Global Scientific JOURNALS

GSJ: Volume 11, Issue 11 November 2023, Online: ISSN 2320-9186

www.globalscientificjournal.com

AN ENSEMBLE MACHINE LEARNING MODEL FOR CRIME CLASSIFICATION AND PREDICTION

Nahum Emmanuel[1],Dr. John Abiodun Oladunjoye[2], Orji Mary[3], Shaimaki Hephzibah Danlami[4],Wella Rande [5]
Taraba State College of Health Technology Takum, Department of Computer Science, Taraba State, Nigeria[2.]
Department of computer science,faculty of computing and information sciences Federal University Wukari, Taraba State, Nigeria[2.]
Department of Computer Science, School of Basic and Applied Sciences,Taraba State Polytechnic Suntai, Jalingo Campus Taraba State, Nigeria[3&4].
Department of Information Science, Scho

ol of Basic and Applied Sciences, Taraba State Polytechnic Suntai, Jalingo Campus, Taraba State, Nigeria[5].

E-mail: Marybless200@yahoo.com

Keywords: Crime, Data mining, Decision trees, Ensemble learning,Machine learning. Evaluation metrics.

1. Abstract

The most serious security challenges we face in these turbulent times are terrorist attacks and the transmission of disease. Length and breadth are measured in hundredths of a centimeter. On a daily basis, we see the most minor offences committed by ordinary citizens. Details of breaches and recurring cases of items should be applied to files to ensure that they are up to date. When it is known that a crime has been committed, people believe that disciplinary action will be taken, even if there is no means of knowing which one. The research of criminology helps to broaden our understanding of who is likely to become a suspect. In the midst of his attempts to identify and deter alleged criminals from reoffending the legal system, he is incorporating both computer science and deep learning. Anyone interested in learning more about the workings of the Chicago Police Force should visit "The Chicago Police Department Site." The Crime Timeline will keep track of all criminal activity as well as the time and date of any incident that occurs. The data collection and modelling have been completed; all that remains is on-line modelling and compilation. To address this question, we must first determine if the case history of K-grooming and other related methods will help with criminal prediction. The invention is typically

used as a testing tool, but it can also be used in conjunction with other technologies. Based on internal or external metrics, an algorithm can estimate how easily law enforcement authorities may be able to track, anticipate, and cope with, or preempt, risks, such as the ratio of those sentenced to those arrested, with a life sentence to those awaiting the risk of life imprisonment

2.Introduction

Crime for decades has been a socio-economic disaster affecting human daily life activities and qualities with a negative impact on economic growth (Kim, Joshi, Kalsi, and Taheri, 2018). On a daily basis, crime rates are increasing as modern technologies and hi-tech emerge, thus, negatively optimized by criminals in achieving illegal activities such as burglary, theft, homicide, arson, murder, etc., (Kim *et al.*, 2018). The specifics of how crimes are conducted change depending on the type of society and community. Previous research conducted by researchers and criminologists in response to crime prediction reveals that some factors such as education, poverty, employment, and climate affect the crime rate (Kshatri, Singh, Narain, Bhatia, Quasim, and Sinha 2021). Hence, for a crime to occur, there must be victims, offenders, and properties of some targeted location. Consequentially, (Hemant, Haresh, and Majid, 2019) defined crime as an act that seems to violate and breach an existing law of a state. This implies that, for any crime to occur, there must be a motivated offender; a suitable target, and the absence of someone or something like close circuit television that can act as a protector. A study conducted by the United Nations Office on Drugs and Crime (UNODC) in 2019 observed that crime fuels corruption infiltrates business, and politics, and destabilizes the growth and development of any society. The UNODC further maintained that crime undermines governance while generating a level of insecurity among individuals with extreme fears instills in each individual (Okpuvwie and Toko, 2020). Crime affects the psychological, financial, physical, and spiritual well-being of the victims (Eidell and Ellis, 2010). Hence, the effects of crime on the socio-economic stability of society cannot be overemphasized and thus demand critical crime analysis.

Understanding the causes of crime has been a longstanding issue on the researcher's agenda. While it is a hard task to extract causality from data, several linear models have been proposed to predict crime through the existing correlations between crime and urban metrics (Luiz *et al.*, 2018). The use of data mining, machine learning, and deep learning techniques has proven significant in predicting crime occurrence from crime report datasets (Kim *et al.*, 2018; Kshatri *et*

al, 2021; Shanjana, and Porkodi, 2021; Ginger and Mihaela, 2017; Luiz *et al.*, 2018). The research to extend its efficiency and effectiveness is thus at its current trend as the need to curtail crime occurrence has been in high demand for decades and centuries. Hence, this study after an extensive review of several kinds of literature and their respective implementation of crime prediction analysis proposes the stacking of some machine learning models in crime prediction and analysis.

2.1. Types of Crimes

According to Charles (2021), there are two basic classifications of crime, namely property crime, and violent crime, and other kinds of crimes can be classified under each of the two basic crimes.

Property Crimes: a crime committed by an individual who damages, destroys, or steals someone else's property. Stealing a car and vandalizing a building are examples of property crimes. Such types of crimes are most common in the United States.

Violent Crimes: this is a crime that occurs when someone harms, attempts to harm, threatens to harm, or conspires to harm someone else. Violent crimes in most cases often involve force or threat of force and include crimes such as rape, robbery, and homicide.

Some crimes can both be property crimes and violent crimes. An instance of such crime includes carjacking someone's vehicle at gunpoint and robbing a convenience store with a handgun.

3.Related work

Jha, Jha, and Sharma, (2019) proposed a big data and Machine Learning technique for behaviour analysis and crime prediction. This paper discusses the tracking of information using big data, different data collection approaches, and the last phase of crime prediction using Machine Learning techniques based on data collection and analysis. The authors conducted a predictive analysis through Machine Learning using RapidMiner by processing historical crime patterns. Sohrab, Ahmed, Imran, Mohammed, and Iqbal, (2020) on crime prediction using spatiotemporal data proposed a data-driven system that predicts crimes by analyzing San Francisco city criminal activity data set for 12 years. The authors utilized the Decision tree (DT) and k-nearest neighbour (KNN) algorithms in predicting crime. They noticed that the two optimized algorithms provide low accuracy in prediction, and thus applied the random forest as an ensemble method with Adaboost

used as the boosting method to increase the accuracy of the prediction model. To evaluate and measure performance, they used log-loss classifiers to penalize false classifications. Because the dataset contains high-class imbalance problems, a random undersampling method for the random forest algorithm to give the best accuracy. Finally, the authors recorded a final accuracy of 99.16% with 0.17% log loss.

Prithi, Aravindan, Anusuya, and Kumar, (2020) developed a graphical user interface-based prediction of crime rates using a Machine Learning approach. The main focus of their study was to investigate machine learning-based techniques with the best accuracy in predicting crime rates and explore their application with particular importance to the dataset. The authors utilized the viabilities of Supervised Machine Learning techniques to analyze the dataset and hence conducted data validation, data cleaning, and data visualization. The results of the different supervised ML algorithms were compared to predict the results. Their proposed systems incorporate methods such as data collection, data preprocessing, and construction of a predictive model, dataset training, dataset testing, and a comparison of algorithms. Zhang, Liu, Xiao, and Ji, (2020) in a research on machine learning algorithms comparison utilized historical data of public property crime from 2015 to 2018 from a section of a large coastal city in the southeast of China as research data to assess the predictive power between the several machine learning algorithms. Results based on their historical crime data analysis suggest that the Long Short-Term Memory (LSTM) model outperformed K-Nearest Neighbor (KNN), Random Forest, Support Vector Machine, Naïve Bayes, and Convolutional Neural Networks. Furthermore, the authors built the environment data of points of interest (POIs) and urban road network density as input into the LSTM model as covariates. They discovered that the model with built environment covariates has a better prediction effect compared with the original model that is based on historical crime data. Therefore, concluded that future crime prediction should take advantage of both historical crime data and covariates associated with criminological theories.

Bandekar and Vijayalakshmi (2020) in their study focused on the analysis and design of ML algorithms to reduce crime rates in India. Machine Learning techniques were applied to a large set of data to determine the pattern relations between them. Their

research was mainly based on providing a prediction of crime that might occur based on the occurrence of previous crime locations. The techniques optimized by the authors include the Bayesian Neural Networks, the Levenberg Marquardt algorithm, and a scaled algorithm was used to analyze and interpret the data, among which the scaled algorithm gave the best result in comparison with the other two techniques. A statistical analysis based on correlation, analysis of variance, and graphs proved that with the help of the scaled algorithm, the crime rate can be reduced by 78%, implying an accuracy of 0.78 as noted in Bandekar and Vijayalakshmi (2020).

Hossain, Abtahee, Kashem, Hoque, and Sarker, (2020) proposed a system that predicts crime by analyzing a dataset containing records of previously committed crimes and their patterns. Their proposed system works mainly on two Machine Learning (ML) algorithms: a Decision Tree (DT) and KNN. Techniques such as the random forest algorithm and Adaptive Boosting were used by the authors to increase the accuracy of the prediction model. Safat, Asghar, and Gillani, (2021) toward the empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques applied different machine learning algorithms, namely, logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbours (KNN), decision tree, multilayer perceptron (MLP), random forest, and eXtreme Gradient Boosting (XGBoost), and time series analysis by Long-Short Term Memory (LSTM) and autoregressive integrated moving average (ARIMA) model to better fit the optimized crime dataset. The performance of LSTM for time series analysis was reasonably adequate in order of magnitude of root mean square error (RMSE) and mean absolute error (MAE), on both data sets as noted by the authors. Exploratory data analysis predicts more than 35 crime types and suggests a yearly decline in the Chicago crime rate and a slight increase in the Los Angeles crime rate; with fewer crimes occurring in February as compared to other months as revealed by the authors.

Matereke, Nyirenda, and Ghaziasgar, (2021) presented a detailed evaluation of three Spatiotemporal deep learning architectures for crime prediction. The network architectures include the Spatio Temporal Residual Network (ST-ResNet), the Deep Multi-View Spatio Temporal Network (DMVST-Net), and the Spatio Temporal

Dynamic Network (STD-Net). The authors trained the architectures using the Chicago crime data set. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used as performance metrics to evaluate the models. Results show that the STD-Net achieved the best results of the three approaches, with an accuracy of 0.89, RMSE of 0.2870, and MAE of 0.2093. The ST-ResNet and DMVST-Net also showed considerable promise as claimed by the authors. The ST-ResNet achieved an accuracy of 0.83, RMSE of 0.4033, and an MAE of 0.3278 while the DMVST-Net achieved an accuracy of 0.79, RMSE of 0.4171, and an MAE of 0.3455.

Saminu, Folorunso, Johnson, Akerele, Ilesanmi, and Ajayi, (2022) on vehicle theft crime prediction optimized the Adaptive Neuro-Fuzzy Inference System (ANFIS) - a computational Artificial Intelligence (AI) technique to develop a model for minimizing investigation time and the number of deployed security operatives towards achieving a high success rate in the prediction, detection, and recovery of stolen vehicles. The authors utilized a collection of vehicle theft and recovery data for (6) six consecutive years with fourteen (14) attributes collated by the Criminal Investigation Department of the Nigeria Police Force, Abeokuta, Ogun state, and analyzed them through the Dimensionality Reduction technique and Routine Activity Theory (RAT) approach to extract the most significant features. The datasets were subdivided into 60%, 20%, and 20% for training, testing, and validating the model respectively. The authors claimed a significant result of 92.91% obtained with the Adaptive Neuro-Fuzzy Inference System (ANFIS) model which signified that it is most efficient in predicting, detecting, and recovering stolen vehicles as compared with other machine learning algorithms such as Random Tree, Naïve Bayes, J48 and Decision Rule of prediction accuracies of 86.51%, 71.24%, 67.68%, and 55.73% respectively as claimed by Saminu *et al.*, (2022).

4.The Proposed System

Research conducted has revealed that the most successful result be it machine learning tasks or others was successfully achieved due to extensive structuring of procedures and strategies needed to achieve the targeted schedule via distinct, effective, and effective methodology. Hence, it becomes very important to engage a methodology to achieve the best performing and precise crime classification system. Therefore, the methodology for this study encompasses

three basic approaches. At first, a data cleansing was conducted to remove Null or Na features from dataset records before the conductance of feature selection. The feature selection approach proposed by this study is the use of the chi-square feature selection techniques. The approach to chi-square involves ranking the features based on their feature importance. To further cleanse the selected feature values, this study encodes the values of the attributes before scaling it to a range between 0 and 1 to reduce the computational overhead for the models while enabling the models to converge at a good result within a finite time. The second phase of the methodology involves the stacking of two machine learning models namely the Decision Tree, Support Vector Machine, and K-Nearest Neighbor Classifier before feeding the stacked models with the filtered dataset. Lastly, to validate the performance of the stacked and individual model for accuracy confirmation accuracy, the third phase entails the evaluation of the model performance based on the accuracy metrics proposed. The three phases of the proposed methodology are shown in Figure 4.1 below.

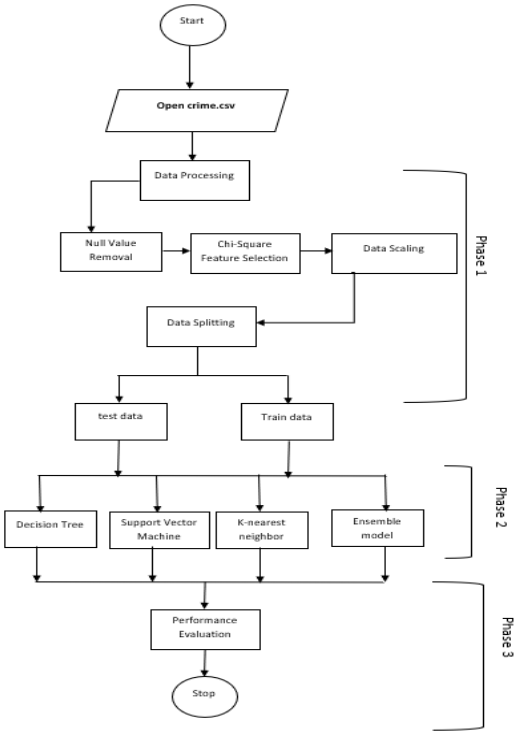

Figure 4.1: Methodology Adopted

4.1 Data Source and Description

The main task in the dataset is to predict the crime category based on a given set of geographical and time-based variables. The content of the dataset contains incidents derived from the SFPD (San Francisco Police Department) Crime Incident Reporting system. The attribute of the dataset:

i. **Dates** - timestamp of the crime incident.
ii. **Category** - category of the crime incident (only in train.csv).
iii. **Descript** - detailed description of the crime incident (only in train.csv).
iv. **Day of Week** - the day of the week.

v. **Pd District** - the name of the Police Department District.
vi. **Resolution** - how the crime incident was resolved (only in train.csv).
vii. **Address** - the approximate street address of the crime incident
viii. X – Longitude
ix. **Y**– Latitude

4.2. Chi-Square Feature Selection

The chi-square feature selection algorithm works by examining the correlation between the features to determine the most significant features that tend to enhance the model's accuracy and performance. In summary, to determine whether features are independent of one another, the chi-square method examines and quantifies the differences or correlations between expected and observed traits. The chi-square formula is as follows:

$$x_f^2 = \sum \frac{(o_i - X_i)^2}{X_i} \dots\dots\dots\dots\dots\dots 3.1$$

Where f is the degree of freedom, o observed values, X expected value.

Algorithm 4.2: Chi-Square Algorithm

Step 1: *Define Hypothesis.*

Step 2: *Build a Contingency table.*

Step 3: *Find the expected values.*

Step 4: *Calculate the Chi-Square statistic.*

Step 5: *Accept or Reject the Null Hypothesis.*

4.3. Support Vector Machine

The Support Vector Machine (SVM) is a powerful machine learning algorithm that is extensively employed in classification problems as in the case of crime prediction adapted by this study. Its efficacy is notably evident in classification tasks, enabling precise categorization or regression of unknown data. The proposed methodology involves the representation of data points as coordinates inside a multi-dimensional space, where each dimension corresponds to a distinct variable. When the algorithm is visualized, it considers the class affiliation of each data point and strategically groups and arranges them to produce distinct patterns and boundaries. In the realm of crime prediction, the support vector machine (SVM) possesses a noteworthy capacity to discover complex interconnections among diverse crime-related variables, rendering it a highly beneficial resource for anticipating criminal behaviors. Figure 4.2.Shows the behavioral framework of the support vector machine algorithm.

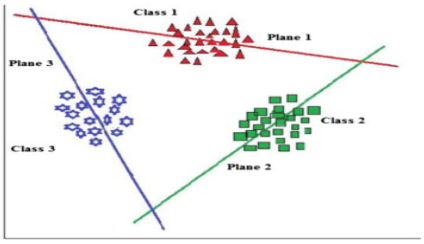

Figure4.2: Support Vector Machine Framework

Algorithm 4.3: Support Vector Machine Algorithm

Step 1:	Data Preparation: Collect and preprocess your dataset. Encode categorical variables if needed.
Step 2:	Feature Selection: Identify relevant features for classification.
Step 3:	Data Splitting: Divide the dataset into training and testing sets.
Step 4:	Training: Choose an appropriate SVM variant (linear, nonlinear, etc.). Train the SVM on the training data. Adjust hyperparameters like C and kernel parameters.
Step 5:	Testing: Evaluate the SVM's performance on the testing data. Use metrics like accuracy, precision, recall, etc.
Step 6:	Prediction: Deploy the trained SVM for making predictions on new data.

4.4. Decision Tree

A decision tree is a predictive modelling approach that utilizes a visual structure reminiscent of a flowchart to make informed decisions by analyzing incoming data. This process involves dividing the data into distinct branches, leading to the assignment of outcomes at the endpoints, known as terminal nodes. Decision trees find widespread application in the realm of machine learning, particularly for tasks involving classification and regression. One of their significant advantages lies in their capacity to create models that are intuitive and easily understandable. This algorithmic framework operates by utilizing conditional control statements. The tree's structure encompasses a central node referred to as the root node, along with branches, internal nodes, and terminal nodes (leaf nodes). Collectively, these components construct a hierarchical arrangement resembling a tree. The functioning of the decision tree algorithm is illustrated in Figure 4.3.

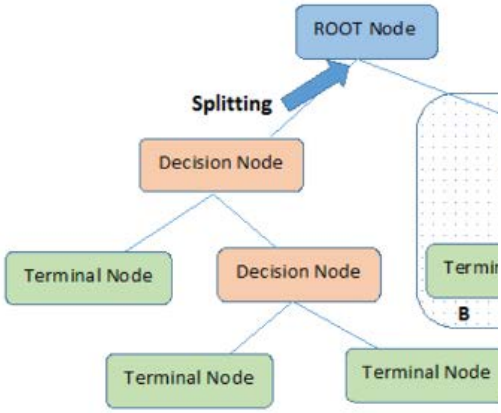

Figure 4.3: Decision Tree Framework

4.5. K-Nearest Neighbor Classifier

The abbreviation KNN stands for "K-Nearest Neighbor". It is a supervised machine-learning algorithm. The algorithm can be used to solve both classification and regression problem statements. KNN is aimed at locating all of the closest neighbors around a new unknown data point to figure out what class it belongs to. Thus, it's a distance-based approach that calculates the distance from all points in the proximity of the unknown variable and filters out the ones with the shortest distance to it. Hence, to correctly classify crimes, the need to first determine the value of K (Number of Nearest Neighbours) is of great importance. Where K denotes the number of nearest neighbours to a new unknown variable that has to be predicted or classified. The concept of selecting K as adopted by the study incorporates the concept of majority voting with a particularly given range of K values and the class with the most votes is then chosen.To select a class from a data point, the distance-based approach utilized is the Euclidean distance between class labels:$(\mathbf{E}) =$

$$\sqrt{(\boldsymbol{X^2 - X^1})^2 - (\boldsymbol{Y^2 - Y^1})^2} \ldots\ldots\ldots\ldots\ldots\ldots \mathbf{4.5}.$$

Where E is the Euclidean distance, X is the independent variable, and Y is the dependent variable.

Algorithm 4.5: K-Nearest Neighbor Classifier Pseudocode

Input: X, D, E, Y

Output: class of X = Y

Step 1: For $(\boldsymbol{X'}, \boldsymbol{l'}) \in \boldsymbol{D}$ do

Step 2: Compute the distance $E(\boldsymbol{X'}, \boldsymbol{l'})$

Step 3: end for

Step 4: Sort the |D| distance by increasing the order

Step 5: count the number of occurrences of each class $\boldsymbol{l'}$ among the k nearest neighbours

Step 6: assign to X the most frequent class

4.6. Justification for Stacking

Stacking is an ensemble machine learning algorithm. The basic idea for stacking involves the use of a meta-learning algorithm to learn how to best combine the predictions from two or more base machine-learning algorithms. The benefit of stacking is that it can harness the capabilities of a range of well-performing models on the

classification task as in the case of this study on crime prediction and make predictions that have better performance than any single model in the ensemble. The architecture involves two or more base models, often referred to as level-0 models and a meta-model that combines the predictions of the base models referred to as a level-1 model. The base models fit the training data and their predictions are compiled and passed to the meta-model, the meta-model learns how to best combine the prediction of the base models. Because the study adapted multiple machine learning models and based on the research conducted, the adapted models have promising skills in crime prediction datasets. This implies that the prediction made by the models or the errors in the prediction made by the models are uncorrelated or have a low correlation.

4.7. Tools and Materials.

For the implementation of the proposed crime classification problem, this study utilized the following frameworks and programming languages:

i. Jupyter Notebook as the programming environment.

ii. Python SDK (3.9) is the library for incorporating Python programming language.

iii. Sklearn, Numpy, Matplotlib, TensorFlow, and Pandas as the dependencies support.

5. Performance Metrics for the Classification Algorithm

The metrics employed to evaluate the performance of the adopted ensemble models in this study are as follows:

Accuracy: is one of the most important metrics of performance evaluation. Accuracy is measured as the percentage of the number of correctly predicted instances to the total number of instances present in the dataset. Thus, the accuracy calculates the ratio of inputs in the test set correctly labelled by the classifier:

$$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots\dots\dots 3.3$$

Precision: measures the classifier's accuracy. It is the percentage of the number of correctly predicted positive instances divided by the total number of predicted positive instances:

$$\boldsymbol{precision} = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots 3.4$$

Sensitivity or Recall: The true positive rate is referred to as sensitivity (also known as recall). It's the number of instances from the positive class that was correct in their predictions. It is a measure of the proportion of initial crime labels and was predicted to

have the same crime label by the model in crime classification.

$$\boldsymbol{Recall} = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.5$$

F-measure (or F-score): defines the harmonic mean of precision and recall. It combines recall and precision metrics to obtain a score.

$$\boldsymbol{F - Measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ldots\ldots\ldots\ldots\ldots 3.6$$

From the defined performance metrics, the description of the TP, TN, FP, and FN is as follows:

i. **True Positives (TP):** a situation where the actual class from a data record is true and thus predicted true by the model.

ii. **True Negatives (TN):** a scenario where a data record was false and hence predicted false by the model.

iii. **False Positives (FP):** an instance where the actual data record class is false but the model predicted true.

iv. **False Negatives (FN):** an instance where an actual data record point is true but the model predicted false.

5.1.Experimental Setup

At the phase of the model implementation and experimentation, the K-Nearest Neighbor, Decision Tree, and the Support Vector Machine algorithm were utilized for the multi-classification of crime from the San-Francisco crime datasets, this study utilized A 64-bit Windows Operating System, with an Intel(R) Corel (TM) i5-3630QM CPU @2.10GHZ with 8.00 GB of RAM for the experimental setup. The programming environment utilized for the program code implementation is the Anaconda environment with Python 3.10 software development kit as the programming language. The application programming interface utilized was Sklearn API with some other Python dependencies such as NumPy for vector operations, pandas for reading files, and the Mat-plot library for data visualization operations.

5.2.Dataset Description and Visualization

i. The data type of each characteristic in the feature set is illustrated in Figure 5.1. Observation can be made that the dataset exhibits variations in its features across multiple types, necessitating the need to convert each feature type to a uniform data type before inputting it into the models. The act of converting a data type to a numeric type is an essential step in data preparation since it guarantees that the dimension of each feature is represented in a numeric format.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 878049 entries, 0 to 878048
Data columns (total 9 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   Dates       878049 non-null  object
 1   Category    878049 non-null  object
 2   Descript    878049 non-null  object
 3   DayOfWeek   878049 non-null  object
 4   PdDistrict  878049 non-null  object
 5   Resolution  878049 non-null  object
 6   Address     878049 non-null  object
 7   X           878049 non-null  float64
 8   Y           878049 non-null  float64
dtypes: float64(2), object(7)
memory usage: 60.3+ MB
```

Figure 5.1: **Dataset data types**

ii. Figure 5.2 shows the analysis of the crimes in the dataset and the number of occurrences for each of the crimes.

```
In [8]: df.Category.value_counts(
Out[8]: LARCENY/THEFT
        OTHER OFFENSES
        NON-CRIMINAL
        ASSAULT
        DRUG/NARCOTIC
        VEHICLE THEFT
        VANDALISM
        WARRANTS
        BURGLARY
        SUSPICIOUS OCC
        MISSING PERSON
        ROBBERY
        FRAUD
        FORGERY/COUNTERFEITING
        SECONDARY CODES
        WEAPON LAWS
        PROSTITUTION
        TRESPASS
        STOLEN PROPERTY
        SEX OFFENSES FORCIBLE
        DISORDERLY CONDUCT
        DRUNKENNESS
        RECOVERED VEHICLE
        KIDNAPPING
        DRIVING UNDER THE INFLUEN
        RUNAWAY
        LIQUOR LAWS
        ARSON
        LOITERING
        EMBEZZLEMENT
        SUICIDE
        FAMILY OFFENSES
        BAD CHECKS
        BRIBERY
        EXTORTION
        SEX OFFENSES NON FORCIBLE
        GAMBLING
        PORNOGRAPHY/OBSCENE MAT
        TREA
        Name: Category, dtype: in
```

Figure 5.2: **Crimes Types Count**

iii. The correlation matrix heatmap, shown in Figure 5.3, offers a visual depiction of the correlation coefficients among variables within the San Francisco crime datasets. This visualization provides valuable insights into the interrelationships and interdependencies among these variables, enabling the study to select features that exhibit strong correlations with the prediction of crime. The correlation matrix was rendered as a heatmap using the Seaborn library. Within the heatmap (Figure 5.3), each cell is colour-coded based on the value of the correlation coefficient. A gradient of colours is employed to convey the magnitude of the correlation, with darker hues signifying stronger correlations (positive or negative), and lighter hues denoting weaker or negligible correlations.

From the heatmap (Figure 5.3), the matrix displays symmetry along its main diagonal. This symmetry arises because the correlations between variables, such as the 'category' and 'resolution' features, mirror those between the 'resolution' and 'category' features. Thus, by scrutinizing the correlation matrix diagram, the study identifies variables characterized by robust positive or negative correlations. Positive correlations imply that as one variable increases, the other tends to increase as well. Conversely, negative correlations suggest that as one variable increases, the other tends to decrease. Variables featuring weak or minimal correlation will exhibit values close to 0.

In particular, observations extracted from the correlation diagram concerning the target label of crime prediction, labelled as 'category,' reveal that certain features like

'resolutions' and 'police districts' exhibit a tendency toward negative correlations with other features in predicting crime categories within the San Francisco crime datasets.

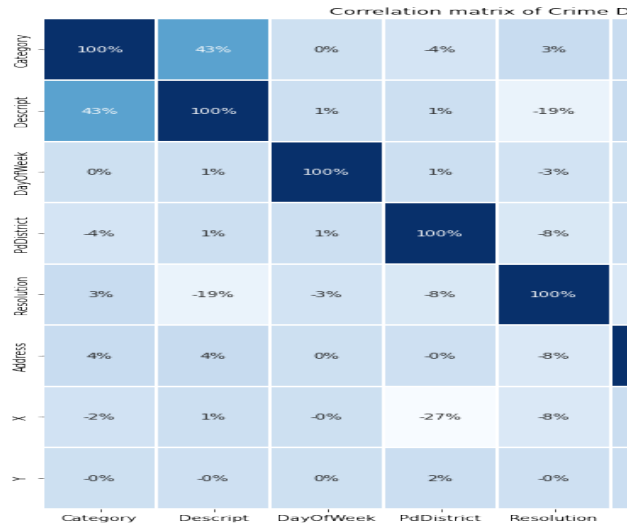

Figure 5.3: **correlation diagrams**

iv. Considering an exploration of the respective crime base on the districts, figures 4.5 show the analysis of the crimes based on the respective districts in the dataset and the rate of crimes within the districts.

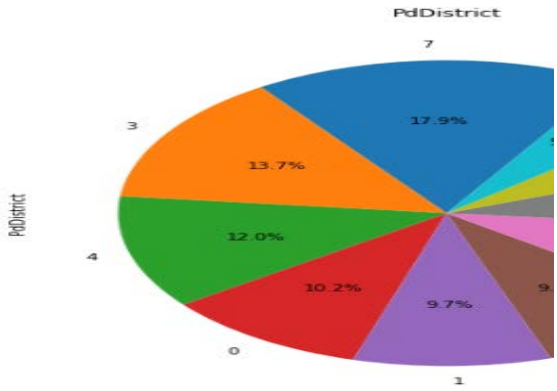

Figure 5.4: **District Crimes**

5.3. Data Scaling

Normalization of the dataset is a crucial component in machine learning experiments, as it contributes to the optimization of model performance. This process involves the expense of data preprocessing. Furthermore, it is imperative to do dataset scaling to normalize the feature range, hence improving the classification capabilities of the model. This is particularly important in the context of crime categorization, as depicted in Figure 4.2, where the field exhibits many types. To achieve feature scaling, this study employed the standard scaler functionality to transform the dataset features into a normalised range of values between zero and one. One advantageous practice in machine learning is to scale the features of a dataset, as it can facilitate the rapid convergence of the model by constraining the range of variables. The code snippet for the dataset scaling is shown in figure 5.5.

```
In [81]: # Feature scaling
         scaler = StandardScaler()
         X[X.columns] = scaler.fit_transform(X)
```

Figure 5.5: Data Scaling

5.4. Percentage Split Technique

In training the developed models, this study divided the dataset into some training and

test proportions, where the training set was used to train the models and the test set to validate the workings of the model. From the total of 138721 sentiment records for the San Franciso dataset, 30% (4, 162) of the dataset was used to test and validate each model's performance using some performance evaluation metrics, whereas, the other 70% (9,709) was used to train the models.

Considering that the study conducted feature selection using the chi-square algorithms. The features that demonstrated a strong association with the outcome of crime prediction labels are presented in Table 5.1.

Table 5.1: **Selected Features**

No	Selected fields
1	DayOfWeek
2	Descript
3	PdDistrict
4	Resolution
5	Address

5.5. Classification Algorithm Result Presentation

The results presented in Table 4.2 offer a comprehensive insight into the performance of various machine learning algorithms employed for the prediction of crime in San Francisco. Each algorithm's accuracy is reported in terms of percentage, showcasing its effectiveness in predicting criminal activities based on historical data patterns.

Starting with the K-Nearest Neighbors (KNN) algorithm, it achieves an accuracy of 87.4%. Next, the Decision Tree algorithm demonstrates a remarkable accuracy of 99%. The Support Vector Machine (SVM) algorithm records an accuracy of 86.5%. The Stacked Model (hybrid), which presumably combines multiple algorithms, showcases the highest accuracy at 99.7%.

Table 5.2: Models Accuracy

Algorithm	Accuracy (%)
KNN	87.4
Decision Tree	99
SVM	86.5
Stacked Model (hybrid)	99.7

In summary, the reported accuracy values in Table 5.2 highlight the varying performance of different machine learning algorithms in predicting crime in San Francisco. The Decision Tree and Stacked Model exhibit superior accuracy, potentially capturing the intricacies of crime patterns well. Moreover, a visual display of the accuracies for each algorithms.

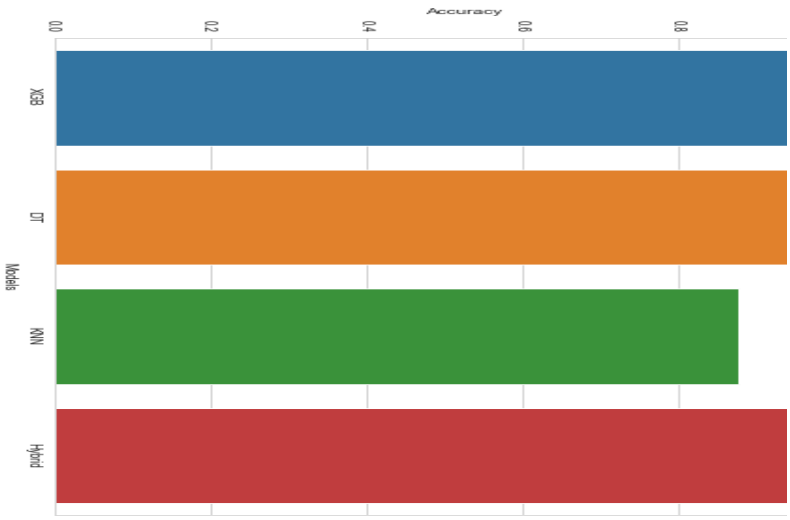

Figure 5.6: **model accuracy graph plot (San Franciso Dataset).**

6. Evaluation Metrics

Table 5.3 presents a comprehensive evaluation of the performance metrics for various algorithms applied to the task of predicting San Francisco crime. The metrics evaluated include Precision, Recall, and F1-score, which collectively provide insights into the algorithms' effectiveness in classifying and predicting criminal incidents. **Table 5.3:** Performance Metrics Confirmation

Algorithm	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Accuracy**
Decision Tree	89	97	93	99
KNN	87	98	93	86.8
SVM	88	97	92	86.5
Stack-Model	98	99	98	99.7

The Decision Tree algorithm demonstrates promising results in predicting San Francisco crime. It achieves a Precision of 89%, indicating that when it predicts a crime, it is correct 89% of the time. Moreover, its Recall stands at an impressive 97%, indicating that it identifies 97% of the actual criminal incidents. This robust recall suggests that the algorithm is successful in capturing a high percentage of true positive cases. The F1-score, a harmonic mean of Precision and Recall, is calculated at 93%, reflecting a balanced performance between precision and recall. These metrics collectively highlight the algorithm's ability to provide a reliable balance between accurate positive predictions and comprehensive coverage of actual crime instances.

The KNN algorithm showcases competitive performance in San Francisco crime prediction. It attains a Precision of 87%,

indicating strong accuracy in its positive predictions. The Recall metric, at 98%, demonstrates that it identifies a significant proportion of actual criminal incidents. The F1 score of 93% suggests a harmonious balance between precision and recall. Overall, KNN's performance showcases its capacity to accurately identify criminal incidents while achieving a commendable level of true positive detection.

The SVM algorithm presents robust results in the prediction of San Francisco crime. It achieves an 88% Precision, signifying its proficiency in making accurate positive predictions. With a Recall of 97%, the algorithm effectively captures a substantial proportion of actual criminal events. The F1 score stands at 92%, highlighting a well-balanced compromise between precision and recall. This combination of metrics indicates that the SVM algorithm achieves a commendable level of both accuracy and comprehensive detection of criminal incidents.

The Stacked Model showcases remarkable performance in predicting San Francisco crime. With an impressive Precision of 98%, the model excels in making precise positive predictions. Additionally, its Recall reaches a remarkable 99%, signifying an ability to capture an exceedingly high proportion of actual criminal incidents. The F1 score of 98% underscores the model's exceptional balance between precision and recall. The stacked model's performance showcases its prowess in achieving a near-perfect equilibrium between accurate predictions and extensive coverage of true positive cases.

In summary, the evaluation results in Table 4.3 demonstrate that all algorithms, namely Decision Tree, KNN, SVM, and the Stacked Model, perform well in predicting San Francisco crime. The Stacked Model, in particular, stands out with superior Precision, Recall, and F1-score metrics, showcasing its potential to provide highly accurate predictions and capture a significant proportion of true positive cases. The other algorithms also exhibit commendable performance, striking a balance between prediction accuracy and comprehensive detection of criminal incidents. These results collectively underscore the effectiveness of machine learning algorithms in aiding law enforcement efforts by providing accurate crime predictions. The precision, recall, and f1-score metrics for each of the algorithms can be seen in the figure 5.7 with their respective scores annotated on each chart.

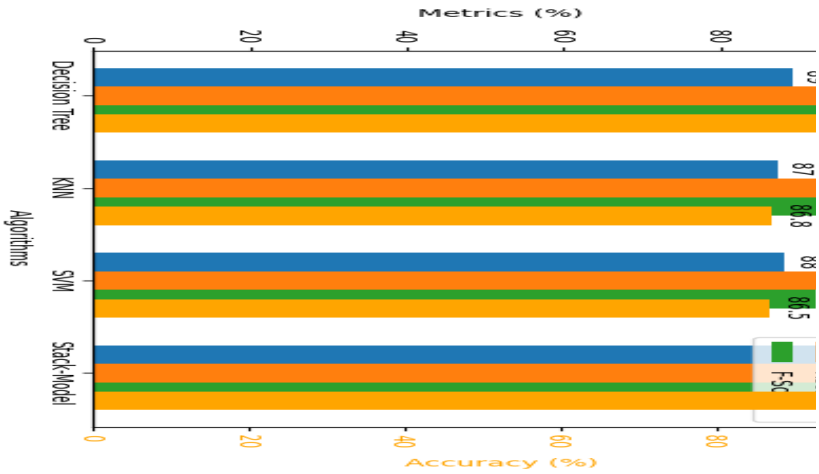

Figure 5.7:Performance Metrics Comparison.

7. SUMMARY, CONCLUSION AND RECOMMENDATION

7.1 Summary

The concept of crime prediction is readily comprehensible, although its practical implementation demands more than a mere comprehension of the concept. The increasing apprehensions regarding human security, as perceived by security personnel, have motivated a range of scholars to actualize crime prediction and effectively deploy modern technologies in practical contexts. While it is true that law enforcement agencies sometimes adopt new technologies like Sting Rays and facial recognition, the integration of such software has the potential to significantly transform police operations, yielding notable improvements by enabling proactive crime prevention.

This study proposed a method aimed at reducing the prevalence of crime rates by utilising machine learning models to develop a more effective system for law enforcement agencies. The models that were produced originated from the utilisation of three machine learning models, specifically the K-Nearest Neighbour, the Decision Tree, and the Support Vector Machine. During the process of developing the model. The study utilised the San Francisco dataset to leverage the prediction capabilities of three machine-learning algorithms. To optimise the performance of the machine learning models, a three-phase methodological approach was employed. The initial stage involves utilising the Python pandas machine learning libraries to read the dataset. Subsequently, data preprocessing is performed to accomplish tasks such as feature selection, elimination of null values, and scaling of the chosen dataset features.

The second phase involves dividing the dataset into training and testing sets before inputting it into three machine learning models: Support Vector Machine, Decision Tree, and K-Nearest Neighbour. Finally, performance evaluation analysis and prediction are conducted using the developed models. It is crucial to acknowledge that the three models were afterwards combined to create a stacked model, which was then assessed for its performance. Moreover, the experimental methodology employed in the development of the crime prediction models utilised the Python programming language and various third-party libraries, including NumPy, Python Open-CV, Matplotlib, Pandas, and Sklearn API.

7.2 Conclusion

In summary, this study has introduced various techniques and methods to enhance crime prediction and support law enforcement agencies. The potential of employing diverse methods for crime prediction and prevention can transform the landscape of law enforcement operations. By leveraging a combination of machine learning algorithms, the efficiency and effectiveness of law enforcement agencies can experience a substantial improvement.

This study specifically focused on the applicability of three distinct machine learning models: Decision Tree, K-Nearest Neighbor, and Support Vector Machine. Moreover, a novel hybrid model was devised by combining these machine-learning models into a stacked ensemble. To assess the performance and viability of these models, the San Francisco crime dataset was employed as the basis for experimentation. The outcomes of the analysis indicated notable accuracy scores, showcasing the effectiveness of these models in crime prediction and prevention.

To rigorously validate the efficacy of each model, an array of performance evaluation metrics were employed, including recall, precision, and the F1-score. The collective results demonstrated the strong performance of the implemented models, with each model achieving accuracy levels surpassing 80% on the San Francisco crime dataset. However, it's worth noting that the stacked model emerged as the most successful, outperforming all individual models with an impressive accuracy of 99%.

In conclusion, the study underscores the substantial potential of machine learning techniques in crime prediction and prevention, showcasing how these models can significantly enhance law enforcement

operations. The outcomes emphasize the significance of leveraging a hybrid approach, as demonstrated by the superior performance of the stacked model, which promises more accurate and efficient crime prediction outcomes.

7.3 Recommendation

The module recommendation proffer two main subjects regarding the application areas of the developed model and the suggestion for possible enhancement of the developed model implemented during this research.

7.3.1 Application Areas

1. A major application of the developed model is in the area of crime prediction within police departments. Thus, law enforcement agents can utilize the models in the prediction of crimes before their occurrence.
2. The models can be utilized during tactical duty deployment enabling the deployment of forces to places prompt to violence and crimes.

7.3.2 Suggestions for Further Research

After an extensive study and review of several kinds of literature relating to the classification of crimes, this thesis suggests the future research areas as follows:

i. The application of deep learning techniques such as the Res-Net, Alex-Net, and the Google-Net algorithm considering that these algorithms over the past decades have proven in hot spotting and classification problem.

ii. To balance the dataset multi-class labels, future studies can apply the Smote algorithm as dataset label balancing, provisioning an equal dataset distribution across the categories of crimes.

iii. Considering that the feature selection technique utilized is the correlation matrix, future studies can apply the viabilities of other feature selection techniques such as information gain, particle swam algorithm, and others.

8. REFERENCE

Charles, M. (2021). What constitutes a crime? https://www.thoughtco.com/what-is-a-crime-970836.

Eidell, W., & Ellis, C.A. (2010). Impact of Crime on Victims. https://doi.org/10.1016/j.procs.2020.05.018

Hemant, K. A., Haresh, M. N., & Majid, S. S. (2020). A Design to Predict and Analyze Crime. https://www.academia.edu/download/87189561/A_Design_to_Predict_and_Analyze_Crime.pdf

Hossain, S., Abtahee, A., Kashem, I., Hoque, M., and Sarker, I.H. (2020). Crime prediction using Spatio-temporal data. *International*

Conference on Computing Science,

Communication and Security, 4(5), 277-289.

Jha, P., Jha, R., and Sharma, A. (2019). Behavior analysis and crime prediction using big data and machine learning. *International Journal of Technology Engineering*, 8(1), 461–468.

Kim, S., Joshi, P., Kalsi, P.S., and Taheri, P. (2018). Crime analysis through machine learning. https://doi.org/10.1109/IEMCON.2018.8614828

Kim, S., Joshi, P., Kalsi, P.S., and Taheri, P. (2018). Crime Analysis Through Machine Learning. *Information Technology, Electronics, and Mobile Communication Conference)*, 3(4), 415-420, DOI: 10.1109/IEMCON.2018.8614828.

Kshatri, S.S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., and Sinha, G. R. (2021). An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization. *International Journal of Electrical Engineering*, 9, 67488-67500. doi:10.1109/access.2021.3075140.

Matereke,T., Nyirenda, C. N., and Ghaziasgar, M. (2021). A Comparative Evaluation of Spatio Temporal Deep Learning Techniques for Crime Prediction. *International Research Journal of Engineering Technology*, 5(4), 1-6, DOI: 10.1109/AFRICON51333.2021.9570858.

Okpuvwie, E.J, & Toko, M.I. (2020). An Appraisal of the Spatial Distribution of Crimes in Ife Central Local Government Area of Osun State in Nigeria. *African Journal of Law and Criminology*, 10(1).

Prithi, S., Aravindan, S., Anusuya, E., and Kumar, A.M. (2020). GUI-based prediction of crime rate using machine learning approach. *International Journal Computing Science and Mobile Computation*, 9(3), 221–229.

Safat, W., Asghar, S., and Gillani, S.A. (2021). Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *International Research Journal of Engineering Technology*, 9,7080-7094. Doi: 10.1109/ACCESS.2021.3078117.

Saminu, A., Folorunso, O., Johnson, F., Akerele, J., Ilesanmi, S., & Ajayi, F. (2021). Adaptive Neuro-Fuzzy Model for Vehicle Theft Prediction and Recovery. *International Conference on Informatics and Intelligent Applications*, 3(4), 20-34.

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(1), 67-99. DOI:10.1186/s42492-021-00075-z

Shanjana, A.S., and Porkodi, R. (2021). crime analysis and prediction using data mining. *International Journal of Creative Research Thought (IJCRT),* 9(2), 465-485. ISSN: 2320-2882

Sohrab, H., Ahmed, A., Imran, K., Mohammed, M.H., and Iqbal, H.S. (2020). Crime Prediction Using Spatio-Temporal Data. *Communications in Computer and Information Science*, 12(4), 277-289.

Tabedzki, C., Thirumalaiswamy, A., Van, V.P., Agarwal, S., & Sun, S. (2018). Yo home to Bel-Air: predicting crime on the streets of Philadelphia. https://www.seas.upenn.edu/~tabedzki/machine-learning-report-final.pdf

Zhang, X., Liu, L., Xiao, L., and Ji, J. (2020). Comparison of Machine Learning Algorithms for Predicting Crime Hotspots. *International Journal of Electronic and Electrical Engineering*, 8, 302-310. DOI: 10.1109/ACCESS.2020.3028420.