International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 09, September 2022)

Machine Learning for Crime Analysis and Prediction

Nikita Jha[1], Prof. Vishal Paranjape[2], Prof. Saurabh Sharma[3], Prof. Zohaib Hasan[4]
[1,2,3,4]*Global Nature Care Sangathan Group of Institutions, Jabalpur (M.P), India*

Abstract:-- **The prevention of crime is a crucial undertaking because it is one of our society's most significant and pervasive problems. Large amounts of crimes are perpetrated often every day. This necessitates keeping note of all crimes and establishing a database for them that may be consulted in the future. In order to forecast and solve crimes in the future, it is currently difficult to keep an accurate record of crimes and analyse it. The goal of this research is to evaluate a dataset that includes many crimes and make predictions about the kinds of crimes that could occur in the future depending on various factors. In this project, we will use data science and machine learning to predict crimes using a set of crime data from Chicago. The Chicago police's official portal is where the crime statistics are taken from. It includes details on the offence, including the time, date, place, and kind of crime. Data preprocessing will be performed prior to training the model, and this will be followed by feature selection and scaling to ensure high accuracy. Several other algorithms, including the K-Nearest Neighbor (KNN) classification, will be examined for their ability to forecast crimes, and the algorithm that performs the best will be used to train others. Dataset visualisation will take the form of graphical representations of various situations, such as when criminal activity rates are highest or which month has the highest criminal activity rates. The main goal of this project is to provide a basic understanding of how machine learning may be utilised by law enforcement organisations to identify, anticipate, and solve crimes considerably more quickly, which lowers the crime rate. This is not only applicable to Chicago; depending on the dataset's accessibility, it may also be used in other states or nations.**

Keywords:-- **K-Nearest Neighbor Support, Vector Machine Autoregressive moving average, recurrent neural network, Recursive Feature Elimination, National Crime Records Bureau**

I. INTRODUCTION

Crime poses a serious threat to humanity. Many crimes occur frequently at regular intervals. Maybe it's growing and dispersing quickly and widely. From small towns and villages to large metropolis, crimes occur. There are many distinct types of crimes, including robbery, manslaughter, rape, assault, battery, and false imprisonment. There is a need to resolve cases much more quickly because crime is rising.

The police agency is responsible for containing and decreasing the crime activities, which have increased at an accelerated rate. Given the vast amount of crime data available, crime prediction and criminal identification are the police department's two biggest issues. Technology is required so that case solving can be completed more quickly. The aforementioned issue prompted me to look at ways to make crime case solving simpler. It was discovered through extensive documentation and cases that machine learning and data science may expedite and simplify the work. The purpose of this project is to forecast crime using the dataset's attributes. The official websites are where the dataset was taken. The type of crime that will occur in a specific region can be predicted with the aid of machine learning algorithms, which use Python as their basic language. To train a model for prediction would be the goal. Utilizing the training dataset, the test dataset will be used to validate the training. Depending on the accuracy, a better algorithm will be used to build the model. For crime prediction, the K-Nearest Neighbor (KNN) classification and other algorithms will be employed. The dataset is visualised to examine potential crimes that may have occurred in the nation. This effort improves the ability of Chicago's law enforcement agencies to anticipate and identify crimes, which lowers the city's crime rate.

II. CONCEPTS OF THE PROPOSED SYSTEM

Predictive Modeling

Building a model that can make predictions is done through predictive modelling. A machine learning algorithm is used in the procedure to create those predictions by learning specific properties from a training dataset.

The two subfields of predictive modelling are regression and pattern categorization. In order to forecast the values of continuous variables, regression models are built on the investigation of relationships between variables and trends.

The goal of pattern classification, in contrast to regression models, is to assign discrete class labels to a specific data value as an output of a prediction.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 09, September 2022)

A pattern classification problem in weather forecasting could be the prediction of a sunny, wet, or snowy day. This is an example of a classification model.

Tasks involving pattern classification can be split into two categories: supervised learning and unsupervised learning. In supervised learning, the classification model's input dataset's class labels are predetermined. In a supervised learning situation, we would be able to anticipate outcomes for unobserved data by knowing which training dataset has the specific output that would be utilised to train.

Types of Predictive Models Algorithms

Classification and Decision Trees A decision tree is an algorithm that uses a tree shaped graph or model of decisions including chance event outcomes, costs, and utility. It is one way to display an algorithm.

Naive Bayes -In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with independence assumptions between the features.

The technique constructs classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

Linear Regression – The analysis is a statistical process for estimating the relationships among variables. Linear regression is an approach for modelling the relationship between a scalar dependent variable Y and one or more explanatory variables denoted X. The case of one explanatory variable is called simple linear regression. More than one variable is called multivariate.

Logistic Regression - In statistics, logistic regression, is a regression model where the dependent variable is categorical or binary.

Data Preprocessing

This process includes methods to remove any null values or infinite values which may affect the accuracy of the system. The main steps include Formatting, cleaning and sampling. Cleaning process is used for removal or fixing of some missing data there may be data that are incomplete.

Sampling is the process where appropriate data are used which may reduce the running time for the algorithm. Using python, the preprocessing is done.

Functional Diagram of Proposed Work

It can be divided into 4 parts:

1. Descriptive analysis on the Data
2. Data treatment (Missing value and outlier fixing)
3. Data Modelling
4. Estimation of performance

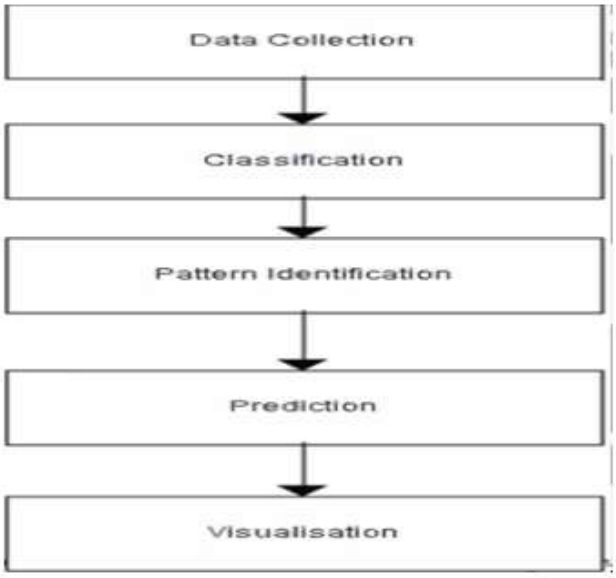

Figure 1- Architecture

Prepare Data

1. In this step we need prepare data into right format for analysis
2. Data cleaning

Analyze and Transform Variables We may need to transform the variables using one of the approaches

1. Normalization or standardization
2. Missing Value Treatment

Random Sampling (Train and Test)

- *Training Sample:* Model will be developed on this sample. 70% or 80% of the data goes here.
- *Test Sample:* Model performances will be validated on this sample. 30% or 20% of the data goes here

Model Selection

Based on the defined goal(s) (supervised or unsupervised) we have to select one of or combinations of modeling techniques. Such as

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 09, September 2022)

- KNN Classification
- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machine (SVM)
- Bayesian methods

Build/Develop/Train Models

- Validate the assumptions of the chosen algorithm Develop/Train Model on Training Sample, which is the available data(Population)
- Check Model performance - Error, Accuracy

Validate/Test Model

- Score and Predict using Test Sample
- Check Model Performance: Accuracy etc.

III. IMPLEMENTATION

The dataset used in this project is taken from Kaggle.com. The dataset obtained from kaggle is maintained and updatedby the Chicago police department.

The implementation of this project is divided into followingsteps –

Data collection

Crime dataset from kaggle is used in CSV format.

Data Preprocessing

10k entries are present in the dataset. The null values are removed using df = df.dropna() where df is the data frame. The categorical attributes (Location, Block, Crime Type, Community Area) are converted into numeric using Label Encoder. The date attribute is splitted into new attributes like month and hour which can be used as feature for the model.

Feature selection

Features selection is done which can be used to build the model. The attributes used for feature selection are Block, Location, District, Community area, X coordinate , Y coordinate, Latitude , Longitude, Hour and month,

Building and Traning Model

After feature selection location and month attribute are used for training. The dataset is divided into pair of xtrain ,ytrain and xtest, y test. The algorithms model is imported form skleran. Building model is done using model. Fit (xtrain, ytrain).

Prediction

After the model is build using the above process, prediction is done using model.predict(xtest). The accuracy is calculated using accuracy_score imported from metrics - metrics.accuracy_score (ytest, predicted).

Visualization

Using mathpoltlib library from sklearn. Analysis of the crimedataset is done by plotting various graphs.

IV. RESULTS AND DISCUSSION

The results are obtained after undergoing various processesthat comes under machine learning

Predictive modelling.

Table 1
Dataset

Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward
HM155213	2006-01-31 12:13:00	066XX N BOSWORTH AVE	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	SCHOOL, PUBLIC, BUILDING	True	False	...	40
HM245080	2006-03-21 19:00:00	062XX S WESTERN AVE	1330	CRIMINAL TRESPASS	TO LAND	PARKING LOT/GARAGE(NON.RESID.)	True	False	...	15
HM171175	2006-02-09 01:44:00	056XX S SHIELDS AVE	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	STREET	True	False	...	20
HM244805	2006-03-21 16:45:00	011XX N SPAULDING AVE	810	THEFT	OVER $500	CHURCH/SYNAGOGUE/PLACE OF WORSHIP	False	False	...	26

Data preprocessing Data preprocessing includes dropping row without any row and converting any value which consist of value as infinity. Converting string variable into numerical so that it can undergo further processing.

Table 2
Dataset after Preprocessing

	ID	Case Number	Date	Block	Type	Location	District	Ward	Community Area	X Coordinate	Y Coordinate	Latitude	Longitude	Hour	Month
0	4647368.0	HM155213	2006-01-31 12:13:00	4748	NARCOTICS	60	24.0	40.0	1.0	1164737.0	1944183.0	42.002478	-87.669297	12	1
1	4647370.0	HM245080	2006-03-21 19:00:00	4581	CRIMINAL TRESPASS	50	8.0	15.0	66.0	1161441.0	1863039.0	41.780598	-87.680676	19	3
2	4647372.0	HM171175	2006-02-09 01:44:00	4323	NARCOTICS	69	7.0	20.0	68.0	1174868.0	1866967.0	41.787955	-87.634037	1	2
3	4647373.0	HM244805	2006-03-21 16:45:00	1015	THEFT	17	11.0	28.0	23.0	1154100.0	1907414.0	41.901774	-87.708415	16	3

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 09, September 2022)

After dividing the data set intro training set and testing setthe model is trained using algorithm as mentioned in the table. The accuracy is calculated using the function score_accuracy imported from metric from sklearn. The accuracy is mentioned in the table below.

Table 3
Accuracy obtained after Testing

ALGORITHM	ACCURACY
KNeighbors Classifier	0.78734858681022879
GaussianNB	0.6460296096904441
MultinomialNB	0.45625841184387617
BernoulliNB	0.31359353970390308
SVC	0.31359353970390308
DecisionTree Classifier	0.78600269179004034

As we can see from the results obtained from the table the algorithm which can be used for the predictive modeling will be KNN algorithms with accuracy of 0.787 highest among therest of the algorithm.

The least which can be used will be SVM. For further modelling using unseen data there is no need for using other algorithm.

Crime Visualization

This section deals with the analysis done on the dataset and plotting them into various graphs like bar, pie, scatter. Analysis done are

1. Types of crimes committed over Time (Month/ Hour).
2. No of crimes of all types of crime over the whole city ofChicago.
3. Arrested ratio.
4. Crimes committed across different location.
5. Details of Major crimes committed in the city.

This graph shows which crimes have occurred most in the city. The y coordinate denotes the Types of crimes committed and x coordinate denotes the no of crimes committed.

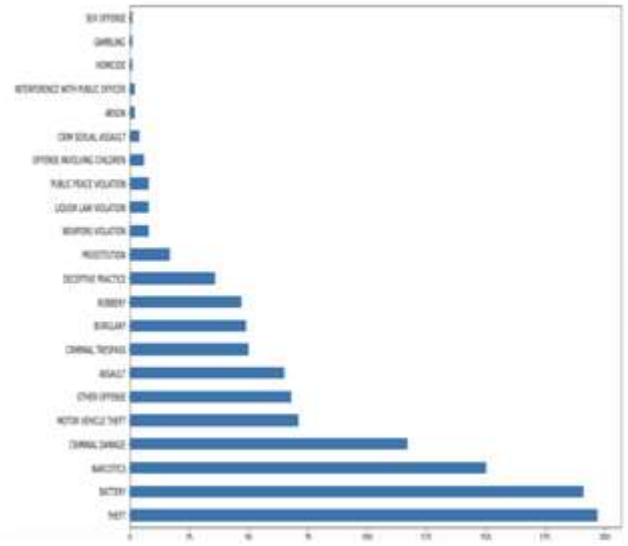

Figure 2- Types of crimes vs No of crimes occurred

The graph below tells us about in which month occurrence of crimes is highest. As we can see the month of march is peakwhere rate of occurrence is high.

The x coordinate denotes the month and y coordinated denotes the crime rate.

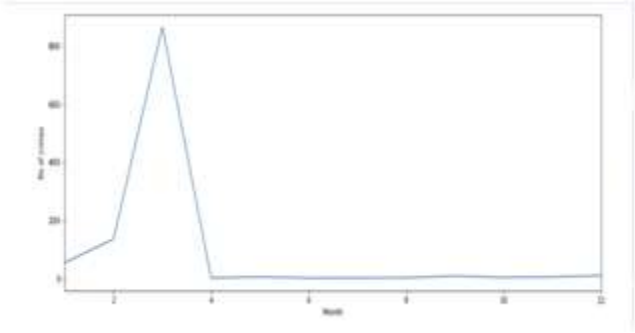

Figure 3 – No of crimes committed over months in ayear

The graph below shows the arrest ratio made in the city. 67.2 % of crimes committed by the criminals are not arrested and rest 32.8 % are arrested.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 09, September 2022)

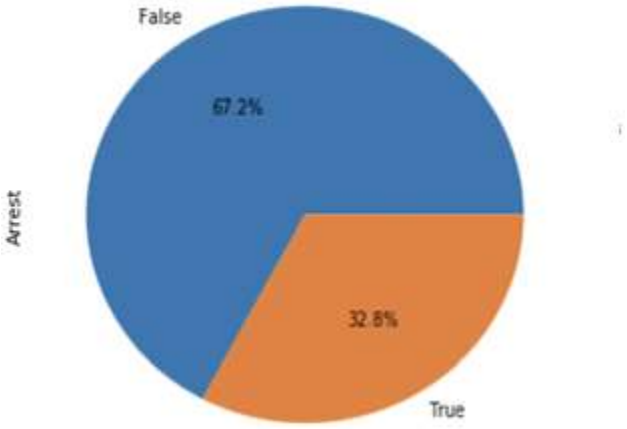

Figure 4 – Percentage of arrest made

The graph below shows crime occurrence over particular hour. The x axis is hour and y axis is rate of crimes

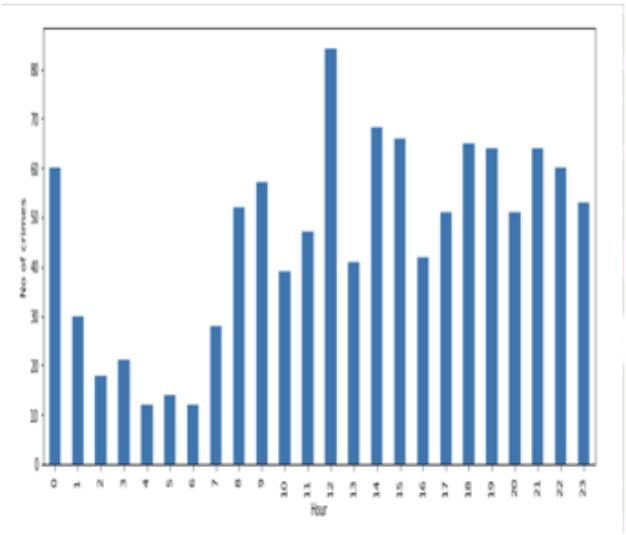

Figure 5 - No of crimes committed over an Hour

The graph below shows the location where crimes are committed. The x axis is the crime rate and y axis is the location.

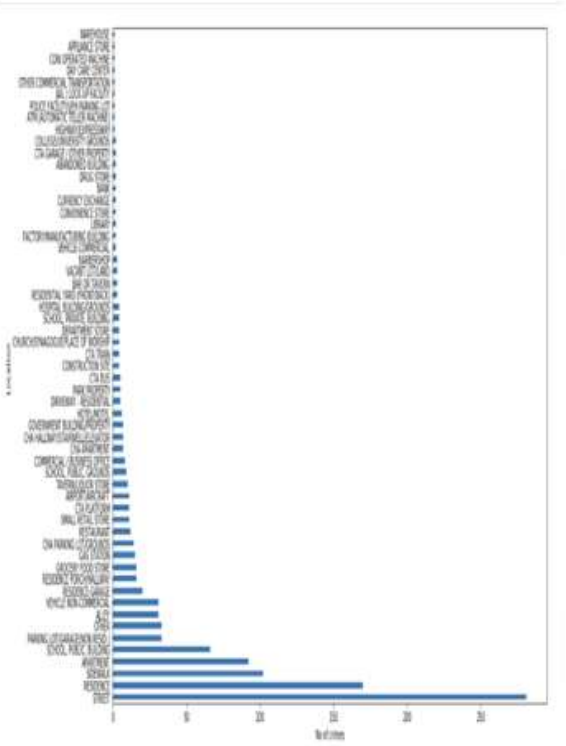

Figure 6- Crimes committed across different location

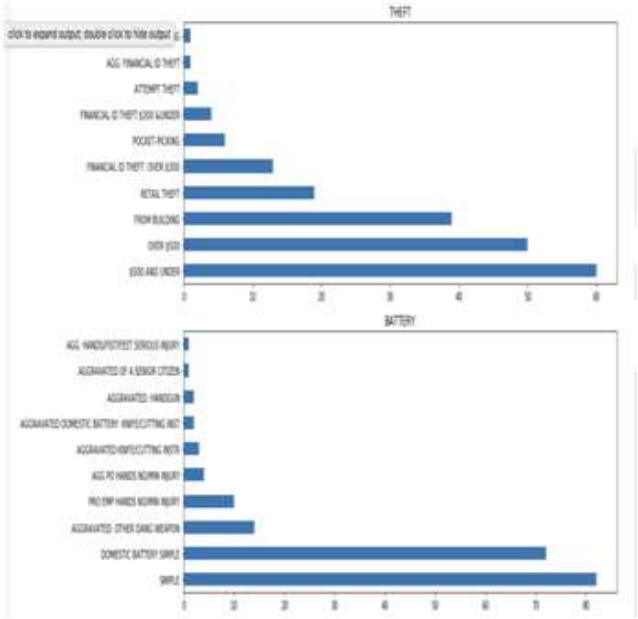

Figure 7- Details of the Major crimes (theft andbattery) committed

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 09, September 2022)

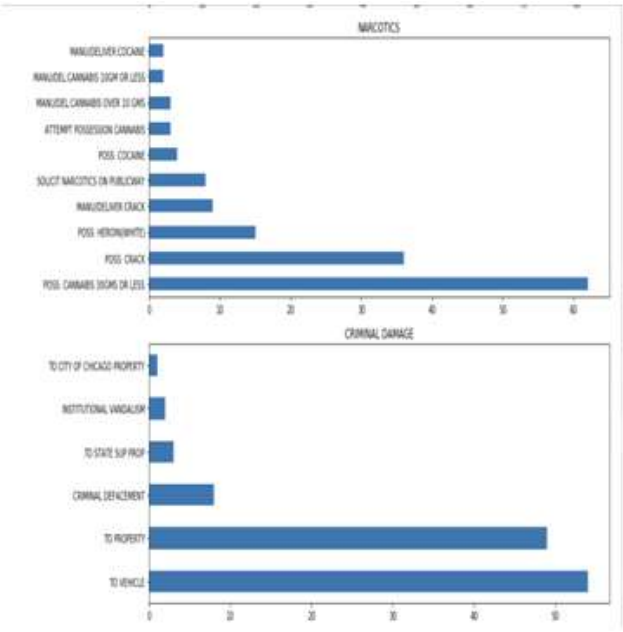

Figure 8- Details of the major crimes (Narcotics andCriminal damage) committed

V. CONCLUSION

Finding relationships and patterns among varied data has become much simpler with the use of machine learning technology. The major task of this research is to determine the type of crime that might occur given the place where it has already happened. Using a training set of data that has undergone data cleansing and data transformation, we have developed a model using the machine learning idea. With an accuracy of 0.789, the model can identify the type of crime. Analyzing a data set is made easier by data visualisation.

The graphs include bar, pie, line, and scatter diagrams, each with their unique features. In order to better comprehend the Chicago crime datasets that can help in capturing the aspects that can help in keeping society secure, we created numerous graphs and discovered intriguing statistics.

REFERENCES

[1] Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R.,& Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In Electronics, Communication and Aerospace Technology (ICECA),2017 International conference of (Vol. 1, pp. 225- 230). IEEE.

[2] Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime prediction methods. InStudent Project Conference (ICT-ISPC), 2017 6th ICT International (pp. 1-5). IEEE.

[3] Sivaranjani, S., Sivakumari, S., & Aasha, M. (2016, October). Crime prediction and forecasting in Tamilnadu using clustering approaches. In Emerging Technological Trends (ICETT), International Conference on (pp. 1-6). IEEE.

[4] Sathyadevan, S., & Gangadharan, S. (2014, August). Crime analysis and prediction using data mining. In Networks & Soft Computing (ICNSC), 2014 First International Conference on (pp. 406-412). IEEE.

[5] Nath, S. V. (2006, December). Crime pattern detection using data mining. In Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on (pp. 41-44). IEEE.

[6] Zhao, X., & Tang, J. (2017, November). Exploring Transfer Learning for Crime Prediction. In Data Mining Workshops (ICDMW), 2017 IEEE International Conference on (pp. 1158-1159). IEEE.

[7] Al Boni, M., & Gerber, M. S. (2016, December). Area-Specific Crime Prediction Models. In Machine Learning and Applications (ICMLA), 2016 15th IEEEInternational Conference on (pp. 671-676). IEEE.

[8] Tayebi, M. A., Gla, U., & Brantingham, P. L. (2015, May). Learning where to inspect: location learning for crime prediction. In Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on (pp. 25-30). IEEE.