understanding

This dataset contains a list of every shooting incident that occurred in NYC going back to 2006 through the end of 2021.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of 2021. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location, and the time of occurrence. In addition, information related to suspect and victim demographics is also included.

The purpose of this exploratory data analysis is to explore the nature of the threat and uncover hidden insights for the detection of potential threat locations, demographics & ethnicity of the victim & the perpetrator, and estimation of the number of casualties.

## Attributes types and descriptions

Table 1: Attributes Description

|  |  |  |  |
| --- | --- | --- | --- |
| Row | Attribute | Type | Description |
| 1 | INCIDENT\_KEY | Ordinal | Randomly generated persistent ID for each arrest |
| 2 | OCCUR\_DATE | String | Exact date of the shooting incident  (mm/dd/yyyy) |
| 3 | OCCUR\_TIME | String | Exact time of the shooting incident  (hh:mm:ss) |
| 4 | BORO | nominal | Borough where the shooting incident occurred |
| 5 | PRECINCT | numeric | Precinct where the shooting incident occurred |
| 6 | JURISDICTION\_CODE | nominal | Jurisdiction where the shooting incident occurred. Jurisdiction codes  (0: Patrol, 1: Transit, 2: Housing) |
| 7 | LOCATION\_DESC | Nominal | Location of the shooting incident |
| 8 | STATISTICAL\_MURDER\_FLAG | Boolean | Shooting resulted in the victim’s death which would be counted as a murder  (False: Survived , True: Death) |
| 9 | PERP\_AGE\_GROUP | Nominal | Perpetrator’s age within a category  (<18, 18-24, 25-44, 45-65, 65+) |
| 10 | PERP\_SEX | Nominal | Perpetrator’s sex description  (M: Male , F: Female) |
| 11 | PERP\_RACE | nominal | Perpetrator’s race description |
| 12 | VIC\_AGE\_GROUP | nominal | Victim’s age within a category  (<18, 18-24, 25-44, 45-65, 65+) |
| 13 | VIC\_SEX | nominal | Victim’s sex description  (M: Male , F: Female) |
| 14 | VIC\_RACE | nominal | Victim’s race description |
| 15 | X\_COORD\_CD | string | Midblock X-coordinate for New York State Plane Coordinate System |
| 16 | Y\_COORD\_CD | string | Midblock Y-coordinate for New York State Plane Coordinate System |
| 17 | Latitude | string | Latitude coordinate for Global Coordinate System |
| 18 | Longitude | string | Longitude coordinate for Global Coordinate System |
| 19 | Lon\_Lat | point | Longitude and Latitude Coordinates for mapping |

# Data Preparation

First of all, we looked for duplicate data which there weren't any duplicates in our dataset.

Then, we checked the dataset and found some cells had no value and their number was high and also we could not remove them, we filled the cells with mode() replacement for each attribute.

Based on our goal, we found that the following attributes are not useful and have no effect on our prediction:

* INCIDENT\_KEY
* PRECINCT
* JURISDICTION\_CODE
* X\_COORD\_CD
* Y\_COORD\_CD
* Latitude
* Longitude
* Lon\_Lat

So we removed these columns.

In the following, we categorized PERP\_AGE\_GROUP and VIC\_AGE\_GROUP to the following age groups:

* Child for age <18
* Adult for 18-24 and 25-44
* Senior for 45-64 and 65+

Also we created Time\_Category for OCCUR\_TIME to the following time groups:

* Day for times between 6 am to 6 pm
* Night for time between 6 pm to 6 am

In addition, we split OCCUR\_DATE into 3 columns which are month, day and year and then removed the day and categorized the months (Month\_Category) into 3 groups:

* Cold Weather for November, December, January, February and March
* Nice Weather for April, may, June and October
* Hot Weather for July, August and September

Finally, changed the values in the STATISTICAL\_MURDER\_FLAG as the following:

* False to 0 which means survived
* True for 1 which means death

At the end, we saved the dataset as a CSV file named NYPD\_ Shooting\_ Incident\_ Data\_\_ Historic\_Processed.

Here is a screenshot of several first rows with their titles:

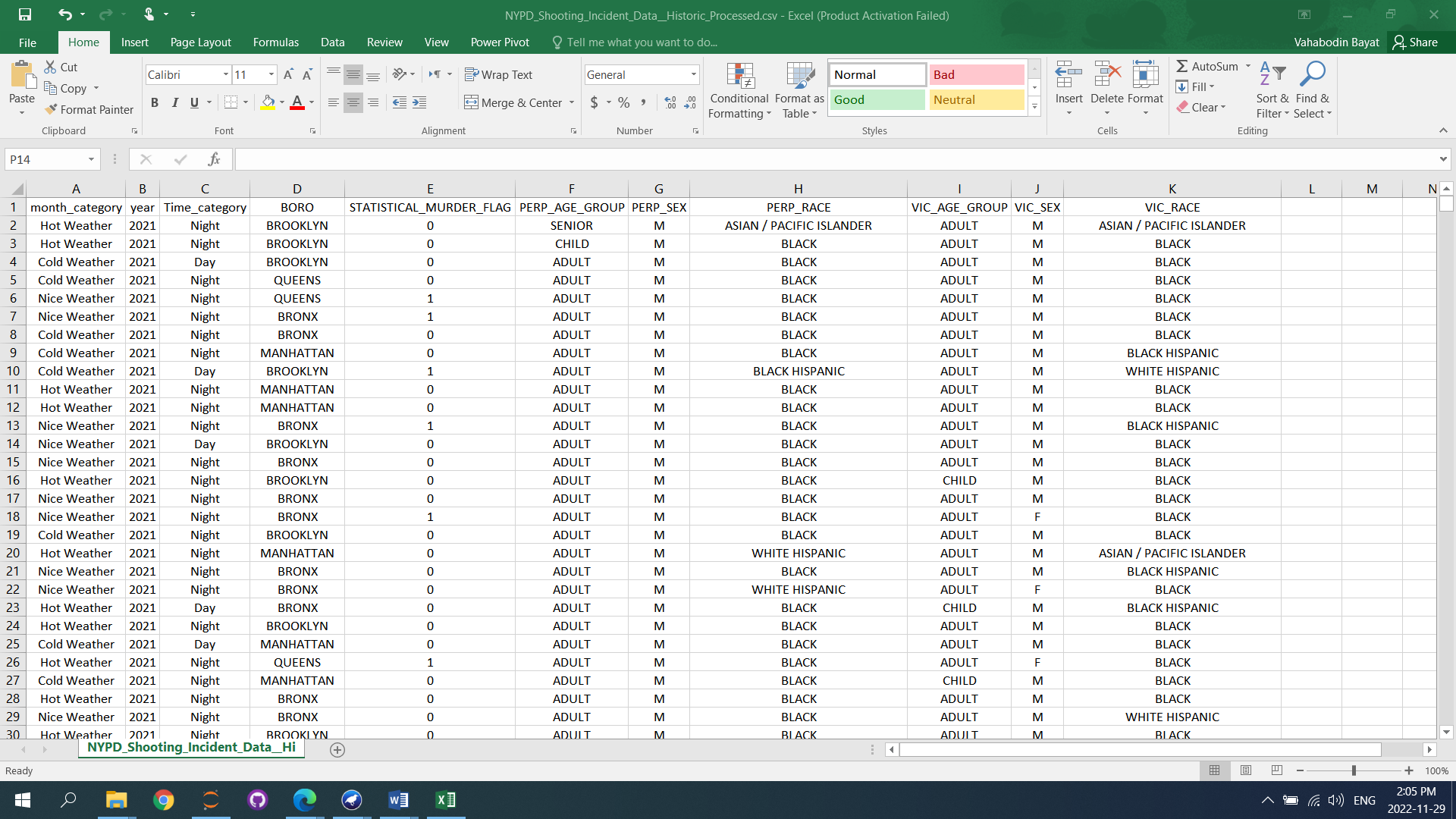


Figure : Dataset

In the next step we loaded the file (NYPD\_Shooting\_Incident\_Data\_\_Historic\_Processed.csv) in Weka for checking each attribute type and we found two attributes such as Year and STATISTICAL\_MURDER\_FLAG have to convert from Numeric to Nominal then we applied filter from the process tab>filters>unsupervised>attribute selected NumericToNominal to convert them and save it by name NYPD\_Shooting\_Incident\_Data\_\_Historic\_Converted.arff.

Here is a screenshot of distribution of the STATISTICAL\_MURDER\_FLAG attribute as our class:

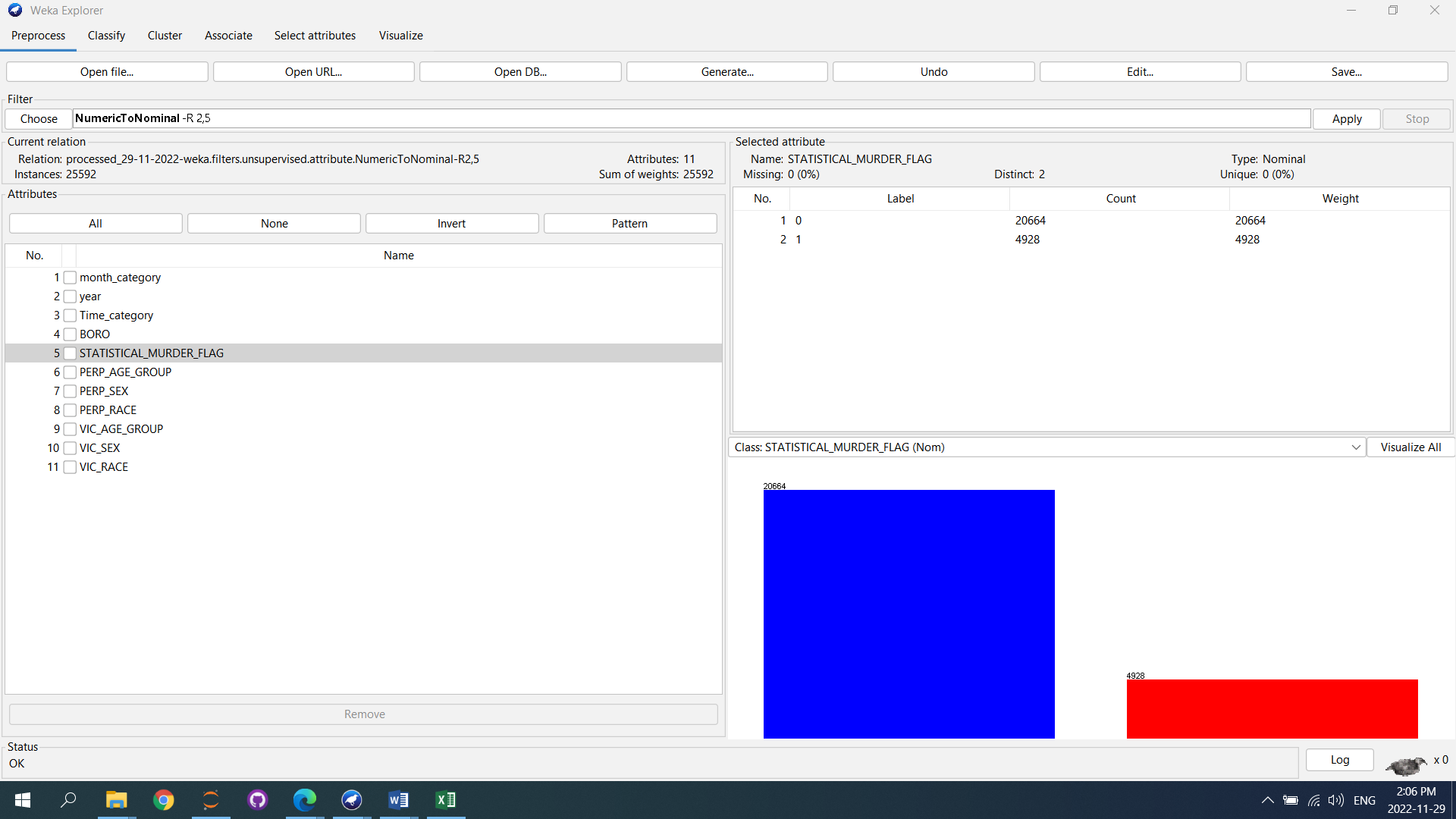


Figure : distribution of the class attribute

## Attribute values

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table : Value of Month\_Category attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | Hot Weather | 8427 | | 2 | Cold Weather | 10280 | | 3 | Nice Weather | 6885 | | Table : Value of Year attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | 2006 | 2055 | | 2 | 2007 | 1887 | | 3 | 2008 | 1959 | | 4 | 2009 | 1828 | | 5 | 2010 | 1911 | | 6 | 2011 | 1939 | | 7 | 2012 | 1717 | | 8 | 2013 | 1338 | | 9 | 2014 | 1464 | | 10 | 2015 | 1433 | | 11 | 2016 | 1208 | | 12 | 2017 | 970 | | 13 | 2018 | 958 | | 14 | 2019 | 967 | | 15 | 2020 | 1948 | | 16 | 2021 | 2010 | |
|  |  |
| Table : Value of Time\_Category attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | Night | 19491 | | 2 | Day | 6101 | | Table : Value of Boro attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | BROOKLYN | 10363 | | 2 | QUEENS | 3828 | | 3 | BRONX | 7400 | | 4 | MANHATTAN | 3265 | | 5 | STATEN ISLAND | 736 | |
|  |  |
| Table : Value of STATISTICAL\_MURDER\_FLAG attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | 0 (Survived) | 20664 | | 2 | 1 (Death) | 4928 | | Table : Value of PERP\_AGE\_GROUP attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | SENIOR | 592 | | 2 | CHILD | 1463 | | 3 | ADULT | 23537 | |
|  |  |
| Table : Value of PERP\_SEX attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | M | 25221 | | 2 | F | 371 | | Table : Value of PERP\_RACE attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | ASIAN / PACIFIC ISLANDER | 141 | | 2 | BLACK | 21812 | | 3 | BLACK HISPANIC | 1203 | | 4 | WHITE HISPANIC | 2162 | | 5 | WHITE | 272 | | 6 | AMERICAN INDIAN/ALASKAN NATIVE | 2 | |
|  |  |
| Table : Value of VIC\_AGE\_GROUP attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | ADULT | 21046 | | 2 | CHILD | 2681 | | 3 | SENIOR | 1865 | | Table : Value of VIC\_SEX attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | M | 23189 | | 2 | F | 2403 | |
|  | |
| Table : Value of VIC\_RACE attribute   |  |  |  | | --- | --- | --- | | Row | Label | Count | | 1 | ASIAN / PACIFIC ISLANDER | 354 | | 2 | BLACK | 18344 | | 3 | BLACK HISPANIC | 2485 | | 4 | WHITE HISPANIC | 3740 | | 5 | WHITE | 660 | | 6 | AMERICAN INDIAN / ALASKAN NATIVE | 9 | | |

# Modeling & Evaluation

## kNN Clustering Method

We ran our dataset in Weka with lazy.IBK classifier to take results from kNN with 10-fold cross-validation and a percentage split of 70%.

Our next step is to modify the k in the Nearest-Neighbors algorithm to check which condition is the best for our data and its class.

Performing of kNN clustering with 10-fold cross validation:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table : kNN clustering with 10-fold cross validation, Seed 1   |  |  | | --- | --- | | **K** | **Percentage of correctly classified instances** | | 3 | 79.57 % | | 5 | 80.18 % | | 7 | 80.42 % | | 9 | 80.56 % | | Table : kNN clustering with 10-fold cross validation, Seed 2   |  |  | | --- | --- | | **K** | **Percentage of correctly classified instances** | | 3 | 79.58 % | | 5 | 80.19 % | | 7 | 80.48 % | | 9 | 80.56 % | |

Performing of kNN classification percentage split of 70%:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table : kNN clustering percentage split of 70%, Seed 1   |  |  | | --- | --- | | **K** | **Percentage of correctly classified instances** | | 3 | 79.51 % | | 5 | 80.33 % | | 7 | 80.48 % | | 9 | 80.60 % | | Table : kNN clustering percentage split of 70%, Seed 2   |  |  | | --- | --- | | **K** | **Percentage of correctly classified instances** | | 3 | 79.77 % | | 5 | 80.20 % | | 7 | 80.47 % | | 9 | 80.47 % | |

As we can see in table 12 to 15 in 4 tables, the best conditions will occur when k is 9 for both 10-fold cross validation and percentage split of 70%.

## Decision Tree Method

In this method, we used j48 model with minimum number object 15 and True for unpruned option for decision tree, the results of which are set according to the following attributes selected such as Time\_category, BORO, STATISTICAL\_MURDER\_FLAG, PERP\_AGE\_GROUP, VIC\_AGE\_GROUP, VIC\_RACE.

|  |
| --- |
| Figure : confusion matrix with 10- fold cross validation |
|  |

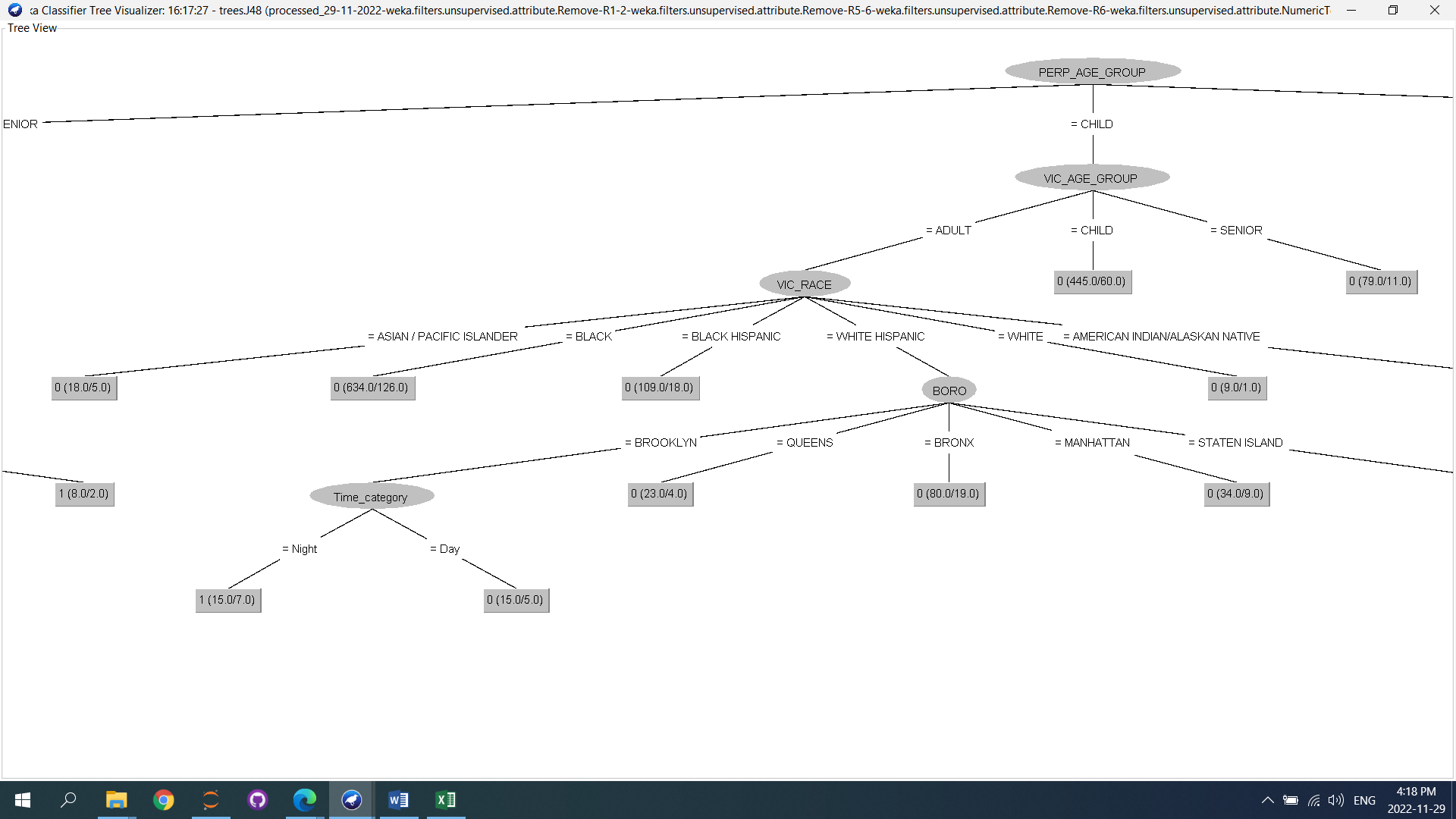


Figure : Visualize decision tree

From the obtained tree (DecisionTree.model) that we show a part of that which is a subset of the child group we can predict which people in which age group and under what conditions survived.

If the person who was shot was white Hispanic and in adult group, and the shooting happened in Brooklyn, if it happened during the day, it is likely that the person survived, but if it happened at night, they would not have survived.

## Random Forest Method

In order to measure accuracy, we used RandomForest with a split of 70.0% between the train and the test for our class attribute, as well as a maximum depth of 5 (RandomForest-S.M.F.model), and found that the accuracy of about 72.2% can be obtained if STATISTICAL\_MURDER\_FLAG is considered as the class. The output is as follows:

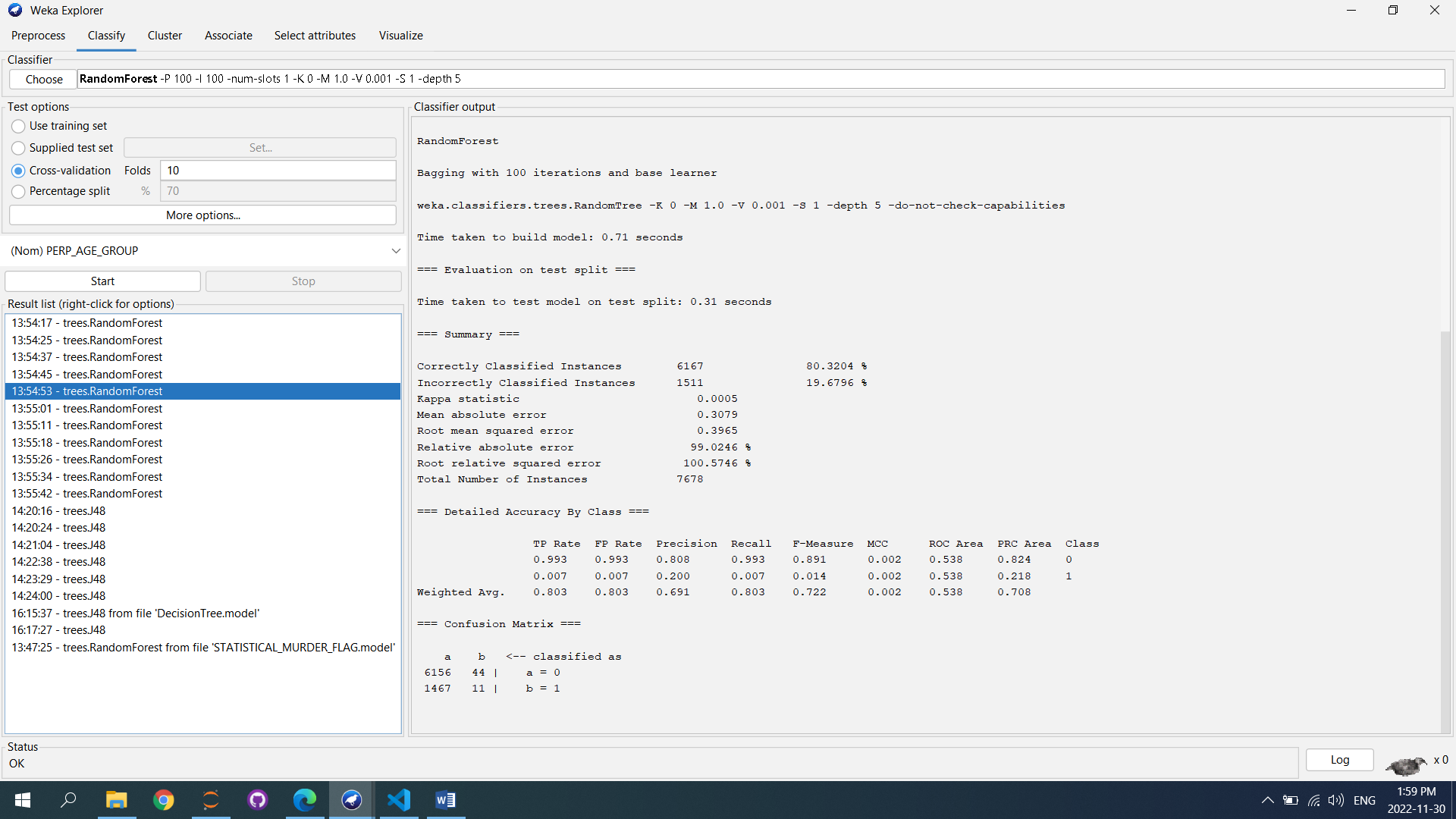


Figure : Random Forest Method for STATISTICAL\_MURDER\_FLAG as the class

In the following, to further examine the data, we plotted different graphs based on different criteria whith python (plot.ipynb), the most important of which are shown:

|  |
| --- |
| Figure : Number of shootings per month |

From the graph above, we can see that shooting has increased in the year's cold months.

|  |
| --- |
| Figure : Number of shootings per year |

Graph shows us 2006 had the most shooting incident and 2018 had the least.

|  |
| --- |
| Figure : Number of shootings per borough |

Brooklyn had the most with 10363 and Staten Island had the least with 736.

|  |
| --- |
| Figure : Number of shootings per borough and time category |

There were 7667 shootings in Brooklyn at night and 176 in Staten Island during the day.

Also we did an analysis of the most shooting incident perpetrators age category by borough.

|  |
| --- |
| Figure : Number of shooting per borough and perpetrator age category |

The results show us that the highest number of shootings by adults was in Brooklyn with 9,608 shootings and the lowest number of shootings by seniors was in STATEN ISLAND with 31 shootings.

In order to investigate more deeply, we increased the number of factors and obtained the following graph based on the borough, the perpetrator age category, victims age category, and also the shooting time.

|  |
| --- |
| Figure : Number of shootings per borough, perpetrator age group, victims age group and time category |

Brooklyn had the most shooting at night for Adult perpetrator and Adult Victim with 6096.

# Discussion of Results

As per our analysis in 2006 we had the most shooting incident and 2018 we had the least shooting incident.

The probability of shooting at night is higher, and if the shooting happened in the borough of Brooklyn and both the shooter and the victim were adult black man, it led to the death of the victim.

In total 20665 victim survived, while 4928 victim passed away.

# Conclusion

We discovered through exploratory data analysis that the majority of gunshot occurrences occur at night and in cold weather. Men make up the majority of both perpetrators and victims. The plots claim that Brooklyn is generally the most hazardous part of New York City.

Random forest performs the best among all algorithms in terms of predictive models. It may be deduced that ensemble algorithms outperform single algorithms on our dataset as the random forest model is an ensemble learning technique that combines a number of decision tree models to create a stronger learner.