ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

CRIME TYPE AND OCCURRENCE PREDICTION USING MACHINE LEARNING ALGORITHM

M.Naresh Babu [1]; S.N.Vyshnavi[2], S.Keerthi[3], P. Divya Sree[4],

D. S. V. Naga Prasad[5], [1]Asst Professor, [2,3,4,5]Students

Dept Of Computer Science And Engineering

Sri Vasavi Institute Of Engineering And Technology , Pedana , A.P, India

ABSTRACT:

In this era of recent times, crime has become an evident way of making people and society under trouble. An increasing crime factor leads to an imbalance in the constituency of a country. In order to analyse and have a response ahead this type of criminal activities, it is necessary to understand the crime patterns. This study imposes one such crime pattern analysis by using crime data obtained from Kaggle open source which in turn used for the prediction of most recently occurring crimes. The major aspect of this project is to estimate which type of crime contributes the most along with time period and location where it has happened. Some machine learning algorithms such as XGBoost,KNN is implied in this work in order to classify among various crime patterns and the accuracy achieved was comparatively high when compared to precomposed work.

Keywords:

Crime, Analyse, Crime patterns, Kaggle, Estimate, XGBoost, AdaBoost, Random Forest, KNN, Accuracy.

INTRODUCTION:

A crime is nothing but an action. It constitutes an offense. It's punishable by law. The identification and analysis of hidden crime is a very difficult task for the police department. Also, there is voluminous data of the crime is available. So, there should some methodologies that should help in the investigation. So, the methodology should help to solve the crime. The machine learning approach can better help in the prediction and analysis of the crime. The machine learning approach provides regression algorithms.

The classification techniques provide help to fulfill the purpose of investigation. Regression techniques such as multilinear regression are a statistical method. This method helps to find the relationship between two quantitative values or variables. This approach predicts the values of the dependent variables based on the independent variables. The classifier techniques such as XGBoost, AdaBoost, Random Forest, KNN . These classifiers are used to classify the multiclass target variables. The neural networks are used to improve the accuracy. The neural network has an input layer dense and has an output layer.

Based on the above algorithms the perpetrator description such as gender, age, and the relationship are predicted. The model is thus expected to help to remove the burden of the police investigation. Thus, it helps to solve homicide cases.

DOI : 10.36893.JK.2022.V12I01N01.879-886

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

LITERATURE SURVEY:

[1]. Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." Power and Energy Engineering Conference (APPEEC), 2011 Asia- Pacific. IEEE, 2011.

Jyoti Agarwal, Renuka Nagpal, et al., (2013) [2] has studied the crime analysis using K-means clustering on the crime dataset. They have developed this model using the rapid miner tool. The clustered results are obtained and analysed by plotting the values over the years. This model gives the result of the analysis that the number of homicides decreased from 1990 to 2011.

ShijuSathyadevan, Devan M. S, et al., (2014) [3] have predicted the regions where there is a high probability of the crime occurred. They have visualized crime-prone areas also. They have classified the data using Naive Bayesclassifiers. This algorithm is a supervised learning algorithm that also gives the statistical method for classification. This classification gives an accuracy of the 90%.

Lawrence McClendon and Natarajan Meghanathan (2015) [4] have used Linear Regression, Additive Regression, and Decision Stump algorithms using the same set of input (features), on the Communities and Crime Dataset. Overall, the linear regression algorithm gave the best results compared to the three selected algorithms. Chirag Kansara, Rakhi Gupta, et al., (2016) [8] proposed a model which analyses the sentiments of the people on Twitter and predicts whether they can become a threat to a particular person or society. This model is implemented using the Naive Bayes Classifier which classifies the people by sentiment analysis.

PROPOSED:

Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster. "***Here we are proposing a new regression based feature section algorithm for avoid the nosie information for feature engineering by using this we can achieve the best accuracy"***.

- The above problem made me to go for a research about how can solving a crime case made easier. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster.
- The aim of this project is to make crime prediction using the features present in the dataset. The dataset is extracted from the official sites. With the help of machine learning algorithm, using python as core we can predict the type of crime which will occur in a particular area.

DOI : 10.36893.JK.2022.V12I01N01.879-886

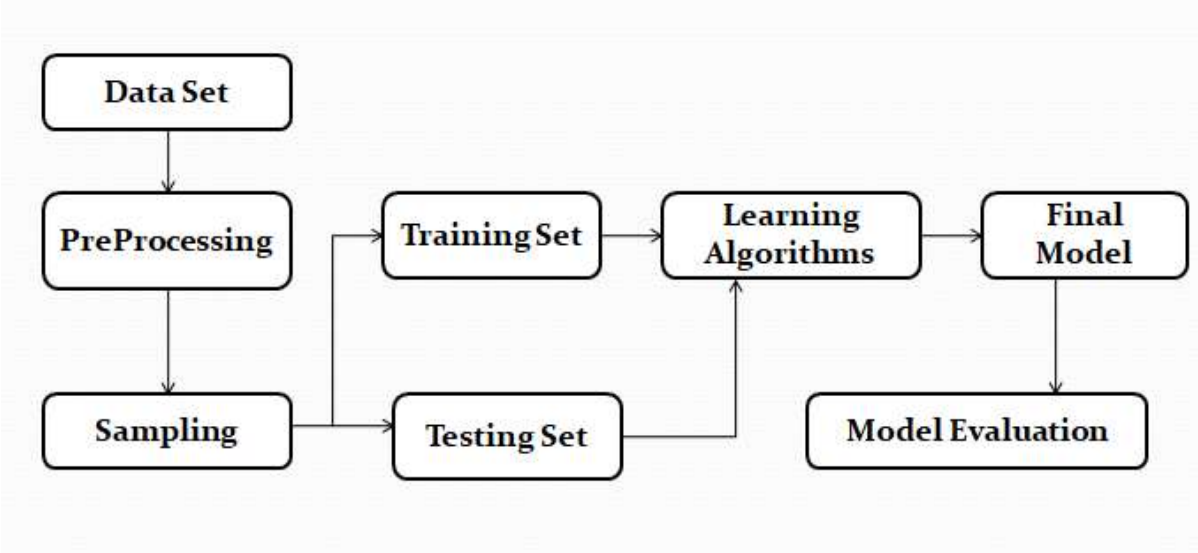
ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

- The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the accuracy. (XGBoost, AdaBoost, Random Forest, KNN)classification algorithms will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in the country. This work helps the law enforcement agencies to predict and detect crimes in Chicago with improved accuracy and thus reduces the crime rate.

Advantages:

- The initialization of optimal value is not required.
- The accuracy has been relatively high when compared to other machine learning prediction model.

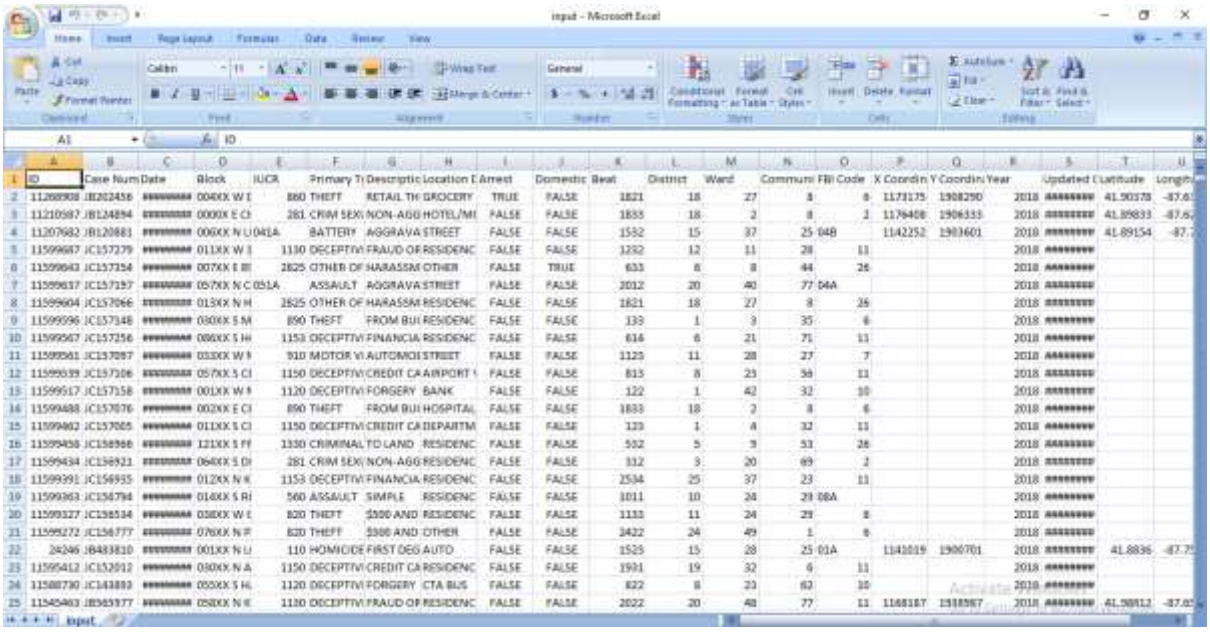
SYSTEM ARCHITECTURE:



Dataset:

DOI : 10.36893.JK.2022.V12I01N01.879-886

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022



IMPLEMENTATION:

- Data Set Reading and Inspection.
- Text Preprocessing.
- Analysis.
- Classification
- Evaluation.

Data Set Reading and Inspection:

The dataset can be taken from the Kaggle repository. The dataset contains homicide entries collected from the FBI's supplementary Homicide Report.From the dataset, the significant features like State, Year, Month, Crime Type, Crime Solved, Victim Gender, Victim Age, Victim Race, Victim Count and Weapon are chosen as the input features for the system. The record collected is almost 63000. The features Perpetrator Age, Perpetrator Gender and Relationship of the perpetrator with the victim are chosen as the target variable to be predicted by the system.

Text Preprocessing:

Once the dataset is collected, it must be pre-processed to get the clean dataset. The pandas and NumPy libraries are available in python for the pre-processing. it is removing of empty values from the dataset or repeated records should be removed.

Analysis:

DOI : 10.36893.JK.2022.V12I01N01.879-886

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

The analysis includes the graphical representation of different values to analyse the dataset property. The different graphs are plotted by Matplotlib libraries. The graphical analysis gives a direction towards the prediction.

The dataset is divided into training and testing. Generally, 70 % dataset is kept for training and 30% for testing. The dataset ratio can be 70: 30 or 80:20.

Classification:

KNN:

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a nonparametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

Random Forest Classifier.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap= =True (default), otherwise the whole dataset is used to build each tree.

XGBoost:

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

AdaBoost:

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

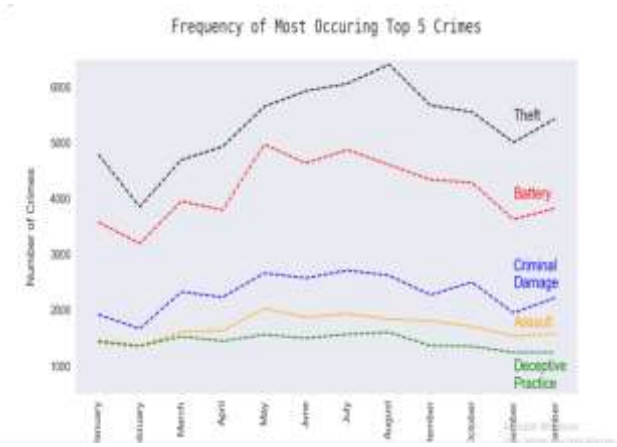
Evaluation:

Once the model is created, it should be validated with the real-time data values. This is called validation. The validation is nothing, but its predicted value and it's also called the output value.

DOI : 10.36893.JK.2022.V12I01N01.879-886

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

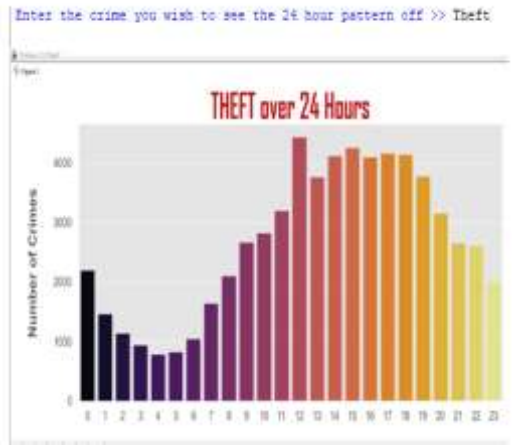
.Analysis:

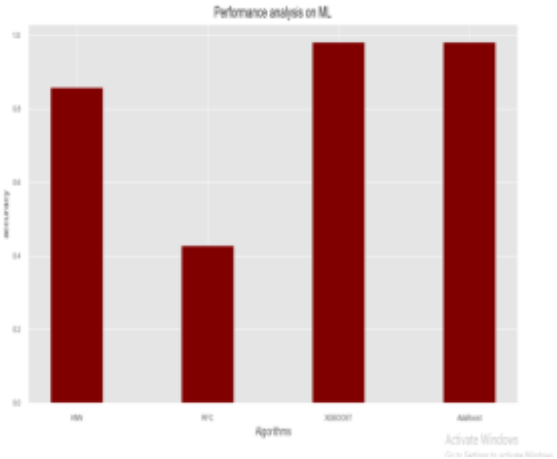


Result:

DOI : 10.36893.JK.2022.V12I01N01.879-886

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022





ACCURACY SCORE

Accuracy score of the KNN model is 0.8577154308617234

Accuracy score of the Random Forest Tree model is 0.4709418837675351

Accuracy score of the XGBoost model is 0.9819639278557114

Accuracy score of the AdaBoost model is 0.981568757114

CONCLUSION:

This model helps to predict crime. The perpetrator's age, perpetrator sex, and relationship can be predicted using a machine learning approach. The regression and classifier are used here give almost 98 % accuracy. The dataset can be enhanced and can be used in other countries if the scenario is almost same. The model gives the overall prediction of any crime. This model can be enhanced by using deep learning techniques.

REFERENCES:

[1]. Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." Power and Energy Engineering Conference (APPEEC), 2011 Asia- Pacific. IEEE, 2011.

[2]. Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using K-means clustering." International Journal of Computer Applications 83.4 (2013).

DOI : 10.36893.JK.2022.V12I01N01.879-886

ISSN: 2278-4632
Vol-12 Issue-01 No.01: 2022

[3]. Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014.

[4]. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyse crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015). [5]. Kiani, Rasoul, Siamak Mahdavi, and Amin Keshavarzi. "Analysis and prediction of crimes by clustering and classification." Analysis 4.8 (2015)

DOI : 10.36893.JK.2022.V12I01N01.879-886