Predicting Characteristics of Suspects - NYPD Data

Shreya Valish Koushik(50495153) , Vidush Bhardwaj(50495155), Aditi Soni(50495122), Disha Bhavesh Maru(50495123), Arjun Srinivasan(50495151)

ABSTRACT

Using a large database, this research study seeks to forecast the demographic characteristics of suspects involved in crime scenes, such as age, race, and gender. The main goals are to increase public safety, make it easier for victims to receive justice, support law enforcement agencies' investigations, and ultimately lower crime rates. In order to identify individuals and follow up on leads, the NYPD uses a variety of current resources, including surveillance cameras, informants, forensic evidence, investigation tactics, and community involvement. This research uses decision trees and random forests as prediction models to produce precise forecasts. The anticipated impacts include faster suspect identification, increased public safety, optimized resource allocation, reduced crime rates, and improved accuracy while minimizing bias. Future endeavors include incorporating additional data sources, refining data preprocessing and feature engineering techniques, ensuring continuous model updates, and collaborating with domain experts. The expansion of this project to other cities hinges on the successful performance of the prediction models and their associated benefits.

INTRODUCTION

In today's rapidly changing world, ensuring public safety and delivering justice to crime victims are paramount concerns. Law enforcement agencies are constantly seeking innovative approaches to aid their investigations and decrease crime rates.Profiling suspects who were present at crime scenes is an important component of criminal investigations. Law enforcement agencies can considerably benefit from the use of suspect demographic factors, such as age, race, and gender, to make predictions based on the information at hand.

The goal of this analysis is to create a prediction model for estimating suspect demographics at crime scenes using data analysis and machine learning approaches. By analyzing various factors such as victim age, crime timing, victim gender, crime location, and crime type, the model aims to provide valuable insights to law enforcement agencies.Existing law enforcement tools, including surveillance cameras, informants, forensic evidence, interviews, and community involvement, have proven somewhat helpful. However, leveraging the vast amount of available data can further enhance the accuracy and efficiency of suspect identification.Two algorithms—Decision Tree and Random Forest—were taken into consideration to meet the research goals. These methods were selected because they can work with non-linear, unbalanced datasets. A number of criteria, such as accuracy, robustness, bias, variance, sensitivity, specificity, and precision, were used to assess the performance of the models.

The paper examines the performance of various methods and offers the outcomes of their application. Additionally, we discuss the interpretability, computational efficiency, and advantages of these algorithms. We also address challenges related to data availability and quality, feature engineering, bias and fairness considerations, model complexity, and deployment and monitoring.

The impact of this research can be significant, as it enables law enforcement agencies to identify suspects more quickly, improve public safety, allocate resources more effectively, and ultimately reduce crime rates. By leveraging the insights gained from the predictive model, law enforcement agencies can enhance their investigative efforts and deliver justice to victims. However, further refinement and improvement of the model are necessary to ensure its reliability and accuracy.

METHODOLOGY

Research Design:

This study uses a quantitative research approach to create a model that predicts the characteristics of suspects at crime scenes. Data analysis and machine learning methods, notably the decision tree and random forest algorithms, are used in the research design.

Data Collection:

The New York Police Department's (NYPD) database provided the information used in this study. The database contains details on previous criminal activity, including elements like the age of the victim, the time of the offense, the victim's gender, the location of the crime, and the type of crime. The information is gathered from a variety of sources, including forensic evidence records, witness accounts, and crime reports.

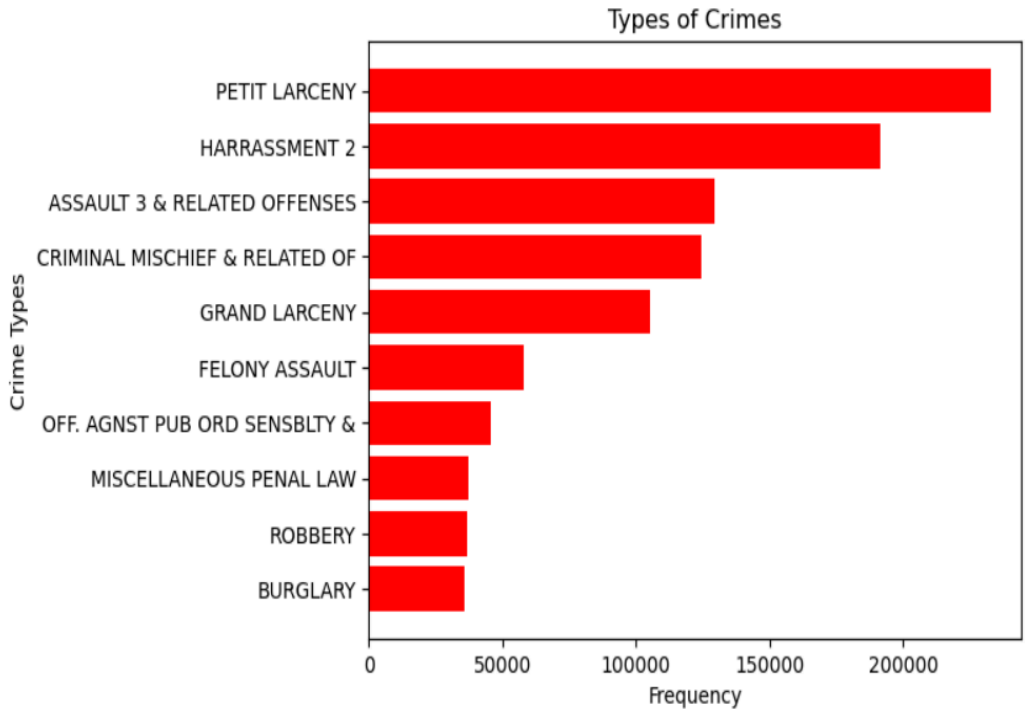
EDA:

Step 1: Data Exploration and Preprocessing

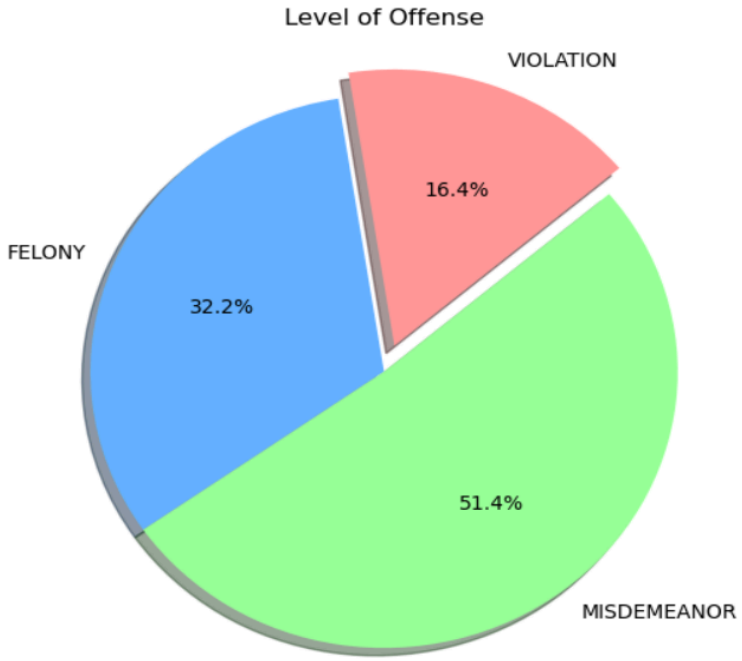
1. Loaded the dataset from "NYPD_Complaint_Data_Historic-5.csv".
2. Examined the structure of the dataset, including the columns and missing values
3. Performed data cleaning by handling missing values, such as imputing them with appropriate methods or removing rows with missing values, depending on the amount and relevance of missing data.
4. Transformed categorical variables, like “SUSP_RACE”, “SUSP_SEX”, “PREM_TYP_DESC” etc into numerical representations suitable for modeling

Strep 2: Data analysis

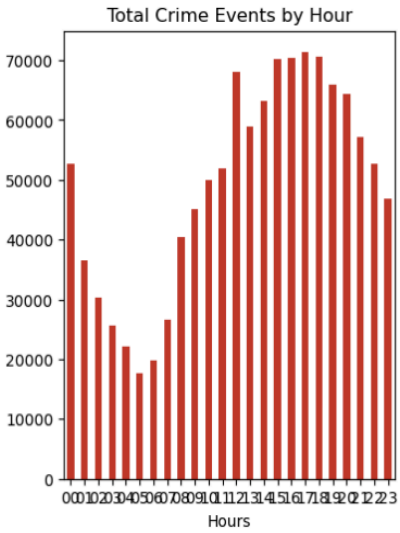
1. Types of crime

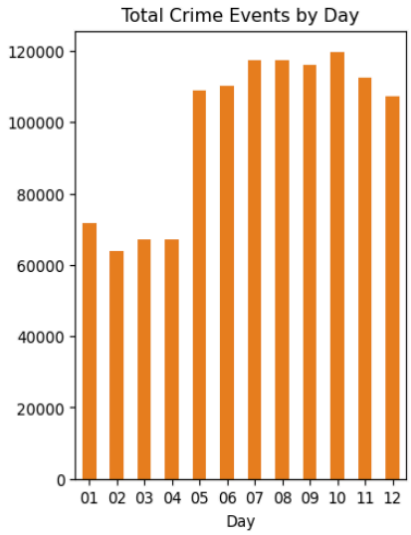


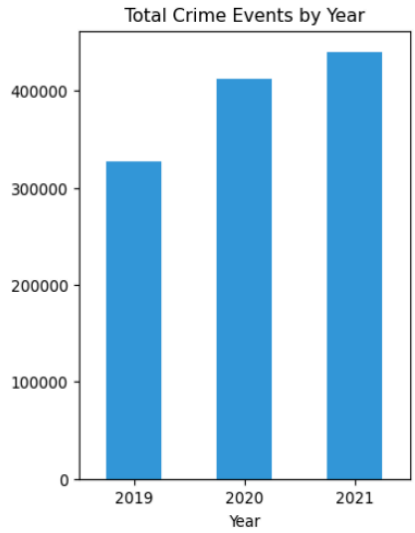
2. Level of offense



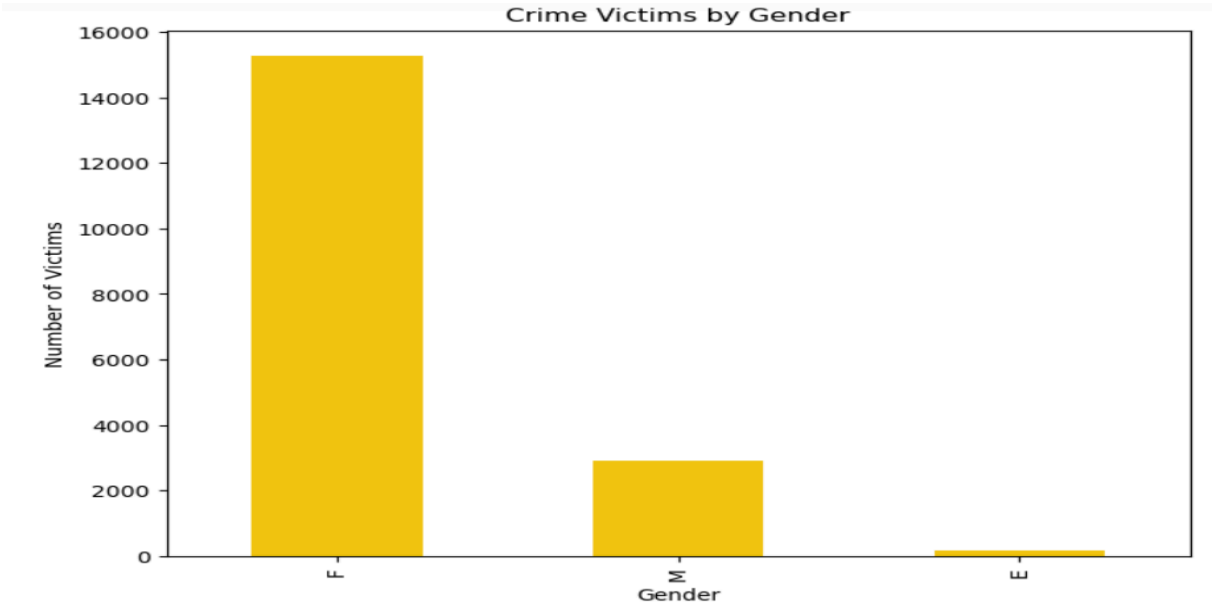
3. Total crime Events by Day, Month, Year



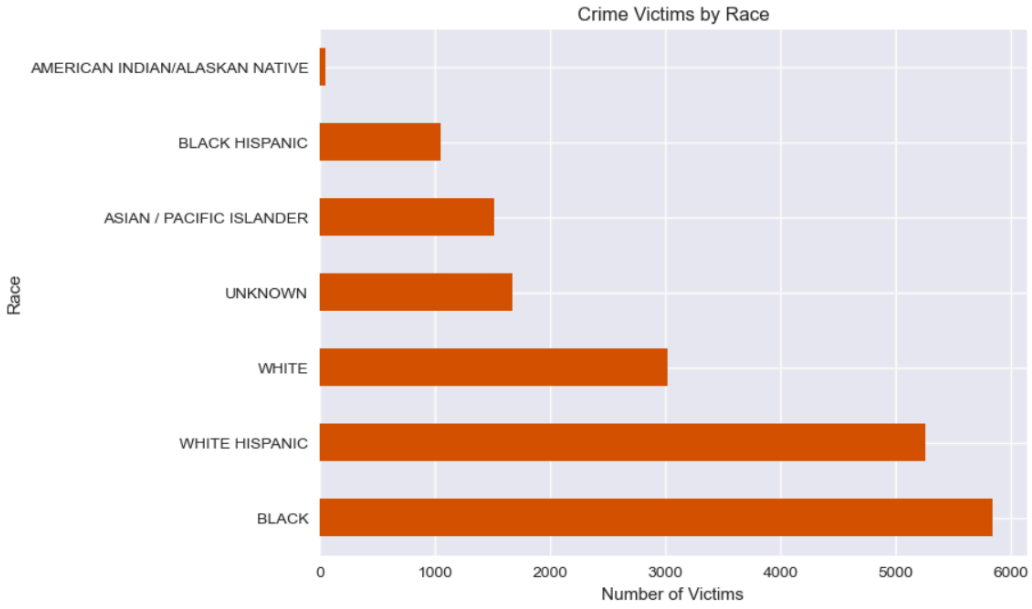




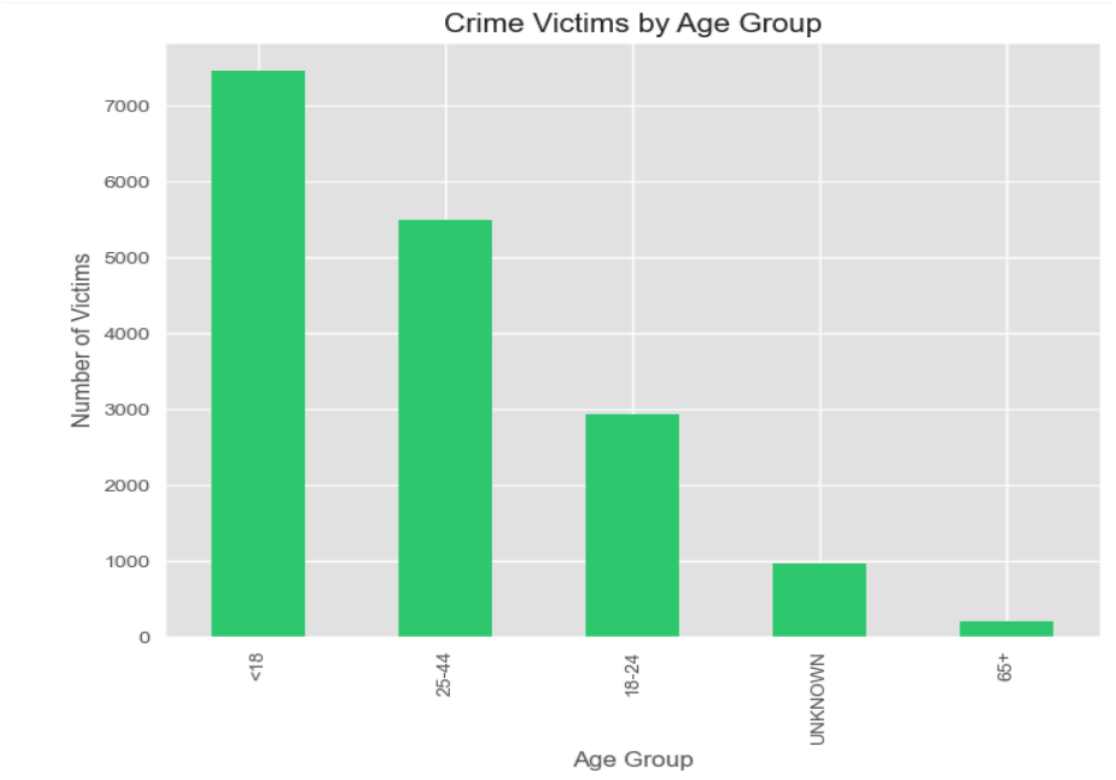
4. Crime by gender



5. Crime by race



6. Crime by Age group

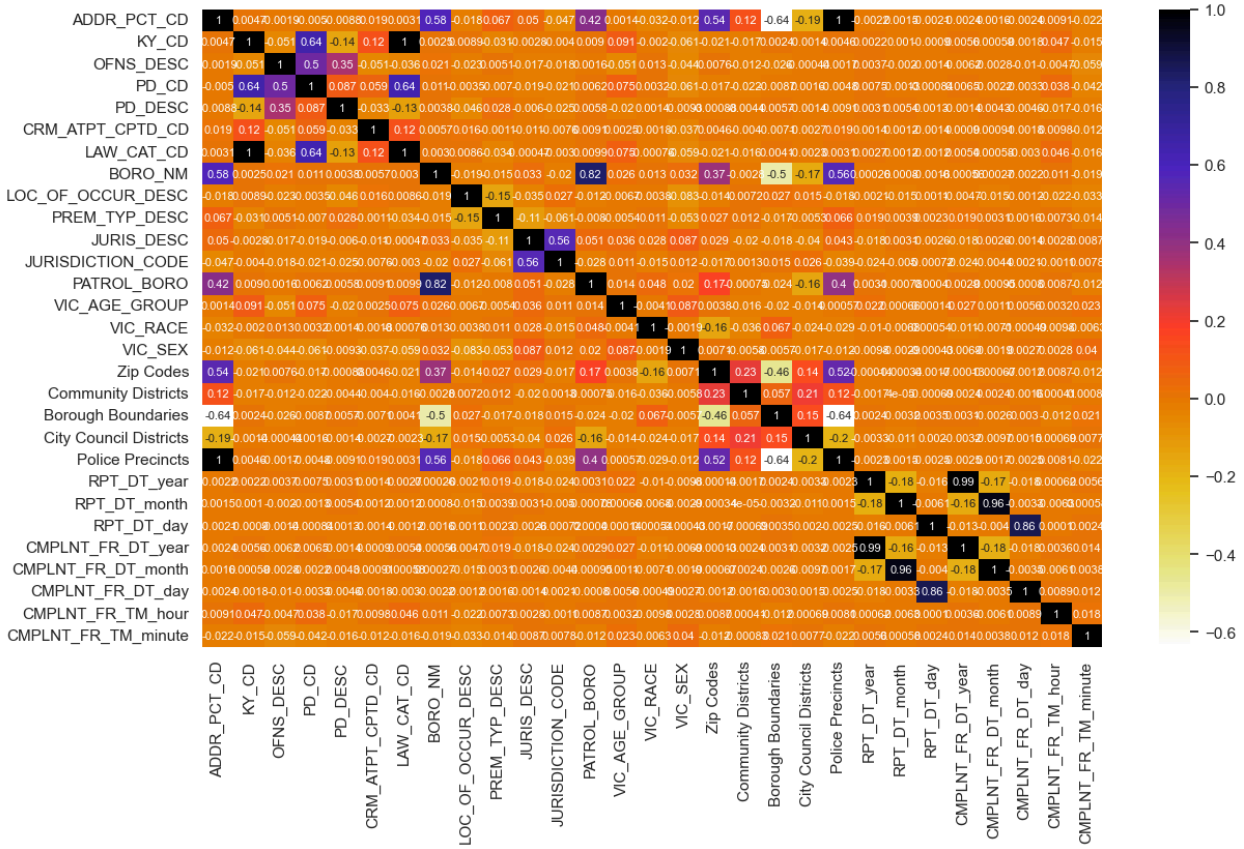


Step 3: Feature Engineering

1. Removed descriptor and ID columns such as
    1.Precinct Code was kept and descriptor were removed
    2. CMPLNT_NUM was removed

2. Columns which had alot of Nan values were removed

3. Used Correlation matrix and Heatmap for removing correlated columns.
Firstly, calculated the correlation matrix of all the features to know which features are highly related with one another. Followed by a plot of all the feature correlation matrix using a Heatmap. The heatmap is as follows :



Variables and Measures:

The key variables of interest in this study are the demographic characteristics of suspects, namely age, race, and gender. These variables serve as the target variables to be predicted by the predictive model. The predictor variables include victim age, time of crime, gender of the victim, premises of the crime, and the type of crime. These variables are selected based on their potential influence on suspect demographics.

Sample Size and Selection Criteria:

The data from the NYPD database are used to calculate the sample size for this study. The dataset contains a sizable number of crime episodes, giving analysts access to a wide variety of cases. The availability of complete and pertinent data for the predictor and target variables forms the basis of the selection criteria for instances to be included in the analysis.

Statistical and Analytical Techniques:
Decision Tree and Random Forest are the two primary machine learning algorithms used in the study. By inferring decision rules from the predictor variables, the decision tree algorithm forecasts the target variables. The majority voting process used by the random forest algorithm creates many decision trees and combines their forecasts. The right tools and programming languages, such Python's scikit-learn, are used to build these algorithms.

Several measures, including accuracy, robustness, bias, variance, sensitivity, specificity, and precision, are used to assess the performance of the predictive models. These measures shed light on how well and how consistently the models forecast questionable demographics. In order to evaluate the models' performance in terms of true positive, true negative, false positive, and false negative predictions, a confusion matrix is also used.

In order to construct and evaluate the predictive model for suspect demographics at crime scenes, this methodology combines data analysis, machine learning techniques, and statistical evaluation. The chosen factors, metrics, and analytical approaches are intended to produce precise and trustworthy forecasts to aid law enforcement organizations in their inquiry-related activities.

RESULTS
The results of the study are presented below, highlighting the performance of the decision tree and random forest algorithms in predicting suspect demographics.

Table 1: Performance Metrics of Decision Tree and Random Forest Algorithms
The results of the study are presented below, highlighting the performance of the decision tree and random forest algorithms in predicting suspect demographics.

Table 1: Performance Metrics of Decision Tree and Random Forest Algorithms

PARAMETERS	DECISION TREE	RANDOM FOREST
ACCURACY	58.08%	58.75%
BIAS	0.2%	0.1%
VARIANCE	0.5%	8.8%
PRECISION	82.19%	79.23%
SENSITIVITY/RECALL	85.94%	85.52%
SPECIFICITY	82.19%	72.23%
ROBUSTNESS	57%	58%
SAMPLE CONFUSION MATRIX FOR SUSPECT'S AGE	Confusion Matrix: [[651 141 1964 68 0] [274 1675 8445 242 1] [273 1604 34796 1162 5] [91 216 11602 1225 14] [6 23 1350 207 28]]	Confusion Matrix: [[538 141 2098 47 0] [161 951 9432 93 0] [107 711 36500 522 0] [30 77 12215 826 0] [2 4 1419 187 2]]

The accuracy of the decision tree algorithm is measured at 58.08%, while the random forest algorithm achieves a slightly higher accuracy of 58.75%. Both algorithms exhibit low bias, with values of 0.2% and 0.1% for the decision tree and random forest respectively. However, the random forest algorithm has a higher variance of 8.8% compared to the decision tree's 0.5%.

In terms of precision, the decision tree algorithm achieves 82.19%, while the random forest algorithm achieves a slightly lower precision of 79.23%. The sensitivity or recall, which represents the true positive rate, is high for both algorithms, with values of 85.94% for the decision tree and 85.52% for the random forest. The decision tree algorithm also exhibits a higher specificity of 82.19% compared to the random forest's 72.23%.

The robustness, which represents the overall performance of the models, is 57% for the decision tree and 58% for the random forest. This indicates that both models have room for improvement in terms of their robustness.

The results indicate that both decision tree and random forest algorithms provide reasonable accuracy in predicting suspect demographics. However, the random forest algorithm shows slightly better performance in terms of accuracy, precision, and sensitivity. It is important to note that further analysis and evaluation are required to fine-tune and improve the models for more reliable predictions.

The accuracy of the decision tree algorithm is measured at 58.08%, while the random forest algorithm achieves a slightly higher accuracy of 58.75%. Both algorithms exhibit low bias, with values of 0.2% and

0.1% for the decision tree and random forest respectively. However, the random forest algorithm has a higher variance of 8.8% compared to the decision tree's 0.5%.

In terms of precision, the decision tree algorithm achieves 82.19%, while the random forest algorithm achieves a slightly lower precision of 79.23%. The sensitivity or recall, which represents the true positive rate, is high for both algorithms, with values of 85.94% for the decision tree and 85.52% for the random forest. The decision tree algorithm also exhibits a higher specificity of 82.19% compared to the random forest's 72.23%.

The robustness, which represents the overall performance of the models, is 57% for the decision tree and 58% for the random forest. This indicates that both models have room for improvement in terms of their robustness.

The results indicate that both decision tree and random forest algorithms provide reasonable accuracy in predicting suspect demographics. However, the random forest algorithm shows slightly better performance in terms of accuracy, precision, and sensitivity. It is important to note that further analysis and evaluation are required to fine-tune and improve the models for more reliable predictions.

DISCUSSION

- Interprets the results and relates them to the research objectives and previous literature.
- Discusses the implications and significance of the findings.
- Identifies any limitations or constraints of the study.

CONCLUSION

In this study, we developed a predictive model to estimate suspect demographics during crime scenes using data from the New York Police Department (NYPD). Decision tree and random forest algorithms were utilized to predict age, race, and gender of suspects. Results showed moderate accuracy, with the random forest algorithm slightly outperforming the decision tree. Both models demonstrated reasonable precision, sensitivity, and specificity.Specific features such as the victim's age, time of the crime, gender of the victim, crime location, and crime type were found to be crucial in identifying and characterizing suspects. Leveraging these features, particularly VIC_AGE_GROUP, VIC_RACE, and VIC_SEX, can enhance the accuracy of suspect demographic predictions.

Implications of this research are significant for law enforcement, expediting investigations, optimizing resource allocation, and ultimately enhancing public safety while reducing crime rates.However, limitations exist, including data availability and quality, imbalanced dataset, non-linear relationships, ethical considerations, and potential biases. Future research should address these limitations and incorporate additional variables like socio-economic factors and fairness measures to improve the model's accuracy and effectiveness.

In conclusion, this study contributes to predictive analytics in law enforcement by developing a model to predict suspect demographics during crime scenes. Further refinement and research are needed to

unlock the full potential of the model, supporting law enforcement agencies in ensuring public safety and delivering justice to victims.

REFERENCES

- Dataset: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i?category=Public-Safety&view_name=NYPD-Complaint-Data-Historic
- Data Dictionary : https://data.cityofnewyork.us/api/views/qgea-i56i/files/ee823139-888e-4ad0-badf-e18e2674a9cb?download=true&filename=NYPD_Complaint_Historic_DataDictionary.xlsx
- https://github.com/jw782cn/New-York-Crime-Analysis
- https://www.kaggle.com/datasets/adamschroederlcrimes-new-york-city