Int Journal of Social Sciences Management and Entrepreneurship 8(2): 1328-1349 2024

ISSN 2411-7323

www.sagepublishers.com

CRIME PREDICTION AND MAPPING USING MACHINE LEARNING ALGORITHMS

Ngoge Lucas[1*], Dr. Ogada Kennedy[2], Dr. Kaburu Dennis[3]

[1] Jomo Kenyatta University of Agriculture and Technology; Kenya; ngogelucas@gmail.com

[2] Jomo Kenyatta University of Agriculture and Technology; Kenya; kodhiambo@scit.jkuat.ac.ke

[3] Jomo Kenyatta University of Agriculture and Technology; Kenya; dennis.kaburu@jkuat.ac.ke

Abstract. One of the major roles of government is to curb crime. Despite the measures the government has taken to counteract criminal activity, the security situation in many urban centers has gotten worse. The goal of this study was to create and assess a machine learning model with the core function of forecasting crime categories and utilizing contextual features found in the datasets to visualize the locations in which they occur. This was achieved by combining time, space, and contextual information with machine learning to improve crime prediction and mapping. The datasets were collected from various sources were subjected to a number of machine learning algorithms to evaluate how well they performed. The random forest algorithm emerged as the best algorithm with a classification accuracy of 97% or 0.973301 using the confusion matrix. The longitude and latitude features were used to tag the specific locations of crime occurrences on a map.

Keywords: Machine Learning Algorithms; Classification; Prediction; Mapping; Data Visualization[1]

Correspondence: Lucas Ngoge, Jomo Kenyatta University of Agriculture and Technology, Kenya, ngogelucas@gmail.com

1. Introduction

Like most urban settings in the world, Kenya battles with all forms of crime. There has been an upsurge in different types of crime as reported in many parts of urban centers, despite the strategies and efforts the government put in place to combat them as presented in (NPS, 2022). This situation is aggravated by the weak social control that operates through formal and informal institutions for reporting crimes that took place. Due to the aforementioned problems, it was necessary to review models that, from technology, have contributed to the improvement of the crime prevention strategies that guarantee public safety. The growth of research methodologies aimed at extracting data from records to better understand criminal patterns and ultimately prevent future occurrences has been brought about by the increase in the recording of crime data, but it may be challenging to resolve any case involving crime if there are no data available beforehand. Therefore, creating a machine-learning model that can classify data is necessary as stated by (Veena et al, 2022*)* are able to forecast classes using the features that are present in it. The increased use of these approaches in crime analysis has enabled significant advances in crime prediction and mapping. According to (Sarker, 2021) machine learning enables computers to automatically learn from experience without being explicitly designed. These learning algorithms are broadly categorized into two major types, namely supervised and unsupervised. In supervised learning, datasets are used to train, test, and produce the intended outcomes for the models, but in unsupervised learning, the models classify or cluster an inconsistent, unstructured dataset. In a dataset by (Kanimozhi et al,2021) multiclass target variables were classified using supervised learning methods such as decision trees, random forests, naive bayes, K-nearest neighbors, and support vector machines. Additionally, with these techniques, the crime predictors such as 'Incident_Number', 'Incident', 'Crime_Decsription', 'Crime_code', 'Crimetype', 'Arrest',' Age', 'Gender', 'Date', 'Year', 'Month', 'Location', 'Location_code', 'Lat' and 'Long' were used to correctly classify and predict a target variable.

A machine learning technique called classification divides data into labels or classes to help with accurate analysis and forecasting. In order to help (Pratibha et al, 2020) forecast results based on a given input, it is utilized to develop patterns that accurately characterize the significant data classes within the data set. Classification algorithms look for connections between attributes that could help forecast the outcome. The algorithms are used to analyze input and produce a prediction as illustrated in (Llaha, 2020). Classification models are expected to input an unseen dataset and correctly predict category labels, as shown in Figure 1 below:

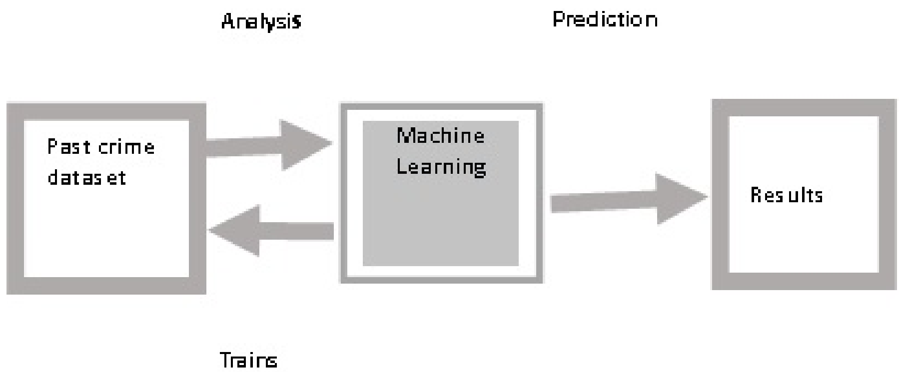

Figure 1: Classification process

Crime prediction is a process where a model uses different algorithms to solve classification problems based on historical data. Using machine learning, these models can predict the likelihood of a crime provided that the required dataset is available as explained in (Mahmud et al, 2021). The crime data used to develop the model was collected from various sources including law enforcement organizations within Nairobi County and various websites sources. It consisted of fifteen (15) predictors (columns) and two thousand and fifty-nine (2059) rows/instances. This data was preprocessed into a suitable form to improve the classification of various types of crime. Various machine learning algorithms were applied to it to determine their performance and effectiveness in predicting different types of crime. Law enforcement agencies are faced with large volumes of data generated every day about a crime that require machine learning and visualization tools to analyze and depict their occurrence locations. By preprocessing this data into a suitable form, law enforcement agencies can use machine learning models with the best performance to get correct predictions of crime categories and their occurrence locations as explained in (Llaha, 2020). However, as demonstrated in (Sarker, 2021), the success and efficiency of a machine-learning solution depend on the accuracy and performance of the learning algorithm Thus, it is through this background, that a machine learning model was developed that centered on predicting crime categories and visualizing their occurrence locations using contextual features present in the datasets. The main objective of this research was to find an effective model for crime prediction by comparing the accuracies of the various classification algorithms to select one with the best performance on the test dataset. From the research findings, the random forest algorithm was selected as the best algorithm with a classification accuracy of 97% or 0.973301. The visualization of crime was done and presented using interactive plots such as bar graphs, line graphs, pie charts, and maps.

This paper addresses the need to combine time, space, and contextual information with machine learning to improve crime prediction and mapping. This model intends to identify the types of crime that are likely to take place in a location at a particular time. This information can be used to distinguish the types of preventive measures to be used for each type of crime. This paper is structured as follows, Section 1 is the introduction; Section 2 reviews the literature on related work in machine learning algorithms used in crime prediction; Section 3 describes the methodology and the dataset; Section 4 presents the results and discussions on the application of machine learning algorithms; and Section 5 presents the overall conclusions.

2. Related Work

Numerous studies have been conducted to address crime reduction and several predictive algorithms have been suggested. These studies have used machine learning models to see how well they predict different types of crime and show where they occur. Machine learning models are predictive models that analyze recent and past data to make accurate predictions about what will happen in the future as explained by (Mahesh & Dipti, 2020). According to (Mohamed et al, 2022) machine learning is the scientific study of the techniques that computer systems employ to carry out a certain task effectively without utilizing several explicit instructions. They are classified into two broad categories namely supervised and unsupervised machine learning in Figure 2 below:

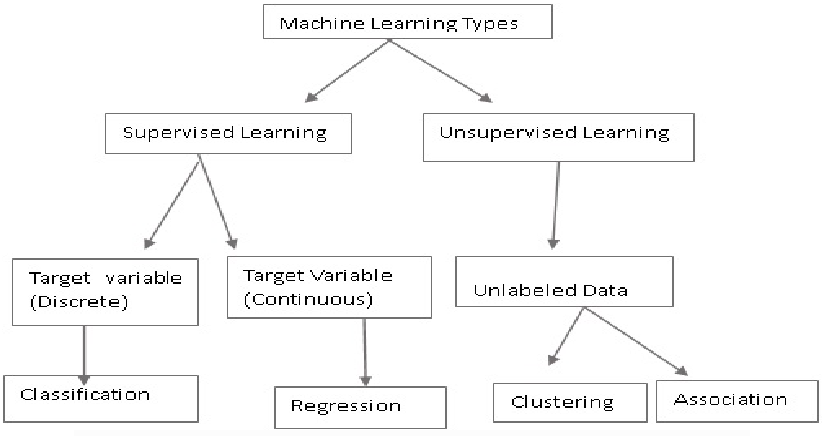

Figure 2: Types of machine learning algorithms

In unsupervised learning, the machine learning algorithms divide an inconsistent, unstructured dataset into classes or clusters while in supervised learning, the machine learning models use datasets to train, test, and get the desired results on them. To find links between attributes that could help predict the outcome, they employ classification algorithms. Regression and classification are the two sorts of tasks that supervised learning is capable of handling. While classification uses the value of a categorical target or categorical class variable to predict similar information, regression uses new, unseen input data to predict a numeric value. According to (Mohamed et al, 2022) it is a beneficial strategy for any kind of statistical data. Although classification is a well-known machine learning technique, it has problems with duplicates and missing data. According to (Yoganand et al, 2020) missing values in the dataset can be problematic during both the training and classification phases. Despite the great range of supervised learning methods that are accessible, classification is the most widely used technique in predictive modeling. According to (Sen, 2021), the most used classification algorithms are Decision Trees/Rules, Random Forest, K-Nearest Neighbors, Gradient-Boosted Machines, Naive Bayes, and Support Vector Machines amongst others.

2.1 Machine Learning used in Crime Prediction

The popular machine learning algorithms used for classification and prediction tasks noted from the above discussion are decision trees (DT), naive bayes (NB), support vector machine (SVM), KNeighbors (KNN), and random forest (RF). The prediction tasks studied for each of the classification algorithms are discussed below:

2.1.1 Decision Trees

A decision tree is a diagram in the shape of a tree that is used to choose a course of action. Every branch of the tree symbolises a potential choice, event, or response. A decision tree is built by breaking down the dataset into smaller units step by step and incrementally. It produces classification models in the form of a hierarchical structure. In the end, a tree with leaf nodes and decision nodes is generated. There are two or more branches on a decision node. A determination or classification is represented by the leaf node. In (Dikananda et al., 2022) the root node of a tree corresponds to the best predictor from the provided datasets. Decision Trees are mostly used in operational analysis to help identify a strategy most likely to reach a goal. They are preferred where the model is easy to understand as demonstrated in Figure 3.

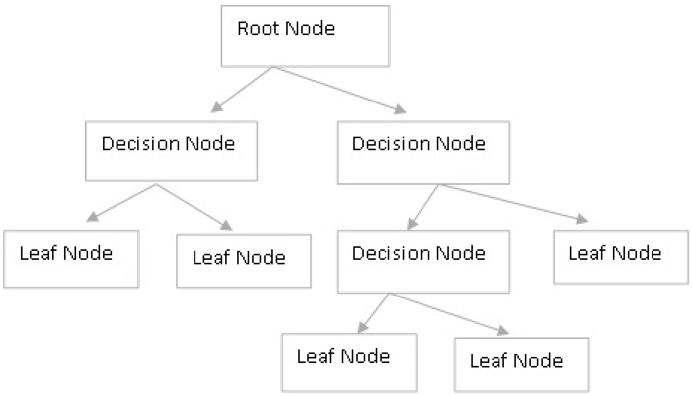

Figure 3: Decision tree structure with tree and subtree

Regression or classification tasks can be performed using decision trees. A tree will decide on a set of logical IF - THEN Conditions to classify problems in the classification task. For instance, making a distinction based on a specific circumstance between three different sorts of crime, such as robbery, theft, or murder. As mentioned by (Patrick & Erin, 2020) when doing regression tasks, a decision tree is employed when the target variable is numerical or continuous and the model is fitted to the target variable using each of the independent variables. To decide which attribute to utilize to divide the records in each node, the decision tree employs information gain. For each node, information gain is calculated, and the attributes with the highest information gain are chosen. Information gain is the anticipated decrease in entropy following the use of a specific property to divide the records in a node. The information gained by the model from dividing the records in a node using particular attributes is measured (Patrick & Erin , 2020). Finding a property that gives the most information gain uniformly in Sen [4]. is the key to developing a decision tree. Entropy, on the other hand, is the expected information needed to classify a record in each node and is calculated using the formula below:

$$\text{Entropy}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (1)$$

Where:

m is the total number of classes and *pi* is the probability that any given record in the node belongs to class *i*.

The decision tree is generated as follows:

- Using the idea of information gain, the tree's root is chosen from the dataset's properties.
- Create subgroups from the training dataset. Each of these subsets is constructed so that the data within it has the same value for each attribute.
- Repeat steps 1 and 2 on each subset until leaf nodes are visible in every branch of the tree.

The main advantage of using a Decision Tree is its capability to deal with a wide variety of classification problems including handling non-linear parameters and missing values efficiently. Secondly, it has less training period and little effort is required for data preparation. However, it gets unstable due to small variations in data which may lead to overfitting.

2.1.2 Naive Bayes

It is a classification technique built on the Bayes Theorem, which makes the assumption that predictors are independent. According to (Viet et al., 2021) the Naive Bayes classifier makes the assumption that the presence of a feature in a class is unrelated to the presence of any other features. According to the Bayes theorem's definition of conditional probability, the classifier operates as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (2)$$

Where:

- P(B|A) = Conditional probability of A given B
- P(A|B) = Conditional probability of B given A
- P(A) = Probability of event A
- P(B) = Probability of event B

Based on prior knowledge of conditions that might be associated to an event, Baye's theorem determines the conditional probability of the occurrence of an event. According to (Viet et al.2021), it is mostly utilized for facial recognition, weather forecasting, medical diagnosis, and news classification. One issue with the Bayesian network classifier is that it needs continuous attributes to be discretized, and the conversion process adds classification problems such noise and missing information that may lead to incorrect results. The Bayesian network has a number of benefits, including the following: (1) it needs less training data; (2) it is extremely scalable with many predictors and data points; and (3) it can fill in missing data by taking into account the overall probability of the missing values.

2.1.3 Support Vector Machine

According to (Zeineddine et al.,2020) support vector machine is a supervised learning technique that divides data points into categories by employing an N-dimensional hyperplane, where N is the total number of attributes that make up a data point. Finding a hyperplane in N-dimensional space (where N is the number of features) that clearly classifies the data points is the goal of support vector machine techniques. Numerous hyperplanes can be selected to divide the two groups of data. According to (Viet et al., 2021) the objective is to identify the plane that has the greatest distance between data points for both classes. The line that separates the data into classes is the support vector classifier or hard margin. The problem with hard margins is that it doesn't work with non-linearly separable data. Hence, soft margins are introduced to accept the new data points and optimize the model for non-linear data points. The soft margins pass through the data points at the border of the classes in (Mohamed et al. 2022). A Support Vector Machine's key benefit is its capacity to handle a wide range of classification problems, including high dimensional and nonlinearly separable issues. In contrast, a support vector machine needs a number of crucial settings to produce outstanding classification results.

2.1.4 K-Nearest Neighbors

K-nearest neighbours is a straightforward method that categorises new cases based on the similarity metric and stores all of the existing examples. It finds out the class of a new data point by finding its nearest neighbors. For example, if there are three data points of class A and two data points of class B near the new data point, then the k-nearest neighbors classify the new data point as class A according to (Sen,2021). K-Nearest Neighbours are calculated using a distance function such the Minkowski, Manhattan, or Euclidean. K is typically used as an odd number solely when making decisions. If K is equal to 1, it is merely put into the class of its closest neighbour. According to (Mohamed et al.,2022) if K is an odd number, it is put into the category of the neighbour with the most votes. Too, if the K in K-Nearest Neighbors is the number of nearest neighbors we are looking for, then we say K=3, this means that we are looking for the nearest 3 neighbors of the unclassified data point. Usually, a K value between 3 to 10 is taken as it leads to better results. The range of K between 3 and 10 is chosen since a smaller number of K indicates that noise will have a stronger impact on the results and a larger value of K makes it computationally expensive. When non-parametric (no set of parameters) techniques are needed, k-nearest neighbours are utilised. It is used in pattern recognition, data mining, and intrusion detection as explained in (Sen, 2021). The advantages of K-Nearest Neighbors include (1) simplicity: it doesn't need a separate training period, (2) transparency: new data can be added seamlessly without affecting the accuracy of the model, (3) robust to noisy training data (4) easy to understand and implement. The disadvantages include (1) computation complexity (2) memory limitation (3) poor runtime performance for a large training.

2.1.5 Random Forest

As demonstrated in (Sen,2022), random Forest is an ensemble decision tree classifier that uses bootstrap sampling to build numerous individual decision trees before assigning a final class. An undetermined collection of alternative trees with K random features at each node are combined to create a random tree. "Random means that there is an equal chance that each tree in the set will be sampled. The decision tree's bad habit of overfitting the training dataset is corrected. A generic meta-approach to machine learning called ensemble learning seeks out the most accurate predicted performance by combining many techniques. According to (Sen, 2021) the employment of many machine learning algorithms together will produce better accuracy because using them alone may not produce the greatest results. Experimentally, ensemble models outperform single models due to the strong differences between them.

Different machine learning algorithms that make up the ensemble learning approach, which is utilized for prediction and classification tasks, are depicted in Figure 4 below.

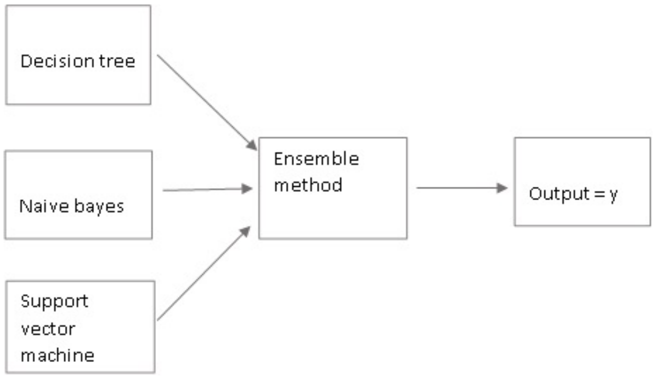

Figure 4: Ensemble machine learning

According to (Mohamed et al., 2022) there are three types of ensemble learning: bagging, stacking, and boosting. While stacking is concerned with fitting numerous different types of models to the same data while using a different type of model to learn the combined predictions, bagging is concerned with making many decisions on various samples of the exact same dataset and obtaining the average prediction. On the other hand, boosting entails the progressive addition of ensemble members in order to correct the earlier forecast made by the other models and then provide the average of the predictions made by (Mohamed et al,2022). The primary distinction between the random forest algorithm and the decision tree method is that the random forest creates root nodes and divides nodes arbitrarily. To produce the necessary prediction, it also uses the bagging approach. Utilising several training data samples as opposed to a single sample is known as bagging. Because random forest operates effectively on big datasets and does not require properly distributed training data, it is advantageous for classification and prediction. However, too many trees could slow down and render the algorithm ineffective for making predictions in real-time.

Several studies on crime prediction have demonstrated that it is easy to identify the location of criminal activity using classification techniques as suggested in the work of (Pratibha et al, 2020). The authors presented research on the opportunities and challenges of machine learning in predicting future crime categories and visualizing their occurrence locations. They conducted an experiment using various machine learning methods such as K-Nearest Neighbors, Decision Trees, Extratress, Artificial Neural Networks, Support Vector Machine, and particular inputs to predict crimes. They used crime datasets collected from many sources to train and test models. They then evaluated the effectiveness of these models in predicting violent crimes occurring in a particular region and the results showed that the decision trees outperformed other algorithms with a classification accuracy of (88%). In their conclusions, the decision tree, K-Nearest Neighbors, and Extratress classifiers worked best with optimal training even though any model that works best is dependent on the dataset that is used.

3. Methodology

The machine learning workflow illustrated in Figure 5 was used as a methodology to develop the model. Machine learning techniques can be applied to data to extract information that can help in making better decisions regarding crime patterns and their occurrence locations. This research aimed at developing and testing machine learning models that predict types of crime and their occurrence locations. Nine (9) experiments were done according to the machine learning workflow illustrated below. Nine (9) datasets of different instances of 200, 400, 600, 900, 1000,1200, 1800, 2000, and 2059 were used to conduct the experiments where each of the five (5) machine learning models was applied to each dataset and their performance accuracy was computed and recorded. Python Jupyter Notebook Integrated

Development Environment (IDE) was used to develop and run the models. The algorithms built were decision tree (DT), random forest (RF), naive bayes (NB), kneighbors (KNN), and support vector machines (SVM).

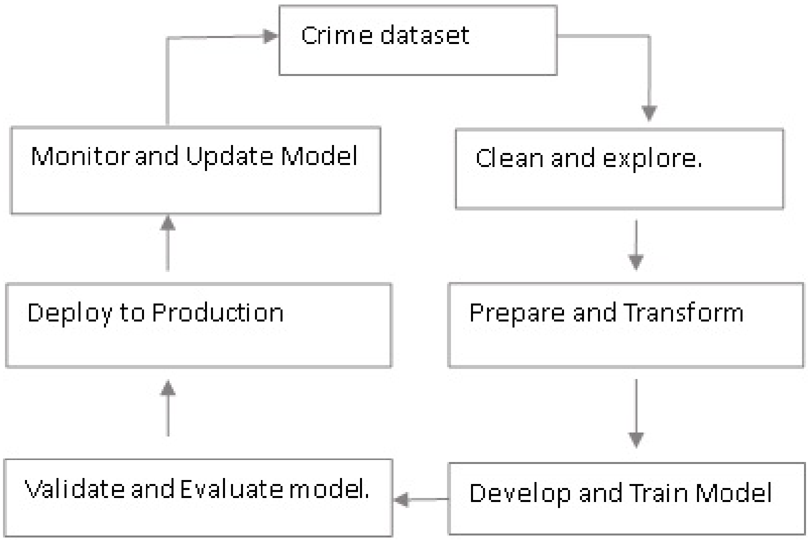

Figure 5: Machine Learning workflow

The data flow diagram in Figure 6 represents various aspects of the models' processes by considering the input, the process conducted, and the output.

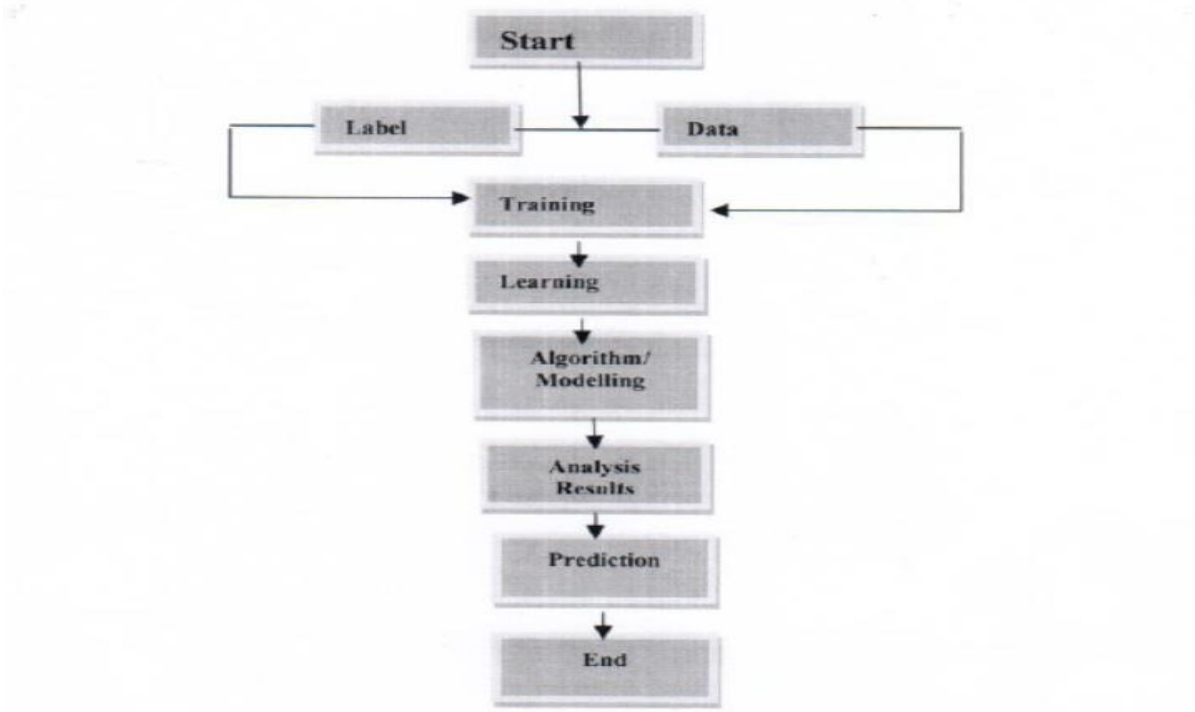

Figure 6: Flowchart diagram of the classification and prediction

3.1 Crime datasets

The model was developed using crime data collected from various sources including law enforcement organizations within Nairobi County and various websites sources through web scraping and stored in a CSV file on the computer's hard disk. It consisted of crime information such as 'Incident_Number', 'Incident', 'Crime_Decsription', 'Crime_code', 'Crimetype', 'Arrest',' Age', 'Gender', 'Date', 'Year', 'Month', 'Location', 'Location_code', 'Lat' and 'Long' as shown in Table 1. This dataset was a subset of a much larger dataset with feature vectors that have a higher degree of correlation for predicting crime.

Table 1: Crime dataset

Field Name	Description
Incident_Number	Identifier
Incident	Reports of crime
Crime_Decsription	Offense description
Crime_code	Offense description code
Crimetype	Class of the crime
Arrest	Whether the suspect was apprehended or not
Age	Age of the offender
Gender	Whether male or female
Date	Occurrence date
Year	Year of offense occurrence
Month	The month of offense occurrence
Location	Areas where crime incidents happened
Location_code	Code of the crime occurrence location
Lat	Latitude of the location
Long	Longitude of the location

The dataset consisted of fifteen (15) predictors (columns) and two thousand and fifty-nine (2059) rows/instances that were read and viewed in Python (Jupyter Notebook IDE) using the Pandas functions ‘PD.read_csv()’. The data provided the necessary information about crime in Nairobi County which assisted in identifying types of crime and their occurrence locations.

3.2 Data Preprocessing

Data cleaning is the process of adding missing values, reducing noise, identifying outliers, and fixing errors in the dataset before the machine learning approach is applied to it. Data preparation is the act of transforming raw data into an appropriate form. The dataset was cleaned and unnecessary data was removed using the following techniques.

3.2.1. Interpolation of missing values

The missing values were replaced through interpolation. This was achieved using the seaborn function ‘sb. heatmap (df, IsNull())’ that checked for any missing values in the DataFrame as shown in Figure 7.

```
In [7]: sb.heatmap(nairobicrime.isnull())
        plt.savefig("Missing_values.pdf")
```

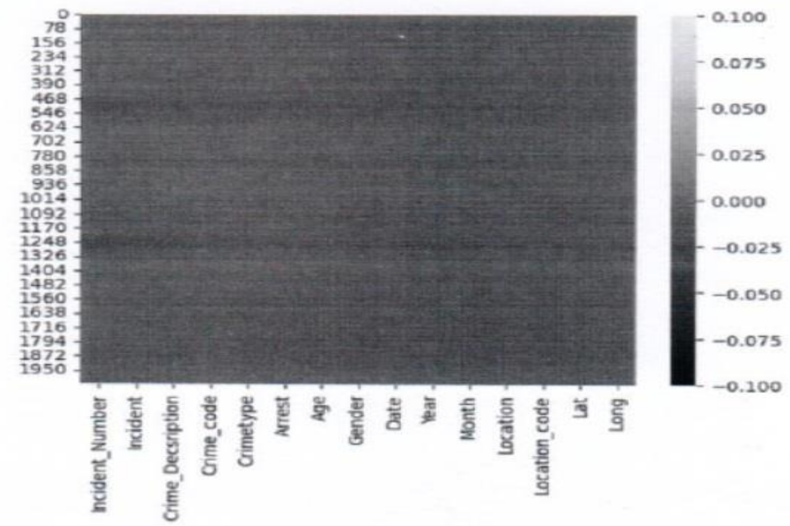

Figure 7: Heatmap for missing values

3.2.2 Dropping of missing values and unnecessary columns.

The necessary columns were retrieved from the data frame by dropping rows that had missing/null values using functions ‘df.dropnat ()’ and ‘df.drop([‘column’],axis=1)’ as shown in Figure 8.

```
In [5]: Arrest1_var=pd.factorize(nairobicrime['Arrest']) # Arrest
        nairobicrime['Arrest']=Arrest1_var[0]
        definition_list_Arrest=Arrest1_var[1]

        gender_var=pd.factorize(nairobicrime['Gender']) # gender
        nairobicrime['Gender']=gender_var[0]
        definition_list_gender=gender_var[1]

        nairobicrime.drop(['Crime_Decsription'],axis=1,inplace=True) # Dropping column Offense
        nairobicrime.drop(['Incident_Number'],axis=1,inplace=True) # Dropping column Offense
        nairobicrime.head(5)
```

Out[5]:

	Crime_code	Crimetype	Arrest	Age	Gender	Date	Year	Month	Location	Location_code	Lat	Long
0	10	Fraud	0	46	0	05-Jul-23	2023	7	Langata	129	-1.335508	36.781960
1	17	Murder	1	51	0	05-Jul-23	2023	7	Kayole	92	-1.266303	36.916927
2	7	Narcotic_Drugs	0	37	1	04-Jul-23	2023	7	Githurai	49	-1.203007	36.916698
3	17	Murder	1	24	0	04-Jul-23	2023	7	Ruaraka	218	-1.243123	36.875125
4	21	Robbery	0	29	0	04-Jul-23	2023	7	Njiru	187	-1.250933	36.933919

Figure 8: Sample dataset with the dropped column

3.2.3 Converting categorical data into numerical values.

This is a process of transforming data by mapping values to concept labels. The LabelEncoder was used to convert the categorical columns into numerical values to extract valuable information from the dataset as shown in Figure 8 above.

3.2.4. Converting the date column to the month

The date column was converted to month (number) and month_name using the function ‘pd.to_datetime()’.

3.3 Classification

The classification was done by categorizing the dataset into two classes namely independent features (X) and dependent features (y). Dependent feature (y) is often referred to as target, label, or category, the column 'Crimetype' was used to hold dependable features as shown in Figure 9.

```
In [9]: col_list=['Crime_code','Gender','Age','Arrest','Year','Location_code','Lat','Long','month','Crimetype']
        DF=nairobicrime[col_list]
        DF.tail(5)
Out[9]:
```

	Crime_code	Gender	Age	Arrest	Year	Location_code	Lat	Long	month	Crimetype
2029	15	0	31	1	2008	112	-1.279815	36.822885	6	Mugging
2030	20	0	36	1	2008	247	-1.287383	36.827336	6	Robbery
2031	1	0	29	1	2008	99	-1.310681	36.790684	4	Assault
2032	21	0	33	1	2001	92	-1.266303	36.916927	10	Robbery
2033	26	0	30	0	2001	77	-1.286059	36.839043	4	Disturbance

Figure 9: Sample of the cleansed dataset.

The classification of data was done to distinguish the types of crime and the prevention measures to be used in each crime occurrence because different crimes require different treatment.

3.3.1 Visualizing target variable

The seaborn function "sns.countplot(DF['Crimetype'])" was used to plot the graph to visualize the target /class variables used for classification and prediction tasks as shown in Figure 10 below.

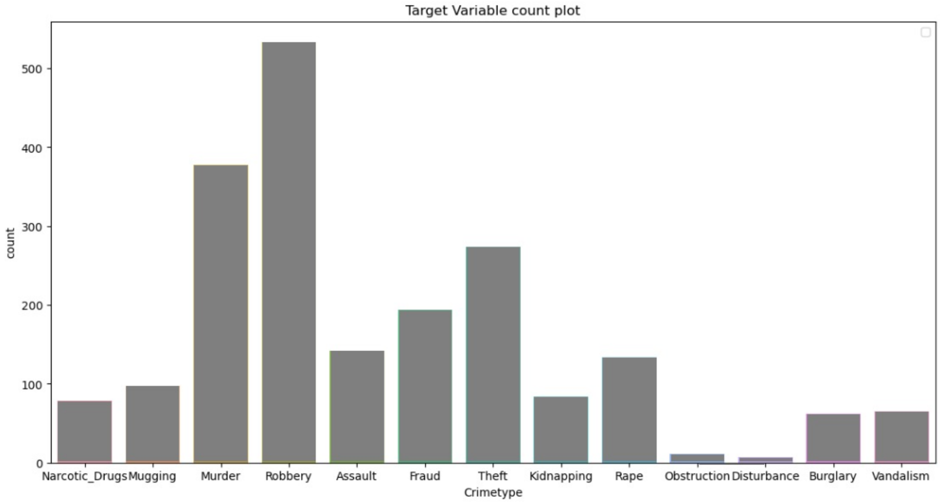

Figure 10: Target variable count

3.3.2 Splitting the dataset into the training and test sets.

The dataset was split into training sets, and test sets using 'train_test_split()' function. The split was done in the ratio of eighty percent (80%) for training and twenty percent (20%) for testing. As a result, the train size was one thousand six hundred and forty-seven (1647) data points while the test size was four hundred and twelve (412) data points as shown in Figure 11. The training was done to teach (train) the algorithm to perform classification and prediction tasks while the validation was done to test the generalization ability of the model on the unseen dataset.

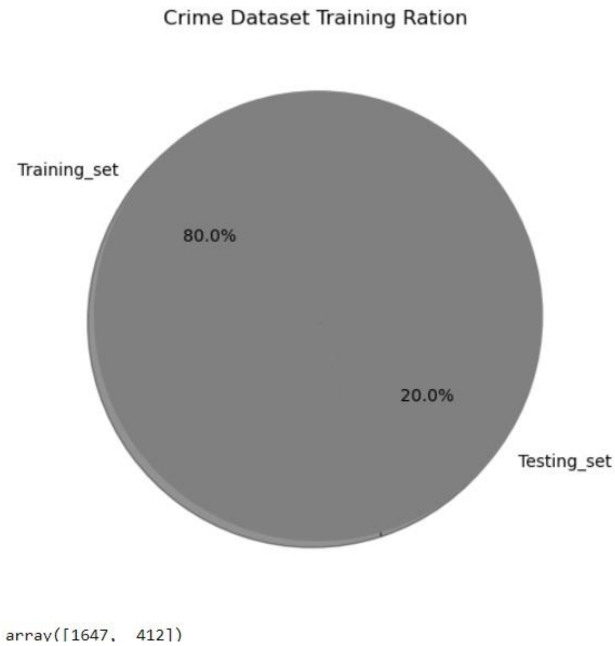

Figure 11: Training and Test set.

3.4 Model Building and Training

Various algorithms were imported from sklearn library to build the models using the training data. The algorithms built were decision tree (DT), random forest (RF), naive bayes (NB), KNeighbors (KNN), and support vector machines (SVM).

3.4.1 Decision Trees

After dividing the dataset into random training and test sets, the Decision Tree was constructed to forecast the target column. The dataset was divided using the splitting criterion "Entropy." This classifier was built by importing the 'DecisionTreeClassifier' from the 'sklearn.tree' library and fitted to the training set using the function 'dt_clf.fit(X_train,y_train)'. The validation was done by predicting the test set results.

3.4.2 Random Forest

With the Random Forest ensemble learning method, many learners are combined to improve the performance of the machine learning model. Decision Trees' propensity to overfit the training dataset is rectified by Random Forest classifiers. It builds several trees throughout training period and produces mean and mode predictions for classification and regression, respectively. The classifier selects the majority judgement of the trees as the final choice. This classifier was built by importing the 'RandomForestClassifier' from the 'sklearn. Ensemble' library and fitted to the training set using the function 'rf_clf.fit(X_train,y_train)'. The validation was done by predicting the test set results.

3.4.3 Naive Bayes

This classification technique makes the premise that predictors are independent based on the Bayes Theorem. It makes the assumption that a feature's inclusion in a class has nothing to do with the inclusion of any other features. This classifier was built by importing the 'GaussianNB' from the 'sklearn.naive_bayes' library and fitted to the training set using the function ‘nb_clf.fit(X_train,y_train)’. The validation was done by predicting the test set results.

3.4.4 K-Nearest Neighbors

This algorithm classifies a data point according to the categorization of its neighbours. A similarity metric is used to categorise new cases and store all of the existing cases. K-Nearest Neighbours is a parameter that specifies how many of the closest neighbours should be considered when casting a majority vote. In K-Nearest Neighbors, the best accuracy was achieved by selecting the right value of K, that is 3 or 5, or 7. This classifier was built by importing the ‘KNeighborsClassifier’ from the ‘sklearn.neighbors’ library and fitted to the training set using the function ‘knn_clf.fit(X_train,y_train)’. The validation was done by predicting the test set results.

3.4.5 Support Vector Machine

The Support Vector Machine algorithms locate a hyperplane that clearly categorises the data points in an N-dimensional space (where N is the number of characteristics). However, the SVM's drawback is that non-linearly separable data cannot be used with it. This classifier was built by importing the ‘SVC’ from ‘sklearn.svm’ library and fitted to the training set using the function ‘ sv_clf.fit(X_train,y_train)’. . The validation was done by predicting the test set results.

3.5. Prediction

The prediction was carried out using ‘model.predict(xtest)’ function. The accuracy was calculated using accuracy_score imported from metrics - metrics.accuracy_score (ytest, predicted).

3.5.1 Model evaluation

The test set was used to evaluate and select the best model for crime prediction. The confusion matrix was deployed to check the performance of each model. A confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. It summarized the number of correct and incorrect predictions yielded by models. The accuracy of the models was determined by looking at the diagonal values for counting the number of accurate classifications. This was obtained by summing all the total numbers in the diagonal and then dividing it by the total number of all observations. For instance, according to the heatmap, the random forest's classification accuracy was determined to be (97%) or 0.973301 as shown in Figure 12.

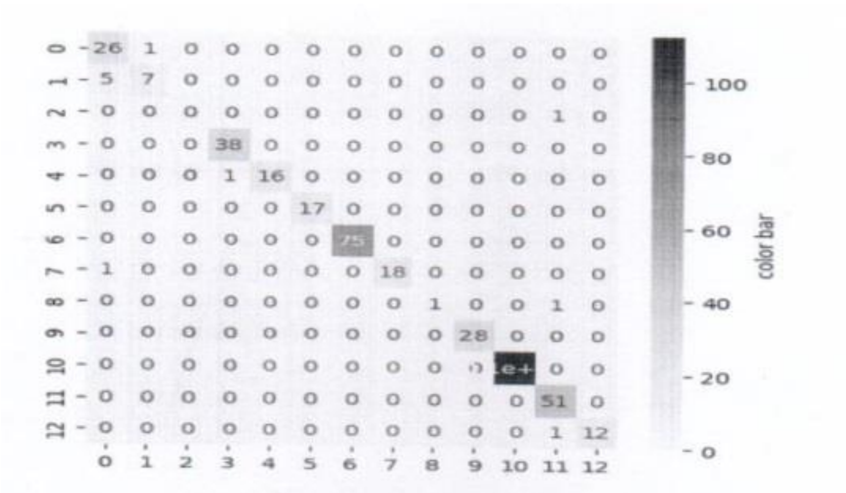

Figure 12 Random Forest (RF) model accuracy heatmap

3.5.2. Data Visualization tools

The matplotlib, seaborn, and folium libraries were utilized to build the data visualization that presented data points in graphs, pie charts, and maps as shown in section 4.

3.6. Ethical Considerations

Given the sensitivity of the subject of crime, this research adhered to ethical principles of secrecy, anonymity, informed consent, and data protection. In all phases of the research procedure, including the voluntary nature of involvement, the freedom to withdraw at any time, and the measures used to preserve and anonymize data, ethical permission was obtained.

4. Results and Discussions

This section presents implementation results and comparative analysis of decision tree (DT), random forest (RF), naive bayes (NB), KNeighbors (KNN), and support vector machine (SVM) on the crime dataset described in the previous sections.

4.1 Comparative Analysis of the Machine Learning Algorithms

The developed algorithms included decision tree (DT), random forest (RF), naïve bayes (NB), kneighbors (KNN), and support vector machines (SVM) were tested for performance accuracy using a confusion matrix. The preliminary experiments were conducted on nine (9) datasets of different instances such as 200, 400, 600, 900, 1000,1200, 1800, 2000, and 2059. Each of the five (5) machine learning models was applied to each dataset and their performance accuracy was computed and recorded as shown in Figure 13 below.

	DATASET	NB	RF	SVM	DT	KNN
0	200	0.775000	0.750000	0.800000	0.725000	0.425000
1	400	0.912500	0.825000	0.950000	0.812500	0.475000
2	600	0.850000	0.908333	0.825000	0.800000	0.458333
3	900	0.927778	0.922222	0.822222	0.755556	0.455556
4	1000	0.945000	0.945000	0.845000	0.775000	0.480000
5	1200	0.933333	0.954167	0.845833	0.816667	0.537500
6	1800	0.941667	0.961111	0.866667	0.797222	0.586111
7	2000	0.915000	0.962500	0.885000	0.802500	0.562500
8	2059	0.922330	0.973301	0.866505	0.815534	0.546117

Figure 13: The models' performance accuracy.

The validation was done by predicting the test set and examining the accuracy score on the existing dataset with two thousand and fifty-nine (2059) rows/instances. The confusion matrix checked and visualized the performance of each model through the heat map as shown in Figure 3.8 above. The performance of individual machine learning algorithms was analyzed and presented as shown in Table 2.

Table 2: Algorithms' performance comparison.

S/No.	Model	accuracy	Score (%)	Limitations
1	Random Forest (RF)	0.973301	97%	It makes predictions with high accuracy for huge datasets, but if there are too many trees, the algorithm may be too sluggish and inefficient to generate predictions in real time. Even when a significant amount of data is absent, accuracy can still be maintained. It requires less time to train than other models.
2	Naive Bayes (NB)	0.922330	92%	The Naive Bayes model will give it zero probability and won't be able to make any predictions if your test data set contains a categorical variable of a category that wasn't present in the training data set. It can, however, outperform other models and needs considerably less training data if its premise about the independence of characteristics is correct. It works better with category input variables than it does with numerical ones.
3	Support Vector Machine (SVM)	0.866505	87%	It has low performance on a large dataset. However, It doesn't perform well when classes are not distinct.It can handle large feature sets efficiently
4	Decision Tree (DT)	0.815534	82%	They are unstable; a minor alteration in the data can result in a significant alteration in the structure thereby affecting its performance. - it causes instability if there is any change in the data.Not suitable for large datasets
5	K-Nearest Neighbors (KNN)	0.546117	55%	Data-filling algorithms need to be added to increase accuracy for example adjusting the value of K. It doesn't work well with large datasets. It doesn't handle categorical features very well

The accuracy of each model was calculated using the function accuracy_score by importing metrics.accuracy_score (y_test, predicted) from sklearn's metrics function and presented the scores using a multiple-line graph as shown in Figure 14 below.

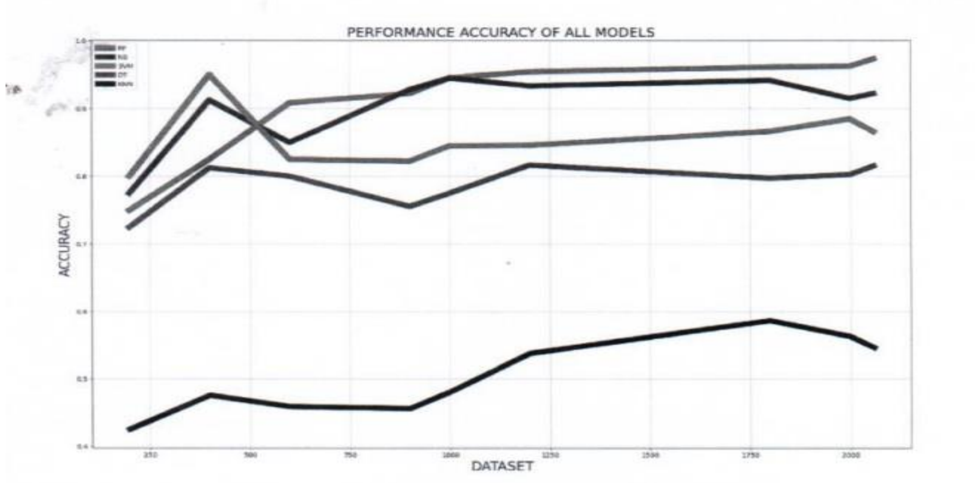

Figure 14: Algorithms performance using a line graph.

The results from the research findings indicate that the random forest model outperformed the other models with a classification accuracy of **0.973301** as shown in Figure 15 below.

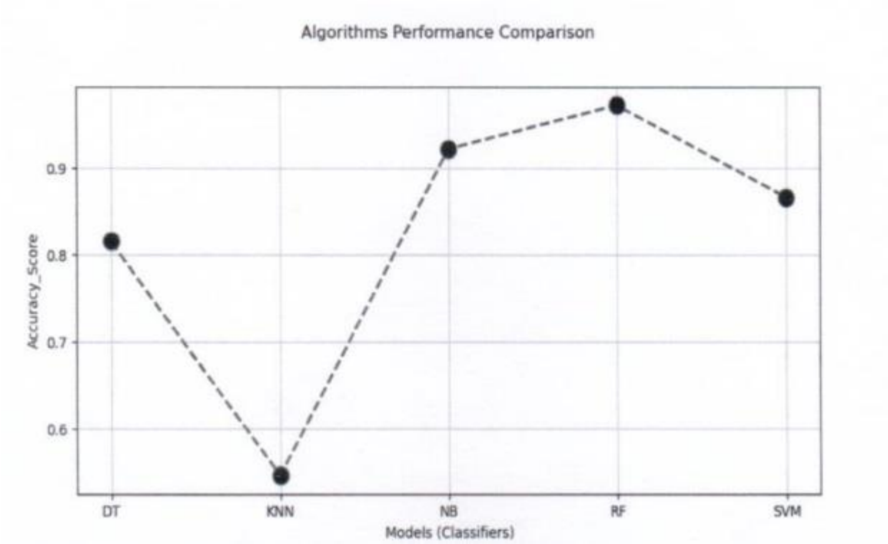

Figure 15: Algorithms performance comparison.

4.2 Model Deployment

Model deployment is the action of implementing machine learning models. The model was applied to an existing dataset of four hundred and seventy-eight (478) instances and nine (9) columns collected from January 2023 to July 2023 as shown in Figure 16 and evaluated under increasing production conditions by iteratively running the tasks.

	Crime_code	Gender	Age	Arrest	Year	Location_code	Lat	Long	month
473	15	1	19	1	2023	194	-1.283308	36.824899	1
474	17	1	35	1	2023	34	-1.278557	36.848633	1
475	17	0	34	0	2023	151	-1.262544	36.860663	1
476	15	1	23	1	2023	252	-1.286906	36.883107	1
477	17	1	30	1	2023	75	-1.186979	36.906021	1

Figure 16: Existing dataset between January 2023 to July 2023.

4.2.1. Crime Prediction

The prediction was done by feeding the following attributes, ‘Crime_code’, ‘Gender’, ‘Age’, ‘Arrest’, ‘Year’, ‘Location_code’, ‘Lat’, ‘Long’ and ‘Month’ into the model to predict and fetch different categories of crime as shown in Figure 17. The ‘model. predict(xtest)’ function was used to carry out the prediction tasks.

```
Out[49]:  array(['Narcotic_Drugs', 'Mugging', 'Murder', 'Robbery', 'Murder',
                 'Assault', 'Robbery', 'Fraud', 'Murder', 'Narcotic_Drugs', 'Fraud',
                 'Murder', 'Robbery', 'Murder', 'Narcotic_Drugs', 'Theft', 'Theft',
                 'Murder', 'Robbery', 'Assault', 'Murder', 'Theft',
                 'Narcotic_Drugs', 'Kidnapping', 'Robbery', 'Kidnapping', 'Rape',
                 'Murder', 'Assault', 'Murder', 'Theft', 'Assault', 'Murder',
                 'Fraud', 'Mugging', 'Robbery', 'Theft', 'Murder', 'Theft',
                 'Vandalism', 'Murder', 'Murder', 'Murder', 'Theft', 'Robbery',
                 'Robbery', 'Theft', 'Assault', 'Murder', 'Murder', 'Assault',
                 'Obstruction', 'Obstruction', 'Obstruction', 'Assault', 'Murder',
                 'Assault', 'Assault', 'Fraud', 'Disturbance', 'Murder', 'Robbery',
                 'Kidnapping', 'Murder', 'Murder', 'Murder', 'Murder', 'Fraud',
                 'Fraud', 'Theft', 'Narcotic_Drugs', 'Robbery', 'Assault', 'Fraud',
                 'Robbery', 'Robbery', 'Disturbance', 'Vandalism', 'Assault',
                 'Fraud', 'Murder', 'Robbery', 'Robbery', 'Robbery', 'Robbery',
                 'Assault', 'Fraud', 'Assault', 'Robbery', 'Theft', 'Theft',
                 'Assault', 'Rape', 'Murder', 'Rape', 'Fraud', 'Murder', 'Murder',
                 'Rape', 'Fraud', 'Murder', 'Murder', 'Robbery', 'Assault',
                 'Assault', 'Narcotic_Drugs', 'Fraud', 'Murder', 'Fraud', 'Murder',
                 'Theft', 'Theft', 'Theft', 'Fraud', 'Murder', 'Murder', 'Assault',
                 'Assault', 'Rape', 'Robbery', 'Kidnapping', 'Murder', 'Murder',
                 'Robbery', 'Robbery', 'Robbery', 'Fraud', 'Murder', 'Assault',
                 'Fraud', 'Fraud', 'Narcotic_Drugs', 'Kidnapping', 'Murder',
                 'Murder', 'Robbery', 'Rape', 'Assault', 'Narcotic_Drugs', 'Murder',
                 'Murder', 'Murder', 'Murder', 'Murder', 'Murder', 'Murder',
                 'Theft', 'Theft', 'Fraud', 'Theft', 'Fraud', 'Robbery', 'Theft',
                 'Murder', 'Theft', 'Fraud', 'Murder', 'Theft', 'Fraud', 'Murder',
                 'Murder', 'Murder', 'Murder', 'Murder', 'Murder', 'Murder',
                 'Murder', 'Murder', 'Robbery', 'Theft', 'Fraud', 'Fraud', 'Theft',
                 'Murder', 'Murder', 'Theft', 'Narcotic_Drugs', 'Fraud', 'Fraud',
                 'Kidnapping', 'Narcotic_Drugs', 'Obstruction', 'Assault',
                 'Kidnapping', 'Murder', 'Robbery', 'Robbery', 'Murder', 'Theft',
                 'Theft', 'Fraud', 'Fraud', 'Murder', 'Theft', 'Fraud', 'Murder',
                 'Murder', 'Murder', 'Murder', 'Narcotic_Drugs', 'Theft', 'Assault',
                 'Assault', 'Robbery', 'Rape', 'Fraud', 'Murder', 'Murder',
                 'Robbery', 'Obstruction', 'Disturbance', 'Fraud', 'Theft',
                 'Murder', 'Murder', 'Murder', 'Murder', 'Rape', 'Rape', 'Rape',
                 'Assault', 'Kidnapping', 'Murder', 'Rape', 'Rape', 'Vandalism',
                 'Fraud', 'Robbery', 'Disturbance', 'Vandalism', 'Assault', 'Fraud',
                 'Kidnapping', 'Mugging', 'Vandalism', 'Assault', 'Murder',
                 'Murder', 'Murder', 'Theft', 'Disturbance', 'Vandalism', 'Assault',
```

Figure 17: Predicted types of crime.

4.2.2 Observed values vs. predicted values.

The observed values are the actual values that are obtained by observation while the predicted values are the values of the variable predicted based on the classification. The predicted values are contained in the column named "Crimetype_predicted" as shown in Figure 18.

	Location	Month	Crimetype_predicted	Frequency	Crime_level
0	Biashara Lane	February	Assault	1	Low
1	City Hall Way	March	Assault	1	Low
2	Dandora	January	Assault	3	High
3	Embakasi	April	Assault	1	Low
4	Embakasi	March	Assault	2	Moderate
5	Fedha Estate	April	Assault	1	Low
6	Githurai	April	Assault	1	Low
7	Githurai	March	Assault	1	Low
8	Githurai	May	Assault	1	Low
9	Huruma	March	Assault	1	Low
10	Kahawa West	January	Assault	1	Low
11	Kamukunji	June	Assault	1	Low
12	Kangemi	March	Assault	1	Low
13	Kariobangi North	May	Assault	2	Moderate
14	Kasarani	March	Assault	1	Low

Figure 18: Observed values vs predicted values.

4.3 Data Visualization

The crime dataset was analyzed using data visualization tools mentioned in the previous section to examine the relationships between attributes that would make it possible to predict crime occurrences. Different categories of crime were depicted using markers of different colors in a heat map. The data analysis was done using matplotlib, seaborn, and folium functions, and the results were presented as follows.

- Types of crime indicators
- Types of crime committed over time between January 2023 to July 2023
- Trends in crime in Nairobi County
- Crimes that are committed across different locations.
- The pattern of crime occurrences by locations

4.3.1. Types of crime indicators

Table 3 shows the overall number of crimes predicted based on new cases reported between January 2023 and July 2023. Murder, Robbery, Theft, Fraud, and assault recorded a high number of cases that occurred within various locations within Nairobi County.

Table 3: Predicted crime indicators

Crimetype	Total_Occurence	Percentage
Murder	83	20.6
Robbery	75	18.61
Theft	56	13.9
Assault	54	13.4
Fraud	49	12.16
Rape	19	4.71
Narcotic_Drugs	17	4.22
Vandalism	16	3.97
Kidnapping	10	2.48
Obstruction	8	1.99
Mugging	7	1.74
Disturbance	5	1.24
Burglary	4	0.99

According to Figure 19 below, Murder led with 20.6 % of all predicted crimes, followed by Robbery at 18.61%, Theft at 13.09 %, Assault at 13.4%, Fraud at 12.16%, and the rest of the crimes were predicted to occur below 5%.

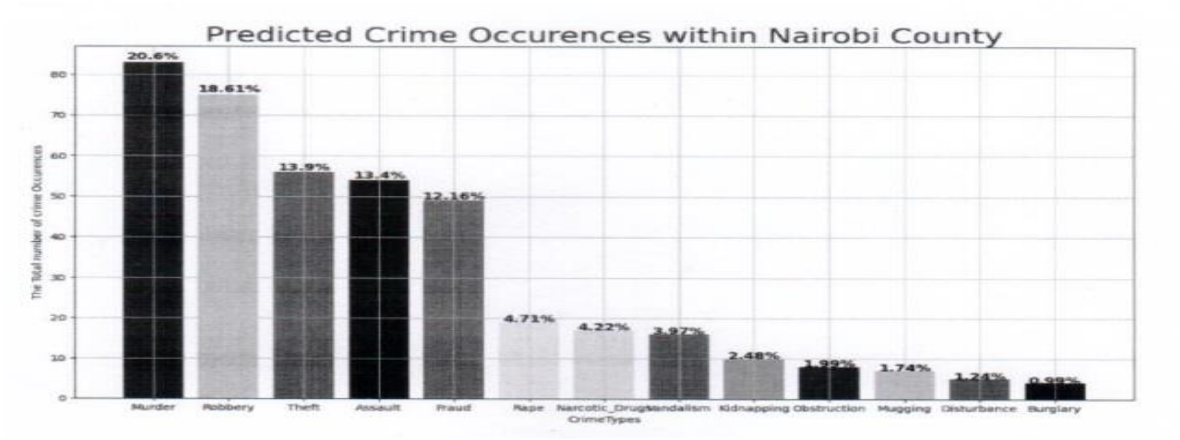

Figure 19: Total number of crime cases predicted.

4.3.2 Types of crime committed over time between January 2023 to July 2023

According to Table 4 and Figure 20 below, the month of April 2023 recorded a high number of crime cases at (88) accounting for 21.84% of all crime cases that were likely to occur between January 2023 and July 2023. It was followed by March with (86) cases accounting for 21.34% of the total crime cases predicted. The month of February had (59) cases at 14.64%, May had (58) cases at 14.39%, January had (54) at 13.4%, June had (41) at 10.17% and July had (17) at 4.22%.

Table 4: Predicted crime occurrences by time

Month	Total	Percentage
January	54	13.4
February	59	14.64
March	86	21.34
April	88	21.84
May	58	14.39
June	41	10.17
July	17	4.22

The other coming months were projected to record less than 15% likelihood of crime occurring.

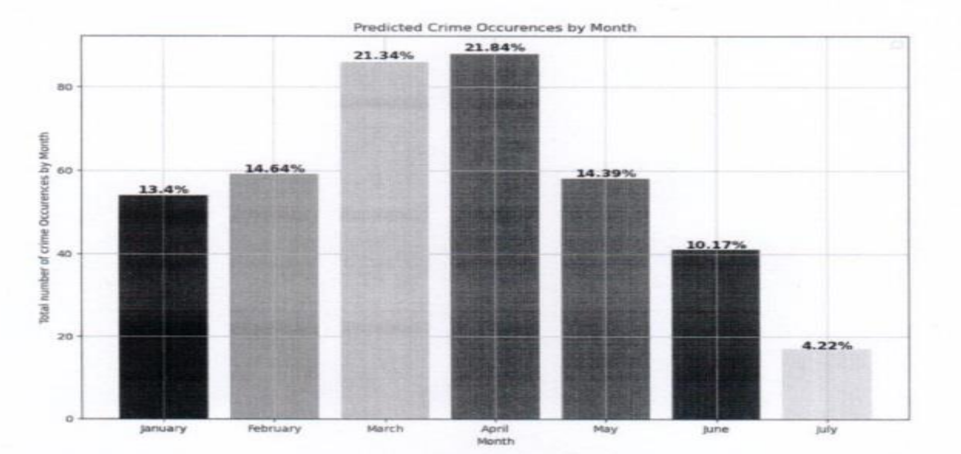

Figure 20: Predicted crime over time.

4.3.3 Trends in Crime in Nairobi County

There was an upward trend from January through to April 2023, this was probably due to the erratic nature of crime then reduced from April 2023 to July 2023 as shown in Figure 21 below. The reduction was probably based on measures the government put in place to mitigate crime from the month of April 2023. If such trends continue, then during the following months after July 2023, many residents of Nairobi County are likely to experience fewer criminal activities.

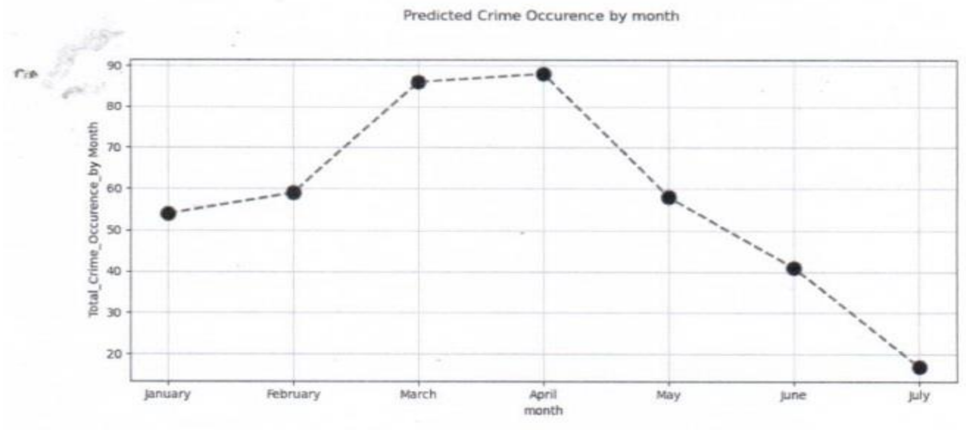

Figure 21: The trend in crime every month

4.3.4 Crimes committed across different locations.

Figure 22 below indicates the crime levels based on place, time, and nature of the crime. The level of crimes such as Assault, Robbery, Murder, Theft, and Vandalism were high in Dandora, Kibera, Mathare, Eastleigh, Kasarani, Kilimani, Huruma, and Mama Ngina Street between the months of January 2023 and April 2023.

	Location	Month	Crimetype_predicted	Frequency	Crime_level
0	Dandora	January	Assault	3	High
1	Kibera	March	Assault	6	High
2	Mathare	March	Assault	3	High
3	Eastleigh	April	Murder	3	High
4	Kasarani	April	Murder	6	High
5	Kasarani	February	Murder	4	High
6	Kilimani	April	Murder	3	High
7	Mama Ngina Street	January	Robbery	3	High
8	Dandora	January	Robbery	3	High
9	Huruma	February	Robbery	3	High
10	Mama Ngina Street	January	Robbery	3	High
11	Kasarani	April	Theft	3	High
12	Kibera	March	Vandalism	3	High
13	Biashara Lane	February	Assault	1	Low
14	City Hall Way	March	Assault	1	Low

Figure 22: Crime levels

From the research findings, areas with high crime levels as shown in Figure 23 require constant Police intervention such as heightened patrols, police crackdowns, neighborhood watch, and community policing to reduce the level of crime to Low or Moderate.

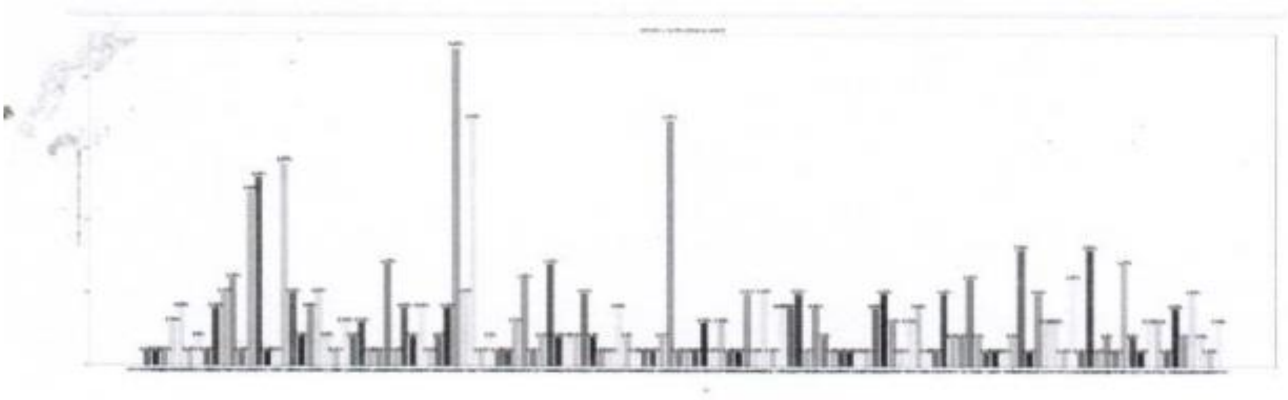

Figure 23: Areas with high crime levels.

4.3.5 The pattern of crime occurrences by locations

A crime pattern is a considerable change in crime occurrences within a given geographical area over time. Crime cases reported between January 2023 and July 2023 were linked with information to a map to present the pattern of crime occurrences. This was done to aid in tracking crime from location to location and recognizing the patterns in them in real time. Markers of different colors were used to represent different types of crime on a map. The markers colors that were added to the map included 'red', 'blue', 'green', 'purple', 'orange', 'dark red',' light red, 'beige', 'dark blue', dark green', 'cadet blue', dark purple', 'white', 'pink', 'light blue', 'light green', 'gray', 'black', 'light gray' as shown in Figure 24 below.

```
In [86]: def color(typeofcrime):
             if typeofcrime== 'Assault':
                 return 'blue'
             elif typeofcrime== 'Burglary':
                 return 'gray'
             elif typeofcrime== 'Mugging':
                 return 'orange'
             elif typeofcrime== 'Murder':
                 return 'red'
             elif typeofcrime== 'Narcotic_Drugs':
                 return 'beige'
             elif typeofcrime== 'Rape':
                 return 'green'
             elif typeofcrime== 'Robbery':
                 return 'purple'
             elif typeofcrime== 'Theft':
                 return 'pink'
             elif typeofcrime== 'Vandalism':
                 return 'black'
             elif typeofcrime== 'Fraud':
                 return 'darkred'
             elif typeofcrime== 'Kidnapping':
                 return 'darkgreen'
             elif typeofcrime== 'Disturbance':
                 return 'darkblue'
             elif typeofcrime== 'Obstruction':
                 return 'cadetblue'
             else:
                 return 'lavender'
```

Figure 24: Marker colors for distinct types of crime

From the research findings, Figure 25 above shows different types of crime committed in different locations and the markers of different colors reveal the patterns of each crime occurrence as shown in Figure 4.12 below. For instance, a murder that occurred in one location might be a match for a murder that would probably occur in another location in the future. The areas with high crime levels provided an enabling environment for crime to thrive. This is probably because it could take a long time to arrest criminals because of the large population. This delay allowed crimes to occur without being detected.

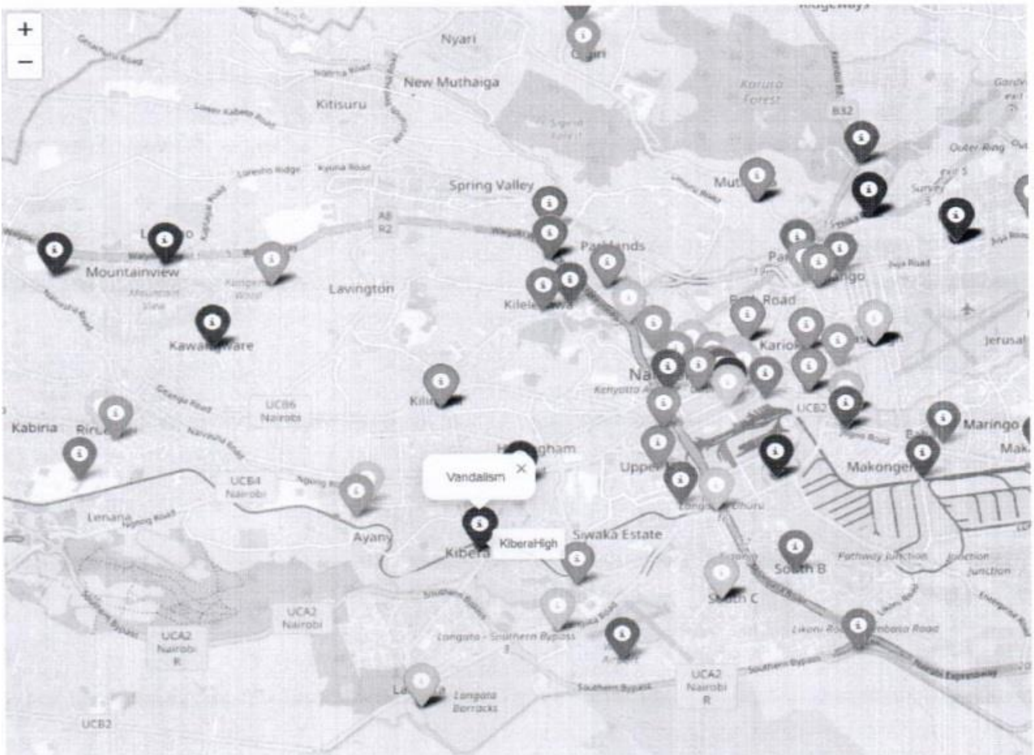

Figure 25: Pattern of crime occurrences.

5. Conclusion

Crime prediction and mapping is one of the current developments in crime prevention and the goal is to lower the number of crimes that occur. This was achieved by collecting raw datasets from various sources including law enforcement organizations within Nairobi County and various websites sources through web scraping and stored in a CSV file on the computer's hard disk. The dataset was then pre-processed and transformed into a proper form for further processing. Different machine learning), naives such as decision tree (DT), Kneighbors (KNN), support vector machine (SVM), naive bayes (NB), and random forest (RF) were developed and evaluated using a confusion matrix and based on the results of the experiments, the random forest model was found to be the best method for estimating the likelihood of a crime occurring in a specific place, with a classification accuracy of 97%. The prediction was done by feeding the following attributes, 'Incident_Number', 'Incident', 'Crime_Decsription', 'Crime_code', 'Crimetype', 'Arrest',' Age', 'Gender', 'Date', 'Year', 'Month', 'Location', 'Location_code', 'Lat' and 'Long into the best model to predict and fetch different categories of crime. The longitude and latitude features were used to tag the specific locations of crime occurrences on a map. Markers of different colors were used to reveal the patterns of each crime occurrence. Data visualization was done to identify types of crime that might have occurred in a specific location under a variety of parameter constraints. Various interactive plots such as bar graphs, line graphs, and pie charts were created to present the data points. The experimental results from the research findings showed that every location is known with a particular crime type and hence, crime prediction and mapping intend to identify the types of crime that are likely to take place in a location at a particular time. This information can be used to distinguish the types of preventive measures to be used for each type of crime.

The future enhancement of the model is to integrate crime with an interactive user interface like a web portal so that the model can be easily accessible by anyone intending to report a crime. The web portal will enable users to report crimes by entering details of the crime in the database in real time as opposed to the current methods of manual input or recordings. The aim is to make this model a centralized system that connects all law enforcement offices across the country to report crime online. This would be quite easier to predict crimes in a location and recognize the patterns in them in real-time.

References

Dikananda et al (2022). Comparison of Decision Tree Classification Methods and Gradient Boosted Trees. TEM Journal. 11. 316-322 https://doi.org/10.18421/TEM111-39

Kanimozhi et al (2021). Crime Type and Occurrence Prediction Using Machine Learning Algorithm, International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India. https://doi.org/10.1109/ICAIS50930.2021.9395953.

Llaha (2020). Crime Analysis and Prediction using Machine Learning, 43rd International Convention on Information, Communication and Electronic Technology *(MIPRO)*, Opatija, Croatia. https://doi.org /10.23919/MIPRO48935.2020.9245120.

Mahesh & Dipti (2020). Machine Learning Algorithms for Predictive Analytics: A Review and New Perspectives.

Mohamed et al (2022). Supervised Machine Learning Techniques.

Mahmud et al (2021). Crime Rate Prediction Using Machine Learning and Data Mining. In: Borah, S., Pradhan, R., Dey, N., Gupta, P. (eds) Soft Computing Techniques and Applications. Advances in Intelligent Systems and Computing, vol 1248. https://doi.org/10.1007/978-981-15-7394-1_5

National Police Service (NPS)(2022),"Annual Report" .

Nguyen et al (2018). Applying Random Forest Classification to map land use/land cover using Landsat 8 oli, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-3/W4, pp 363–367,2018. https://doi.org/10.5194/isprs-archives-XLII-3-W4-363-2018

Patrick and Erin (2020). Optimizing the C4.5 Decision Tree Algorithm using MSD-Splitting, International Journal of Advanced Computer Science and Applications(IJACSA), pp 11(10). http://dx.doi.org/10.14569/IJACSA.2020.0111006

Pratibha et al (2020), Crime Prediction and Analysis, 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India. https://doi.org /10.1109/IDEA49133.2020.9170731.

Sarker (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. Sn Comput. Sci. Pp 2. https://doi.org/10.1007/s42979-021-00592-x

Sen (2021). Machine Learning - Algorithms, Models and Applications. IntechOpen . https://doi.org /10.5772/intechopen.94615.

Saraiva et al (2022). Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics. ISPRS Int. J. Geo-Inf,. https://doi.org/10.3390/ijgi11070400.

Tahir et al (2021). Crime Prediction using Naïve Bayes Algorithm .International Journal of Advance Research, Ideas, and Innovations in Technology, Volume 7, Issue 4 - V7I4-1713. www.IJARIIT.com.

Veena et al (2022). Cybercrime: Identification and Prediction Using Machine Learning Techniques. Computational Intelligence and Neuroscience, 1-10. https://doi.org/10.1155/2022/8237421

Viet et al (2021). The Naïve Bayes Algorithm for Learning Data Analytics, Indian Journal of Computer Science and Engineering. 12. 1038-1043. https://doi.org/10.21817/indjcse/2021/v12i4/211204191

Wasim et al (2020). Crime Analysis And Prediction Using The K-Means Clustering Technique. Epra International Journal of Economic and Business Review. 05. 277-280.

Yoganand et al (2020). An User-Friendly Interface for Data Preprocessing and Visualization using Machine Learning Models.

Zeineddine et al (2020). Enhancing Prediction of Student Success: Automated Machine Learning Approach. Computers & Electrical Engineering. 89.