Exploring Crime data for Prediction and Prevention

Madhuri Muppa
Data Analytics and Engineering
George Mason University
Fairfax, USA
mmuppa@gmu.edu

Jason Rayen
Data Analytics and Engineering
George Mason University
Fairfax, USA
jrayen@gmu.edu

Abhinav Chandoli
Data Analytics and Engineering
George Mason University
Fairfax, USA
achandol@gmu.edu

***Abstract*— Crime has always been a major considerable factor for whatever need we have in our life. It must be reduced no matter what, But how? This research explains why crimes are being caused and who causes the most crimes. Considering these factors, how crimes can be reduced have also been found out in this paper. The project aims to answer several key questions, including predicting crime occurrences per census tract based on available features, determining the significance of these features, and identifying the best prediction algorithms. Additionally, we aim to explore the distribution of data, conduct exploratory data analysis, and create a dashboard with relevant statistics and visualizations. In addition, the study investigates correlations between unemployment rates, labor force participation, and crime incidence by simulating data to confirm crime accuracy. With in-depth modeling techniques and by performing Hyperparameter tunning, the research aims in predicting crime and classifying whether the place is a high crime or a low crime area. Overall, this project aims to contribute to the understanding of crime prevention strategies by providing insights into the factors influencing crime occurrences and the effectiveness of various preventive measures.**

Keywords*— *prediction, statistics, visualizations, correlation analysis, machine learning, deep learning, statistical analysis, Random Forest Algorithm, Hyperparameter tunning, Grid Search, Classification, Regression, Deployment.

I. Introduction

The escalating incidents of crime over the years, ranging from petty theft to heinous acts of violence, have had profound and unsettling impacts on individuals and communities alike. The imperative for implementing stringent preventive measures to curb these occurrences cannot be overstated. It is within this context that the process of data analysis of crime emerges as a pivotal approach. When selecting a dataset and delving into the intricacies of a problem, individuals often ponder: Why this dataset? Why this problem? What is the underlying need? These questions are integral, particularly in the context of crime prevention, where the overarching goal is to mitigate crime rates and enhance public safety. Understanding the underlying reasons behind crime occurrences is paramount to safeguarding people's well-being and fostering secure communities. Crime, while unsettling, is often rooted in identifiable factors. By scrutinizing data and discerning patterns, it becomes possible to pinpoint these underlying causes and devise targeted interventions aimed at prevention. Crime prevention is something which is important to reduce several impacts it could create in the later years. Most of the people find it difficult to live in big cities because of the crime being more, now what is the reason for it? and can we find out what is causing that many crimes? If a new area has been added to the dataset and I need the approximate crime incidents. Can it be found out? Everything is expected to be answered here.

The essence of our project lies in this endeavor: to delve into the depths of crime data, unravel its complexities, and decipher actionable insights that can inform preventive measures. By interrogating the dataset and comprehensively understanding the problem, we find out the reason for crimes and find ways to prevent it from growing in the future. Variables that play a vital role here are found out and suitable analysis is done proving the reason for it to occur. Crimes have been committed and we all know, what has been done is done and there are no ways in reducing what is being caused but there is a way to prevent it from happening in the future, that's the primary goal of this analysis. Being a random person, If I want to move into an area, I'll be looking at the crime rate for sure. Things can't be changed but we can put in analysis which might help the government to be aware of what is coming in the future and prevent so that the increase will stop with the existing number, if it is 535 crime rate for a region in 2022, it has to be less which can be told to people who are moving in. Yes, The description being clear, we are analyzing the current trends in prevention of crimes for the future. The aim of our research is to explore the following questions:

- Can we predict the number of crime occurrences per census tract based on the features available and which features play a vital role in the prediction?
- Can we find out the significance of each of the features to see how the algorithm behaves and calculate the relative risk and the p-value?
- Which algorithm gives the best prediction accuracy, and which are the features involved in the prediction?
- How is the data distributed and how does it affect the number of crime incidents? Can you provide some Exploratory Data Analysis on the dataset?
- Can you prepare a dashboard which gives information about the features of the dataset with suitable statistics and visualizations?
- There are limited records in the dataset. Can we simulate data and cross verify the accuracy to check if there is any effect because of the simulated data?
- Are there observable relationships between unemployment rates, workforce participation rates, and the incidence of various types of crimes?
- How do levels of educational attainment and employment status among residents impact the prevalence of crime within communities?
- How do demographic factors such as age, gender, race, and ethnicity correlate with the frequency and types of crime incidents in different census tracts?

II. Literature Review

RELATED WORK 1: PREDICTION OF CRIME OCCURRENCE FROM MULTI-MODAL DATA USING DEEP LEARNING

This paper proposes a comprehensive approach to crime prediction, leveraging multi-modal data and deep learning techniques to achieve notable accuracy improvements. By integrating data from various sources like the City of Chicago Data Portal, it offers valuable insights for law enforcement agencies to enhance public safety measures. [1]

RELATED WORK 2: CRIME HOTSPOT IDENTIFICATION USING SVM IN MACHINE LEARNING

Using SVM-based techniques, this study focuses on identifying crime hotspots to aid law enforcement strategies, emphasizing the significance of machine learning in crime data analysis. By visualizing and analyzing crime data, the research facilitates more effective resource allocation for crime prevention efforts. [3]

RELATED WORK 3: CHICAGO CRIME ANALYSIS USING R PROGRAMMING

Employing the K-Nearest Neighbor algorithm, this study analyzes historical crime trends in Chicago, achieving an 83.2% accuracy in crime prediction. Through exploratory data analysis, it reveals patterns in crime occurrences, offering valuable insights for law enforcement and urban planning. [2]

RELATED WORK 4: USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA

Comparing violent crime patterns using WEKA software, this research demonstrates the effectiveness of Linear Regression with 92% accuracy. It showcases the potential of machine learning algorithms in predicting violent crime patterns and enhancing law enforcement efforts. [3]

RELATED WORK 5: BIG DATA ANALYTICS AND MINING FOR EFFECTIVE VISUALIZATION AND TRENDS FORECASTING OF CRIME DATA

Examining crime statistics from various cities, this study employs big data analytics and machine learning for trend forecasting. By utilizing deep learning methods like Keras stateful LSTM and Prophet model, it offers insights for improving decision-making and public safety in law enforcement. [3]

RELATED WORK 6: CRIME ANALYSIS AND PREDICTION USING MACHINE LEARNING

This paper evaluates data mining methods for crime detection and prevention, identifying Decision Tree as the most effective approach. It underscores the importance of data mining in understanding criminal behavior and developing strategies for crime prevention. [4]

RELATED WORK 7: SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING AND MACHINE LEARNING TECHNIQUES

Investigating the impact of data mining and machine learning on crime prevention, this survey emphasizes AI's role in evaluating unstructured data for crime forecasts. It aims to integrate data sources and utilize cutting-edge technologies for effective crime prediction tactics. [5]

RELATED WORK 8: CRIME PATTERN DETECTION, ANALYSIS & PREDICTION

Applying data mining tools to analyze crime data, this research highlights the importance of understanding crime patterns for prevention. By employing supervised, semi-supervised, and unsupervised learning algorithms, it aims to provide local police departments with valuable insights for dealing with crime trends. [6]

RELATED WORK 9: CRIME DETECTION USING DATA MINING

Using machine learning approaches, this research aims to forecast trends and patterns in crime, assisting law enforcement organizations in proactive measures. By analyzing extensive data on criminal activity, it identifies potential criminal occurrences and aids in enhancing public safety measures. [7]

RELATED WORK 10: CRIME PREDICTION AND ANALYSIS

This initiative utilizes machine learning algorithms to forecast and prevent crime, emphasizing the importance of an extensive and up-to-date crime database. By employing various techniques including Support Vector Machine and Decision Trees, it aims to better understand crime trends and pinpoint hotspots for preventive action. [8]

RELATED WORK 11: ADVANCING CRIME ANALYSIS AND PREDICTION: A COMPREHENSIVE EXPLORATION OF MACHINE LEARNING APPLICATIONS IN CRIMINAL JUSTICE

This study looks at how protecting individual rights and improving the efficiency of the criminal justice system can be achieved through machine learning. To extract insights, a variety of methods are applied, such as deep learning and statistical models. The study underlines machine learning's technological features and how it might increase efficiency in criminal investigations. Additionally, it talks about privacy and bias as ethical issues, highlighting how crucial it is to have accountable and transparent algorithms. [11]

RELATED WORK 12: CRIME ANALYSIS MAPPING, INTRUSION DETECTION - USING DATA MINING

The study investigates the use of data mining techniques in crime investigation, including VID, pruning method, SVM, and Apriori. It highlights the application of ANN and KNN algorithms to COPS-funded Crime Mapping study. The study highlights the value of crime mapping in identifying high-crime zones and promotes the use of evidence-based research to analyze and lower crime rates. It talks about how quantitative and qualitative data are integrated in crime analysis and highlights how data mining helps law enforcement prevent and reduce crime. [12]

RELATED WORK 13: ENSEMBLE MACHINE LEARNING FOR BETTER CRIME DETECTION AND PREVENTION

The application of ensemble machine learning, such as AdaBoost, Gradient Boosting, and Random Forest, for high-crime prediction and crime detection is the main goal of this work. To create accurate models, it makes use of socioeconomic variables, geographic location, historical crime data, and temporal patterns. To optimize ensemble models and show their superiority over alternative techniques for crime detection, the research places a strong emphasis on feature engineering and hyperparameter tweaking. Overall, the study demonstrates how the use of ensemble models can enhance resource allocation for crime prevention as well as public safety. [13]

RELATED WORK 14: MACHINE LEARNING BASED CRIME IDENTIFICATION SYSTEM USING DATA ANALYTICS

The goal of this research is to apply data mining techniques to create criminal recognition and crime identification systems for Indian cities. It uses 35 preset criminal qualities to address the issue of rising crime rates in India, especially because of poverty. To support investigating authorities, the method consists of data retrieval, pre-processing, clustering, Google Maps integration, and classification. While K-Nearest neighbor classification is utilized for criminal identification and forecasting, K-Means clustering is used for crime detection. The overall goal of the strategy is to aid in the identification and solving of crimes, which could result in a decrease in crime rates. [14]

RELATED WORK 15: A MACHINE LEARNING AND DEEP LEARNING INTEGRATED MODEL TO DETECT CRIMINAL ACTIVITIES

In order to combat the rising crime rate in Bangladesh, this study suggests a real-time crime detection and prevention system. Using Deep Learning and Machine Learning algorithms, the system consists of three modules: criminal activity identification, weapon detection, and criminal recognition. By warning authorities before to crimes, it seeks to increase the effectiveness of policing and maybe lower crime rates. Through experiments with self-collected datasets, the study shows how effective the suggested strategy is in identifying criminal actions and identifying criminals. [15]

RELATED WORK 16: REAL-TIME CRIME DETECTION USING CUSTOMIZED CNN

To stop crime before it starts, this study suggests a real-time criminal detection technique based on deep learning that uses CCTV data. Using real-time video feed analysis, the model categorizes motions as violent or peaceful and notifies a supervisor of any aggressive activity. Reducing human error and efficiently monitoring numerous displays at once are two ways to improve security. The system's proactive detection and deterrent of criminal conduct is intended to increase safety in both public and private areas. [16]

RELATED WORK 17: INTRUSION DETECTION AND ATTACK CLASSIFICATION LEVERAGING MACHINE LEARNING TECHNIQUE

Using a publicly available dataset, this research compares Naive Bayes and Decision Tree classifiers for intrusion detection. Full data and specific feature sets are included in WEKA simulations. Decision Trees excel in terms of accuracy, error rate, f1-score, and recall, but Naive Bayes performs better in terms of computing time. In terms of accuracy, specificity, recall, precision, f1-score, error rates, and response time, the study assesses classifier performance. All things considered; Decision Tree outperforms Naive Bayes in important measures. [17]

RELATED WORK EXPLANATION

The related work focuses only on prediction of crime numbers for each census and nothing interesting. This was done using a direct multimodal deep learning algorithm and no comparisons between other algorithms were implemented. There is no reason the paper explains why this algorithm was selected. No analytics and visualizations were also done here, the only visualization which was posted was the crime rate per census which is easy to understand and nothing new. The paper has lack of analytics, visualizations, predictive modelling and stuffs. [1]

Most of the other papers do the same mistake here as well. All literature reviews have not done enough of in-depth analytics which is one of the reason the paper looks so vague with no information. In one of the papers WEKA has been used which gives us enough results but the results are not compromising as no various analysis or cleaning have been done on the data. [2] [7] [8]

One of the paper's future work says of implementation of multivariate visualizations which we will be implementing in our current proposed approach as well. EDA has been done in this paper but prediction modelling is not done which will be done in our proposed approach. [3]

III. Proposed Approach

With several papers and research work focusing on only predictive analytics or Exploratory Data Analysis, we get information about what is happening currently but not what is to be done in the future.

Yes, predictive analytics gives us certain predictions about the future and stuff but what does it actually give? Just the prediction of crime rate which is not enough.

The proposed approach by us combines various methods used in other research papers to come up with a collaborative approach where the paper will help us predict the crime occurrences of any census with suitable information being provided. This paper will also help us provide suitable Exploratory Data Analysis, Statistical Analysis as well which helps us identify the combination of variables and under what circumstances crime occurrences are increasing so that suitable actions can be taken in the future by the officials.

The predictive analytics part tells us the important variables which play a vital role in the prediction of crime occurrences and in-depth Exploratory Analysis are to be

performed on these variables to get information about the circumstances when the crime is increasing. This will help us prevent crimes in the future.

The best prediction model is chosen, and the selected important variables are separated into categories and being entered as factor variables so that the relative risk, Upper Confidence Interval, Lower Confidence Intervals can be calculated and the highest significance can be taken into account which are the reason for the highest crime rate in the country.

The regression modelling might give us results which involve RMSE, r-squared and stuffs which might be confusing to people whether the solution for the problem is an acceptable one or not. We have extended this research by applying classification algorithm to the problem and identified the place as high crime or low crime rate regions.

Let's have the steps broken down which will be the methodology for the research.

A. DATA CLEANING

- Begin by acquiring the crime dataset and ensuring its integrity and quality.
- Perform data cleaning procedures to address missing values, outliers, and inconsistencies.
- Standardize data formats and address any data entry errors or inconsistencies.

B. EXPLORATORY DATA ANALYSIS

- Conduct exploratory data analysis to gain insights into the distribution, patterns, and relationships within the dataset.
- Visualize key statistics, such as crime incident frequencies, demographic distributions, and temporal trends.
- Identify potential correlations and dependencies between different features and crime occurrences.

C. MODEL IMPLEMENTATION

- Several Models are implemented and the RMSE value along with the r-squared metric is derived and are tabulated.
- It is a Regression problem; Regression modeling has been performed and the results are got.
- Classification modeling is also done to check and confirm if the area is a heavy crime area or a low crime area.

D. FEATURE ENGINEERING AND SELECTION

- Utilize insights from EDA to engineer new features or transform existing ones that may enhance predictive modelling.
- Explore demographic, socioeconomic, and environmental variables that could influence crime rates.
- Select relevant features based on their significance and predictive power, employing techniques such as correlation analysis and feature importance ranking.

E. RESULT ANALYSIS

- Implement and tabulate all the prediction model results and select the best model with the highest accuracy. This will help people to get the latest crime rate for any census that has not been loaded.
- Get future trends of data by finding the variables and the categories under which the crimes get influenced which might serve as a helping measure for officials to reduce the crime.

IV. IMPLEMENTATION

As per Fig 1, The whole data science process which was the main concept of this research started off with Data Cleaning followed by the Exploratory Data Analysis, Predictive Modeling and finally the Result Analysis.

Data Cleaning was done on the dataset where all the null values are removed or replaced with the mean of the whole column. The dataset at start had 801 rows with 79 variables and post data cleaning the records where reduced 799 rows with 79 variables. More than 40 columns had null values and all the values have been treated and transformed.

```
Number of variables with null values: 40
Percent of male population                                                                3
Percent of female population                                                              3
Median age                                                                                3
Percent of under 16 years                                                                 3
Percent of over 65 years                                                                  3
Percent of occupied housing units                                                         3
Percent of vacant housing units                                                           3
Median housing value                                                                      9
Percent of less than 9th grade                                                            3
Percent of 9th to 12th grade                                                              3
Percent of high school graduate                                                           3
Percent of some college                                                                   3
Percent of associate's degree                                                             3
Percent of bachelor's degree                                                              3
Percent of graduate or professional degree                                                3
Percent of high school graduate or higher                                                 3
Percent of bachelor's degree of higher                                                    3
Percent of less than high school graduate                                                 3
Percent of less than bachelor's degree                                                    3
Percent of over 16 years                                                                  3
Percent of population in labor force                                                      3
Percent of population not in labor force                                                  3
Percent of employed population                                                            3
Percent of unemployed population                                                          3
Percent of population not in labor force and unemployed population                        3
Percent of population of agriculture, forestry, fishing, hunting, and mining              4
Percent of population of construction                                                     4
Percent of population of manufacturing                                                    4
Percent of population of wholesale trade                                                  4
Percent of population of retail trade                                                     4
Percent of population of transportation, warehousing, and utilities                       4
Percent of population of information                                                      4
Percent of population of finance, insurance, real estate, rental, and leasing             4
Percent of population of professional, scientific, management, administrative, and waste management services   4
Percent of population of educational services, health care, and social assistance         4
Percent of population of arts, entertainment, recreation, accommodation, and food services   4
Percent of population of public administration                                            4
Percent of population of other services                                                   4
Median income                                                                             5
Mean income                                                                               3
```

Fig. 1. Null valued columns

Post the in-depth Data Cleaning, The Analytical part was performed which involved exploring the data and finding out as much of insights which derives the model implementation. The exploratory data analysis was done in 3 different tools to get the results in an advanced and visually appearing manner. R-studio, Jupyter Notebook, Tableau were used for the analysis part. The analysis involved dealing with a lot of variables as all variables played a vital role in determining the prediction of the result. All the results have been shared and explained in the results section.

Predictive Modeling is the main part of the research which involves confirming whether the prediction got is accurate or not. Several algorithms were involved in implementing this research. The problem which was dealt with is a regression problem. Regression algorithms which were implemented includes Multiple Linear Regression, Support Vector Machine, Decision Trees Algorithm,

Random Forest Algorithm, XGBoost, Bagging and Boosting techniques as well.

The modeling was done with the training and testing data split and k-fold cross validation as well setting the k-folds to 10 folds and the train/test split being 80-20 split. The RMSE values along with the r-squared are got and evaluated for the best fit of the data.

Classification models are also being performed to categorize the data on high crime rate, low crime rate and average crime rate. Algorithms performed includes Random Forest Model, Bagging Classified and K-Nearest Neighbors.

The specific algorithms which are used and implemented are explained below along with the results. Result Analysis include tabulation of all the results derived from algorithms and best of all is being chosen as the one needed for model deployment.

The pickle package from Python is used for model deployment, Both the regression and the classification model are downloaded and saved to be used for further future deployment in any of the servers from any system. This creates 2 models each for regression and classification having a file seize of 200 kb which has information of a pretrained model which was trained for more than 2 hours.

V. Results and Discussion

EDA was performed with the help of R, Python and Tableau and the results are illustrated below.

i. RACE AGAINST CRIME OCCURENCES:

The crime rate have been grouped under 4 categories which are 200 to 500, 500 to 800, High Crime rate and less number of crimes. Several bar charts have been done to find out the highest crimes caused by every set of race as shown in figures 2,3,4 and 5.

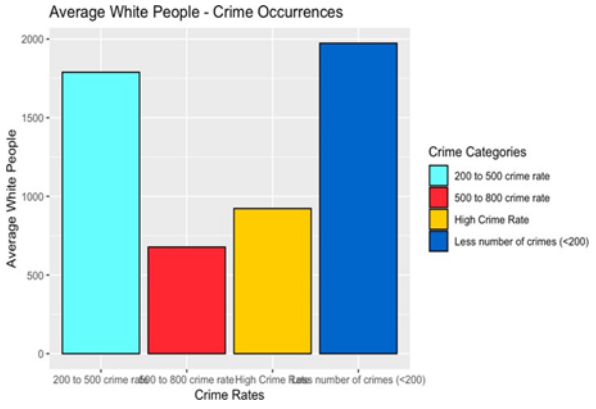

Fig. 2. Average white population against Crime incidents

Fig 2 explains the distribution of white people across all regions who have committed low crimes which range below 800 and high crimes as well. We can see from the beautiful graphical representation that white people have committed most of the crimes ranging from 200 to 500. The crime rate is considerably less here.

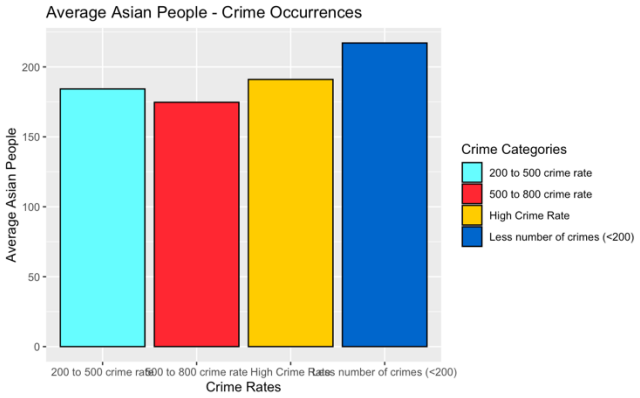

Fig. 3. Average Asian population against Crime incidents

Fig 3 explains the effect of Asian people on the rate of crime being done and this interpretation gives us the information that Asian people have an equal distribution on all sectors meaning they are bound to cause higher number of crimes as well. We can conclude from this figure that Asian people may cause higher number of crimes.

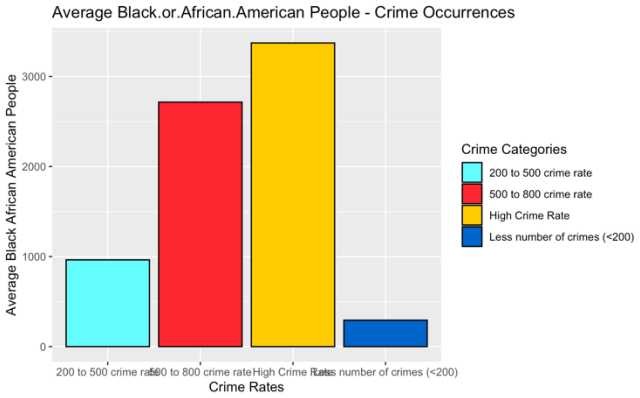

Fig. 4. Average African American population against Crime

Fig 4 gives us information about the distribution of African American people with the crime occurences. High crime occurences have been observed in this distribution.

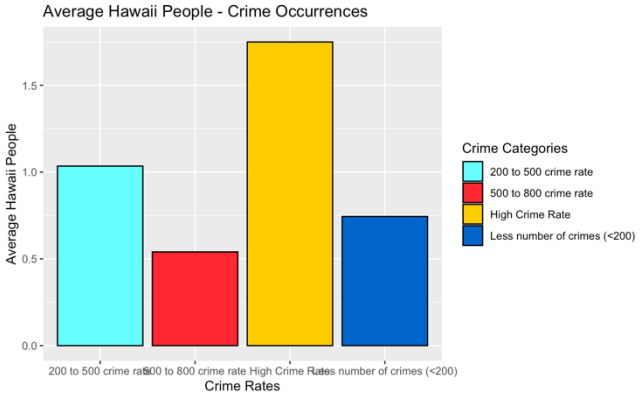

Fig. 5. Average Hawaii population against Crime incidents

Fig 5 explains the effect of Hawaii population on the crime rate. It is obvious and straight that Hawaii people have been one of the major effects for the occurrence of crime in the

Chicago region as the crime rate is much higher with an average crime occurrence of about 1500.

ii. EMPLOYMENT EFFECT ON CRIME RATE

As similar to Race, the employment sector is also being categorized with the crime rate and visualizations have been done to detect the changes. Separate density charts have been done for Employed and Unemployed population as shown in Fig. 6.

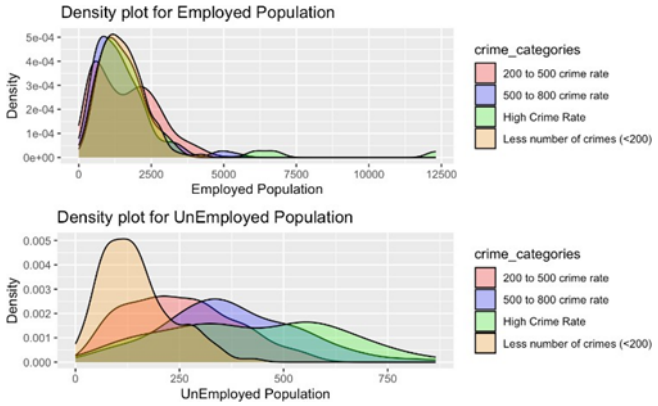

Fig. 6. Employed and Unemployed population density

Fig 6 gives us a clear view that the employed population are expected to perform equal number of crimes and as the number increases the crime rate decreases to a much lesser number.

When the count of unemployed population increases, the crime rate also increases and we have to be aware of this effect while taking insights.

iii. GENDER EFFECT ON CRIME RATE

The effect on Gender has also been done and results have been posted in the Fig. 7.

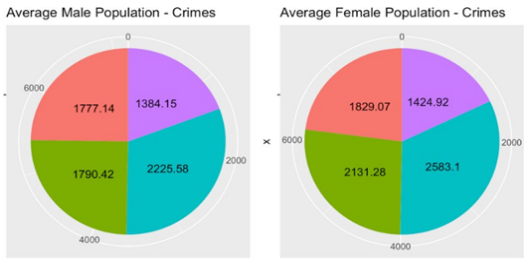

Fig. 7. Gender against crime

Fig 7 proves us the fact that neither being a male or female the crime rate has no effect on that. If the population is more the crime rate effect is going to be more. The numbers gives us a brief view that for both male and female as the count increases, the crime rate increases as well.

iv. EFFECT OF AGE AND EDUCATION ON CRIME RATE

The median age for each census tract has been grouped into age bins to visualize the overall age groups and their number of crime incidents as shown in Fig. 8. To explore the effect of education on the number of crime incidents, scatter plot is visualized for each of the level of education as shown in the Fig. 9.

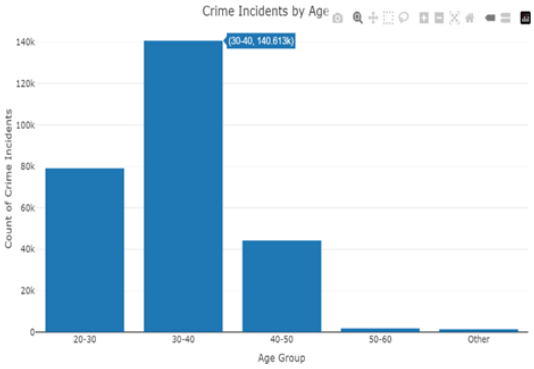

Fig. 8. Age group against crime incidents

Fig 8 gives us a clear idea that the age group ranging between 30 to 40 have committed a lot of crimes. This interpretable visualization was done using plotly in R.

Fig. 9. Education level against Crime incidents

Fig 9 is developed to interpret information for the education level against the crime incidents. The colour legend shows the tabulation of education column against the number of crime incidents.

v. HOUSING VALUE EFFECT ON CRIME

Median housing value for each census tract has been categorized into 3 levels namely 'low', 'medium', and 'high' and plotted against the number of crime incidents to see how different housing values have different crime rate distribution as shown in Fig. 10.

Fig. 10. Housing value and crime incidents

Fig 10 gives us a clear picture that we have some outliers with the medium housing level which might be a major

effect when the data science modelling is performed. People who have high housing levels are expected to cause less crimes in the Chicago region.

vi. CORRELATION ANALYSIS

Another dataset for the purpose of model fit has been created where the correlated variables have been removed to avoid biases in the data. The list of correlated variables are shown below in Fig. 11. A total of 40 variables have been identified has highly correlated and have been removed. The correlation threshold being considered here is 0.7.

```
Total population - Male population : correlation = 0.9760193250538579
Total population - Female population : correlation = 0.9748166773759532
Total population - Over 25 years : correlation = 0.9561844655805099
Total population - Over 16 years : correlation = 0.9797757158651133
Total population - Population in labor force : correlation = 0.9083976773372419
Male population - Female population : correlation = 0.9028947188138949
Male population - Over 25 years : correlation = 0.9343491448970911
Male population - Over 16 years : correlation = 0.9626040652492975
Percent of male population - Percent of female population : correlation = -1.0
Female population - Over 25 years : correlation = 0.9309831609766871
Female population - Over 16 years : correlation = 0.9486230065177955
Percent of under 16 years - Percent of over 16 years : correlation = -1.0
Total housing units - Occupied housing units : correlation = 0.9873782587114923
Total housing units - Over 25 years : correlation = 0.9030277976037915
Occupied housing units - Over 25 years : correlation = 0.9273128396887107
Occupied housing units - Population in labor force : correlation = 0.927762537814297
Occupied housing units - Employed population : correlation = 0.920537099146989
Percent of occupied housing units - Percent of vacant housing units : correlation = -1.0
Over 25 years - Over 16 years : correlation = 0.9828079942809218
Over 25 years - Population in labor force : correlation = 0.9518415754674034
Over 25 years - Employed population : correlation = 0.9293353496955552
Percent of bachelor's degree - Percent of bachelor's degree of higher : correlation = 0.9600288712446148
Percent of bachelor's degree - Percent of less than bachelor's degree : correlation = -0.9600288712446146
Percent of graduate or professional degree - Percent of bachelor's degree of higher : correlation = 0.9461836602046
586
Percent of graduate or professional degree - Percent of less than bachelor's degree : correlation = -0.9461836602046
588
Percent of high school graduate or higher - Percent of less than high school graduate : correlation = -1.0
Percent of bachelor's degree of higher - Percent of less than bachelor's degree : correlation = -1.0000000000000002
Over 16 years - Population in labor force : correlation = 0.9425803716110462
Over 16 years - Employed population : correlation = 0.9132580882417736
Population in labor force - Employed population : correlation = 0.9910758217505763
Percent of population in labor force - Percent of population not in labor force : correlation = -0.999999859766694
Percent of population in labor force - Percent of employed population : correlation = 0.929796814570342
Percent of population in labor force - Percent of population not in labor force and unemployed population : correla
tion = -0.9298840240645645
Population not in labor force - Population not in labor force and unemployed population : correlation = 0.983172230
0878416
Percent of population not in labor force - Percent of employed population : correlation = -0.9297954134891193
Percent of population not in labor force - Percent of population not in labor force and unemployed population : cor
relation = 0.9298828240180159
Percent of employed population - Percent of population not in labor force and unemployed population : correlation =
-0.9999622071472473
Median income - Mean income : correlation = 0.9497181632212526
```

Fig. 11. Correlation analysis

vii. SKEWNESS CHECK

The skewness check was done having a threshold value of and the results were satisfactory and acceptable as there not much skewness in the dataset. Fig 12 shows the results.

Variable Name	Skewness
American Indian or Alaska Native	4.174821
Asian	6.932746
Native Hawaiian and Other Pacific Islander	7.280884
Total housing units	3.429634
Vacant housing units	3.632187
Percent of population of agriculture, forestry, fishing, hunting, and mining	4.433094

Fig. 12. Skewness Analysis

viii. MODELLING RESULTS

A lot of modelling has been done involving 10 models. All the results will be discussed in brief below and the reasons why other models were formed are also discussed. The best model's results along with the accuracy metric are described below. All model results are tabulated at the end.

MULTIPLE LINEAR REGRESSION – ALL VARIABLES

The algorithm was implemented with all variables having number of crimes as the outcome variable and the result got had a RMSE value of 323 which is high for this data frame and the r-squared is very much low with -0.58. Results are got and are illustrated in figure 13. The RMSE value got for this model is the highest of all.

```
Mean Absolute Error: 139.49385630933546
Root Mean Squared Error: 323.6028419432187
Relative Absolute Error: 0.40528915056499515
Root Relative Squared Error: 0.9402042814039565
R-squared: -0.582178429361939
```

Fig. 13. Result from Multiple Linear Regression

MULTIPLE LINEAR REGRESSION IMPORTANT CV

Fig 14, MLR was again implemented with just the important variables got from the initial model fit and the accuracies improved a lot with an RMSE being just 220 with excellent improvement. The r-squared got increased as well.

```
Mean Absolute Error: 130.1940844572043
Root Mean Squared Error: 220.49416396550964
Relative Absolute Error: 0.37826934672474355
Root Relative Squared Error: 0.6406295931768544
R-squared: 0.2654438429445195
```

Fig. 14. Result from Multiple Linear Regression

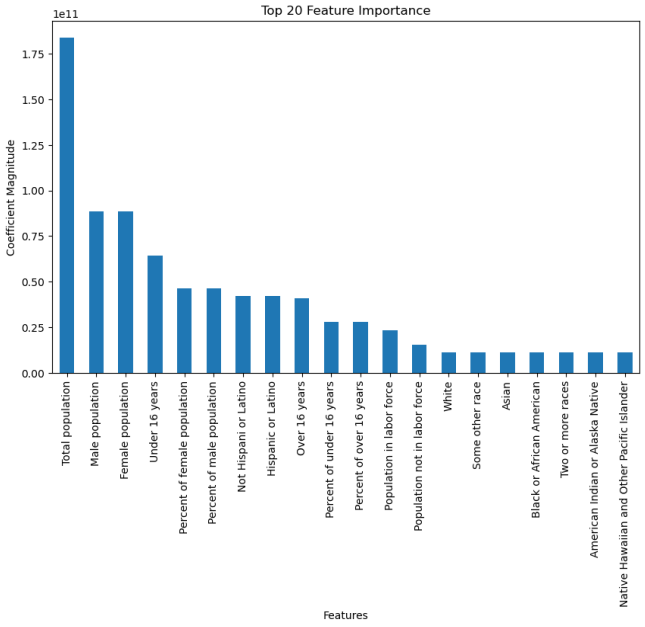

Fig. 15. Important Variables

As per Fig 15, The important variables got from the Multiple Linear Regression model are illustrated above. The 20 variables have been considered as important variables and used in various models.

SUPPORT VECTOR MACHINE

The SVR was implemented as a linear model where the RMSE value posted was 254 which is higher than that of MLR proving that the problem we are facing is not a linear problem but a nonlinear one. Fig 16 shows the results.

```
SVR Model Metrics:
Mean Absolute Error: 174.3684236943464
Root Mean Squared Error: 254.4018884174016
Relative Absolute Error: 0.5066146437856371
Root Relative Squared Error: 0.7391459952915439
R-squared: 0.022151741328030106
```

Fig. 16. Result from Support Vector Machine

SUPPORT VECTOR MACHINE - HYPERTUNNED

The same SVR was hyper tunned with several parameters and allowed to execute for more than 2 hours and got stunning results. The hyper tunned version of SVR reduced the RMSE value to a greater extent bringing it down to 136 and a r-squared value of 0.71. This model not only improved the RMSE but also reduced the error rate bringing the r-squared value high to about 0.71. The best metrics got are svr_C:100 and the kernel as rbf. Fig 17 shows the results.

```
Improved SVR Model Metrics:
Mean Absolute Error: 95.99666491207594
Root Mean Squared Error: 136.82866363914673
Relative Absolute Error: 0.2789112568012357
Root Relative Squared Error: 0.3975456290797372
R-squared: 0.7171312469657409
```

Fig. 17. Result from Support Vector Machine - Hypertunned

XGBOOST MODEL

XGBoost Model was applied on the dataframe with a thought and aim of improving the accuracy with the help of Boosting techniques and it improved the accuracy a bit but not as better as SVM. Results are shared in Fig 18.

```
XGBoost RMSE: 220.84071186007637

XGBoost Model Metrics:
Mean Absolute Error: 106.27302896523777
Root Mean Squared Error: 149.41816431862392
Relative Absolute Error: 0.3087684775290555
Root Relative Squared Error: 0.4341234983237265
R-squared: 0.6626834899558307
```

Fig. 18. Result from XGBOOST Model

XGBOOST - HYPERTUNNED

The same model was hyper tunned heavily to get the best of it and the results were spectacular as well. Fig 19 shows us the result of the hyper tunned model and the parameters selected for the best fit were learning_rate as 0.2, max depth with 5 and the number of estimators being 200.

```
Improved XGBoost Model Metrics:
Mean Absolute Error: 97.800773862042
Root Mean Squared Error: 137.10157579601747
Relative Absolute Error: 0.28415296280323343
Root Relative Squared Error: 0.39833855530002676
R-squared: 0.7160017276449098
```

Fig. 19. Result from XGBOOST Model

ADABOOST ENSEMBLE MODEL

We have done and played with the most important models and now what else. The ensemble trick must do the magic, The ensemble model was implemented and interpreted to get the results and the results were not that satisfactory as well. Fig 20 gives us the results for this model.

```
AdaBoost Ensemble Model Metrics:
Mean Absolute Error: 113.68843622724984
Root Mean Squared Error: 155.79097395073748
Relative Absolute Error: 0.33031339850141545
Root Relative Squared Error: 0.45263922848451715
R-squared: 0.6332962172312487
```

Fig. 20. Result from AdaBoost Ensemble Model

RANDOM FOREST ALGORITHM - HYPERTUNNED

The best of all, Random Forest Algorithm was implemented with all parameters being provided and hyper tunning was performed. The hyper tunned model gave great results providing the lowest RMSE value recorded until present. The RMSE value recorded was just 127 with a r-squared value of 0.75 being the best model recorded so far. (Fig 21)

```
Tunned Random Forest Model Metrics:
Best parameters found: {'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 300}
Mean Absolute Error: 89.67887378391195
Root Mean Squared Error: 127.91571029056293
Relative Absolute Error: 0.2605553788613318
Root Relative Squared Error: 0.3716496979810768
R-squared: 0.7527828595330363
```

Fig. 21. Result from Random Forest Algorithm

The best metrics are got with a max depth of 20, min leaf sample being 4 and estimators being 300.

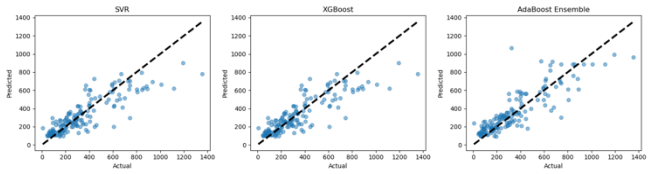

Fig. 22. PREDICTED vs ACTUAL Fit

BAGGING MODEL – RF + SVR

We have got the best of the best from SVR and RF with low RMSE values and high r-squared values. The model was not only a combination but a hyper tunned model as well and so the best of the best was combined to get a stunning result.
The RMSE reported here is the lowest of all being just 117 and the r-squared value reported is 0.78 which is the highest. Declaring this model as the best one. Fig 23 shows the results got from the model.

```
Bagging Ensemble Model Metrics:
Mean Absolute Error: 86.99358306350126
Root Mean Squared Error: 119.53423627246862
Relative Absolute Error: 0.2527534639678049
Root Relative Squared Error: 0.3472979410281172
R-squared: 0.7841184997851715
```

Fig. 23. Results from BAGGING MODEL – RF + SVR

RESIDUAL PLOT

The residual plot shows us the fact that there is no overfitting or underfitted values in the model. The prediction is accurate. The actual vs predicted results also proves that the predictions are present on the line which proves the predictions are very accurate. Fig 24 gives us the results.

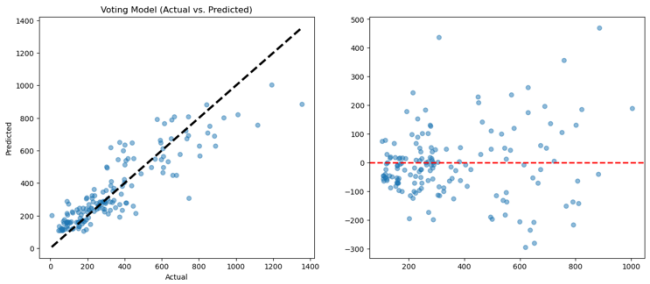

Fig. 24. Residual Plot & Actual vs Predicted Fit

EXAMPLE RESULTS

The example results show that the prediction is almost near to the expected result except for some big errors. The errors are ranging around 100 which is the best RMSE value got as well. The r-squared is 0.78 hence getting a result which is almost correct and accurate. Fig 25 gives us the results.

There are some errors which must be fixed. For which, A classification model was developed which will be explained in the later part of the research.

Actual Value	Predicted Value
207	237.314918
615	715.604217
110	164.536219
744	693.871582
150	133.352737
400	514.630534
659	450.535622
119	218.761097
407	538.854005
458	244.707115

Fig. 25. Actual vs Predicted Examples

TABULATION OF REGRESSION MODELS

All the model's results are tabulated in Fig 26, for an easy and visual view. These results are got through the implemented models and all models are tabulated here.

Models	RMSE	R-squared
Multiple Linear Regression	323.6	-0.58
Support Vector Machine	254.4	0.22
MLR - Important Variables	220.5	0.28
XGBoost Model	149.4	0.66
LASSO Model	144.3	0.71
Support Vector Machine - Hypertunned	136.8	0.72
XGBoost - Hypertunned	137.1	0.72
Random Forest Algorithm - Hypertunned	127.9	0.75
Bagging Model - RF + SVR	119.1	0.78

Fig. 26. Actual vs Predicted Examples

From the figure, it is easy and obvious that the best model is the Bagging model which was implemented at the end. A combination of RF and SVR being the best model as the RMSE value is just 119 and the r-squared value being 0.78.

CLASSIFICATION MODELING

Regression results seems not satisfactory thinking maybe the RMSE value is too high, or we could have got a better r-squared value. The classification models were done to prove it right with a higher accuracy prediction. (Fig 27)

The major aim of the project is to suggest people whether the place is expected to be a higher crime occurrence place or a low crime occurrence place and this classification model with a very good accuracy proves us that as well. We get either 3 predictions which are High Crime Occurrence, Low Crime Occurrence or Averaging around 400 crime occurrences. The best 3 algorithm results have been posted.

	ACCCURACY
Random Forest Algorithm - Hypertunned	0.95
SVR - Classifier	0.60
K - Nearest Neighbors - Hypertunned	0.67

Fig. 27. Tabulation of Classification Models

Fig 27 shows that the Random Forest Hyper tunned model is the best model with an accuracy of 0.95 being the highest, lets discuss this model in detail.

RANDOM FOREST MODEL - HYPERTUNNED

The Random Forest classification model was hyper tunned with different combination of parameters such as number of estimators, maximum depth, min samples split, min samples leaf. The results are extraordinary with great accuracy rates.

```
Tuned Random Forest Model Metrics:
Best parameters found: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Accuracy: 0.9556962025316456
Precision: 0.9571178188899708
Recall: 0.9556962025316456
F1 Score: 0.954994729829167
```

Fig. 28. Random Forest Classification Model - Hypertunned

The parameters which produced the best model are or minimum samples leaf of 1, minimum samples split of 2 and the number of estimators being 100. The accuracy got is 0.95 with precision (number of accurate true positive rate) being 0.95 and the F1 and recall being the same 0.95 as well. Fig 28 has all information about the results.

CONFUSION MATRIX – RESULT ANALYSIS

The confusion matrix has 160 accurate predictions with 7 incorrect predictions. The incorrect predictions are 5 wrong predictions which had to be predicted as Low Crime rate but got predicted as Average Crime Rate (400).

The 5 incorrect predictions can be easily fixed by changing the split of the data and performing suitable analysis accordingly. The accuracy rate can be improved easily giving a great accuracy in an easy and efficient way. The 1 incorrect prediction is Average Crime Rate (400) which was

predicted as Low Crime Rate. The Fig 29 gives the confusion matrix of Random Forest Algorithm.

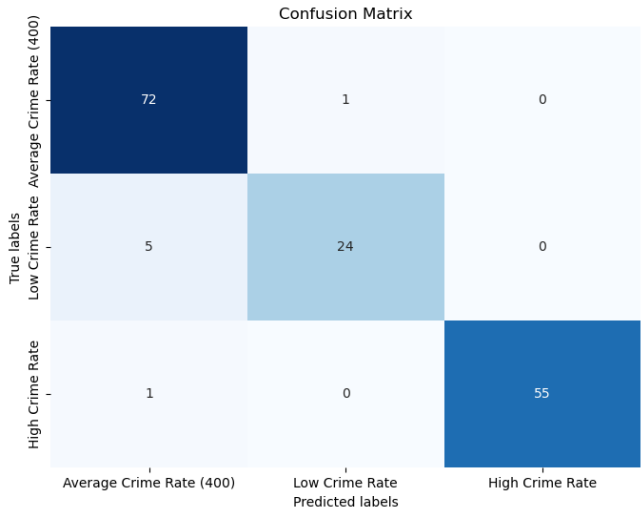

Fig. 29. Confusion Matrix – Random Forest Algorithm

EXAMPLE RESULTS – RF CLASSIFICATION

The results derived here as per Fig 30 are almost correct and makes sense better than that of the regression modeling. This classification model can be used to identify which class the crime occurred is present.

Actual Value	Predicted Value
Average Crime Rate (400)	Average Crime Rate (400)
High Crime Rate	High Crime Rate
Low Crime Rate	Low Crime Rate
High Crime Rate	High Crime Rate
Low Crime Rate	Low Crime Rate
Average Crime Rate (400)	Average Crime Rate (400)
High Crime Rate	**Average Crime Rate (400)**
Low Crime Rate	Low Crime Rate

Fig. 30. Example Results – Classification Model

ix. EDA – TABLEAU DASHBOARD

A tableau dashboard was developed with all the crime happened at different places of Chicago over the years. Each year's crime can be innovatively visualized and viewed on the dashboard. Fig 31 represents the dashboard that was developed.

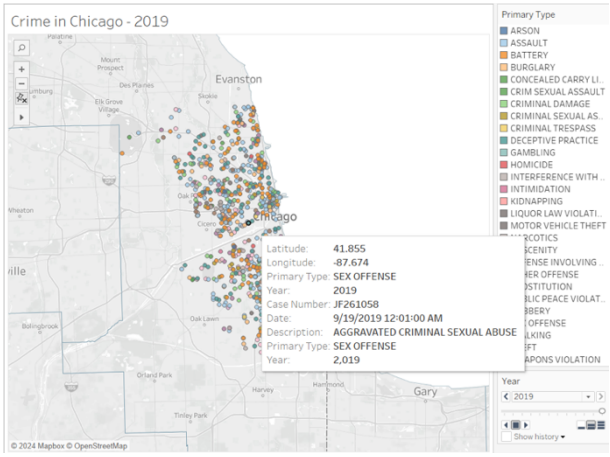

Fig. 31. Tableau Dashboard

VI. MODEL DEPLOYMENT

The best regression model which is the bagging model (RF+SVR) and classification model (RF – Hyper tunned) was downloaded using the pickle package and was deployed.

REGRESSION DEPLOYMENT PKL: voting_model.pkl
CLASSIFICATION DEPLOYMENT PKL: best_rf.pkl

VII. CONCLUSION

To find patterns and trends, the research used data science approaches to perform a thorough analysis of crime data. Important factors impacting crime rates were found by carefully cleaning and analyzing exploratory data. Regression and classification algorithms were two of the many predictive modelling techniques used to effectively forecast crime rates and categorize locations according to the frequency of crimes. The models' efficacy was confirmed by model evaluation and hyperparameter tuning, with the Bagging model which combines Random Forest and SVM proving to be the best performer. This thorough approach not only offered insightful information about the patterns of crime but also set the stage for further research, especially when it came to investigating deep learning methods for crime analysis and prediction. Classification modeling was also performed to identify crime categories and the analysis proved to answer all the research questions that were posted.

VIII. FUTURE WORK

- At first the number of records analyzed were less which can be improved by focusing more on data collection. We had just 800 records which played a vital role in the RMSE. If more records where analyzed, we might have got a better accuracy.
- PCA can be done to limit the number of variables to be analyzed. This can be a better future work so that we can have less variables and provide visualizations if less variables have been used in the model.
- One important future work for classification models is we saw some wrong predictions in the bracket of Low Crime/Average Crime which can be dealt with by trying different crime combinations to improve the accuracy.

REFERENCES

[1] H.-B. K. Hyeon-Woo Kang, "Prediction of crime occurrence from multi- modal data using deep learning," [Online]. Available:

[2] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.017 62 44.

[3] L. M. a. N. Meghanathan*, "USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA," [Online]. Available:

[4] https://www.researchgate.net/profile/Natarajan-Meghanathan/publication/275220711_Using_Machine_Learning_Algori thms_to_Analyze_Crime_Data/links/571dc8ae08ae408367be5de8/Usin g-Machine-Learning-Algorithms-to-Analyze-Crime-Data.pdf.

[5] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li, Y. Xi and Q. Liu, “BIG DATA ANALYTICS AND MINING FOR EFFECTIVE VISUALIZATION AND TRENDS FORECASTING OF CRIME

[6] DATA,” [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8768367.

[7] O. Llaha, “Crime Analysis and Prediction using Machine Learning,” [Online]. Available:

[8] https://ieeexplore.ieee.org/abstract/document/9245120.

[9] K. Vinothkumar, K. S. Ranjith, R. R. Vikram, N. Mekala, R. Reshma and

[10] S. Sasirekha, “Crime Hotspot Identification using SVM in Machine Learning,” [Online]. Available: https://ieeexplore.ieee.org/document/10104689.

[11] N. T. Singh, M. Mehra, I. Verma, N. Singh, D. Gandhi and M. Ahm, "Advancing Crime Analysis and Prediction: A Comprehensive Exploration of Machine Learning Applications in Criminal Justice," [Online]. Available: https://ieeexplore.ieee.org/document/10467221.

[12] B. Panja, P. Meharia and K. Mannem, "Crime Analysis Mapping, Intrusion Detection - Using Data Mining," [Online]. Available: https://ieeexplore.ieee.org/document/9140074.

[13] B. Dhanwanth, R. A. Roshan, C. H. Bhargavi, G. V. Shri and S. Raja, "ENSEMBLE MACHINE LEARNING FOR BETTER CRIME DETECTION AND PREVENTION," [Online]. Available: https://ieeexplore.ieee.org/document/10425908.

[14] S. M. R. D. o. C. C. (. t. b. University), I. Chiranmai and N. Jayapandian, "Machine Learning Based Crime Identification System using Data Analytics," [Online]. Available: https://ieeexplore.ieee.org/document/10370717.

[16] K. F. Arpa, T. Mittra, T. Ferdous, N. Jahan, R. A. K. Tayna and M. H, "A Machine Learning and Deep Learning Integrated Model to Detect Criminal Activities," [Online]. Available: https://ieeexplore.ieee.org/document/10272002.

[17] H. U. T. P. Shahtaj Shaukat Department of Electrical Engineering, A. Ali, A. Batool, F. Alqahtani, J. S. Khan, Arshad and J. Ahmad, "Intrusion Detection and Attack Classification Leveraging Machine Learning Technique," [Online]. Available: https://ieeexplore.ieee.org/document/9299093.

[18] K. Muthumanickam, B. Selvalakshmi, P. Vijayalakshmi and P. Nareshkumar, "An Effective Method for Forecasting Crime Analysis using Deep Learning Classifiers," [Online]. Available: https://ieeexplore.ieee.org/document/10425673.

[19] V. V. Nojor, J. A. C. Austria, A. A. Galit and J. T. B. Guevarra, "Design of a Deep Learning-based Detection System for Criminal Activities," [Online]. Available: https://ieeexplore.ieee.org/document/9998276.