Original Article

Crime Prediction and Analysis

R. Karthik Sriraam[1], S.M. Keerthivasan[2], K. Sukant[3], A. Krishnamoorthy[4*]

[1,2,3]UG Student, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

[4*]Assistant Professor (Senior), School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

Abstract

The systematic technique used to detect and analyze crime patterns and trends is called crime analysis and prevention. The proposed model will be able to predict regions that have a high chance of crime occurrence. Crime data specialists can assist police officers in stopping crime more quickly. Our goal is to develop methods that can forecast homicide activities based on demographic and economic data from a specific area. In this project, we use a classification and clustering algorithm to build an effective crime prediction model. Support vector networks, multivariate time- series data, and artificial neural networks are some of the prediction approaches that can be used to test the efficiency of prediction models. We use multiple algorithms for classification and clustering and then we compare with accuracy. To measure the out-of-training effectiveness of classifiers, we use a ten- fold cross-validation approach.

Keywords: Decision Tree Classifier, k-means, GaussianNB.

DOI: 10.47750/pnr.2022.13.S03.009

Introduction

Crime is growing substantially these days, creating a great threat to a capital's law enforcement. Various law enforcement agencies have gathered a massive amount of data over the last few decades. Data does not relate to a specific criminal act; rather, it encompasses data on several kinds of criminal acts that have occurred in various states and municipalities all over the country. The best path to justice for the perpetrator has been to solve the offenses. Both for scientists and analysts, data presents an enormous number of obstacles and possibilities. Investigators can use the qualities of the data to uncover associations that will help law enforcement identify that offender. For discovering trends and intricate details in reported crimes, a different approach to analyse the data is required. The fundamental goal of crime prediction is to figure out what kind of criminality will occur in whatever location. Forecasting crime gets tough and complicated, particularly when evidence about the crime committed is included in the study, such as the convict's background, the convict's social network of operation, and any form of spatial information, such as the date of the occurrence and the area of the illegal activity.

Background

Here we have used classification and clustering algorithms to build an effective crime prediction model. These are all the processes in conducting a criminal investigation:

1. Data Collection
2. Classification
3. Pattern Identification
4. Prediction
5. Visualization

We use 3 metrics. Those are accuracy, precision, and recall.

Accuracy - The simplest obvious evaluation method is accuracy, which is just the proportion of properly appropriate patterns to all sightings. One can believe that our model is more appropriate if it is accurate. Yes, accuracy is a useful statistic, but only when the records are neutral and the ratios of false positives and false negatives are nearly equal. Therefore, you have to look at other parameters to evaluate the performance of your model.

Precision - The proportion of accurately predicted positive sightings to overall expected positive sightings is known as precision. The query that this measure answers is that many of the passengers who were identified as having saved actually did. The low false positive rate is related to high precision.

Recall (Sensitivity) - Yes, recall is defined as the proportion of accurately correctly predicted observations to all observations in the category. How many of the passengers who genuinely survived were labeled, according to the question?

Literature Survey

1. This paper uses grunt shell algorithm to analyze the data. First, they find out the crime area wise in the city. Second, they order crimes by their types. Third, the crime happened based on the age group. The efficiency pace of the Grunt shell method is raised by up to twenty percent, so time utilization plays a significant part in processing such massive volumes of data. As a result, this method clearly describes the crime being committed, the perpetrator, and the type of individual affected. The data set used is Crime data collected in the whole of Chennai city for one year.
2. This paper uses the K-means algorithm. The K-means method divides data into clusters according to their mean. The expectation-maximization method is an add-on of the K-means approach that partitions data depending on its variables. It is simple to identify criminal offense locations using the grouped data, that can be used to develop future prevention techniques. The coordinates of every offense recorded in the online system are used to generate these crime statistics. Data analysis is primarily used to differentiate the types of precautionary actions that should be taken with each incident. Various offenses demand different punishments, which may be easily accomplished with this software.
3. They investigate the application of attribute data and classification methods for crime data using data analysis in this study. To ease the process of adding the required data and identifying crime trends, they employed the MV approach and the Apriori approach with certain additions. They used genuine crime statistics from metropolitan police to test their ideas. They also employ nearly fully teaching techniques to assist them to gain an understanding of criminal convictions and improve prediction performance.
4. They have implemented an Automatic Crime Detector Structure analysis and Network Mapping are different analytical methodologies that are combined in this architecture. This methodology can uncover patterns of crime from enormous amounts of criminal information, rendering crime intelligence organizations' jobs easier. The dependence matrix can help a researcher by presenting a complete image of how reliant a given point is on others and how reliant others are on that specific element. The database contains criminal backgrounds from 2005 to 2014, which were gathered from the Los Angeles City Deputy's Agency's official web page.
5. To judge the efficiency of each found cluster, they employed a clustering method. Kernel density estimation (KDE), the most suited spatial analytical method for visualizing crime data, was applied for the investigation of crime groupings. This stage employs a systematic technique to determine the minimum number of homicide rates necessary for a clustering to be considered salient. The crime data used for this study consists of 100,000 illegal occurrences during five years periods across a region of 457, 23400,040 m2.
6. In Tamilnadu, crime inspection and prediction is a process for recognizing and dissolving instances and trends of misbehavior against women. Our methodology may predict criminal occurrences in urban areas with a large number of people living in areas, offices, and corporate entities, as well as crime point regions. By gaining greater access to data, crime in series specialists can assist rule implementation chiefs and special evaluation police groups in pursuing additional investigations into understanding infractions. We can use the origination of in process that focuses to extract previously unknown, useful facts from large amounts of data. In this work, we discuss how to improve an information accumulation system that will help solve crimes against women more quickly by incorporating computer programming and criminal justice. We tend to concentrate mostly on the crime components of each day rather than specializing in obstacles of crime frequency such as a crook's criminal record.
7. Computer security is on the rise in this rapidly evolving technology arena, putting investigators' capabilities to the test. The compilation of crime statistics, which is primarily electronic, has also increased significantly in recent years. Conventional analysis techniques can no longer effectively handle today's created datasets. For that large amount of data, Big Data Analytics would just be preferable to standard data processing approaches. Essentially, acquired data will be dispersed across a territorial place, and groupings will be formed as a result. The formed groups are then evaluated utilizing Big Data Analytics in phase two. Ultimately, the examined groups are fed into an Artificial Neural Network, which generates a predictive sequence. Safety agencies can utilize this forecast pattern to assist them to allocate resources and reducing crime. Synopsis - Big data analytics is being used to analyze criminal offense data in order to construct a criminal forecasting model.
8. Criminal behavior is one of our society's most serious issues. With the resurgence of this kind of action around the world daily basis, homicide investigation organizations are finding it harder and harder to manage and investigate events, either due to a lack of police or just because offenders are outsmarting the investigative process. The conventional police investigative procedure takes a very long time to anticipate criminal characteristics, anticipate the next prospective crime place, or understand the crime trend. As a result, hence the need to investigate historical crime trends more

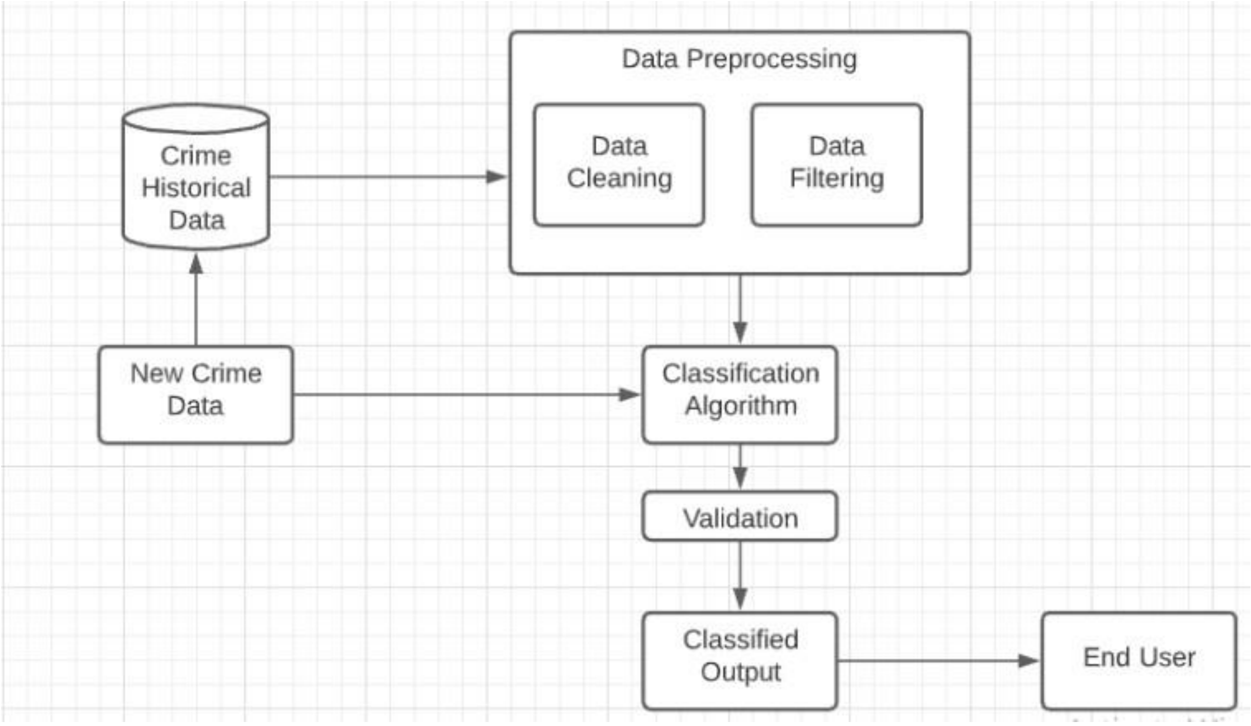
accurately in a smaller duration of time, as well as anticipate future crime location and size. Law enforcement involves a methodical method for quickly criminological investigation profiles and identifying offenders who may be linked to that crime. For criminal behavior tracking, an advanced statistical system is also required to track other information such as vehicle sensing devices, voice message, recordings, and law enforcement service messages among many other things. We highlighted how Big Data-based data analysis techniques might be used to avoid dealing with such situations in this paper. Furthermore, we have examined various data-gathering methodologies, including Volunteered Geographic Information (VGI), Geographic Information System (GIS), and Web 2.0. The judgment derived from data gathering and examination will be the final phase. It will be accomplished through the use of Deep Learning to identify and prevent future crimes.

9. Big data is a term that refers to describe the collecting of massive amounts of data made by various applications including social media, e-commerce, and so on. The storing and processing of such enormous amounts of data proved to be time-consuming. To overcome this difficulty, numerous tools and techniques have emerged in recent years. One of the applications where a large quantity of info is rapidly growing is an offense, which poses a significant challenge for the administration in making informed decisions while adhering to the rules and regulations. Big Data Analytics was used to keep track of this massive amount of data. When compared to the conventional method of analysis, the computation of this massive amount of data collected can be done quickly. The administration and the citizens will be supplied with protection and a healthy country as a result of these analyses.

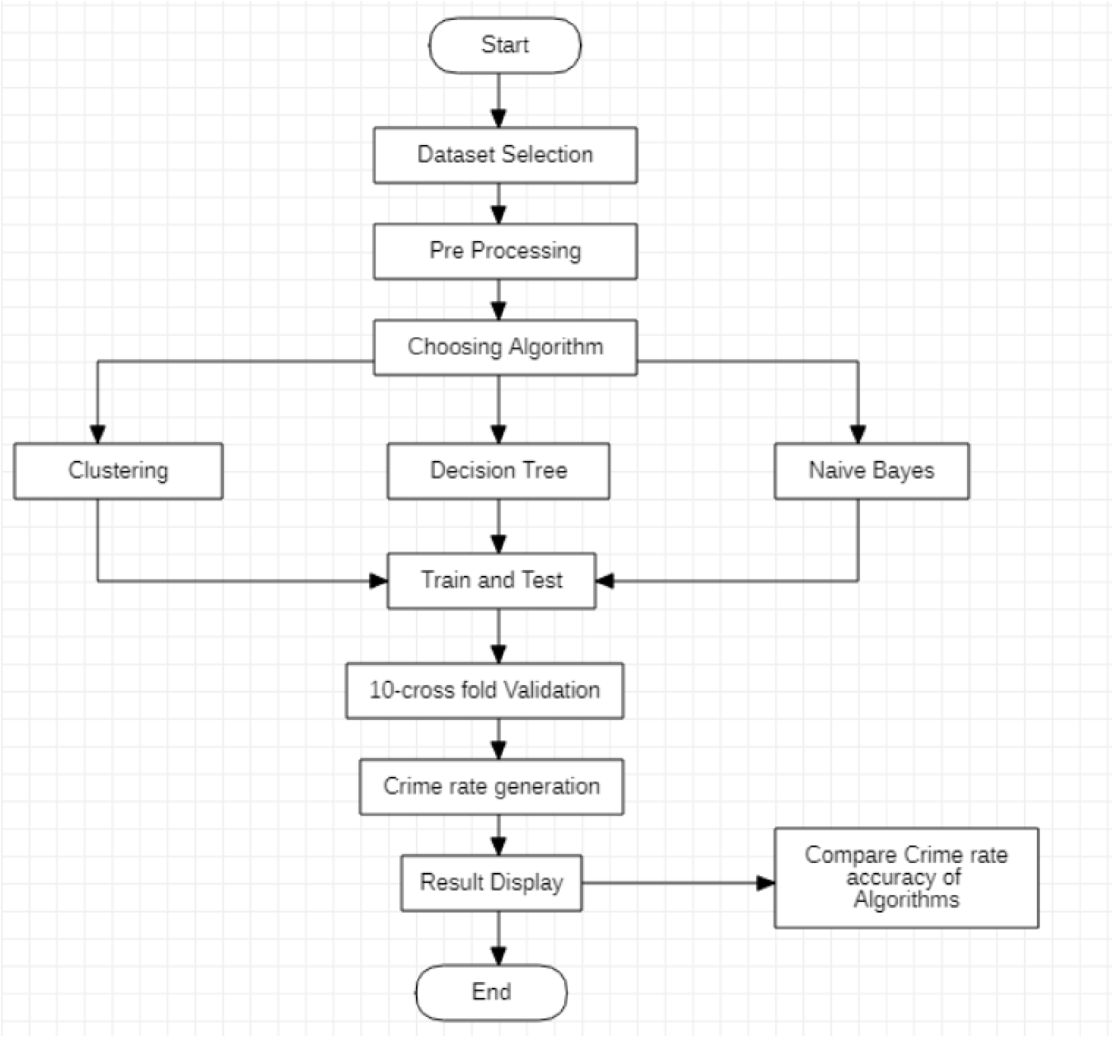
Proposed Algorithm with Flowchart Proposed System

In this project, we used a classification and clustering algorithm to build an effective crime prediction model. The justification for using this technique is that we should only have statistics on reported offenses, thus we'll receive a criminal trend for a certain location. As a result, a categorization system based on previously solved offenses will not be able to anticipate future attacks with high accuracy. Also, because the type of crimes changes over time, clustering approaches are more effective in detecting fresh and unexpected patterns in the upcoming. We are trying to use multiple algorithm to classify and cluster. then we compare the algorithm according to the accuracy. For classification, we use decision tree classifier and naives bayes. for clustering we use k-means. we perform ten-fold cross- validation to measure the out-of-training accuracy for classifiers.

a. Architecture Diagram


Data Preprocessing
Crime Historical Data
Data Cleaning
Data Filtering
New Crime Data
Classification Algorithm
Validation
Classified Output
End User

b. Flow Chart



c. Modules

1. Selection of data and preprocessing
   In our dataset, we have clean data and dirty data. We train and test both clean and dirty data and compare the results. In preprocessing we replace the missing values by null. We drop a few columns that are irrelevant.
   Make a new variable called "highCrime" that is real if the violent crime rate per population (ViolentCrimesPerPop) is more than 0.1 and false elsewhere. After adding this new field to the data, we calculate the percentage of positive and negative instances in the dataset.

2. Using decision tree classifier to predict
   We perform this using decisionTreeClassifier() method.we predict using d.predict(features). This module calculates accuracy, precision and recall score. Also, it visualizes the tree with the help of graphviz method and pydotplus for exploring the graph and it produces image output of the decision tree. This module lists out the main feature used for classification by checking feature importance, the higher the value, more important the feature. To measure the out-of-training accuracy of decision tree training for this problem, it uses ten cross-fold verification.
   Using naives bayes too predict
   Uses gaussianNB for linear classification. Uses the columns and high crime as x and y parameters and perform the algorithm. Applies 10 cross fold and calculate accuracy, precision and recall. Calculates 10 most predictive features.

3. Using k-means cluster
   Predicts crime using kmeans() method. Does the prediction for both dirty and clean data. Applies 10 cross fold validation and find out accuracy, recall and precision.

4. Comparison
   Plots graph for metrics and model. on x axis we have decision tree and naïve bayes. on y axis we have metrics [numerical value] and we plot the respective accuracy, recall and precision and choose the best model.

d. Experiments and Results

1. Data Preprocessing

We train both clean data and dirty data. we use imputer package for doing preprocessing. Imputer is a sci-kit-learn class that can be used to deal with incomplete information in a good predictive dataset. It substitutes a placeholder for the NaN values.

sklearn. preprocessing is a class.

missing values='NaN', strategy='mean', axis=0, verbos e=0, copy=True) Imputer(missing values='NaN', strategy='mean', axis=0, verbos e=0, copy=True)

for scaling purpose we use fit_transform. The deletion of whole rows and/or columns having variables is a basic method for using imperfect databases. Sadly, this comes at the price of potentially valuable data being lost (even though incomplete). Imputing the incomplete data, i.e. inferring them from the known part of the data, is a preferable method. It's a technique in the sklearn.preprocessing module. When scaling or standardizing our training and testing data, StandardScaler() and are nearly always used jointly. The mean and variance of every one of the characteristics in our data are calculated using the fit approach. All of the features are transformed using the relevant mean and variance in the transform method.

fit_transform(X, y=None, fit_params)

Fits transformer to X and y with optional variables fit_params and produces a changed version of X.

We also drop some of the columns from the data.

As discussed earlier, we implemented 3 models for both dirty data and clean data.

2. Decision Tree Classifier

Uses method called DecisionTreeClassifier().The DecisionTreeClassifier module of the Scikit-learn toolkit is used to conduct multiclass grouping on datasets.

Pseudocode:

```
1.      Tree = {}
2.      MinLoss = 0
3.      for all Attribute k in D do:
3.1.    loss = GiniIndex(k, d)
3.2.    if loss<MinLoss then
3.2.1.  MinLoss = loss
3.2.2.  Tree' = {k}
 4.     Partition(Tree, Tree')
5.      until all partitions procressed
6.      return Tree
```

OUTPUT: Optimal Decision Tree

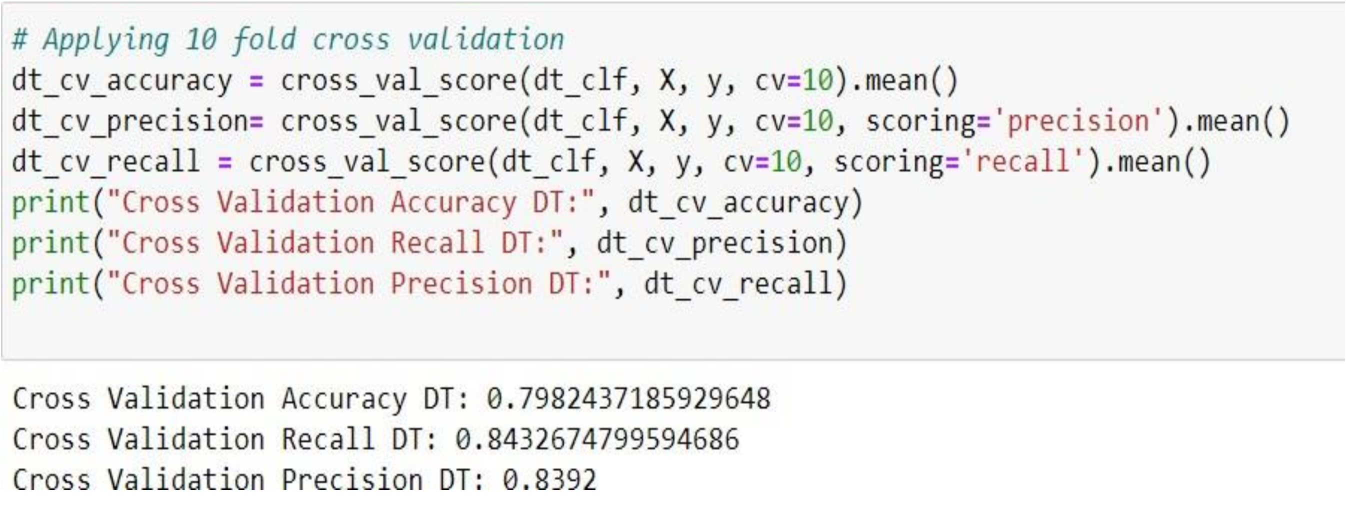
Method used : decisiontreeclasssifier(). It has parameters:

criterion,splitter,max_depth,min_samples_split,max_features,max_leaf_nodes.

dt_clf = DecisionTreeClassifier(max_depth=3)

dt_clf.fit(X,y)

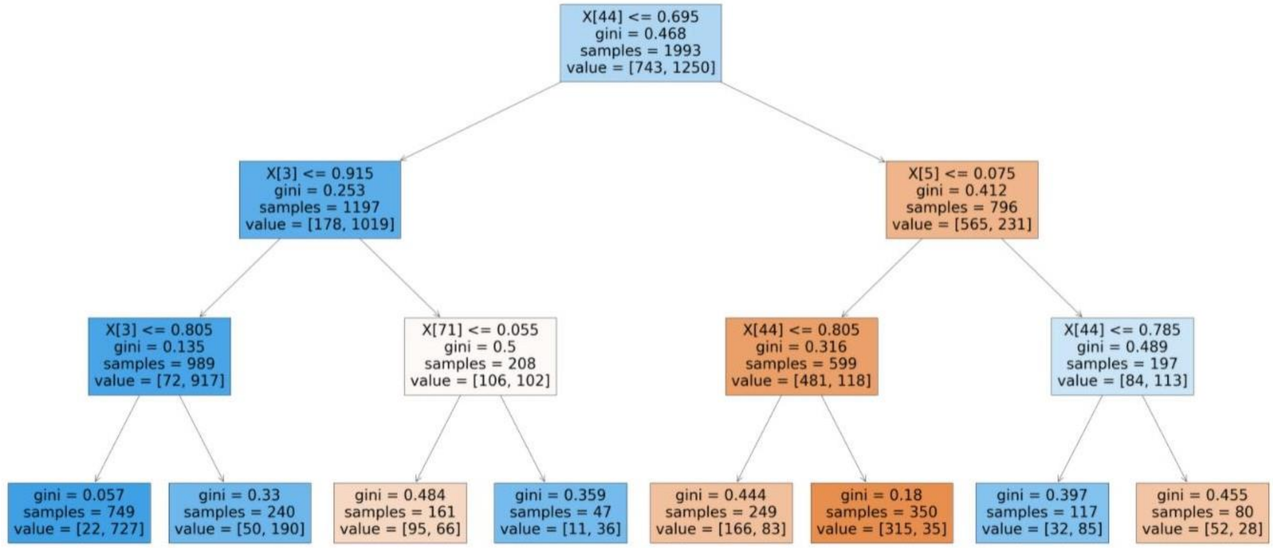
Performance metrics for decision tree:[dirty data]



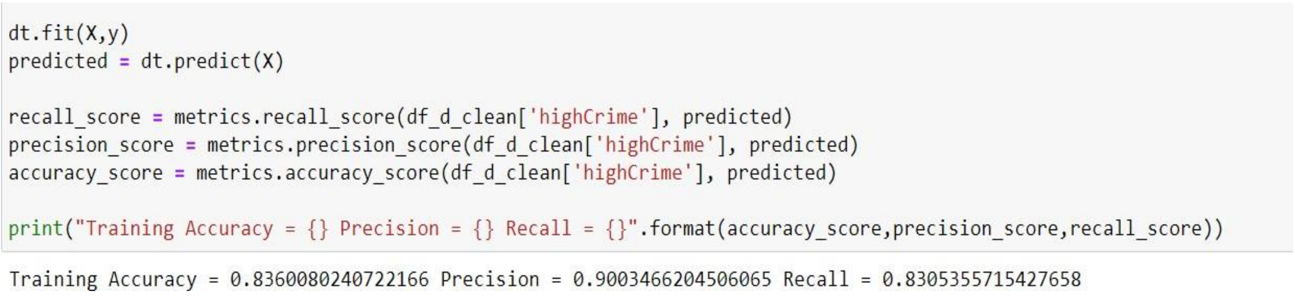
```
# Applying 10 fold cross validation
dt_cv_accuracy = cross_val_score(dt_clf, X, y, cv=10).mean()
dt_cv_precision= cross_val_score(dt_clf, X, y, cv=10, scoring='precision').mean()
dt_cv_recall = cross_val_score(dt_clf, X, y, cv=10, scoring='recall').mean()
print("Cross Validation Accuracy DT:", dt_cv_accuracy)
print("Cross Validation Recall DT:", dt_cv_precision)
print("Cross Validation Precision DT:", dt_cv_recall)
```

```
Cross Validation Accuracy DT: 0.7982437185929648
Cross Validation Recall DT: 0.8432674799594686
Cross Validation Precision DT: 0.8392
```

Decision Tree



Clean Data Metrics



```
dt.fit(X,y)
predicted = dt.predict(X)

recall_score = metrics.recall_score(df_d_clean['highCrime'], predicted)
precision_score = metrics.precision_score(df_d_clean['highCrime'], predicted)
accuracy_score = metrics.accuracy_score(df_d_clean['highCrime'], predicted)

print("Training Accuracy = {} Precision = {} Recall = {}".format(accuracy_score,precision_score,recall_score))
```

```
Training Accuracy = 0.8360080240722166 Precision = 0.9003466204506065 Recall = 0.8305355715427658
```

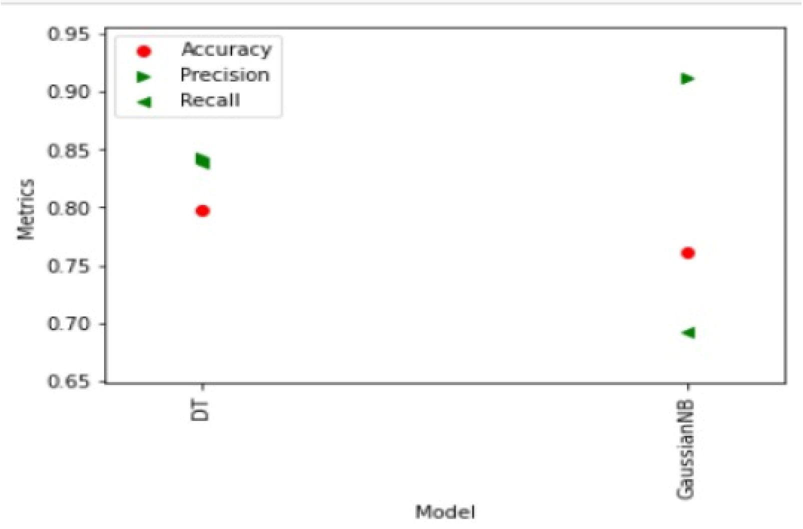
Main features after classification[top 5]

Feature ranking:

The main features used for classification

Index(['PctKids2Par', 'racePctWhite', 'racePctHisp', 'HousVacant', 'LemasPctOfficDrugUn'],dtype='object')

Top main feature is PctKids2Par

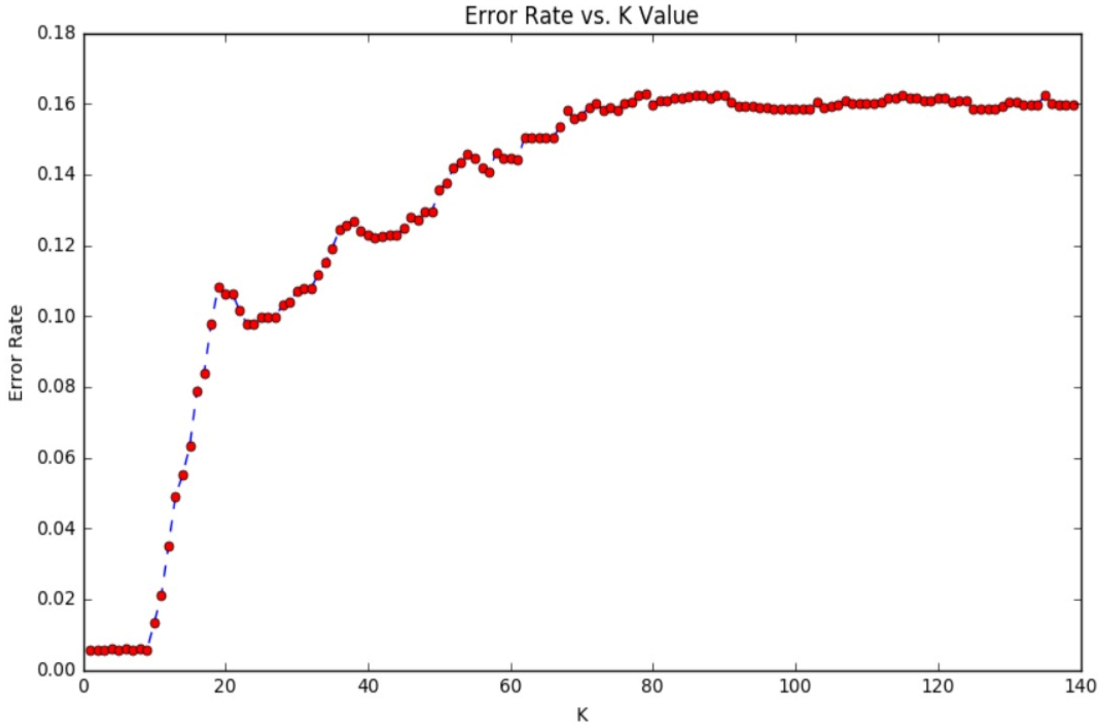


3. GaussinNB

Gaussian Naive Bayes agrees continuous numerical variables and analyzes them all as Gaussian (normal) distributions. Assume the data is defined by a Gaussian function with no covariance (independent dimensions) between parameters to develop a basic model.

4. KNN Algorithm

For misbehavior prediction, the K-Nearest Neighbor (KNN) order is used. The proposed methodology can predict areas with a high risk of crime and assess locations prone to misbehavior. K-Value vs Error rate.



Comparative Study

In this paper, the proposed framework is graph-based knowledge representation. Graphs can describe collected information in a way that allows for knowledge-based analysis.

Methods used:

1. tree.export_graphviz This method constructs a GraphViz depiction of the decision tree, which will then be saved to the out file variable. Visual projections can be created once the data has been received.
2. pydotplus. graph from dot data Data in DOT format is used to define the load chart. It is expected that the data is in DOT format. It will be translated and a Dot class describing the graph will be generated.
3. plt.legendLegend() is a method in the matplotlib library that is used to place a legend on the axes. The legend's position is specified by the attribute Loc in legend(). loc="best" is the default setting for loc (upper left). Comparing the decision tree classifier and gaussianNB.

Algorithm used	**Dataset**	**Accuracy**	**Precision**	**Recall**
Decision tree	Clean	83.6	0.900	0.830
Decision tree	Dirty	79.82	0.843	0.839
Gaussian NB	Clean	76.16	0.692	0.911
Gaussian NB	Dirty	71.11	0.587	0.93
KNN	Clean	75.89		0.85
KNN	Dirty	72.56		0.823
k-means	Clean	48.42	0.684	0.581
k-means	Dirty	42.86	0.627	0.505

Conclusion

We have applied various models Decision Trees, Gaussian NB, and K means. We also performed the 10-fold cross-validation. The results are different and we have plotted results based on the metrics for the different models. We got main features that are used for feature classification based on values. It can also be evaluated on a variety of models to find the best one for predicting crime rates. We conclude that, on basis of accuracy, the decision tree classifier works better than the rest two algorithms and gave 83 percent accuracy. Thereby, the prediction of crime was successful.

References

Data Prediction on Crime Detection, GRD Journals | Global Research and Development Journal for Engineering | National Conference on Computational Intelligence Systems (NCCIS'17) | March 2017 e-

ISSN: 2455-5703

K. Zakir Hussain, M. Durairaj and G. Rabia Jahani Farzana, "Application of Data Mining Techniques for Analyzing Violent Criminal Behaviour by Simulation Model", International Journal of Computer Science and Information Technology & Society, Vol. 02, No. 01, ISSN: 2249-9555, 2012

A. Malathi, Dr.S. Santhosh Baboo, "Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters", Global Journal of Computer Science and Technology Vol. 11, No. 11, pp. 139-145, 2011.

Automatic Crime Detector: A Framework for Criminal Pattern Detection in Big Data Era Md Ileas Pramanik, City University of Hong Kong, mpramanik2-c@my.cityu.edu.hk

A Predictive Model for Mapping Crime Saoumya, Anurag Singh Baghel, IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police. S. 22 Lavanyaa, D. Akila International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

Using Big Data Analytics for developing Crime Predictive model Tirthraj Chauhan, Rajanikanth Aluvalu Proceedings of RK University's First International Conference on Research & Entrepreneurship (Jan. 5th& Jan. 6th, 2016) ISBN:978-93-5254-061-7

Lenin Mookiah, William Eberle, Ambareen Siraj, Survey of Crime Analysis and Prediction, Proceedings of the twenty-Eighth International Florida Artificial Intelligence Research Society Conference, 2015.

(PDF) Using Big Data Analytics for developing Crime Predictive Model. Available from:

https://www.researchgate.net/publication/302026832_Using_Big_Data_Analytics_for_developing_Crime_Predictive_Model [accessed Feb 21 2020].

Behavior Analysis and Crime Prediction using Big Data and Machine Learning Pranay Jha, Raman Jha, Ashok Sharma International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.

Big Data Prediction on Crime Detection, GRD Journals | Global Research and Development Journal for Engineering | National Conference on Computational Intelligence Systems (NCCIS'17) | March 2017 e-ISSN: 2455-5703

Ibrahim, S. (2022). Mathematical Modelling and Computational Analysis of Covid-19 Epidemic in Erbil Kurdistan Using Modified Lagrange Interpolating Polynomial. International Journal of Foundations of Computer Science, 1-17.