8-2024

CRIME DATA PREDICTION BASED ON GEOGRAPHICAL LOCATION USING MACHINE LEARNING

Sai Bharath Yarlagadda

CRIME DATA PREDICTION BASED ON GEOGRAPHICAL LOCATION USING MACHINE LEARNING

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Computer Science

---

by

Sai Bharath Yarlagadda

August 2024

CRIME DATA PREDICTION BASED ON GEOGRAPHICAL LOCATION USING MACHINE LEARNING

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Sai Bharath Yarlagadda

August 2024

Approved by:

Dr. Yan Zhang, Advisor, Computer Science and Engineering

Dr. Jennifer Jin, Committee Member

Dr. Qingquan Sun, Committee Member

ABSTRACT

This project employs machine learning methods like K Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Decision Tree algorithms to monitor crime data based on location and pinpoint areas with risks. The project implements and tunes the four models to improve the precision of predicting crime levels. These models collaborate to offer a trustworthy evaluation of crime patterns. K Nearest Neighbors (KNN) categorizes locations by examining the proximity of data points considering coordinates and other factors to identify trends linked to increased crime data. Logistic Regression gauges the likelihood of crime incidents by studying the connection, between factors (like location and time) and the crime activity, assisting in forecasting crimes in various regions. Decision Tree Classifier uses a tree structure to make decisions based on feature values dividing the data into branches representing decision paths. This approach is particularly useful for identifying high-risk areas using crime data. Random Forest Classifier constructs decision trees and combines their results for classification purposes, resulting in enhanced prediction accuracy and robustness by merging outcomes from multiple trees, thus reducing the risks of overfitting and improving generalization to unseen data.

The system's efficiency is assessed using a crime dataset that includes information, about crime occurrences, geographical locations, and time-related data. Metrics, like accuracy, precision, and recall are employed to assess the model's ability to anticipate crimes and identify hotspots accurately.

ACKNOWLEDGEMENTS

I sincerely extend my thanks to my project committee, Dr. Yan Zhang (Advisor), Dr. Jennifer Jin (Committee Member), and Dr. Qingquan Sun (Committee Member). I want to thank my friends and family for their continuous support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Background

In today's rapidly advancing world, protecting public safety and reducing crime data are the most important concerns for communities and law enforcement agencies. Traditional methods of monitoring crime, such as periodic reports and statistical analyses, often lack the immediacy and accuracy required to effectively address crime-related challenges. However, with the advancement of technology and data-driven approaches, there are now opportunities to develop innovative solutions that can enhance crime prevention and awareness.

One effective solution involves implementing a crime data prediction that is based on geographical location. This system harnesses the capabilities of real-time data collection, analysis, and dissemination to deliver timely and location-specific crime information to both individuals and authorities. With the use of this technology, communities can remain updated about criminal activities in their area, thus empowering them to take proactive measures to safeguard themselves and their neighborhood.

K Nearest Neighbors (KNN) predicts crime likelihood by analyzing the similarity between a new location and existing crime data. Other effective models

for crime prediction include Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Logistic Regression forecasts crimes in various regions, while the Decision Tree Classifier identifies high-risk areas. The Random Forest Classifier enhances prediction accuracy by merging outcomes from multiple decision trees.

Motivation

We aim to enhance safety by equipping users with information, about crime incidents occurring in their local area. We want to empower individuals to make informed decisions about their safety by providing them with localized crime data. We strive to encourage community engagement and raise awareness about crime, in areas. Implementing crime prediction initiatives can help build trust in law enforcement organizations by showcasing their dedication to proactive crime prevention tactics. Open and clear communication regarding the creation and application of models can also strengthen the bond of trust, between law enforcement and the neighborhoods they protect.

Problem Statement

Predicting criminal activities involves creating models that can anticipate behaviors by analyzing data. These models examine patterns and trends, in incidents to offer insights for law enforcement to efficiently allocate resources and implement measures. While traditional statistical methods have their limitations in capturing relationships within crime data the use of machine learning techniques and methods, like K Nearest Neighbors (KNN), Logistic Regression, Decision

Tree Classifier, and Random Forest Classifier hold potential for enhancing the accuracy of predictions. Our goal is to boost the precision and dependability of the system, in forecasting crime risk levels by utilizing these models. Our approach includes steps; initially processing crime datasets by cleaning, standardizing, and refining features to align with machine learning models; then creating and training customized K Nearest Neighbors (KNN), Logistic Regression, Decision Tree Classifier, and Random Forest Classifier structures for predicting crime models undergoing training on crime data focusing on adjusting hyperparameters and applying regularization techniques to address overfitting concerns. After training, we evaluate model performance using confusion matrices and assess metrics like accuracy, precision, recall, and F1 score to determine ability and adaptability to data sets. Additionally, we utilize visualization tools such as bar graphs and pie charts to analyze the crime patterns according to statistical data from the crime dataset.

Challenges

This initiative is centered around creating a system that notifies users about crime data using machine learning techniques like K Nearest Neighbors (KNN), logistic regression, decision tree, and random forest classifier. These methods analyze crime data based on location and alert individuals about crime rates in their area. The primary aim is to furnish users with timely information to aid them in making informed choices regarding their safety and security. Obtaining trustworthy and comprehensive crime data poses difficulties due to

incompleteness, inconsistency, or bias in the data. Identifying attributes, such as types of crimes, location characteristics, and time of incidents, is essential for accurately predicting crime data. Selecting machine learning models, like K Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest, and refining them for improved accuracy, efficiency, and scalability is crucial. Ensuring the system can swiftly process and assess data in time for delivering notifications to users. The crime data must be capable of managing vast amounts of information and accommodating a growing user community without compromising efficiency. It should facilitate accurate results forecasting while ensuring user privacy and compliance with data protection regulations. Additionally, the models built on this data must be thoroughly assessed and confirmed to provide reliable forecasts.

Proposed System

Analyzing information about criminal activities includes specifics such as where and when the crimes occurred, the nature of the offenses, and potential socio-economic influences. This phase entails refining the data addressing any missing details handling outliers and formatting the data appropriately for analysis purposes. Determining which characteristics (or factors) are most crucial for forecasting behavior. This process might involve methods like crafting features or reducing dimensionality.

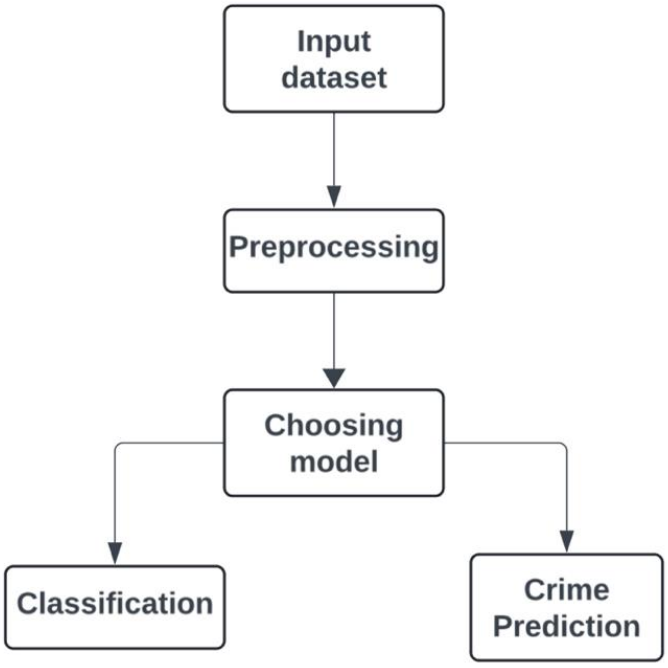

Figure 1. Classification of crime incidents as per data records

The process of analyzing crime data, which involves acquiring an input dataset, preparing the data for analysis through preprocessing, and selecting a suitable model based on the specific requirements of the task. This selected model is then applied for functions, like categorizing crimes or forecasting crime incidents.

Objectives of the Paper

Analyzing crime data involves gathering, refining, and converting information into a usable form. Key factors for forecasting crimes include selecting appropriate machine-learning methods. The algorithms are trained

using crime data and assessed using accuracy, precision, recall, and F1-score evaluation metrics. Time-based assessment is performed to identify high-crime areas, recognizing a distribution with more prevalent low-crime zones. This step helps in selecting the best-performing model. Once the model is trained and evaluated, it can be used to predict future crime incidents based on crime data inputs. Monitoring the performance of the system over time and updating the model as new data becomes available or as the characteristics of crime change.

CHAPTER TWO

LITERATURE REVIEW

Feng M, Zheng J, and Han Y Investigate the use of big analytics and mining for the analysis, visualization, and prediction of crime data. Their research focuses on applying advanced computational techniques to manage and analyze large volumes of crime-related data [12]. This approach aims to reveal patterns and trends that are essential for understanding criminal activities. By enhancing crime data analysis through big data analytics, the study provides law enforcement agencies and policymakers with the tools to make well-informed decisions and develop proactive crime prevention strategies. Additionally, the study emphasizes the importance of visualization of insights derived from the analysis. Overall, the research by Feng, and colleagues significantly contributes to the field of crime analytics by leveraging big data and computational methods to enhance crime prediction and prevention efforts.

Dash, Safro, and Srinivas Murthy propose an innovative method for predicting crimes using a network analytic approach that integrates spatial and temporal dimensions. Their research delves into the complex relationship between where and when crimes occur, employing network analysis techniques to uncover hidden patterns and correlations within the crime data[9]. By incorporating these spatial and temporal elements into their predictive models, the study aims to improve the accuracy and effectiveness of crime forecasting.

This approach allows law enforcement agencies to better anticipate and prevent criminal activities. The research offers a comprehensive framework for predictive analytics in crime prevention, providing valuable insights that help develop proactive strategies to tackle crime-related issues.

Chung, Hisen Yu's study underscores that crime patterns, though not entirely random, exhibit discernible trends rather than complete unpredictability. Understanding these patterns is crucial for developing targeted crime prevention strategies[28]. The study details efforts with a northeastern U.S. police department to create a crime forecasting model. It involved organizing extensive datasets derived from police records, encompassing various crime types, incidents, locations, and timings. Temporal features extracted from raw data were also considered. Using data mining classification techniques, the study evaluated several methods to forecast areas prone to high crime rates or potential increases in criminal activities. Ultimately, the research recommends a forecasting approach that integrates explicit spatial and temporal data to enhance the accuracy of predicting future criminal occurrences.

Baloian Nelson's research contributes to crime prediction by exploring the utilization of patterns and contextual factors in crime analysis [5]. The study focuses on employing computational methods to identify patterns and contextual cues associated with criminal activities. By examining environmental, social, and situational factors, the research aims to develop predictive models capable of pinpointing potential crime hotspots and trends. Integrating patterns and context

in crime prediction provides a holistic approach to understanding the dynamics of criminal behavior, enabling law enforcement agencies to implement targeted interventions and preventive measures effectively. The study underscores the significance of incorporating contextual information alongside pattern recognition techniques to enhance the accuracy and reliability of crime prediction systems.

Cesario, Catlett, and Talia investigate the enhancement of time series data prediction accuracy by integrating ARIMA with Daubechies wavelet transformation functions[9]. Their study assesses real-world datasets to evaluate the effectiveness of this method compared to other forecasting approaches. The results of their experiments emphasize the method's benefits and efficiency, demonstrating its potential for accurately forecasting data trends in practical scenarios.

.

CHAPTER THREE

DATA COLLECTION AND PREPROCESSING

Data Collection

We deal with data.detroitmi.gov to a single large CSV file [8]. The initial phase involves gathering information, like crime rates, demographics, and Incident patterns. Following that the data goes through a cleaning and transformation process to make it usable. The process involves cleaning and handling missing values in the data to ensure its integrity[13]. Data preprocessing involves identifying and prioritizing pertinent attributes to enhance model efficiency and predictive accuracy, ensuring robust analysis of the dataset. Once the features are chosen different machine-learning algorithms can be employed to train and predict outcomes. Lastly, we assess the accuracy and effectiveness of the trained models by utilizing performance accuracy in predicting crime[1].

Preprocessing

To begin should bring in the datasets I have collected for the machine learning project. To start the machine learning project, I need to import the collected datasets. sorting the dataset is an aspect of data preprocessing in machine learning. However, before you proceed with importing the dataset/s it's necessary to set the directory as your working directory. We will make a column in the new panda’s data frame. Data_set=pd.read_csv(“Dataset.csv”) to import the dataset in the directory to mount the file.

To begin we need to import the data that will be used in the machine learning algorithm. This is a step, in the preprocessing of machine learning. We import the collected data, for evaluation. After loading the data, it is essential to check for any missing content that may have occurred. Preparing data or input before analyzing or presenting it is known as preprocessing. This step may involve activities, like tidying up refining, or adjusting the data to enhance its accuracy, significance, or suitability for the desired application.

```
# prompt: suggest a model that gives good accuracy for the data in dataframe
from sklearn import preprocessing

label_encoder = preprocessing.LabelEncoder()
label_encoder_off_ca = preprocessing.LabelEncoder()


# Encode labels in column 'species'.
data['offense_ca']= label_encoder_off_ca.fit_transform(data['offense_ca'])
#data['zip_code']= label_encoder.fit_transform(data['zip_code'])
data['day_of_wee']= label_encoder.fit_transform(data['day_of_wee'])
data['hour_of_da']= label_encoder.fit_transform(data['hour_of_da'])
#scout_car_= label_encoder.fit_transform(data['scout_car_'])
#data['precinct']= label_encoder.fit_transform(data['precinct'])
data['Cities'] = label_encoder.fit_transform(data['Cities'])
data['council_di']= label_encoder.fit_transform(data['council_di'])
data['zip_code']= label_encoder.fit_transform(data['zip_code'])



# Separate features and target
X = data.drop('offense_ca', axis=1)
y = data['offense_ca']

# Get the categories associated with each label encoder value for each column
categories_offense_ca = label_encoder_off_ca.classes_
```

Figure 2. Preprocessing refining and optimizing

Preparing data frames, for machine learning involves cleaning, transforming, and structuring data. This process includes tasks such as managing missing values standardizing features, categorizing variables, and dividing data into training and testing sets. The objective is to refine the data to ensure its suitability and usefulness, for the modeling task.

Table 1. Data set segregation

Crime Category	Training Set	Testing Set
Property Crime	69,163	19,274
Violent Crime	42,615	11,827
Drug Crime	16,107	4,696
Misc Crime	9,361	2,327

The data is initially divided into two sets: a training set and a test set following an 80-20 split. The dataset is divided into a training set and a testing set to facilitate model development and evaluation. The training set comprises 137,246 records, categorized into four types of crime that is 69,163 records of property crime, 42,615 records of violent crime, 16,107 records of drug crime, and 9,361 records of miscellaneous crime. This set is used to train the model, allowing it to learn and identify patterns within the data. The testing set consists

of 38,124 records, also divided into the same four crime categories. Specifically, it includes 19,274 records of property crime 11,827 records of violent crime 4,696 records of drugs crime, and 2,327 records of miscellaneous crime. This set is employed to evaluate the model's performance, ensuring its accuracy and reliability in predicting and classifying various crimes based on the learned information. By using a separate testing set, we can objectively assess how well the model generalizes to new, unseen data, which is crucial for validating its real-world applicability.

CHAPTER FOUR

METHODOLOGIES

K Nearest Neighbors

The K Nearest Neighbors (KNN) technique stands out as a method, of learning that operates without parameters relying on instances to make predictions in classification and regression tasks within machine learning. Examining the 'K' nearest data points from the training set determines the class label or value of a data point. KNNs simplicity lies in its reliance on the idea that comparable data points often share characteristics or values. Unlike methods, KNN does not demand a training phase. Instead, it retains the entire training dataset for future reference during prediction tasks. The crucial factors of KNN encompass 'K' (the number of neighbors to consider) and the metric used to gauge similarity, between data points[29]. The KNN algorithm offers an advantage by not requiring a training phase meaning it doesn't need data preparation before use[25]. Instead, it stores the training data, in memory to make predictions. Analyzing crime data involves gathering information about activities, refining it, and utilizing advanced algorithms like K Nearest Neighbors (KNN) and Logistic Regression to predict and prevent crime[27]. However, the performance of the KNN algorithm is influenced by the choice of K value and distance metric used. The specific problem, at hand, will dictate the distance metric and value for K to achieve optimal results.

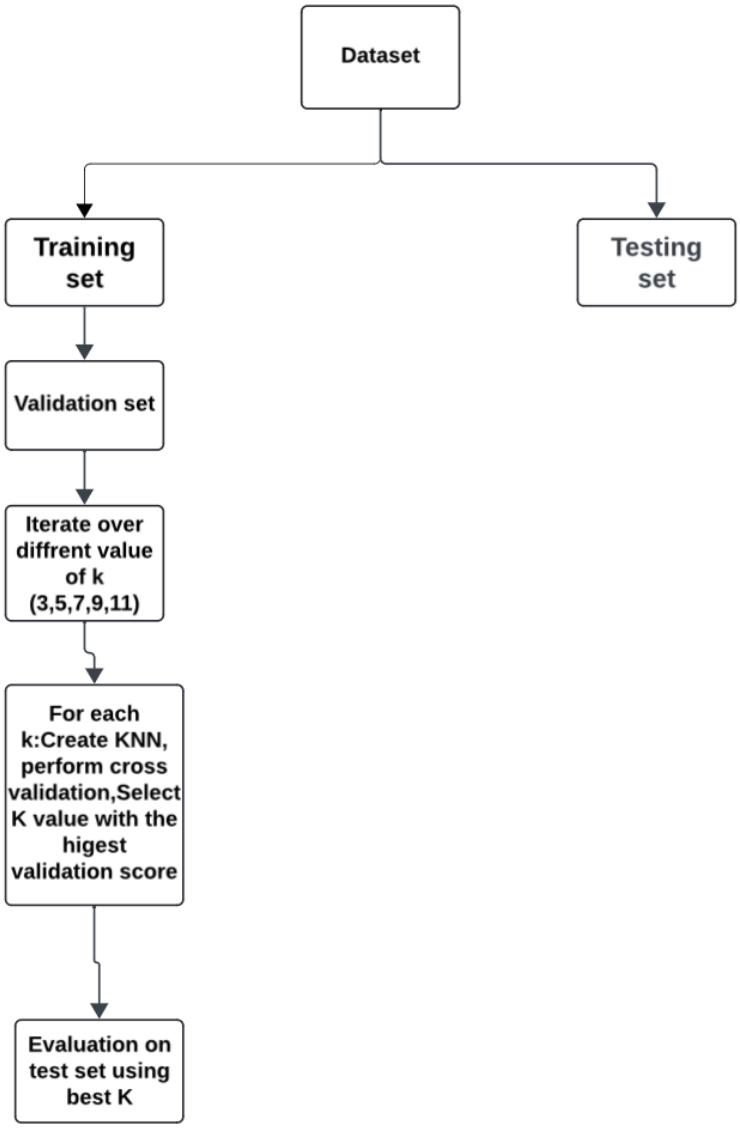

Figure 3. KNN crime prediction model

The original dataset is divided into three sets: training, validation, and testing. The training set is utilized to train the KNN model and conduct cross-

validation. In the process of model selection, we evaluate the KNN model to determine the optimal value K that maximizes its effectiveness. The test set is reserved for the final assessment of the model[31]. Experiment, with values of k (3, 5,7, 9, 11) to identify the model. For each k value carry out cross-validation on the training set train the model and assess its performance on the validation set. Utilize the performing model (selected based on validation results) to make predictions on the test set and calculate accuracy along, with performance measures. This structured approach ensures that the model's performance is rigorously assessed across different parameters and validated against independent data before final deployment.

Logistic Regression

Logistic regression is a predictive modeling employed to predict the likelihood of a result by considering one or more influencing factors. The process begins with the input data, which is subsequently fed into the model preparation and training phase. This phase involves training multiple models with different polynomial degrees: Degree 2, Degree 3, and Degree 7. Each model variant undergoes a cross-validation step, where cross-validation scores (CV scores) and the mean scores (Mean CV scores) are computed. This step helps in assessing the model's performance and stability[4]. Following cross-validation, the models proceed to the evaluation phase. This phase is divided into two parts: validation and testing. For the validation data, the evaluation includes metrics

such as accuracy, class report (precision, recall, F1-score), and confusion matrix. Similarly, for the testing data, the review consists of accuracy, classification report, and confusion matrix [22].

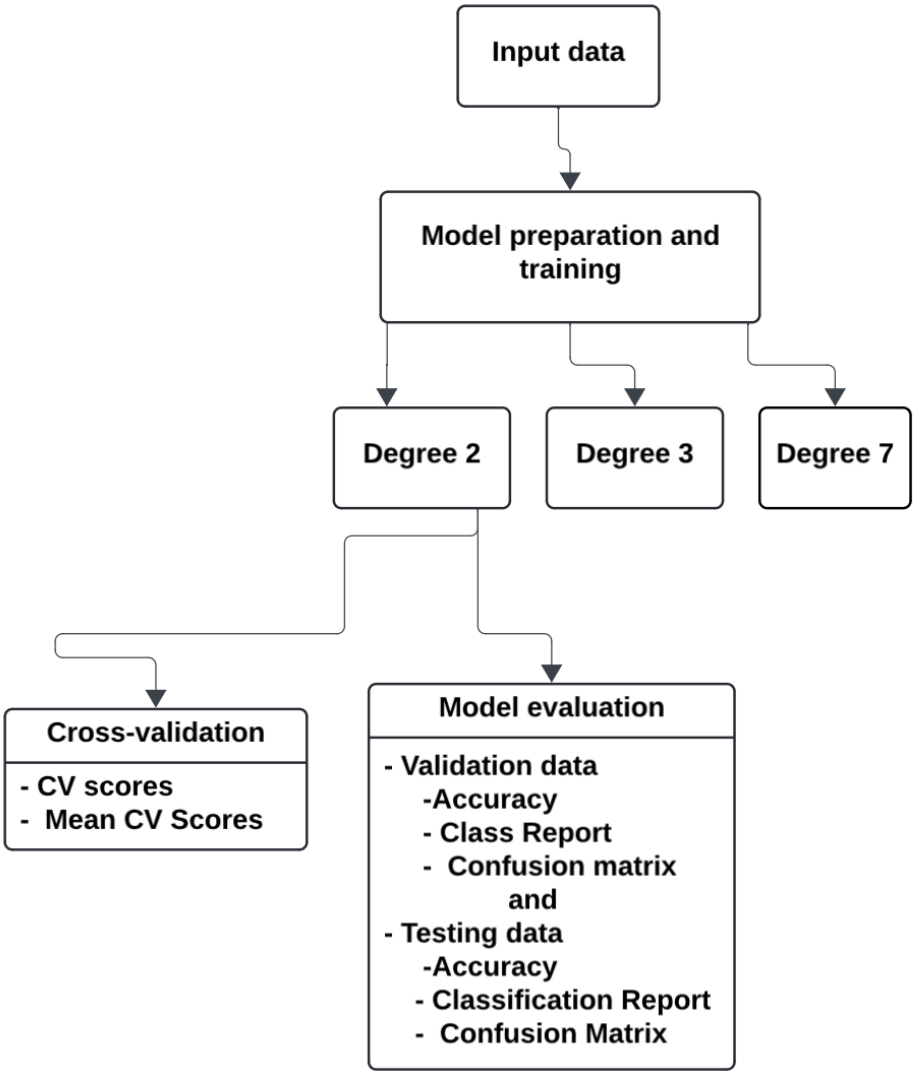

Figure 4. Logistic regression workflow

Decision Tree Classifier

Decision Tree Classifier workflow for implementing and evaluating machine learning models, focusing on both models. The process starts with data preprocessing, which involves cleaning, normalization, and splitting the data into training validation, and test sets [14]. The model training phase consists of training a Decision Tree model, evaluating its effectiveness using cross-validation, and fitting it on the training data ('x_train','y_train'). The model then predicts the validation set ('x_val') to assess its generalization capability [15]. Eventually, the workflow concludes with calculating and printing validation accuracy, classification report, and confusion matrix, which provide detailed insights into the model's performance. This structured approach ensures a thorough evaluation process to select the best-performing model. By iterative refining and validating the model, the workflow aims to achieve optimal predictive performance by identifying and leveraging the most effective Decision Tree classifier for the given dataset [7].

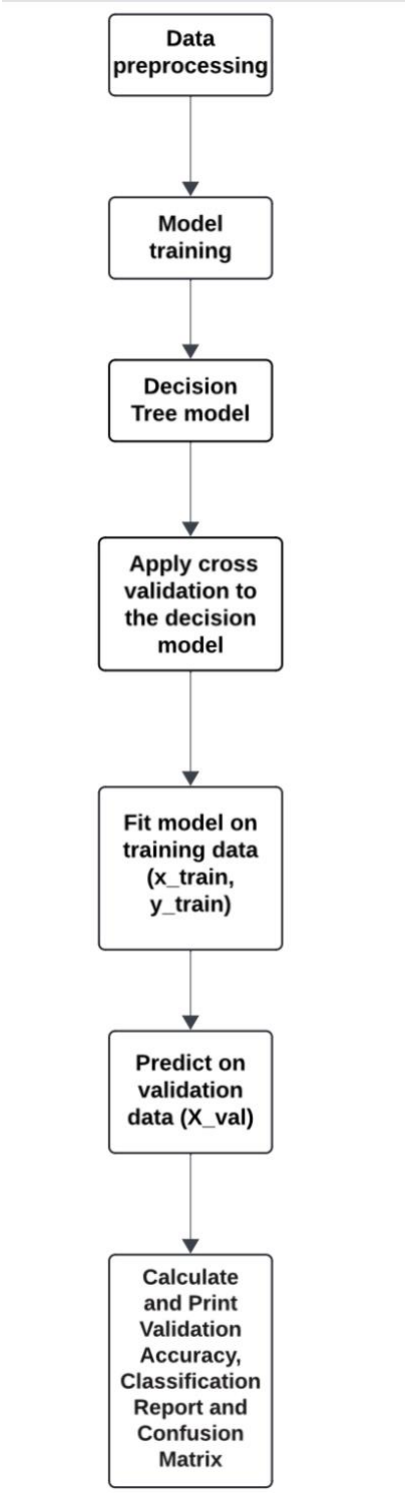

Figure 5. Schematic diagram for decision tree

Random Forest Classifier

Random Forest utilizes multiple decision trees to enhance accuracy and control overfitting with large, high-dimensional datasets. The process begins by splitting the data into training and test sets and fine-tuning the model parameters through grid search[7]. The best model setup is chosen based on performance metrics followed by training the model with the training set and assessing its accuracy by predicting labels, for the test set and creating metrics such, as a confusion matrix and classification report. It optimizes hyperparameters using grid search with cross-validation for model evaluation. Initially, the dataset is divided into training and testing sets, and then hyperparameter optimization is performed using grid search with cross-validation. The resulting model is evaluated using a confusion matrix and a classification report to measure predictive accuracy and other important metrics on the test set[28]. By systematically optimizing hyperparameters and rigorously evaluating model performance, the workflow ensures robustness and reliability in leveraging the Random Forest Classifier for effective predictive modeling[15].

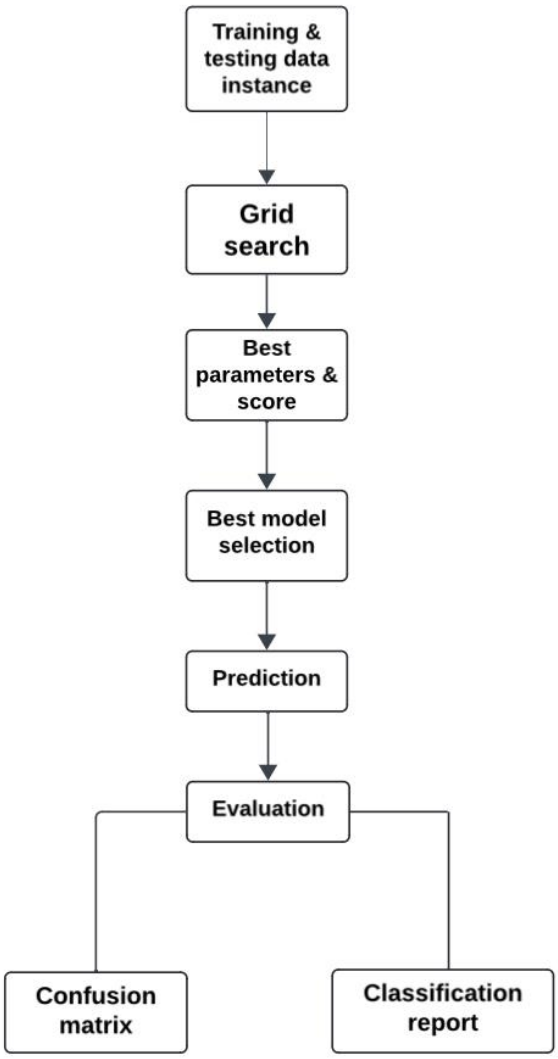

Figure 6. Random forest algorithm workflow

CHAPTER FIVE

EXPERIMENTAL RESULT

Evaluation Metrics

Assessing a trained model's performance, in machine learning entails using metrics and methods to gauge how effectively it can adapt to unseen data. Assessing and comparing are steps, in determining how well various models, algorithms, or methods perform[2]. These steps assist data scientists and professionals in making choices regarding which models to use, how to adjust parameters, and which methods are best suited for a task or dataset[32]. Different standards are utilized to assess machine learning assignments. In classification scenarios metrics such, as precision, recall, accuracy, and the F1 score are frequently employed [3].

Accuracy

The accuracy of the model is defined as the number of correctly predicted outputs out of all ground truths. It gives the percentage of how accurate the proposed model will be on testing [18].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The formula represents the accuracy formula for a classification model, which is the ratio of the sum of true positives (TP) and true negatives (TN) to the total number of instances (TP+ TN + FP + FN) [18].

Precision

Precision for a particular class c is calculated as the ratio of the number of true positives(correctly predicted instances of class c) to the sum of true positives and false positives instances incorrectly predicted as class c precision measures the ratio of predicted observations to all predicted positives [18].

$$Precision = \frac{TP}{TP+FP}. \qquad (2)$$

The formula for precision in a classification model is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (TP +FP). It measures the accuracy of positive predictions[18].

Recall

Recall for a particular class c measures the proportion of correctly predicted positive instances of class c among all actual positive instances of class c. It focuses on the ability of the classifier to find all positive instances of class c Recall represents the ratio of predicted observations to all actual positives [2].

$$Recall = \frac{(TP)}{(TP+FN)} \qquad (3)$$

The formula for recall in a classification model is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). It

measures the model's ability to correctly identify all relevant positive cases[2].

F1 score

F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it useful for imbalanced datasets [18].

$$F1score = \frac{(2*Precision*recall)}{(Precision+Recall)} \qquad (4)$$

The formula for calculating the F1 score is a metric commonly used in classification tasks to evaluate the accuracy of a model. The F1 score is the harmonic mean of precision and recall, which gives a balance between the two metrics [18].

Model Evaluation

In the crime prediction task and reviewing the accuracy results of the models I have compiled a summary table evaluating the methods used. Evaluating machine learning models involves using metrics, such as accuracy, precision, recall, and F1 score, which are specific to their intended task.

Logistic Regression

Table 2 shows the validation accuracy for models of different polynomial degrees. A model with a polynomial degree of 2 achieves a validation accuracy of 0.74, which is the highest among the degrees tested. A polynomial degree of 3 results in a lower validation accuracy of 0.68. The model with a polynomial degree of 7 has the lowest validation accuracy of 0.58. This indicates that as the

degree of polynomial increases, the validation accuracy tends to decrease, suggesting potential overfitting with higher-degree polynomials.

Table 2. Validation accuracies of logistic regression for degrees

Degree	Validation Accuracy
2	0.74
3	0.68
7	0.58

Table 3. Classification report of logistic regression

Crime Category	Precision	Recall	F1-score	Support
Property Crime	0.20	0.02	0.04	4741
Violent Crime	0.00	0.00	0.00	2369
Drug Crime	0.74	0.92	0.82	19169
Misc Crime	0.72	0.84	0.78	11845
Accuracy	0	0	0.73	38124
Macro avg	0.42	0.45	0.41	38124
Weighted avg	0.62	0.73	0.66	38124

Table 3 shows a model performance across different crime categories using precision, recall, F1-score, and support metrics. The model performs poorly

for property crime and violent crime, with very low precision, recall, and F1 scores. In contrast, it performs well for Drug Crime and Miscellaneous Crime, showing high recall and F1-score. The overall macro average metrics are modest, with precision at 0.42, recall at 0.45, and an F1-score of 0.41. The weighted averages are higher, with precision at 0.62, recall at 0.73, and an F1-score of 0.66.

Table 4. Confusion matrix for logistic regression

Crime Category	Property Crime	Violent Crime	Drugs Crime	Misc Crime
Property Crime	92	0	2980	1669
Violent Crime	36	0	1305	1028
Drugs Crime	316	0	17721	1132
Misc Crime	20	0	1893	9932

Table 4 shows the confusion matrix for a model predicting four crime categories: property crime, violent crime, drug crime, and miscellaneous crime. The model accurately predicts 92 instances of property Crime but misclassifies many as drug crime (2,980) and miscellaneous crime (1,669). Violent crime predictions are poor, with no accurate predictions and all instances misclassified, primarily as drug crime (1,305). The model performs well for drug crime, correctly

predicting 17,721 instances, though it misclassifies some as property crime (316) and miscellaneous Crime (1,132). Miscellaneous crime predictions are fairly accurate, with 9,932 correct predictions, but there are misclassifications into property crime (20) and drug crime (1,893). The performance highlights high misclassification rates for certain categories, particularly Violent Crime, while drug crime predictions are more accurate.

Decision Tree Classifier

Table 5 displays the validation accuracy of a model at different depths. At a depth of 5, the validation accuracy is 0.9031, which is the highest among the tested depths. As the depth increases to 10,15, and 20, the validation accuracy slightly decreases to 0.9021, 0.9023, and 0.9020, respectively. This indicates that the model performs best at a depth of 5, and further increases in depth do not improve validation accuracy. Among the evaluated depths (5,10,15,20) depth 5 achieved the highest 10-fold cross-validation accuracy of 0.9031 followed by depth 15 with an accuracy of 0.9023.

Table 5. Optimized decision tree performance and validation accuracy

Depth	Validation Accuracy
5	0.9031
10	0.9021
15	0.9023
20	0.9020

Table 6 shows a classification report evaluating a machine learning model's performance in predicting different crime categories. The model achieves high accuracy for drug crime with a precision of 0.96, recall of 0.96, and F1-score of 0.98, while it performs poorly for violent Crime, scoring 0.00 in precision, recall, and F1-score. Property crime and miscellaneous crime predictions are moderate, with respective F1-score of 0.73 and 0.90. The overall model accuracy is 0.89, but its effectiveness varies across crime categories, with a macro average F1-score of 0.65 and a weighted average F1-score of 0.86.

Table 6. Classification report of decision tree

Crime Category	Precision	Recall	F1-score	Support
Property Crime	0.82	0.66	0.73	4741
Violent Crime	0.00	0.00	0.00	2369
Drugs Crime	0.96	0.96	0.98	19169
Misc Crime	0.82	0.82	0.90	11845
Accuracy	0	0	0.89	38124
Macro avg	0.65	0.66	0.65	38124
Weighted avg	0.84	0.89	0.86	38124

Table 7. Confusion matrix for decision trees

Crime Category	Property Crime	Violent Crime	Drug Crime	Misc Crime
Property Crime	3118	0	0	1623
Violent Crime	671	0	800	898
Drugs Crime	0	0	19169	0
Misc Crime	16	0	0	11829

Table 7 shows a confusion matrix where the model correctly predicts Property Crime in 3118 cases but misclassifies 1623 as miscellaneous Crime. It fails to predict violent crime accurately, misclassifying instances as Property

crime, drug crime, or miscellaneous Crime. The model performs exceptionally for drug crime, correctly predicting all 19169 instances, and accurately predicting 11829 instances of miscellaneous crime, with minor misclassification errors. This highlights the model's high accuracy for drug crime and its significant struggles with violent Crime.

Random Forest Classifier

Table 8 shows the effectiveness of different parameter setups for a machine-learning model. it shows the depth, number of leaf, split criteria, and number of estimators, along with their corresponding validation accuracy. Notably, the model with a depth of 14, 2 leaf, 8 splits, and 120 estimators achieved the highest validation accuracy of 0.9053. Other setups, such as a depth of 12 with 3 leaf, 2 splits, and 150 estimators, and a depth of 18 with 4 leaf,9 splits, and 160 estimators, both reached a validation accuracy of 0.9049. The model with a depth of 20, 4 leaf, 10 splits, and 200 estimators resulted in a slightly lower accuracy of 0.9044.

Table 8. Optimized model parameter values and validation accuracy

Depth	Leaf	Split	Estimator	Validation accuracy
20	4	10	200	0.9044
14	2	8	120	0.9053
12	3	2	150	0.9049
18	4	9	160	0.9049

Table 9 shows the performance metrics of a machine-learning model for different crime categories. For property crime, precision is 0.84, recall is 0.65, and the f1-score is 0.73 based on 4741 instances. Violent crime has a precision of 0.69, recall of 0.19, and an f1-score of 0.30 from 2369 instances. Drug crime shows a precision of 0.96, recall of 1.00, and an f1-score of 0.98 from 19169 instances. miscellaneous crime has a precision of 0.84, recall of 0.98, and an F1-score of 0.91 from 11845 instances. The overall accuracy is misrepresented as 0.00. The macro average precision is 0.83, recall is 0.71, and the F1-score is 0.73. The weighted average precision is 0.89, the recall is 0.90, and the F1-score is 0.88. The support column indicates the number of instances for each crime category.

Table 9. Classification report of random forest classifier

Crime Category	Precision	Recall	F1-score	Support
Property Crime	0.84	0.65	0.73	4741
Violent Crime	0.69	0.19	0.30	2369
Drug Crime	0.96	1.00	0.98	19169
Misc Crime	0.84	0.98	0.91	11845
Accuracy	0.00	0.00	0.90	38124
Macro avg	0.83	0.71	0.73	38124
Weighted avg	0.89	0.90	0.88	38124

Table 10. Confusion matrix for random forest classifier

Crime Category	Property Crime	Violent Crime	Drug Crime	Misc Crime
Property Crime	3072	71	0	1598
Violent Crime	567	447	698	657
Drug Crime	0	51	19118	0
Misc Crime	36	80	0	11729

Table 10 presents the confusion matrix for different crime categories as predicted by the machine learning model. For property crime, the model correctly predicted 3072 instances, misclassified 71 instances as violent crime, none as

drug crime, and 1598 instances as miscellaneous crime. For violent crime, the model correctly predicted 447 instances, misclassified 567 as property crime, 698 as drug crime, and 657 as miscellaneous crime. For drug crime, the model correctly predicted 19118 instances, misclassified 51 as Violent Crime, and none as Miscellaneous or property crime. For miscellaneous crime, the model correctly predicted 11729 instances, misclassified 36 as property crime, 80 as violent crime, and none as drug crime.

K Nearest Neighbors

Table 11. Optimized model K Nearest Neighbors and validation accuracy

N_Neighbors(K)	Validation Accuracy
3	0.8864
5	0.8964
7	0.9008
9	0.9024
11	0.9027

Table 11 shows among the evaluated numbers of neighbors (K) in the KNN model, K=11 achieved the highest validation accuracy at 90.28%, demonstrating its superior predictive performance compared to K = 3,5,7, and 9. There is a consistent improvement in accuracy, highlighting K = 11 as the optimal choice for maximizing predictive performance in this scenario.

Table 12. Classification report of K Nearest Neighbors

Crime Category	Precision	Recall	F1-score	Support
Property Crime	0.79	0.66	0.72	4741
Violent Crime	0.63	0.21	0.31	2369
Drugs Crime	0.96	1.00	0.98	19169
Misc Crime	0.85	0.97	0.90	11845
Accuracy	0	0	0.90	38124
Macro avg	0.81	0.71	0.73	38124
Weighted avg	0.88	0.90	0.88	38124

Table 12 shows KNN model demonstrates strong accuracy across all crime categories, with each achieving 90% or higher. Drug crime is stands out with perfect precision and recall. The macro average shows robust performance, with precision, recall, and F1-score averaging 81%,71%, and 73% respectively. The weighted average reflects high overall performance metrics. This model

attains an accuracy of 0.90 without the need, for training. However, it comes with the drawback of being computationally intensive during prediction lacking scalability, with datasets, and demanding thorough preprocessing for accurate distance calculations

Table 13. Confusion matrix for K Nearest Neighbors

Crime Category	Property Crime	Violent Crime	Drugs Crime	Misc Crime
Property Crime	3138	108	31	1464
Violent Crime	573	488	690	618
Drugs Crime	4	63	19100	2
Misc Crime	279	119	15	11432

Table 13 shows a confusion matrix for the K Nearest Neighbors (KNN) property crime that occurred 3,138 times and included 108 incidents of violent crimes, 31 incidents of drug crimes, and 1,464 instances of miscellaneous crimes. Violent crime happened 573 times and involved 488 property crimes, 690 drug crimes, and 618 miscellaneous crimes. Drug crime occurred 4 times and included 63 incidents of violent crimes, 19,100 drug crime occurrences, and 2 miscellaneous crimes. Lastly, miscellaneous crime happened 279 times,

involving 119 violent crimes, 15 drug crimes, and 11,432 miscellaneous crime occurrences.

Model Comparison

Table 14 provides a comparative analysis of four different classification methods of Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K Nearest Neighbor across four effectiveness metrics accuracy, precision, recall, and F1-Score. The Logistic Regression model achieved an accuracy of 0.73, with a precision of 0.42, recall of 0.45, and F1-score of 0.41. The Decision Trees model demonstrated a higher accuracy of 0.89, with a precision of 0.65, recall of 0.66, and F1-score of 0.65. The KNN model performed similarly to Decision Trees, with an accuracy of 0.90, precision of 0.80, recall of 0.71, and F1-score of 0.73. The Random Forest model also showed an accuracy of 0.90, with a precision of 0.83, recall of 0.71, and F1-score of 0.73. These results indicate that both the KNN and Random Forest Classifier outperformed the Logistic Regression and Decision Tree Classifier in terms of accuracy, precision, recall, and F1-score.

Table 14.Comparison of various models during training

Method	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.73	0.42	0.45	0.41
Decision Tree Classifier	0.89	0.65	0.66	0.65
Random Forest Classifier	0.90	0.80	0.71	0.73
K Nearest Neighbor	0.90	0.83	0.71	0.73

CHAPTER SIX

SYSTEM DESIGN

This is essential since it allows for logical system understanding overall, of the system's circumstances. This section establishes all prerequisites. It would have an impact both internally and externally on the system. This illustrates how a user interacts with the system. As a result, we generate requirements analyze the system create use cases, and identify entities. It encompasses all of how users engage with the system-breaking down the system into components. Developing independent use cases is part of the process, for building use cases. This diagram eliminates redundancy. Users are responsible, for performing functions within the system environment.

Whenever accessing the application, it will retrieve the geographic location or zip code[11]. The system could be designed to give users real-time updates or regular updates of crime data so that they can always have the up, to date information, at their disposal. The app could utilize APIs (Application Programming Interfaces) offered by these sources to retrieve crime data. This might entail submitting a request containing the user’s location details to obtain crime information, for that area [5].

Component Diagram

The diagram illustrating the crime data prediction system consists of elements, including data processing. The data processing element readies the dataset, for forecasting. Classifier elements employ algorithms to estimate crime data using the processed information. This structured design guarantees expandable crime data forecasts prioritizing prediction accuracy.

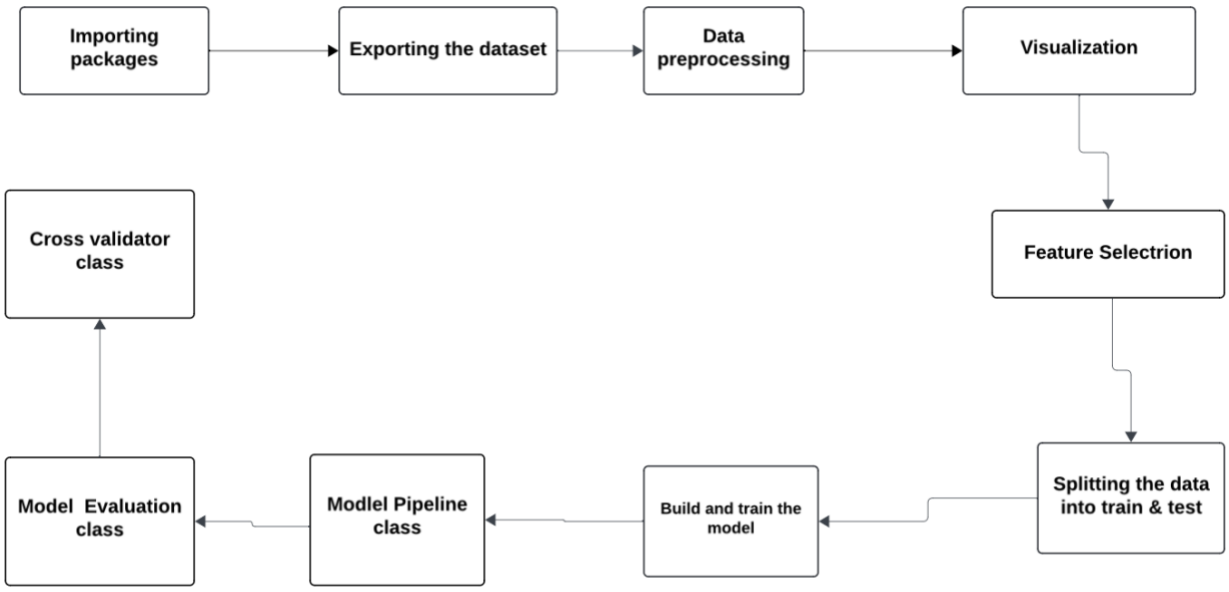

Figure 7.Component diagram

Class Diagram

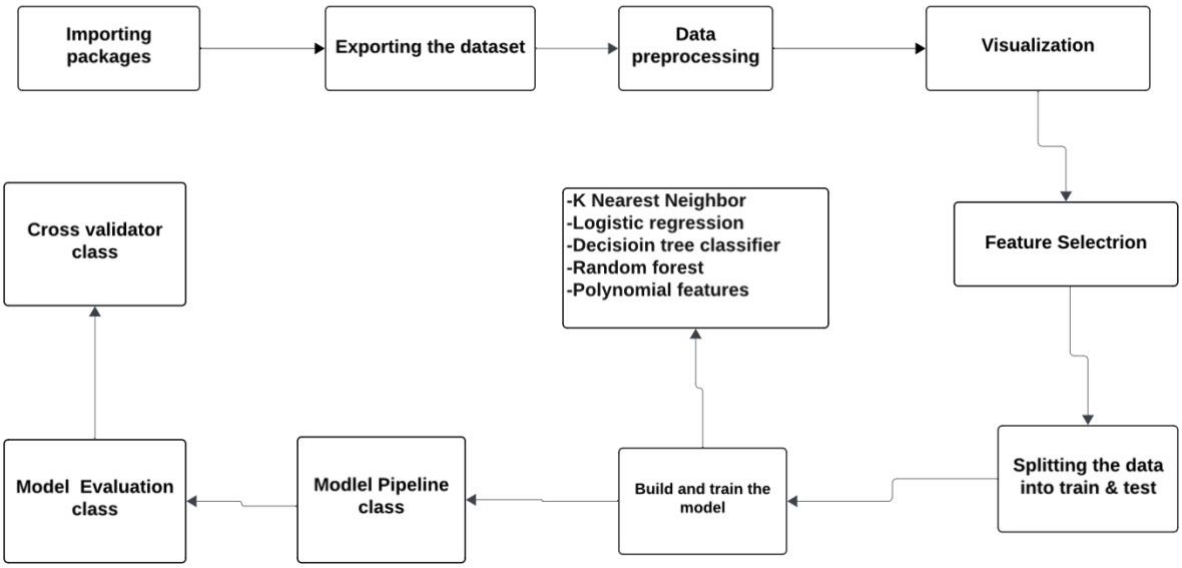

Figure 8. Class diagram

The class frame shows a total evolution plan, Creating and refining machine learning models, like K Nearest Neighbor, Logistic Regression, Decision Tree Classifier, Random Forest, and Polynomial Features. Implementing steps or setting up a pipeline for model input. The outcome of the machine learning process following assessment and validation. Selecting the most important features, for enhancing the model's performance and simplifying its complexity.

Privacy Concern

When it comes to predicting crime using zip code information it's important to make sure the data is anonymized and grouped to safeguard individual's identities ensure data protection and adhere to rules. Being open, about how the data is used, getting consent, and dealing with biases in the models are aspects. Conducting assessments on privacy impacts helps handle risks and safeguard people's privacy. Additionally, this system encourages an exchange of information that builds trust and collaboration, between the community and law enforcement agencies. Through mapping tools, it displays crime data geographically enabling users to pinpoint high crime areas monitor crime trends and make choices regarding their safety.

CHAPTER SEVEN

CONCLUSION

In our research, we assessed how well the K Nearest Neighbors (KNN) Logistic Regression, Decision Tree Classifier, and Random Forest Classifier models performed in predicting crime using machine learning methods. We gained insights into each model's effectiveness by analyzing the evaluation metrics from the confusion matrices and the classification report. Below is a comparison and summary based on our discoveries.

The K Nearest Neighbors (KNN) model demonstrated capabilities for crimes. Although it was easy to implement and understand its accuracy and F1 score were lower when compared to models. KNN accurately identified several actual crime cases. However, its precision was not as high indicating a likelihood of positives.

Logistic Regression excelled by striking a balance between precision and recall. It boasted accuracy with high precision suggesting fewer false alarms. This model proves beneficial in situations where the cost of positives is significant such as in legal proceedings. Moreover, the interpretability of Logistic regression simplifies understanding the connection, between features and outcomes.

The Decision Tree classifier offered interpretability and straightforward visualization of decision-making processes. Nonetheless, its performance metrics

slightly lagged those of random forest. While it did well in accuracy it showed some ups and downs. Tended to get too fixated leading to lower precision and recall compared to the other methods working together. This makes it less dependable when dealing with data.

The Random Forest classifier stood out as the performer among the models. It scored the highest, in accuracy and F1 score showing a balance between precision and recall. Its collaborative approach helped tackle overfitting issues. Delivered results across various measures. With its precision and recall Random Forest proves effective in predicting crimes while minimizing errors in both directions. This quality makes it a great fit for real-world applications, in crime forecasting, where these factors are critical.

Future Work

Predicting crimes involves analyzing data and using machine learning methods to anticipate the locations and times of activities. This method usually relies on crime records, societal factors, population demographics, and environmental conditions to build models. By recognizing patterns and trends, in incidents law enforcement organizations can distribute their resources more efficiently and prevent crime in advance. Nonetheless, it is crucial to acknowledge the ethical issues and prejudices present, in the data and algorithms applied for crime prediction.

REFERENCES

1. A. Almaw and K. Kadam. “Crime Data Analysis and Prediction Using Ensemble Learning”. Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS).Madurai, India June 2018.
2. Altameem.T and M. Amoon."Predicting Crime Activity by Combining the Firefly Optimization Technique with Map Neural Networks”. Neural Computing and Applications. vol 31, May. 2019, pp. 1234-1245, doi: 10.1007/s00521-018-3561-7.
3. A. Azhari and P. E. P. Utomo. “Prediction of the Crime Motorcycles of Theft using ARIMAX-TFM with Single Input". Proceedings of the 2018 Third International Conference on Informatics and Computing (ICIC), Oct. 2018, pp. 1-6.
4. Azeez. J, and Aravindhar, D. J. "Hybrid approach to crime prediction using deep learning". In 2015 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), Kochi, India, pp. 1364-1370, doi: 10.1109/ICACCI.2015.7275858.
5. Baloian. N, F. Rodriguez, J. Smith, L. Johnson, and R. Patel. “Crime prediction using patterns and context”. Proceedings of the 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in

Design (CSCWD), Wellington, New Zealand, Apr. 2017, pp. 2-9, doi: 10.1109/CSCWD.2017.8066662.

6. Borowik.G et al. "Time series analysis for crime forecasting". Proceedings of the 2018 26th International Conference on Systems Engineering (ICSEng),Sydney, NSW, Australia,2018, (pp. 1-10),doi: 10.1109/ICSENG.2018.8638179.
7. B. Mwaniki, T. Mwalili, and K. Ogada."Crime prediction using decision trees, random forests, and hybrid algorithm: A comparative analysis". Proceedings of the 2023 7th International Conference on New Media Studies (CONMEDIA), Bali, Indonesia, Oct. 2023, (pp. 99-104),doi: 10.1109/CONMEDIA60526.2023.10428374.
8. City of Detroit. (2020). RMS Crime Incidents. Retrieved from https://data.detroitmi.gov/datasets/rms-crime-incidents (Accessed June 2, 2020).
9. Dash. S.K, I. Safro, and R.S. Srinivasa Murthy."Spatio-temporal prediction of crimes using network analytic approach". Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), December 2018, pp. 1912-1917,doi: 10.1109/ICSENG.2018.8638230.
10. E. Cesario, C. Catlett, and D. Talia. "Forecasting crimes using autoregressive models". Proceedings of the 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing (DASC), 14th International Conference on Pervasive Intelligence and Computing (PICom), 2nd

International Conference on Big Data Intelligence and Computing (Big Data), Auckland, New Zealand, Aug. 2016, (pp. 104-110).

11. Esan O. A. and I. O. Osunmakinde. “Crime Prediction Linked to Geographical Location with Periodic Features for Societal Security”. Proceedings of the 2023 15th International Conference on Computer Research and Development (ICCRD), Hangzhou, China, (pp. 6-14), doi:10.1109/ICCRD56364.2023.10080235.
12. Feng, M., Zheng J, Han Y. ”Big data analytics and mining for crime data analysis, visualization, and prediction”. In International Conference on Brain Inspired Cognitive Systems, 2018, pp. 605–614.
13. I. S.,Saini, & Kaur, N . “The power of predictive analytics: Forecasting crime trends in high-risk areas for crime prevention using machine learning”. In 2023 14th International Conference *on* Computing Communication and Networking Technologies (ICCCNT) pp. 1-10. Delhi, India. doi: 10.1109/ICCCNT56998.2023.10306731.
14. Li, A ,M. Y. Shalaginov, A. Tao, and T. H. Zeng. “Investigation of Racial Bias in Property Crime Prediction by Machine Learning Models”. Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, Dec. 2023, pp. 2253-2256,doi: 10.1109/ICMLA58977.2023.00340.
15. Mandalapu. V, R. Kumar, J. Smith, L. Johnson, and R. Patel. “Crime prediction using machine learning and deep learning: A systematic review and

future directions". IEEE Access, vol 11, pp. 60153-60170,2023, doi: 10.1109/access.2023.3286344.

16. Mao.L and W. Du. "A method of crime rate forecast based on wavelet transform and neural network". International Journal of Embedded Systems, vol 11, pp. 731-737, 2019, doi: 10.1504/IJES.2019.103990.
17. McClendon.L and N. Meghanathan."Using Machine Learning Algorithms to Analyze Crime Data". Machine Learning and Applications: An International Journal (MLAIJ), vol 2, pp. 1-52,2015, Mar. 10, 2019.
18. M. Mudgal, D. Punj, and A. Pillai. "Theoretical and Empirical Analysis of Crime Data" Journal of Web Engineering vol 20, pp. 113-128, January 2021, doi: 10.13052/jwe1540-9589.2016.
19. M. Saraiva, I. Matijoˇsaitienˇe, S. Mishra, and A. Amante."Crime prediction and monitoring in Porto, Portugal, using machine learning, spatial and text analytics".ISPRS International Journal of Geo-Information, Vol 11, p. 400,2022,doi : 10.3390/ijgi11070400.
20. Mekala, Sasirekha, S. P., & Reshma, R. "Predicting high-risk areas for crime hotspots using hybrid KNN machine learning framework". In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 848-852, Coimbatore, India. doi: 10.1109/ICIRCA57980.2023.10220738.
21. Pratibha. A, A. Gahalot, S. Uprant, S. Dhiman, and L. Chouhan."Crime Prediction and Analysis". Proceedings of the 2nd International Conference on

Data, Engineering and Applications (IDEA),Bhopal, India, Dec. 2020, pp. 1-6,doi: 10.1109/IDEA49133.2020.9170731.

22. S. Abdullah, F. I. Nibir, S. Salam, A. Dey, M. A. Alam, and M. T. Reza. “Intelligent crime investigation assistance using machine learning classifiers on crime and victim information”. Proceedings of the 2020 23rd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, Dec. 2020, pp. 1-4,doi: 10.1109/ICCIT51783.2020.9392668.
23. S. Biswas, S. Ghosh, S. Roy, R. Bose, and S. Soni. “A study of stock market prediction through sentiment analysis".Mapana Journal of Sciences, vol 22, pp. 89-103, Feb. 19, 2023,doi:10.12723/js.64.6.
24. N. Tasnim, I. T. Imam, and M. M. A. Hashem, “A Novel Multi-Module Approach to Predict Crime Based on Multivariate Spatio-Temporal Data Using Attention and Sequential Fusion Model", IEEE Access, vol 10, pp. 48009-48030,2022, doi: 10.1109/ACCESS.2022.3171843.
25. R. M. Alfian and K. M. Lhaksmana, “Classification of Student Work Readiness Using the Decision Tree and KNN Methods,” ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, 2024, pp. 1-5, doi:10.1109/ICETSIS61505.2024.10459355.
26. Wawrzyniak, Z. M, J. Kowalski, and A. Nowak. “Data-driven models in machine learning for crime prediction”. Proceedings of the 2018 26th

International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, Dec. 2018,pp. 1-8, doi: 10.1109/ICSENG.2018.8638230.

27. W. Safat, S. Asghar, and S. A. Gillani. "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques", IEEE Access, Vol 9, pp. 70080-70094,2021,doi: 10.1109/ACCESS.2021.3078117.
28. Yu.C. H. “Crime Forecasting Using Data Mining Techniques”, Proceedings of the 11th International Conference on Data Mining Workshop,2011, pp. 779-786.
29. Y. Sukhdeve, A. Kumar, A. Verma, G. Shinde, and N. Lal. “Crime Prediction Using K-Nearest Neighboring Algorithm”, Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India,2020,doi: 10.1109/ic-ETITE47903.2020.155.
30. Y. -T. Lee, M. -S. Baek, W. Park, J. Park, and K. -H. Jang. "Smart Policing Technique with Crime Type and Risk Score Prediction Based on Machine Learning for Early Awareness of Risk Situation", IEEE Access, vol 9, pp. 131906-131915,2021, doi: 10.1109/ACCESS.2021.3112682.
31. X. Zhang, J. Smith, L. Johnson, and R. Patel. “Comparison of machine learning algorithms for predicting crime hotspots", IEEE Access, vol 8, pp. 181302-181310,doi: 10.1109/ACCESS.2020.3028420.

32. Z. Liu and H. Chen. “A predictive performance comparison of machine learning models for judicial cases”, Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI),2017.