CIND820: FINAL REPORT

Experimental Approaches in the Use of Crime Data

AIDA RADONCIC
500348520
Supervisor: Ceni Babaoglu
Submitted on: Wednesday, April 5, 2023

Ryerson University

Table of Contents

Final Report

Introduction

There has been a notable increase in the use of crime data to make geospatial predictions surrounding crime—when and where crimes occur (Shah et al., 2021). The issue stems from how the data is being used, more specifically, how it is being used within law enforcement settings. The resulting knowledge gathered from crime data is being used to prepare police for training and responses, upholding the belief that using this knowledge will increase the effectiveness of policing (Walzack, 2021). The problem with this approach, however, is that it limits our understanding of how and why these crimes occur in the first place.

These tools fail to both acknowledge and account for the complexities of social systems—how people exist in these spaces, how social, economic, political, religious, and a multitude of other factors intersect and contribute to existing dynamics (Jany, 2022). It fails to account for the bias in not only law enforcement, but also social biases, as well as the personal biases of those who are collecting the data, processing the data, and modelling the data (Dean, 2018). These approaches do not consider how neighbourhoods report crime, why some report more and others choose not to, overlooking the social and historical nuance of tensions between groups of people and police presence (Chattopadhyay, 2022). They also do not address the fact that those in the machine learning community and others dealing with the data may not have sufficient training in the social sciences and critical theory, creating “blind spots” that result in a failure to acknowledge the problematic assumptions our models have the capacity to produce (Dean, 2018). These assumptions consequentially spill into law enforcement decisions, which may further perpetuate bias to the point at which it eventually unravels on a grassroots level in potentially devastating ways, not considering real world implications (Chattopadhyay, 2022).

Research Questions

Given the previous line of questioning, I believe this is a worthy area of exploration. We understand that bias exists in crime data, but are we able to pinpoint them and the moments in which they are produced and reproduced? Is it possible to locate the intersections of social variables that create these biases, as a first step towards an interdisciplinary approach with the goal of minimizing negative consequences that stem from a lack of understanding of the results and how they come to be? Will this allow us to find some nuance to better understand the complexity of crime prediction?

Contribution of the Work

These questions are not necessarily new amongst pre-existing literature surrounding crime prediction using machine learning algorithms. A number of the studies examined use classification algorithms to make predictions, including K-Nearest Neighbours, Random Forest, and Decision Trees (Iqbal, 2013). These methods are made use of in my own work; however, they are used to achieve different ends. In much of the existing research, the focal point is to predict crime and evaluate the model's performance to successfully make these predictions, and establish how this can contribute to law enforcement efficiency. Although I am using comparable methods, my focus is the process itself, similarly to the study done by Bhonsle et al., whose focus is assessing how accurate machine learning algorithms in data mining can be when it comes to making predictions about crime patterns, with a focus on data pre-processing and classification (Bhonsle, 2022). These methods will be used for exploration, taking an approach that maintains openness and curiosity. A number of other studies within the existing body of work address the ideas I have proposed, and offer non technical ways of achieving this. There is an emphasis on the importance of mitigating harmful consequences from the use of machine learning in decision making processes through seeking social awareness,

interdisciplinarity, and a level of reflexivity as we approach the data and how we examine biases (Dean, 2018). Many of these questions have already been considered, and attempts have been made to address them in the existing literature. This manifests through imparting these ideas onto the use of machine learning models, as well as creating a space for dialogue around these issues within the machine learning community (Dean, 2018). There is a clear call for interdisciplinary thinking within technical spaces—we need to move beyond the quantitative data and also examine the implications of the data: What does it mean and what are the dynamics between the data and the real world? What is the data's context, where is it embedded, what complex social processes is it intertwined with (Chattopadhyay, 2022)? How does this affect how we approach the data itself? Because of the exploratory approach I am taking with a focus on curiosity rather than definitive answers, this work may create the space for more experimental approaches that value openness and discovery. This study can act as a step in this direction: an interdisciplinary approach that amalgamates existing thought and technical methodology to explore bias within machine learning in the spaces of law enforcement.

Applied Methodology and Study Design

My methodology and study design replicates the process of preparing data from an unprocessed format, to a state where it is ready to be used for predictive modelling. Similar to existing work in crime prediction, I will be using 3 classification models: K-Nearest Neighbours, Decision Tree, and Random Forest, to predict whether a specific crime will occur, or the likelihood of it occurring. The idea here is to assess which model can most accurately predict crime. However, it is not the evaluation of the predictive power of the model that is my focus, but of the entire process itself. This process includes data cleaning, statistical analyses, correlation analyses, data transformation, etc. The results from our models will reveal which features are

selected, which features contribute the most information to the models, as well as which features may have been dropped early on.

The methodology I will be applying revolves around data mining and data exploration to extract as much information as possible. The focus here is an attempt to find the points where bias is located, and the intersections of social variables that create. On the quantitative side, the focus of the applied methodology is on the data exploration itself—the process of examining our variables, cleaning the data, and allowing any information to come to the surface. Simple technical applications cannot be overlooked, as even the most obvious applications can shed light on important information. The key is to be attentive and open to this information as it comes to light. We can learn about how the data is collected by drawing our attention to where missing values are located, and the ratios of missing values per attribute. We can discover what exactly is being represented within our instances, and consequentially, what is not, and consider what this means and why it may be the case. Finding these points will be done through an iterative process, requiring close engagement, meticulousness, and critical thought throughout. These points of bias do not come out in obvious ways, but rather, they appear in moments where technical processes meet critical thought. Throughout my work, I focus on bringing together the technical process and methodology with a reflexive frame of thinking, as well as an engaged and attentive approach. I attempt to apply these ideas and philosophies, and reflect throughout, utilizing my own background and training in the social sciences to pinpoint areas where bias is found, produced, and reproduced. It is not so much through the methodology, but instead, through thinking about the problem and questions we might ask ourselves during the process within the frame of interdisciplinary thinking (Vestby, 2021).

Conducted Analysis

Preliminary Statistical Analysis

This project begins with a preliminary statistical analysis: first, our needed libraries are imported. The main libraries this project draws from will be Pandas and NumPy. Throughout this repository, it will also draw from sklearn, and matplotlib for more specific functions. The preliminary analysis is meant to provide an opportunity to familiarize oneself with the dataset: its organization, column names, data types, missing values, and shape. The missing values will be prepared here for further processing later on. Missing values in the dataset are represented by the '?' symbol. They will be changed to 'NaN' so they may be dealt with appropriately. It is at this point I will begin an initial exploration of the missing values in each column. In order to do this, I will first check the sum of NaN values per attribute, separate the columns that contain any NaN values, and display the sum of each. This gives us an initial idea of the ratios of missing values for each attribute. We immediately see that a number of attributes are missing 1872 out of 2215, or 85% of its total values.

Upon closer inspection of the attributes, each one of the columns with 85% of missing data are the columns that specifically deal with police demographics. This is the first instance in which we can make note of bias being introduced. As previously mentioned, there are a number of factors that intertwine in spaces that produce certain social dynamics, crime being one of them. The rate at which arrests are made and the ways crime is reported does not exist in a vacuum, but is part of a larger context (Chattopadhyay, 2022). That context includes police presence. Some of these attributes include specific information about the number of full time police officers in a community, the number of full time officers in field operations (on the street rather than administrative positions), total requests for police presence, measures of racial matches between the community and existing police force, percentages of different racial profiles in the police body in these specific communities, number of police cars in the community, and the

police operating budget. It is interesting to note that the most significant number of missing values appear only in police demographics. From a social perspective, I'd like to consider the possibility that police demographics in communities have a massive impact on how crime occurs and how crime is reported. It also tells us about whether a community is over policed or under policed based on total officers per 100k of the population. The total number of requests for police presence, along with the percentages of racial profiles in a population vs. that of the existing police body may offer insight into the dynamics between the community and police presence, and perhaps serve as a starting point for further investigation into social and historical tensions between police and communities, which, in my opinion, would undoubtedly affect how crime is reported.

Moving onto the rest of the preliminary analysis, I have conducted an examination of the summary statistics to further understand the nature of each attribute, such as observing the mean, standard deviation and ranges of each to get a sense of each distribution. The preliminary analysis concludes with sub-setting and aggregation. The dataset is aggregated by the attribute STATE, with an examination of the number of rows represented by each state, and a general statistical summary of how the attributes behave under this aggregation. This is where we begin our initial observations of the potential target variables: each crime-PerPop (per 100k of the total population), by state. Here we can observe which states have the minimum and maximum numbers for each crime. Out of 8 crimes, the state of DC has the highest number of 5 of the 8 crimes reported. Upon further investigation, we come to find that the state DC is represented by only 1 row in the dataset, by the community 'Washington City', with a population of 606,900. We begin to see here that the representation of states by their communities is unbalanced. Some states, such as California ('CA') are represented by a total of 279 communities which represent rows in our dataset.

It is at this juncture we may ask ourselves the following questions: why do some states have significantly more representation than others? Is this based on population number? Economy? Are some neighbourhood types over-represented in the data? What characteristics do these neighbourhoods have in common? It is at this juncture we might also require a deeper and more robust analysis. For the sake of this project, I have taken the first steps into this investigation, however, more would be required to make any definitive claims. I have taken the top represented states, and bottom represented states, and turned them into their own subsets. On both subsets, I have examined the summary statistics in hopes I can find any leads that may give us insight as to why certain states have more representation than others. I have focused on comparing the mean values of each column between the 'minrep_subset' (states with minimum representation) and 'maxrep_subset' (states with maximum representation). A few observations I have made is that the states with maximum representation have a lower overall Land Area (in square miles) and a higher Population Density than those states with minimum representation. Our maximum representation states also have an overall significantly lower numbers of: people in poverty, number of those in shelters, number of vacant houses in their communities, number of people known to be foreign born, and the number of children born to unmarried parents. These values are significantly less than the overall average of these attributes in states with minimum representation. These observations would require further investigation to make any meaningful conclusions, but may offer a starting point to answering the previous questions regarding state representation.

Body of Analysis

In the body our conducted analysis, we will begin to deal with the missing values discovered in the preliminary statistical analysis. A threshold of 0.5 has been set: any attributes that are missing more than 50% of its values will are dropped. Here, we are losing information through dropping attributes because of the ratio of missing values. We are unable to appropriately

impute these attributes without introducing bias into the data. However, by removing an attribute which represents a social factor that may contribute to how crime is committed, we may also be losing some nuance in the data as well. I have also chosen my target variables based on missing values: MURDERS and ROBBERIES. Many of my potential dependent variables are missing values, and are unable to be imputed as it may introduce bias. For this reason, I have chosen these two categories as my dependent variables, as they have the least number of missing values. There are still a number of columns behaving as our independent variables that remain with missing values. The number of these missing values were below the threshold of 0.5, so they will be used in the analysis. To deal with them, we will turn to imputation. For columns with a total of 1-15 missing values, I have chosen to impute them with the overall mode of the column. Because the distribution of our attributes is skewed, I have chosen to impute with the mode value per each individual column.

There remains a number a columns with a significant amount of missing values, yet fall below the given threshold. These columns will have to be dealt with differently. For these values, I chose to aggregate the data first, and then decided on my imputation method. The organizational nature of the crimes is by neighbourhood, each found within a given state. I felt that aggregating by STATE was the most suitable choice, as my initial assumption was that we may find more geographical, social, and economic similarities overall (but not entirely) within each state and its population. While considering that this aggregation method may oversimplify the economic nuance of different neighbourhoods within a state, for the purposes of choosing an imputation method for this analysis, this method felt to be appropriate. Because there are a different number of neighbourhoods represented by each state, the representation is unbalanced. As a result, I thought the most appropriate way to impute these missing values would be to use the mean of the individual column per state, and impute with this number. The

ratio of missing values to the length of rows represented by each state was assessed. If the missing values reach or exceed a threshold of 0.65, the state will be dropped. Otherwise, they were imputed with the mean of the column of the specific state in which these missing values are located. I have achieved this through the following: go through each state and aggregate them one by one into their own data frames. For each state's data frame, I checked for missing values. Rather than checking for the total number of missing values, I chose to instead find the particular column name where these values were located. Then, I confirmed the total number of rows in the state subset, and then checked the total number of missing values in each of these columns to get an idea of how much information is missing per attribute in the particular state. If the ratio was not significant, I moved forward to imputing the missing values as discussed previously.

A number of states came up that had a significant number of missing values per column. They were the same attributes for each of the states: 'rapes', 'rapesPerPop', 'arsons', 'arsonsPerPop', 'ViolentCrimesPerPop', and 'nonViolPerPop'. These were marked as significant because they exceeded the missing value threshold of 0.65. Some of the states were missing 100% of their values in these specific columns. Because of the nature of missing values here, I was unable to impute these values with the mean. For states missing 100% of its values in a column, there is no mean to impute the data with. That being said, the approach to their imputation had to be handled differently. I considered taking the 3-4 surrounding states of the particular state in question, aggregate them, find the mean of the column in question, and then impute the mean into the missing values. This, however, would run the risk of introducing bias to the dataset through introducing inaccurate values. The other option I considered would be to drop these rows entirely, which meant dropping 6 entire states. These states included: MI, AL, IL, IA, VT, KS. Ultimately, I decided to drop all rows with missing values. This decision may introduce bias

into the dataset in a representational sense; we began with an unbalanced representation of states, and now are entirely missing 6 of them. Although this may now be the case, I believe this option will ultimately affect the predictive models less and introduce considerably less bias than the other option discussed.

The dataset has now been appropriately cleaned, rows dropped, and missing values dealt with. It is at this stage where the data is ready to be transformed in preparation for predictive modelling. The first step I took was to explore correlations between the attributes. First, I ran a correlation matrix on the entire dataset. This allowed me to explore the attribute pairs that were most highly correlated, both positively and negatively. Many of these correlations I felt were straight forward. For example, population and number of those living in poverty had one of the higher positive linear relationships. As the population grows, the number of those living in poverty is likely to also grow. There were others, however, that may be worth noting.

Table 1: Correlation Comparison

ATTRIBUTE 1	ATTRIBUTE 2	CORRELATION CO-EFFICIENT
autoTheft	NumUnderPov	0.984159
NumKidsBornNeverMar	robberies	0.982958
NumKidsBornNeverMar	NumUnderPov	0.982756
NumKidsBornNeverMar	murders	0.979956
NumUnderPov	murders	0.975271
NumStreet	robberies	0.954813
HousVacant	larcenies	0.943457
NumImmig	assaults	0.942954
NumKidsBornNeverMar	NumInShelters	0.942867
NumImmig	autoTheft	0.941317

The number of those under poverty seems to have a high positive correlation with autoTheft (total number of automobile thefts in the year 1995) as well as murders (total number of murders in the year 1995). The number of children born to parents that never married also have a high positive correlation with the number of people under poverty, robberies, and murders (total number in the year 1995). It also highly correlates with the number of those in shelters. The number of people found in the streets has a relationship to the total number of robberies that year. The number of vacant houses found in the community correlates to the number of larcenies that occurred. And the total number of people that are foreign born (NumImmig) has a positive correlation with both the number of assaults and automobile thefts that occurred. These highly correlated attributes specifically caught my eye because I believe this is another area where bias has the potential to be reproduced, and it is crucial that we investigate these relationships further and in a more meaningful way before drawing any conclusions. Although strong numbers can seem convincing, it may be at the expense of nuance, and we must be weary of steamrolling complex social dynamics that may contribute to these outcomes.

Next, I created a dataset for each target variable, giving us two separate datasets: one for our ‘murders’ (total number of murders in 1995) target and one for ‘robberies’ (total number of robberies in 1995) target. To create each dataset, I have dropped all other crime columns except for the target crime. This decision was made because I wanted to focus on social and economic attributes about the community population, to see if they might behave as determinants of specific stand-alone crimes. I examined the correlation of all the attributes to ‘murders’ and then to ‘robberies’, as my target variables, and then to ‘murdPerPop’ and ‘robbPerPop’ (murders/robberies per 100k of the population) as my target variables. I chose to examine both for the sake of exploration, and to find any potential differences. The correlation

coefficients for both ‘murdPerPop’ and ‘robbPerPop’ did not have any highly positively or negatively correlated attributes. However, upon examination of ‘murders’ and ‘robberies’ (total number in the year 1995), I found both had no noteworthy negative correlations with any attributes, however, they shared strong positive correlations with the same attributes.

Table 2: Correlation Comparison

‘MURDERS’ TARGET VARIABLE		‘ROBBERIES’ TARGET VARIABLE	
ATTRIBUTE	**CORRELATION COEFF**	**ATTRIBUTE**	**CORRELATION COEFF**
NumStreet	0.899517	NumStreet	0.954813
HousVacant	0.900746	HousVacant	0.860459
NumInShelters	0.918830	NumInShelters	0.948761
NumImmig	0.925304	NumImmig	0.959762
numbUrban	0.964525	numbUrban	0.964597
population	0.966415	population	0.966789
NumUnderPov	0.975271	NumUnderPov	0.974603
NumKidsBornNeverMar	0.979956	NumKidsBornNeverMar	0.982958

The total number of both murders and robberies in a community in 1995 both share the strongest positive linear correlations with: number of people found living in the streets, number of houses vacant in the community, number of people living in shelters, number of people foreign born, number of people living in areas classified as urban, total population, number of people living under poverty, and number of children born to parents never married. The higher values of these attributes, the higher total number of each crime.

Lastly, I created a correlation matrix for both data frame categories, with ‘murdPerPop’ and ‘robbPerPop’ as my dependent variables. I removed the upper triangle from both, and dropped

any variables with a correlation of 0.80 or higher. At this point, the datasets are ready to be normalized. I went through the remaining attributes and reviewed each column summary, as well as the distributions in order to see how balanced each attribute was. This was achieved by checking both the skew values of the attributes, as well as plotting a histogram for each. If the attribute was normally or close to normally distributed, I normalized the column numerically, within a range of 0-1. If the attribute was not normally distributed and clearly skewed or unbalanced, I turned it into a categorical variable with different levels. The levels were established based generally on the min, max, and quartile values in order to create the most balance within the categories. The last attributes to transform are the target variables. These were turned into categorical attributes. I chose to use binary classification on the MURDERS dataset. This decision was made based on the distribution of the data, with an attempt to achieve balance within the levels, as well as create the levels in a way that made sense. The split categories were “YES” when a murder occurred (a value greater than 0) and “NO” when a murder did not occur and the value was 0. The ROBBERIES dataset used multiclass classification. The 3 levels were: “UNLIKELY”, “LIKELY” and “VERY LIKELY” reflecting the likelihood in which a robbery would occur. Both datasets ultimately resulted in the same mix of attributes through the feature selection process. The only differences between the two datasets are their target variables: for the MURDERS dataset, it is ‘murdPerPop’, which is the value for total number of murders per 100k of the population, and for the ROBBERIES dataset, it was ‘robbPerPop’, the total number of robberies per 100k of the population.

The data is one step away from being ready to be used in predictive modelling. In order to be fully prepared, our data values must be able to be converted into floating point numbers by our models. As it stands, most of the attributes in both datasets are of the categorical data type. These attributes must be encoded. There were multiple methods of encoding available. As I

considered each, I weighed the pros and cons of each approach. Because my dataset is already quite large with a high number of columns, taking a one-hot encoding approach, or using dummy variables, would add a colossal number of extra columns. Each categorical attribute had 4 levels, being the majority of the 51 total columns. The ‘state’ column also contained 47 unique values. The next issue came up with the ‘community name’ column, where we had over 1000 unique values. This simply would not work. I chose to omit the ‘community names’ attribute, as I felt it did not offer any information that would be useful in the models. However, the loss of this attribute does not come without consequence. Each community examined in each state has its own unique characteristics and dynamics, and should be taken into consideration during deeper examination.

Ultimately, I chose to go with a label encoding method. This method would maintain the number of columns that already exist in my dataset, and label each category in each column with numbers that corresponded to a level (Joshi, 2018). Perhaps we may lose some information at this juncture, as our categories may be oversimplified and have already lost some nuance in the process of normalization and binning. They values are no longer as precise as they once were, and because of the amount of skew in the distribution, the binning method may not have been ideal in terms of preserving a level of nuance that seems necessary when dealing with social complexities for the sake of balancing the data. Information in our dataset seems further rendered crude by the encoding process, in addition to the loss of precise information in the form of omission of attributes as well as the normalization process.

The last part of my conducted analysis involved the prediction models. I chose to run 16 models in total. 8 were given to the MURDERS dataset, and 8 were given to the ROBBERIES dataset. There are 4 unique models in total, each were run twice for each dataset. The models I chose to use for this project were 3 classification algorithms: K-Nearest Neighbours, Decision Tree, and

Random Forest. The aim of my project was to predict whether a crime would occur based on a set of social variables, so I felt that classification was the best option. I chose KNN because of its simplicity, quick calculation time, and high generally high accuracy rate (Chattergee, 2022). Decision tree was chosen because I felt it could offer more information in regards to information gain, where I could examine the GINI purity value of each variable, and discover which variables contribute the most to predictions. Random Forest was chosen not only for its robust nature, but also because of its ability to handle high dimensionality within datasets: features are automatically reduced in the algorithm, giving us more insight into what features were deemed important and what features were not (Molina, 2021). The last model I used was a logistic regression model. I will not be comparing this regression model to the classification models, but chose to include it in my analysis simply from a place of curiosity and exploratory purposes. I wondered if it would be worth trying to approach classification of my target variables from a probabilistic framework, and might also offer additional insight into understanding the relationships between my variables.

I ran each model twice for both datasets: 4 under a regular train/test split and the same 4 models using k-fold cross validation. A regular train/split will divide the dataset into two parts: 70% of the data will be used to train, and 30% of the data will be used to test. The potential issue that may arise is that the split may be unbalanced between categories, skewing our results. K-fold cross validation, however, will take multiple subsets of the data (10 splits) ensuring that every point in our data has the chance to be used for both training and testing purposes. This will allow for more balance in our predictions, and a lower selection bias in our results (Great Learning, 2022). I created a confusion matrix for each model, and evaluated them using Accuracy, Precision, and Recall. I used the same evaluation metrics for my set of cross validation models, but took the mean of the 10 folds as my final metric for comparison.

Results: Findings and Evaluation

To evaluate the performance of all 4 classification models, I chose to use Accuracy, Precision, and Recall as my metrics. Only the 3 classification models; KNN, Decision Tree and Random Forest will be compared to one another. The logistic regression model will be assessed on performance, but treated as its own entity. Because I have run all 4 models on both a standard train/test split as well as with k-fold cross validation, there will be two sets of metrics examined per each dataset.

Murders Dataset
Table 3: TRAIN/TEST SPLIT CONFUSION MATRICES

LOGISTIC REGRESSION

	Pos (+)	**Neg (-)**
Pos (+)	206	53
Neg (-)	73	244

DECISION TREE

	Pos (+)	**Neg (-)**
Pos (+)	197	62
Neg (-)	74	243

K-NEAREST NEIGHBOURS

	Pos (+)	**Neg (-)**
Pos (+)	183	76
Neg (-)	72	245

RANDOM FOREST

	Pos (+)	**Neg (-)**
Pos (+)	195	64
Neg (-)	74	243

In terms of true positives, where the model predicted an occurrence of a murder when a murder did, in fact occur, the Decision Tree model performed best out of three: Decision Tree, KNN, Random Forest. For true negatives, where the model predicted that a murder did NOT occur, and a murder did not, in fact occur, the K-Nearest Neighbours model outperformed the other

two models, by only 2 points more. Otherwise, all 3 models performed similarly with incorrect predictions of murders NOT occurring. The Decision Tree, although not by much, outperformed the other models when predicting murders that did occur, incorrectly. It had the least number of incorrect guesses when predicting that a murder did not occur, when it did in fact occur. Otherwise, just by examining the resulting confusion matrices, at face value it seems that there are no notably significant differences in model performance.

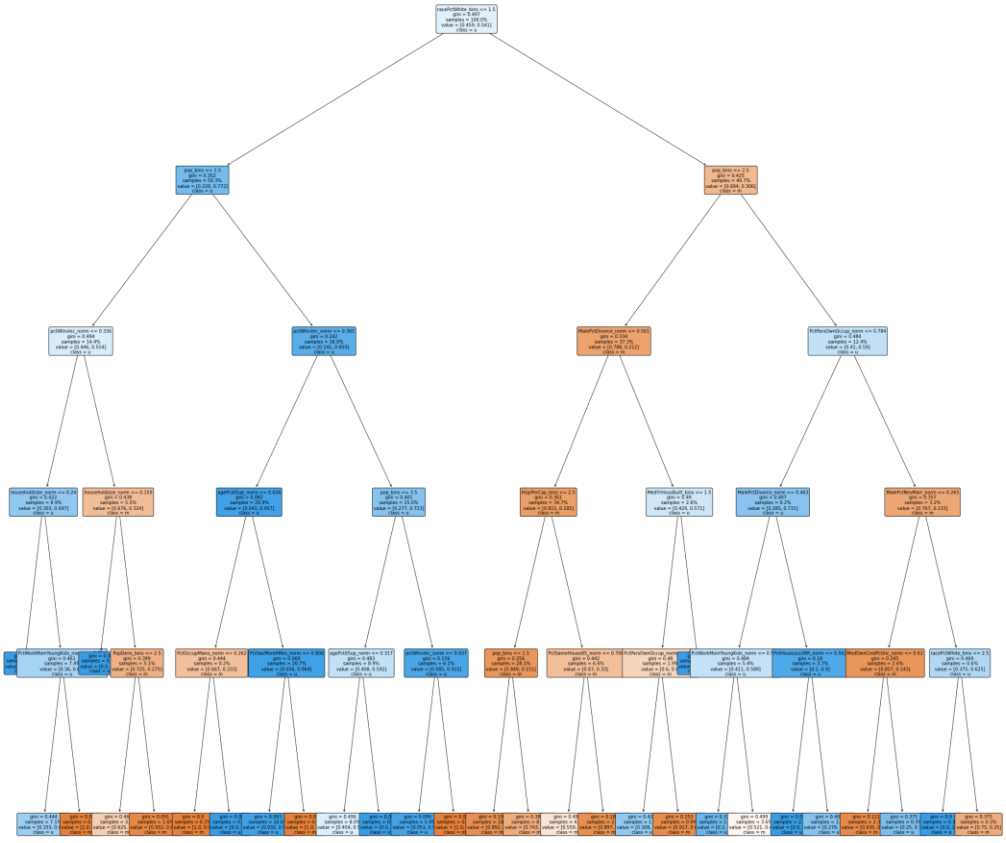
Table 4: TRAIN/TEST EVALUATION METRICS

MODEL	ACCURACY	PRECISION	RECALL
LOGISTIC REGRESSION	0.781	0.738	0.795
K-NN	0.743	0.717	0.706
DECISION TREE	0.763	0.726	0.760
RANDOM FOREST	0.760	0.760	0.760

Moving onto the evaluation metrics, we can see more nuance in performance differences. The Accuracy metric shows us the number of samples that were predicted correctly out of the total number of samples (Iqbal, 2013). Decision Tree surpassed the other models, sitting at 0.763. Precision shows us the total of accurate positive predictions out of the total of positive predictions (Iqbal, 2013). Here, Random Forest outperformed the other models. Lastly, the Recall metric shows us the number of accurate positive predictions out of all actual positives (Iqbal, 2013). Random forest tied with Decision Tree for the highest value of recall, of 0.760. Overall, the Random Forest model performed best with the regular train/test split using the murders dataset. Although not compared to the other classification models, the Logistic Regression model outperformed each classification model and seems to be a generally robust model, as it maintained high performance measurements.

The Decision Tree, aside from evaluating it based on its metrics, can also be examined as a study in feature selection. This Decision Tree model was creating using Gini Index as the measurement of impurity. The nodes this algorithm has chosen can tell us more about the attributes, such as their level of predictive power. The below figure is the visual plotting of the entire tree.

Chart 1: TRAIN/TEST SPLIT DECISION TREE PLOT



Found below, are the first three levels of the tree, starting at the root. The attributes that appear contribute the most information towards classification in this scenario, in terms of which variables best split the data (Iqbal, 2013). The attribute found at the root is ‘racePctWhite’, the percentage of the population of Caucasian race. To make note, the race options are highly

limited, and taken from census data; people may identify their races differently. The Decision Tree splits the root at 1.5 bins out of 4 bins total, with a value sitting at 82.5%. Depending on whether the value falls above or below this value, we move onto the next split, which is 'pop_bins', population. If below, our next split is above or below a population under 16,250. If above, the next split is a population above or below 29,000. The next level of splits on the left divides 'pctWInvInc', percentage of households with investment/rent incomes. On the right of the same level, we have 'MalePctDivorce' (percentage of males who are divorced, and 'PctPersOwnOccup', (percentage of people in owner occupied households). Percentage of the population of white race, the size of the population itself, percentage of households with investment incomes, percentage of divorced males, and percentage of people in owner occupied households are the socio-economic variables that offer the most information for classification of the data into whether a murder occurred or not. How these variables relate to each other, however, can offer us deeper understanding regarding why these particular social and economic attributes contribute to a murder occurring or not. A decision tree can tell us what the variables are, but otherwise can only give us the first step into furthering our knowledge. In this way, a richer understanding is limited, as we can only look at attributes at face value and consequently risk make incorrect assumptions about demographics of communities.

Chart 2: Enlarged Decision Tree Plot

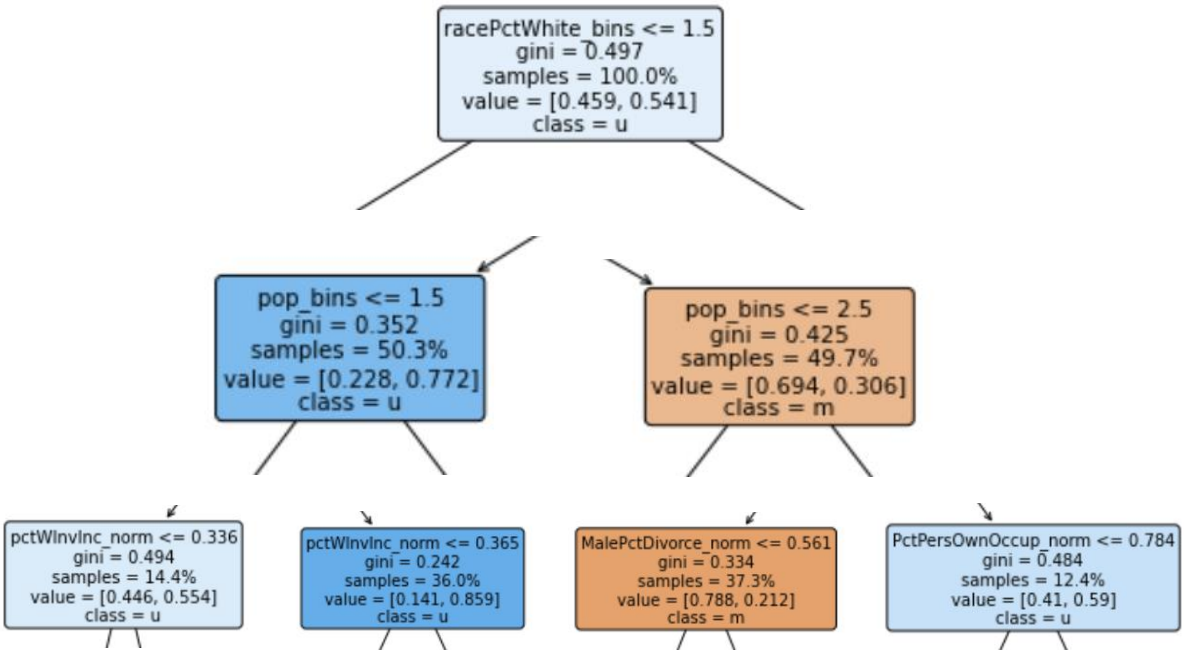


Table 5: K-FOLD CROSS VALIDATION EVALUATION METRICS

MODEL	ACCURACY	PRECISION	RECALL
LOGISTIC REGRESSION	0.776	0.785	0.806
K-NN	0.752	0.759	0.795
DECISION TREE	0.727	0.767	0.713
RANDOM FOREST	0.768	0.794	0.786

When evaluating the models with K-Fold Cross Validation, the scores slightly change. With the use of cross validation, Random Forest rather than Decision Tree, maintains the highest Accuracy value of 0.768. Random Forest also has the highest value for Precision, of 0.794. K-Nearest Neighbours outperformed both tree models when it came to evaluating Recall. In the standard train/test split, KNN was the poorest performing model. However, with cross validation,

it outperforms the Decision Tree. Overall, the Random Forest model, once again, was the best performing predictive model. The scores above are the calculated mean from all 10 folds. Logistic Regression also performed very well here. We can conclude that the models generally performed better when making predictions when using k-fold cross validation in its train/test split.

Robberies Dataset

Table 6: TRAIN/TEST SPLIT CONFUSION MATRICES

LOGISTIC REGRESSION

	Likely	**Unlikely**	**Very Likely**
Likely	112	52	34
Unlikely	43	141	3
Very Likely	31	4	156

K-NEAREST NEIGHBOURS

	Likely	**Unlikely**	**Very Likely**
Likely	100	56	42
Unlikely	55	125	7
Very Likely	37	6	148

DECISION TREE

	Likely	**Unlikely**	**Very Likely**
Likely	120	43	35
Unlikely	64	119	4
Very Likely	32	5	154

RANDOM FOREST

	Likely	**Unlikely**	**Very Likely**
Likely	127	41	30
Unlikely	39	147	1
Very Likely	27	2	162

Because the ROBBERIES dataset was a multiclass classification problem, the confusion matrices look a bit different. The three classes of whether a robbery would occur or not were “LIKELY”, “UNLIKELY” AND “VERY LIKELY”. One of the more crucial actual/predicted values, in my opinion, is whether the model was able to correctly predict “UNLIKELY”. An incorrect prediction would have the highest consequence here. Out of the three classification models, KNN, Decision Tree and Random Forest, the Random Forest model outperformed the others. It had the highest number of correct “UNLIKELY” predictions and the lowest number of incorrect “UNLIKELY” predictions, with a total of 40 incorrect. The Decision Tree model had the highest number of incorrect “UNLIKELY” predictions, which would result in the highest number of robberies as a result of its predictions, with a total of 68 incorrect.

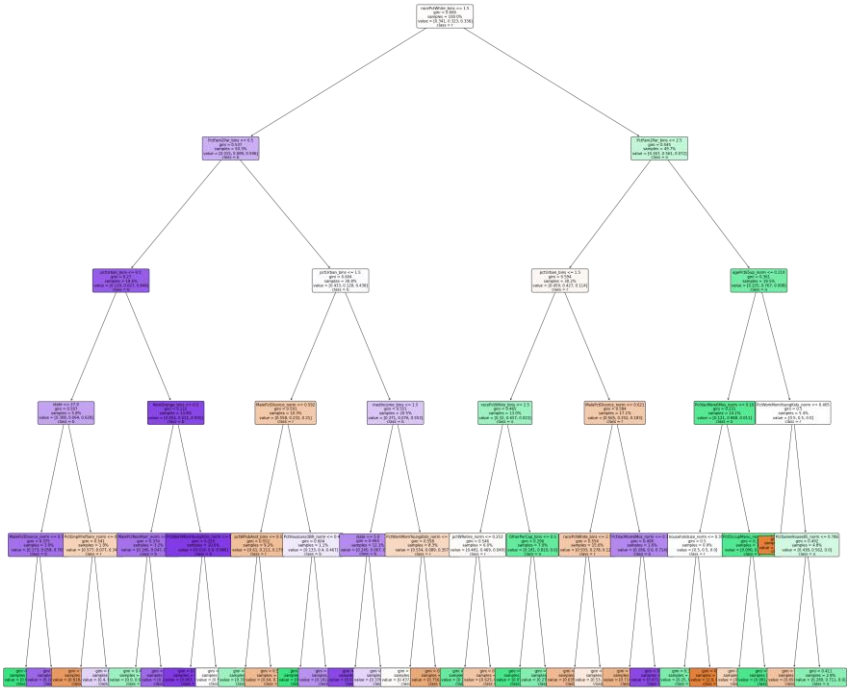
Table 7: TRAIN/TEST EVALUATION METRICS

MODEL	ACCURACY	PRECISION	RECALL
LOGISTIC REGRESSION	0.710	0.708	0.712
K-NN	0.647	0.646	0.649
DECISION TREE	0.682	0.688	0.682
RANDOM FOREST	0.756	0.756	0.756

Examining the evaluation metrics will give us a more accurate assessment of the results than a first glance observation at the confusion matrices. The Random Forest model, out of the 3 compared models (KNN, Decision Tree, and Random Forest), maintains the highest values in all metrics. It has outperformed all other models across the board. It is worth noting that this dataset was a multiclass classification model, using three different classes for prediction. In the regular train/test split, Random Forest retained the highest value for all 3 metrics. However, in the MURDERS dataset, which was a binary classification problem, the Decision Tree model outperformed Random Forest in Accuracy, with the highest proportion of correctly predicted samples out of the total number of samples.

This decision tree model was created the same way as the previous tree model used in the MURDERS dataset, using GINI as the measurement of impurity. The below figure is the visual plotting of the entire tree.

Chart 3: TRAIN/TEST SPLIT DECISION TREE PLOT



Below, are the first 3 levels of the resulting Decision Tree and the attributes that were chosen to represent it. They contribute the most information to how the data is split, to eventually classify our data into 3 classes of robberies: robberies that are likely to happen, unlikely to happen, and very likely to happen. At the root of the tree, we find the same attribute that was found at the root of the MURDERS binary classification decision tree: ‘racePctWhite’ (percentage of the population of white race), which splits at 82.5%. Depending on whether the value is above or below, we move onto the next split, which is ‘PctFam2Par’, percentage of families with kids headed by two parents. If below, our next split is above or below 33.5%. If above, the next split is above 76%. The next level of splits on the left divides ‘pctUrban’, percentage of people living in areas classified as urban. On the right of the same level, we have ‘pctUrban’, and ‘agePct65up’, percentage of the population that is 65 and over. In this tree, the socioeconomic variables that contribute to major splits to predict robberies are: the percentage of the population

that is white, the percentage of families with kids headed by 2 parents, percentage of people living in areas classified as urban, and percentage of the population that is 65 and over. The interpretation of these variables can become problematic, we cannot assume that these splits will accurately predict whether a robbery will occur or not.

Chart 4: Enlarged Decision Tree Plot

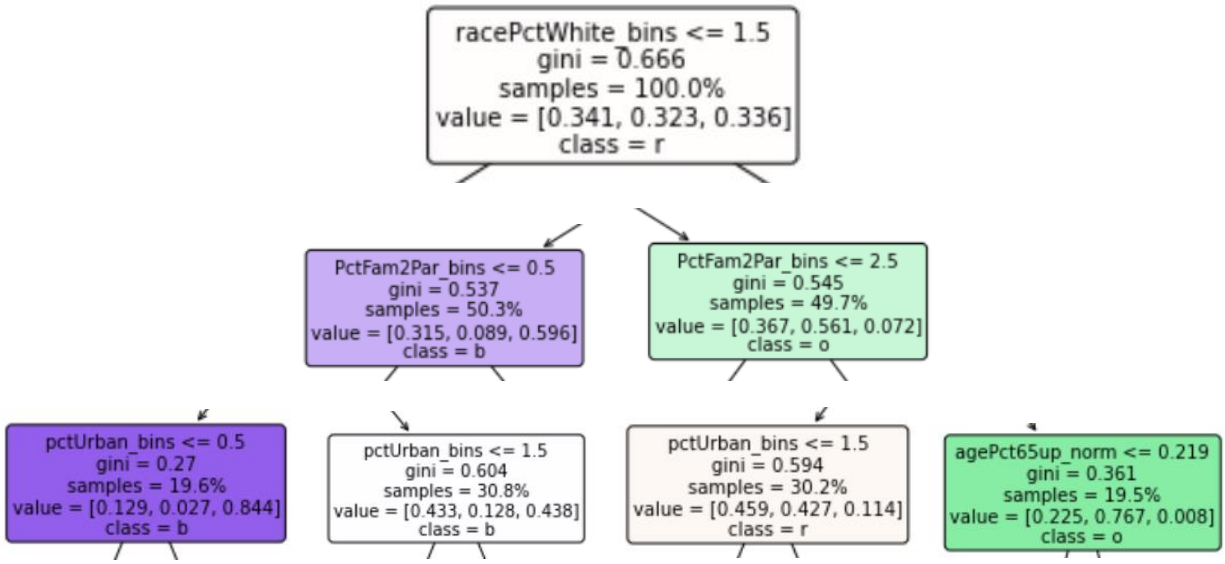


Table 8: K-FOLD CROSS VALIDATION EVALUATION METRICS

MODEL	ACCURACY	PRECISION	RECALL
LOGISTIC REGRESSION	0.705	0.708	0.712
K-NN	0.646	0.646	0.649
DECISION TREE	0.663	0.688	0.682
RANDOM FOREST	0.714	0.757	0.758

The cross-validation scores for the ROBBERIES dataset also produced better results than the regular train/test split. Running both a regular split as well as evaluating with folds on both a

binary classification problem and a multiclass classification problem allowed me to explore different types of classification problems that may have potentially altered the quality of performance and the results. Similarly, in the binary classification problem of the MURDERS dataset, the Random Forest model was also the best performing model using cross-validation, in Accuracy, Precision, and Recall. However, all of the models resulted in a better performance when dealing with the binary classification problem of the MURDERS dataset. Overall, it appears that the Random Forest algorithm proved to be the most robust predictive model with both binary and multiclass classification problems. It performed even better with cross validation than the standard train/test split. In our cross-validated murders dataset, Random Forest gave us performance rates between 75 and 80%, where as the cross-validated robberies dataset only gave us 71-76%. Although performing generally well, we are still seeing an error rate of between 20-30%. This number is significant enough that it should raise concern when being used to predict crime which may result in decision making in law enforcement settings.

Shortcomings and Concluding Remarks

There were a number of instances in the process of cleaning, transforming, and applying data mining techniques where I found, what may be, easily dismissed and misunderstood moments in preparing our data where the production and reproduction of bias could fall through the cracks. There were a multitude of limitations in the way the data was collected, in which we observed an over simplification of social and economic variables that contribute to crime. There was a clear imbalance of representation across the states: some states were represented by more communities than others. This also begs the question, how were these communities chosen? Why were they included in the dataset over others? Because this data came from the census, it leads me to consider the possibility that there were likely a number of people without a permanent address who were not able to partake in the census, but rather were counted as

"number of homeless people counted in the street". These observations are inexact, and have no explanation or methodology offered behind how this took place. The demographics of attributes chosen to be included in the dataset itself is not free from bias: how were they chosen? What was the methodology behind these decisions? Were these choices based on studies and legitimate research, or on reproduced assumptions that are rooted so deeply in our thinking that we come to consider them as objective truth or common knowledge?

It is crucial to consider these questions as we deal with our data. The results our of models *do* tell us something. But, if we don't understand the reality of our data, how it exists in the world, the ways it was collected, and the ways in which our data becomes over-simplified, loses nuance and information through the process of preparing it so our models can even use them for prediction, then what really do these results even mean to us? (Vestby, 2021) How can we consider these results objective in any way and use them to make decisions around policing that could potentially be devastating to communities and the people within them? If the models are wrong, at best, 20% of the time, how can we trust them to make such significant decisions? How can law enforcement accept the use of a model to make decisions that influence the way communities are policed if they are making incorrect predictions almost a quarter of the time?

There were a number of other methods I hoped to use in this project, however, they were outside the scope of my knowledge and level of expertise. In terms of continuing the work, I think this specific undertaking deserves a more in-depth analysis, both in technical terms but also in critical theory. I think this area of study is highly significant in the sense that the consequences of irresponsible undertakings are immense. Being able to spend more time understanding how the data was collected, understanding the methodology behind collecting census and FBI data would, in my opinion, greatly contribute to being able to understand the biases that are produced through the use of this particular dataset, or any crime dataset, in

machine learning models for decision making. Because of the nature of the data, this would require a more comprehensive interdisciplinary approach, as suggested in previous literature. This level of interdisciplinary thinking in this particular study was limited to my own background and training in social sciences and theory. Having a number of perspectives from other humanities disciplines, such as understandings of political history and dynamics between particular groups of people, how these dynamics emerged, an understanding the history of the communities, the geography and how perhaps the physicality of the land area itself may contribute to particular dynamics unfolding, etc., could contribute to the work in meaningful ways (Ji, 2020). These contexts are highly complex and deserve the time and attention to understand them, especially when using this information to make such substantial decisions with potentially massive consequences. I anticipate that research in this area continues in a productive way, where there are efforts towards achieving further interdisciplinarity in these approaches and influencing how much weight we really can put onto machine learning models to make such predictions. It seems there is a disconnect between the technical and social when it comes to approaching problems like this, and I think in order to continue this work, this gap needs to be closed (Vestby, 2021).

In terms of continuing my own work, I would spend more time experimenting with multiple subsets of the entire dataset: I only used murders and robberies, and I chose to omit all other crimes as my focus was on contributing social and economic variables. However, I believe it would have been interesting to also consider how other occurring crimes may contribute to the likelihood of a target crime occurring in tandem. I hope to spend more time experimenting with other data mining approaches as well, such association rules in order to explore which variables contribute most to murders or robberies being committed, and how this differs from the resulting feature selection of the decision tree models. I would also consider more feature selection

methods through decision trees, and comparing the use of GINI, Entropy and Information Gain on both a binary and multiclass problem, and comparing the results.

This work was not without its shortcomings. As mentioned previously, it was limited to a certain level of expertise and experience. As I moved forward in the iterative process of cleaning and transforming the data, I also progressed in my own learning; there were many moments in which I would go back and reconsider a choice, change something in my code, and consider a different approach. This was also limited by time and experience. There were many ways where I felt that I could streamline my code, after the fact. There were many details I would have liked to go back and change, things I felt I needed a deeper understanding of to make certain technical decisions, and things I felt I needed to learn more in depth. I believe this undertaking may have potentially been outside the scope of my own capabilities, however, I think that this is an imperative conversation to have within the machine learning community. I hope there will continue to be opportunities for those with more experience, understanding and expertise in the areas where I may lack, to continue this work in a more meaningful way, and potentially impact policy making in regards to what capacity machine learning models can, and should be used to make decisions in law enforcement settings.

References

Bhonsle, D., Kshatri, S.S., Moyal, V., Pillai, A.G., & Verma, R. (2022). Crime Detections Approach Using Big Data Analytics and Machine Learning. NeuroQuantology, 20(8), 1480-1495.
doi: 10.14704/nq.2022.20.8. NQ44162

Walczak, Steven. (2021). Predicting Crime and Other Uses of Neural Networks in Police Decision Making. Frontiers Psychology. 12. https://doi.org/10.3389/fpsyg.2021.587943

Jany, Libor. (2022, July 4). Researchers use AI to predict crime, biased policing in major U.S. cities like L.A. Los Angeles Times. https://www.latimes.com/california/story/2022-07-04/researchers-use-ai-to-predictcrime-biased-policing

Dean, S., Dobbe, R., Gilbert, T., & Kohli, N. (2019, Jul. 6). A Broader View on Bias in Automated Decision Making: Reflecting on Epistemology and Dynamics [Conference Session]. Workshop on Fairness, Accountability and Transparency in Machine Learning during ICML 2018, Stockholm, Sweden.

Chattopadhyay, I., Evans, J., Huang, Y., Li, T., & Rotaru, V. (2002). Event-Level Prediction of Urban Crime Reveals a Signature of Enforcement Bias in US Cities. Nature Human Behaviour, 6, 1056-1068. https://doi.org/10.1038/s41562-022-01372-0

Vestby, A., & Vestby, J. (2021). Machine Learning and the Police: Asking the Right Questions. Policing: A Journal of Policy and Practice, 15(1), 44-58. https://doi.org/10.1093/police/paz035

Ji, J., Liu, L., Xiao, L., & Zhang, X. (2020). Comparison of Machine Learning Algorithms for Predicting Crime Hotspots. IEEE access, (8), 181302-181310. doi: 10.1109/ACCESS.2020.3028420

Joshi, P., Kalsi, P.S., Kim, S. & Taheri, P. (2018). Crime Analysis Through Machine Learning [Conference Session]. IEEE 9th Annual Information, Technology, Electronics and Mobile Communication Conference, Vancouver, British Columbia, Canada.

Iqbal, R., Murad, M.A.A., Mustapha A., Panahy, P.H.S., and Khanahmadliravi, N. (2013). An Experimental Study of Classification Algorithms for Crime Prediction. Indian Journal or Science and Technology, 6(3), 1-7. DOI: 10.17485/ijst/2013/v6i3.6

Shah, N., Bhagat, N. & Shah, M. (2021). Crime Forecasting: A Machine Learning and Computer Vision Approach to Crime Prediction and Prevention. Visual Computing for Industry, Biomedicine, and Art, 4(9). https://doi.org/10.1186/s42492-021-00075-z

Molina, E. **(2021,** February 2**5)**. *A Practical Guide to Implementing a Random Forest Classifier in Python.* **Towards Data Science.**
https://towardsdatascience.com/a-practical-guide-to-implementing-a-random-forest-classifier-in-python-979988d8a263

Great Learning Team. (2022, June). ***What is Cross Validation in Machine learning? Types of Cross Validation*****. Great Learning.**
https://www.mygreatlearning.com/blog/cross-

validation/#:~:text=The%20purpose%20of%20cross%E2%80%93validation,generalize%20to%20an%20independent%20dataset

Chatergee, M. (2022, December 13). *A Quick Introduction to KNN Algorithm*. Great Learning. https://www.mygreatlearning.com/blog/knn-algorithm-introduction/