SUP'COM
Higher School of Communication of Tunis

New York City Crimes Detection using Machine Learning

Realised by:

Asma Abidalli

Mariem Mezghani

Sarra Hammami

Wissal Weslati

Under the supervision of:

M. Riadh Tebourbi

Table of content

Table of figures

Abstract

Crimes pose a significant challenge worldwide, impacting the quality of life, economic growth, and a nation's reputation. With a noticeable surge in crime rates, there is an imperative need for sophisticated systems and innovative approaches to enhance crime analytics for community protection. Real-time crime prediction, though complex, proves crucial in mitigating crime rates. This project employs diverse visualization techniques and machine learning algorithms to forecast crime distribution across New York City.

The initial phase involved processing a raw dataset, employing multiple visualization techniques to gain deeper insights into the data and understand the relationships between different variables. Subsequently, various machine learning algorithms were applied to predict crime types based on user input, considering geographic locations of users. The final step encompassed the development of a user-friendly interface using Streamlit, enhancing user interaction.

The complete code for this project is available at: https://github.com/sara-hammami/NY_crime_prediction.git

Keywords: Crime Analysis; Crime Prediction; Data Visualization; Crime Maps; Machine Learning; Streamlit

I. Introduction:

Generally, crimes are rather common social issues, influencing a country's reputation, economic growth, and quality of life. They are perhaps a prime factor in influencing several critical decisions in a person's life, such as avoiding dangerous areas, visiting at the right time, and moving to a new place[1]. Crimes define and affect the impact and reputation of a community while placing a rather large financial burden on a country due to the need for courts and additional police forces. With an increase in crimes, there is an increased need to reduce them systematically. In recent times, there has been a record increase in crime rates throughout the world. It is possible to reduce these figures by analyzing and predicting crime occurrences. In such a situation, preventive measures can be taken quickly. Crime forecasting in real-time is capable of helping save lives and prevent crimes, gradually decreasing the crime rate. With a comprehensive crime data analysis and modern techniques, crimes can be predicted and support can be deployed without delay. Machine learning has acquired significant attention in the past few years due to its potential and implications. It has begun to be tested even in forecasting and predicting crime rates and previous studies have attested to its potential. In this study, a comparative analysis is carried out of previous studies on crime prediction through machine learning to identify the current techniques and schemes being used for crime predictions.

II. Related work:

Analyzing and predicting crime is an important activity that can be optimized using various techniques and processes. A lot of research work is done by multiple researchers in this domain, but most of the existing work only focuses on finding where crimes happen using data. They usually ignore important details like the type of crime and when it happened. For instance, Yu, R et al. made maps, but they were not interactive. To fix these issues, our new method uses visualization techniques to show where specific types of crimes are more likely to happen..

Some papers talked about using decision trees[2] to predict crime, like works by Ahishakiye et al. and Iqbal et al. They used data like the country's population, average income, percentage of unemployed people over 16, and the type of crime. But their predictions only said if an area would likely have a high, medium, or low percentage of violent crimes in the future. They didn't predict the specific type of crime. Nasridinov et al. also had a method to say if the crime rate is high, medium, or low, but none of these methods said what type of crime might happen and how likely it was. Our method aims to improve this by giving a more detailed prediction about the type of crime.

III. Dataset

This work relies on NYPD Complaint Data Historic dataset [3]. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2019. The dataset contains 6901167 complaint and 35 columns including spatial and temporal information about crime occurrences along with their description and penal classification.

IV. PROPOSED METHODOLOGY

1. Data Pre-Processing

The dataset comprised many invalid data, negative values, and a lot of data was missing. For illustration, DateTime data included many negative values along with class indicator keys for crimes were negative, etc. In the beginning, we check the percentage of missing values for all features. Features that have higher than 10% Not a Number (NaN) values were removed from the dataset to get rid of those rows containing NaN data. Moreover, numerical data replaced by NaN values such as invalid values. Invalid values include, for example, negative or infeasible values. Additionally, Invalid age ranges for suspect and victim were also replaced by NaN values

2. Exploratory data analysis

In the initial stages of our exploratory data analysis, we delve into three crucial aspects of crime in New York City. First, we examine the distribution of crime types, providing a comprehensive overview of prevalent offenses across different categories. Simultaneously, we explore the success rates of various crimes, offering insights into law enforcement and community response effectiveness. Additionally, our analysis focuses on identifying and ranking the Top 10 common crimes in NYC, enabling a prioritized understanding of prevalent offenses and guiding targeted interventions. These initial insights pave the way for a more nuanced understanding of crime dynamics, setting the stage for further analysis and predictive modeling.

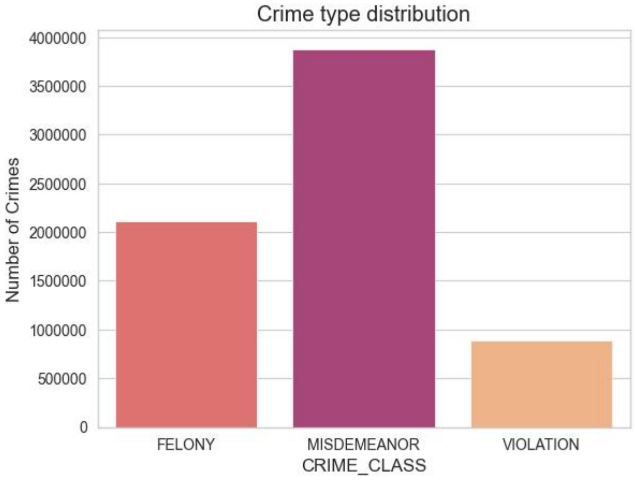


Figure 1: Crime type distribution

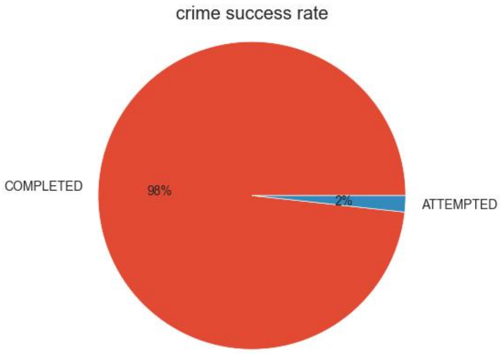


Figure 2: Crime success rate

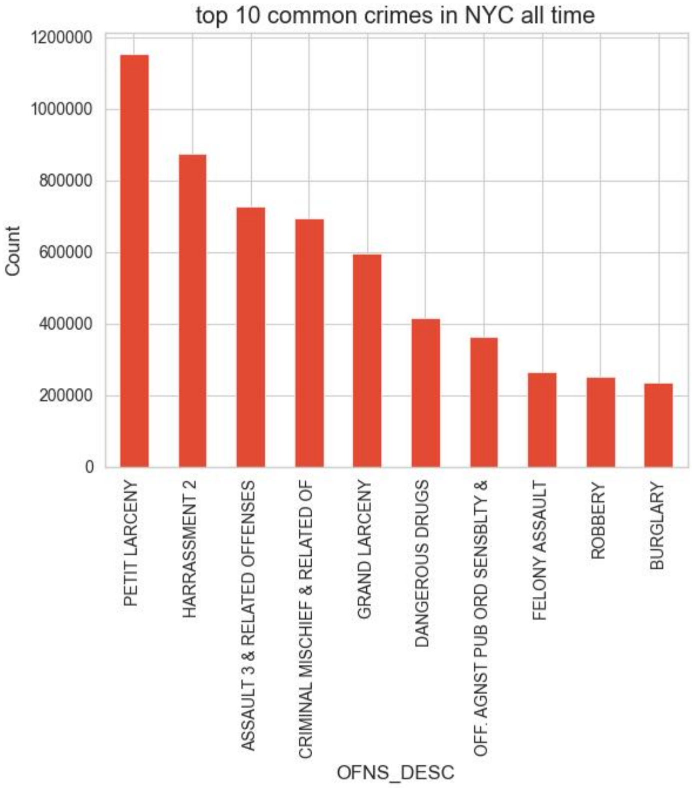


Figure 3: Top 10 common crimes in NYC

In the subsequent phase of our exploratory data analysis, our attention turns to discerning the demographic patterns of crime victims in New York City. We aim to unravel insights related to the distribution of crime victims based on their sex, age, and race. Examining the sex distribution provides a lens into potential gender-based victimization trends, aiding in the identification of any disparities between genders. Understanding the age distribution allows us to pinpoint age-specific vulnerabilities, facilitating the tailoring of preventative measures and support services. Additionally, analyzing the race distribution sheds light on potential racial disparities in victimization, enhancing our understanding of the social dynamics influencing crime patterns. By delving into these demographic dimensions, we seek to unveil nuanced insights into the varied experiences of crime victims in NYC, guiding the development of targeted interventions and policies to address the distinct needs of different demographic groups impacted by crime.

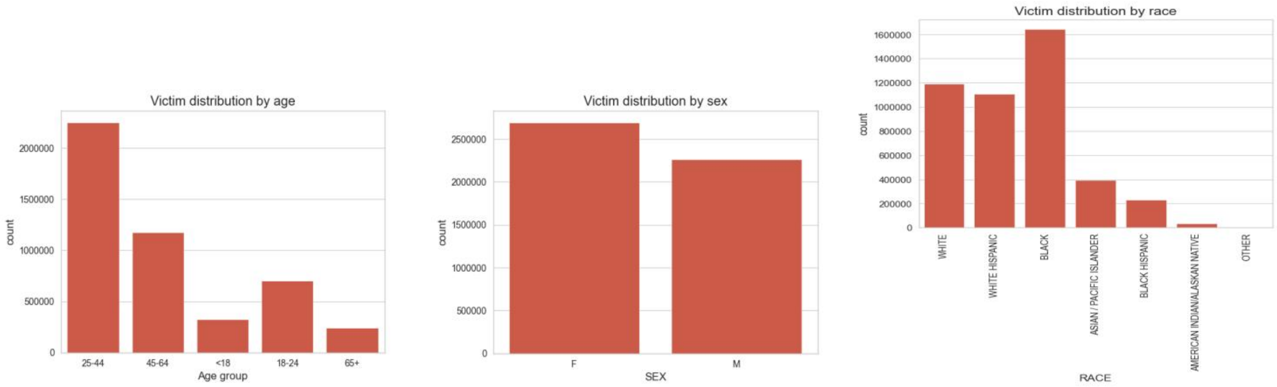


Figure 4: Crime distribution by age, sex and race

In our analysis, we'll examine the distribution of crimes across the boroughs of New York City. This localized perspective will reveal variations in crime frequency and types, aiding in the identification of borough-specific trends. The insights gained will be crucial for tailoring effective crime prevention strategies and resource allocation, ensuring a targeted and responsive approach to public safety in each administrative division. [7]

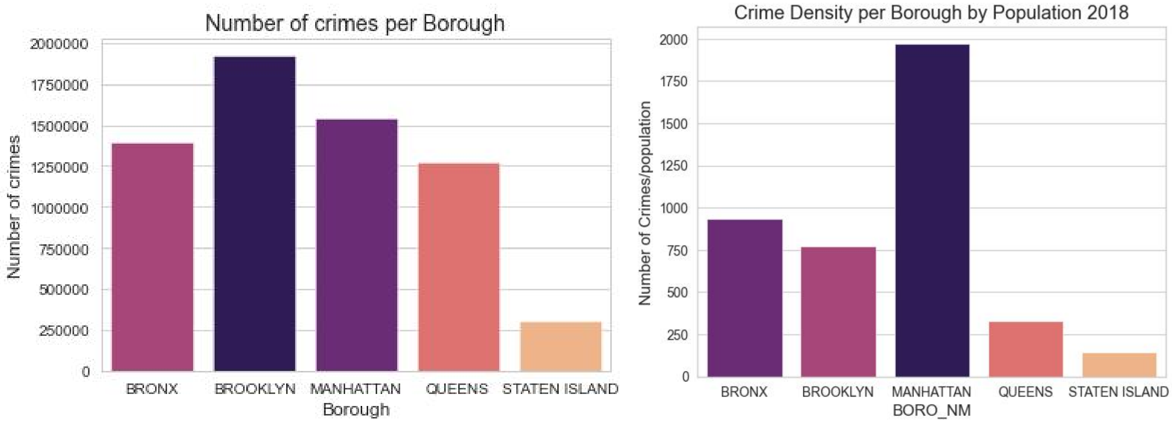


Figure 5:Crime distribution per borough

In our analysis, we examined the distribution of crimes over time, exploring patterns on both a yearly and monthly basis. This allowed us to identify long-term trends and monthly variations in criminal incidents. These insights are essential for tailoring interventions and policies that align with the evolving nature of crime in New York City.

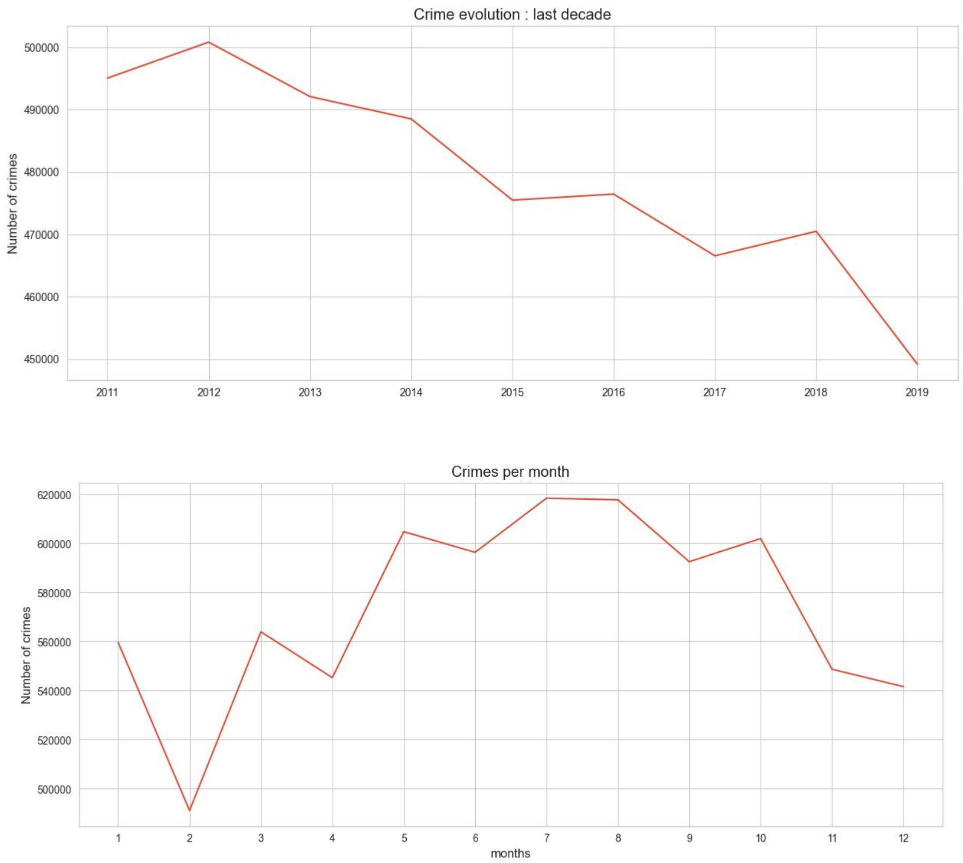


Figure 6: Crime distribution per year and per month

3. Model Building:

In order to predict the most occurrence criminal locations in a specific time in New York City neighbourhoods. Therefore, after the data pre-processing, in order to classify different types of crimes several machine learning algorithms were applied in order to compare their results including, Support Vector Machine (SVM), Random Forest (RF), and XGboost classifiers. The use of the classification is mainly to recognize the labeled classes by knowing a set of their features in the dataset, thus to predict the class label for instances with known features. Hence, the main intention of using the classifiers in crime prediction is to construct a future-oriented model to identify the criminal type location within a specific time.

3.1. LightGBM:

LightGBM is recognized for its efficiency, particularly in scenarios with large datasets, such as predicting crime patterns. This gradient boosting framework is designed for speed, employing gradient-based learning strategies that contribute to faster convergence during training. One of its significant advantages is low memory usage, making it suitable for situations with limited memory resources. While LightGBM excels in terms of efficiency, it might require parameter tuning to achieve optimal results. Its efficiency, however, comes at a trade-off with interpretability, as the model's complexity may reduce its transparency compared to simpler algorithms.

3.2. Random Forest:

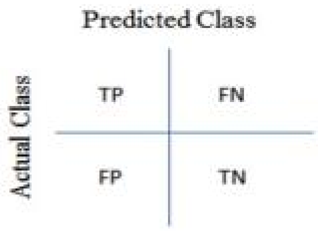
Random Forest[4], another widely used ensemble learning algorithm, excels in predicting crime patterns with its robustness to overfitting and adept handling of noisy data. This algorithm constructs an ensemble of decision trees, each trained on a different subset of the data. Random Forest's strength lies in providing insights into feature importance, enabling a better understanding of the factors influencing predictions. It is also highly parallelizable, making it efficient for processing large datasets. However, in terms of overall model interpretability, Random Forest may have limitations compared to simpler models.

3.3. XGboost:

In the realm of predicting crime patterns, XGBoost[5] stands out as a powerful ensemble learning algorithm known for its exceptional predictive accuracy. It operates by creating a collection of decision trees and iteratively refining them to enhance overall model performance. One of XGBoost's notable strengths is its regularization techniques, which prevent overfitting and make the model robust against noise in the data. Additionally, its flexibility in handling various data types and the ability to customize the objective function contribute to its effectiveness in crime prediction. However, it's important to note that training XGBoost models can be computationally intensive, particularly with large datasets.

V. Results

The classification task was done using three classifiers. For classification tasks, the confusion matrix is an appropriate metric to measure model performance. [6]



	Predicted Class	
Actual Class	TP	FN
	FP	TN

- True Positive (TP),True Negative (TN): represents the number of the prediction, correctly predicts as a given class.
- False Positive (FP), False Negative (FN): represents the number of the prediction, falsely predicts as a given class.

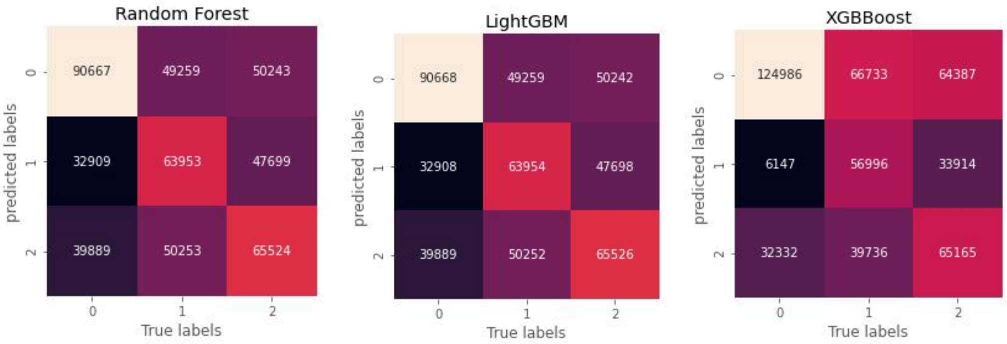


Figure 7: Confusion matrix results

Hence, the f1 score was measured by calculating precision and recall values. Furthermore, the comparison was done among the three models, where accuracy was also measured for each of the models. Formulas for the confusion matrix are shown below:

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ Positive\ Predictions}$$

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ Total\ Positive}$$

$$f1score = \frac{2 * precision * recall}{precision + recall}$$

$$Accuracy = \frac{Number\ of\ True\ Positives}{Total\ Samples}$$

The table below presents the results after training the 3 models:

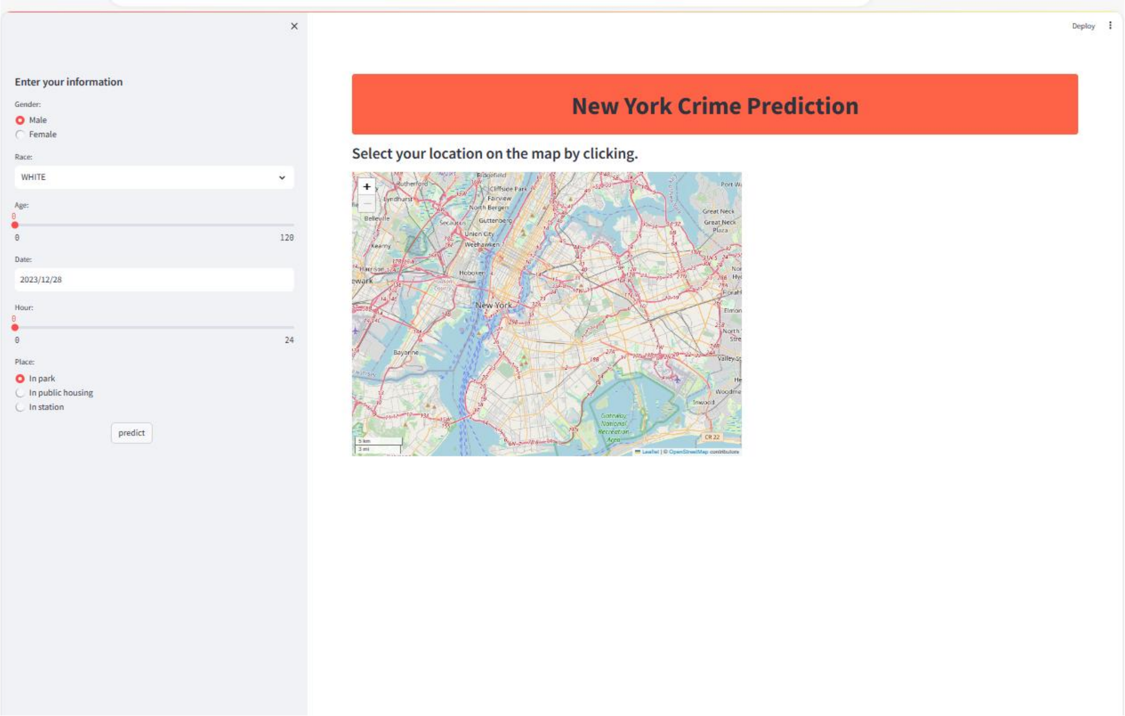
	Precision	Recall	F1-score
Random Forest	0.52	0.51	0.49
LightGBM	0.59	0.58	0.57
XGBoost	0.62	0.60	0.59

Table 1: Classification Results

Hence, for the classifier scores using the confusion matrix. However, LightGBM and XGBoost models tend to have very close validation scores which are respectively 0.59 and 0.6, whereas is relatively low for the Random Forest model with accuracy scores of 0.52. However, XGBoost model tend to be the best classification model for correctly predicting the crime classes with accuracy scores of 0.6040. Furthermore, from the confusion matrix heatmaps we have a multi-class classification task with 3 classes from 0 to 2 where it represents class labels that replaced actual class names. From what we can see from the confusion matrix, We can find that XGBoost model outperforms the other classifiers with the highest predictions in each class. Yet, they all tend to have a low numbers of true classifications for each class.

VI. User Interface

After training the model and saving the weights file, we built a web application using Streamlit and Folium to allow the user to interact with the map and predict the type of crime that could happen. The user can enter his gender, race, age, the date and hour in which he wants to predict the type of crime, the location on the map and finally, the place. This information is then transformed to fit the model input, and then, using the loaded model weights file, we predict the type of crime and send it back to the user along with the potential subtypes of that crime.



Enter your information

Gender:
● Male
○ Female

Race: WHITE

Age: 0 (0 – 120)

Date: 2023/12/28

Hour: 0 (0 – 24)

Place:
● In park
○ In public housing
○ In station

New York Crime Prediction

Select your location on the map by clicking.

Figure 8: User interface

Conclusion

Crimes have been a major problem in many cities, hence lots of researchers tried to solve it and predict the most criminal hot-spots in order to increase the understanding of dangerous places at certain times. In this paper, we have analysed the data of New York city in order to recognize the Spatio-temporal patterns for criminal incidents. we proposed a methodology to classify and predict crimes type by classifying the spatial-temporal locations using three machine learning algorithms; Random Forest, LightGBM, and XGBoost classifiers. As a result, LightGBM and XGboost classifiers were very close in prediction accuracy for 0.59 and 0.6, respectively. However, they fail in getting decent predictions in general. Finally, We have created a user interface to enable users to enter their information and get the class of the crime that can happen in a particular location at a specific time.

As future work, for better classification, it seemed that Deep Learning like Deep Artificial Neural networks or Deep Auto-Encoders might be used. Deep Learning has been in driving condition over traditional ML models in recent past years. Thus, it has good potential for better classification performance in predicting criminal types and hotspots.

References

[1] ToppiReddy, H. Saini, B. & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. Procedia Computer Science, 132, 696-705. https://doi.org/10.1016/j.procs.2018.05.075

[2] Ahishakiye, E., Taremwa, D., Omulo, E. O., Nairobi-Kenya, G. P. O., & Niyonzima, I. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. analysis, 6(03)

[3] New York Police Department (NYPD). Nypd complaint data historic. https://data. cityofnewyork.us/Public-Safety/ NYPD-Complaint-Data-Historic/qgea-i56i, 2016.

[4] Mariana Belgiu and Lucian Dragut¸. Random forest in remote sensing: A ˘ review of applications and future directions. volume 114, pages 24–31. Elsevier, 2016.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[6] Amit Gupta, Azeem Mohammad, Ali Syed, and Malka N Halgamuge. A comparative study of classification algorithms using data mining: crime and accidents in denver city the usa. Education, 7(7):374–381, 2016.

[7] In Kwon Choi. Geo-temporal visualization for tourism data using color curves. 2019.