# CRIME DATA ANLYSIS

[1] MRS. UDAYASREE,[2] K. SHASHIDHAR,[3] G. SAKETH,[4] K. RAJU,[5] B. MANI CHARAN
[1]Assistant Professor, Department of DS, Sri Indu College Of Engineering & Technology.
[2,3,4,5]U.G.Scholor, Department of DS, Sri Indu College Of Engineering & Technology, Hyderabad

***Abstract***— Criminal cases are increasing rapidly in society, leading to a growing backlog of pending investigations. Controlling the rise in crime has become essential, as it can create major challenges for law enforcement agencies. Authorities maintain detailed records of every crime after it occurs because there may be hidden patterns behind these incidents. Identifying such patterns can help in preventing future crimes. In this project, we propose a machine learning model to analyze and predict crime patterns using historical crime data. The model is developed using the K-Nearest Neighbour (KNN) algorithm to improve prediction accuracy compared to existing approaches such as Naïve Bayes and Decision Tree algorithms. Previous studies using the KNN algorithm achieved prediction accuracy of around 75%. In this work, we aim to improve the accuracy to more than 85% by refining the model and optimizing the implementation while also reducing code complexity. The system predicts the probability of occurrence of the next two crime types based on the analyzed dataset. The dataset used in this study was collected from data.gov.in and includes crime records from January 2024 to January 2025. By identifying patterns and predicting possible crime occurrences, the proposed model can assist law enforcement agencies in planning preventive actions and improving crime control strategies.

***Keywords***— KNN, Machine Learning, Crime Prediction, Data Visualization, Accuracy, Pattern Recognition.

## I. INTRODUCTION

Crimes are harmful actions that lead to threats to human lives. Crimes might be Robbery, Murder, Rape, Women trafficking, etc. As the population increases the rate of crime also increases day by day. The increasing cases lead to a backlog of pending cases to the police department, The crime activities have increased at a faster rate and it is the responsibility of the police department to control and reduce the crime activities. the department tries to solve the cases according to the evidence they got but in major cases, it is not as much possible to solve and decrease the crime rate as they think.

This analysis leads us to research the crimes to make them complex and free for solving the cases. The main thing here we are going to work on is predicting the occurrence of the next crime. It might be helpful to the Law Enforcement agencies and police departments to control and be aware of the respective situation. It will only be possible by collecting the previous information. So, we get the information stored in dataset format in which the dataset contains the relative features like crime type, place, time, arrest or not, victims, and whether the case is solved or not, etc… We could extract the dataset from official site data.govern.in. The prediction of the occurrence of crime can happen by working with a machine learning model and one optimal algorithm, here we are going to work with K Nearest Neighbour which is well-suited algorithm for both classification and regression and can also get good

prediction accuracy. Visualization of occurrence of crimes are well-known thing to the normal people hence we could also implement the work with visual graphs.

## II. EXISTING SYSTEM

In existing Systems, they used Naïve Bayes algorithm which is a supervised learning algorithm which is used for classification. Mainly used for text classification based on the training dataset. Naïve Bayes algorithm assumes all features were independent to each other. It depends on the conditional probability.

Formula :

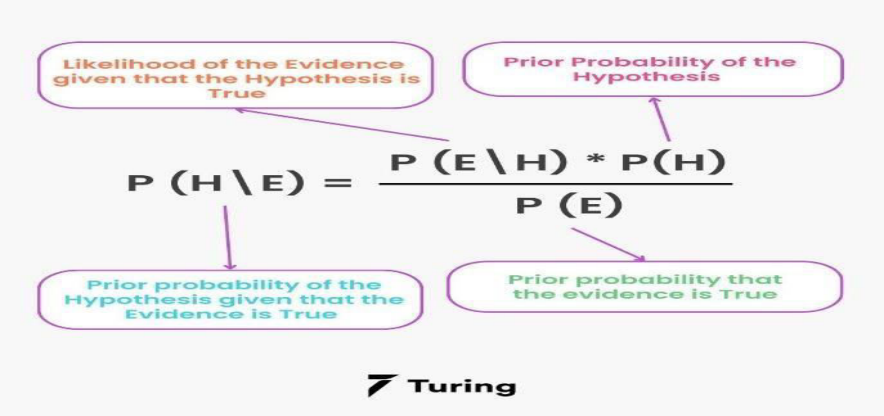


Fig. 1 Naïve Bayes formula.

***Disadvantages:***

- Shows lower performance compared to other classification models.
- Require large Data records to achieve a good accuracy result.
- Features are independent to each other therefore it results in low accuracy.

## III. CONCEPTS OF PROPOSED SYSTEM

### *A. Predictive Modelling*

The concept that we have that is predictive modelling that is any model that we want to build is used to predict the results in order that based on how it had trained. In the process that includes machine learning algorithm that trained from fed dataset. The modelling has divided into two types classification and model regression which describes the analysis of tremendous research between the trends and variables. When it becomes to regression tasks it allows you to assign the class labels to different classes that is assumes a group or a class named as class A, we can simply state to predict that whether a boy can enjoy the sport under different kind of weather conditions.

In the other hand Pattern classification can divided into two parts those are Supervised Machine Learning model and Unsupervised machine learning model. In supervised machine learning model, the dataset can well know with its features and data with also what type of data we are feeding and trains to the model to get accurate predictions can be made for unknown data. when we come and talking about Unsupervised learning the scenario is quite opposite to supervised learning model.

### B. Types of predictive modelling

Generally, we all aware on decision tree which emits the possible number of outcomes as graph or tree shaped liked structured, which used as a classification algorithm. It is a chance to show the algorithm. In the phenomenon which we get possible number of predictive outcomes as the result just like the algorithms like decision tree. Here the problem's features assign to the actual algorithm class labels to make a line of bond to construct the algorithmic

approach, class labels can get from the known set. Naïve Bayes algorithm which is very closest approach algorithm uses probabilistic classifier based the applied bayes theorem with independent factors among the classes. We can also state the Naïve Bayes is family of probabilistic classifier. Linear Regression more aware of a supervised machine learning algorithm approach used to maps the datapoints to the most optimized linear functions. It involves only one dependent and one independent variable. logistic regression, it is a regression model where the dependent variable is categorical, or we can say binary.

*1.Dataset:*

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CASE# | DATE OF ( | BLOCK | IUCR | PRIMARY | SECONDA | LOCATIOI | ARREST | DOMESTIC | BEAT | WARD | FBI CD | X COORDII | Y COORDII | LATITUDE | LONGITUI | LOCATION | | |
| 2 | JG497095 | ######## | 025XX N K | 810 | THEFT | OVER $50( | STREET | N | N | 1414 | 35 | 6 | 1154609 | 1916759 | 41.92741 | -87.7073 | (41.927407329, -87.70729439) | | |
| 3 | JG496991 | ######## | 0000X W C | 560 | ASSAULT | SIMPLE | STREET | N | N | 1832 | 42 | 08A | 1176106 | 1905725 | 41.89667 | -87.6286 | (41.896671699, -87.628635323) | | |
| 4 | JG497145 | ######## | 019XX W 4 | 051A | ASSAULT | AGGRAVA | SIDEWALK | N | N | 931 | 15 | 04A | 1164331 | 1873509 | 41.80853 | -87.6728 | (41.808525157, -87.672792896) | | |
| 5 | JG496701 | ######## | 025XX W E | 502P | OTHER OF | FALSE / ST | STREET | N | N | 2011 | 40 | 26 | 1158314 | 1935772 | 41.97951 | -87.6932 | (41.979505088, -87.693158103) | | |
| 6 | JG484195 | 10/28/202 | 067XX S PA | 810 | THEFT | OVER $50( | APARTMEI | N | N | 722 | 6 | 6 | 1173732 | 1860233 | 41.77189 | -87.6387 | (41.771890947, -87.638705659) | | |
| 7 | JG483131 | 10/28/202 | 057XX N K | 1320 | CRIMINAL | TO VEHICL | STREET | N | N | 1711 | 39 | 14 | 1152676 | 1937956 | 41.98561 | -87.7138 | (41.985611859, -87.713834343) | | |
| 8 | JG498494 | ######## | 089XX S CI | 560 | ASSAULT | SIMPLE | SIDEWALK | N | N | 413 | 7 | 08A | 1193055 | 1846244 | 41.73305 | -87.5683 | (41.733053891, -87.568330657) | | |
| 9 | JG496575 | ######## | 037XX N S( | 860 | THEFT | RETAIL TH | SMALL RE | Y | N | 1922 | 44 | 6 | 1166304 | 1924930 | 41.94959 | -87.6641 | (41.949586612, -87.664085689) | | |
| 10 | JG427641 | 09/17/202 | 001XX W 1 | 820 | THEFT | $500 AND | STREET | N | N | 512 | 9 | 6 | 1177160 | 1835662 | 41.70439 | -87.6269 | (41.704388397, -87.626879123) | | |
| 11 | JG365961 | ######## | 002XX W M | 530 | ASSAULT | AGGRAVA | SMALL RE | N | N | 122 | 34 | 04A | 1174636 | 1900346 | 41.88194 | -87.6342 | (41.881944424, -87.634195294) | | |
| 12 | JG496115 | ######## | 076XX S M | 820 | THEFT | $500 AND | STREET | N | N | 621 | 17 | 6 | 1170966 | 1854231 | 41.75548 | -87.649 | (41.755481563, -87.649019949) | | |
| 13 | JG496955 | ######## | 049XX N M | 320 | ROBBERY | STRONG A | SIDEWALK | N | N | 1623 | 45 | 3 | 1139338 | 1932332 | 41.97043 | -87.763 | (41.970433391, -87.763029002) | | |
| 14 | JG541270 | 12/14/202 | 072XX S C( | 486 | BATTERY | DOMESTIC | APARTMEI | Y | Y | 324 | 7 | 08B | 1189632 | 1857146 | 41.76305 | -87.5805 | (41.763052784, -87.580521082) | | |
| 15 | JG501047 | ######## | 008XX E H | 620 | BURGLAR\ | UNLAWFU | APARTMEI | N | N | 223 | 20 | 5 | 1182731 | 1871378 | 41.80227 | -87.6054 | (41.802269632, -87.605372566) | | |
| 16 | JG496779 | ######## | 013XX W 9 | 051A | ASSAULT | AGGRAVA | SCHOOL - | N | N | 2213 | 21 | 04A | 1169194 | 1841762 | 41.7213 | -87.6559 | (41.721303358, -87.655873595) | | |
| 17 | JG496296 | ######## | 0000X E R/ | 890 | THEFT | FROM BUI | SPORTS AF | N | N | 111 | 34 | 6 | 1176904 | 1901295 | 41.8845 | -87.6258 | (41.884497529, -87.625838595) | | |

Fig. 2 Dataset Image.

*C. Data Preprocessing*

The information involves any null values are unnecessary duplicates that might cause of misleads to the Target. It also affects the work accuracy or algorithm accuracy, So the process involves in removing null values and replacing with respective values in it, simply we can say that handling of null values and duplicates. The process can also move forward with some probabilistic approaches like mean, median and mode. The main mechanism involves in three steps they are Cleaning, Sampling and Formatting. By using these steps in python as preprocessing we reduce the running rate get optimal time of run.

D. *Functional Diagram of Proposed Work*

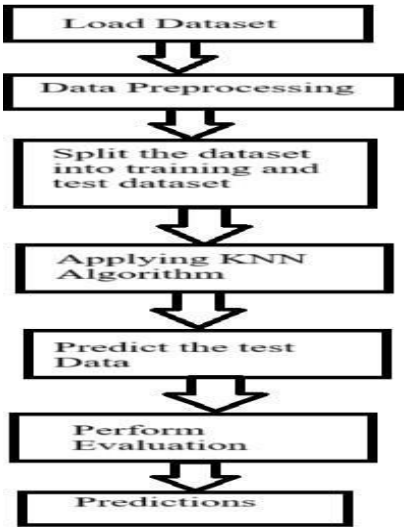


Fig. 3 Functional Diagram of Proposed Work

*E. Prepare Data*

- Have to collect the data from right sources without lack of necessary information.
- Must of data cleaning
- Study and covert or transform the variables
- We can transform the variables by using any approach from the below.

1) Standardization or normalization.
2) Missing value operation.

*F. Random Sampling(Train or Test)*

*1.Training sample* :

Training data is used to train the model for future accuracy prediction with information of 70% to 80% data

*2.Test sample* :

Testing data is used to test and validate the data that we were we fed to the model that is to check whether our model works well on train data or not. The test data maybe about 20% to 30%.

*G. Model Selection*

According the problem we have, we have to choose the appropriate model or approach to deal the situation is necessary. Based on the problem we have to choose either one or combination of modelling techniques that we have such as,

Decision Tree
Logistic Regression
Super Vector Machine (SVM)
KNN classification
Bayesian Methods
Random Forest

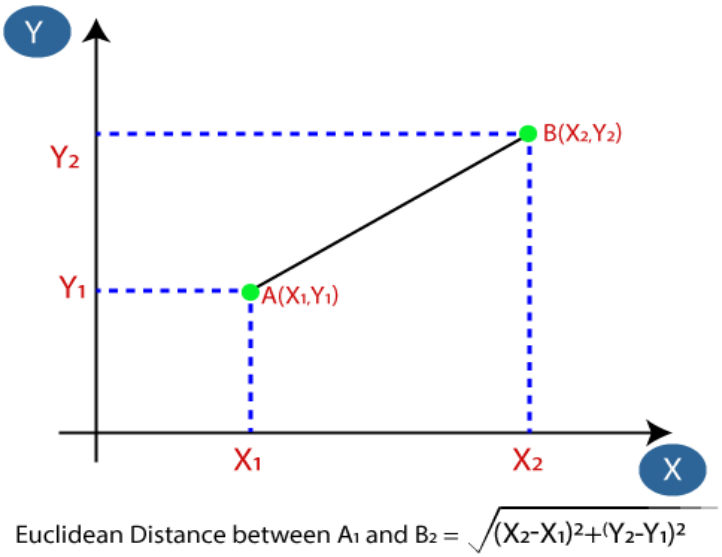


Fig. 4 Proposed KNN Algorithm Formula

***Advantages of proposed algorithm*:**

- Features are dependent which able to give good accuracy.
- It is very simple.
- Able to work even with large Datasets.
- Easy to implement.
- New Data can be added seamlessly.

*H. Build/Develop/Train the model*

- Validating all possibilities of picked algorithm.

- Based on the available of data it is necessary to train the model sufficiently.
- Validating the model performance like Error and Accuracy.

*I. Validate or test model*

- By using test data we have to validate test accuracy and final the score.
- Checking model performance that is model accuracy

## IV. IMPLEMENTATION

We have collected the dataset from officially at data.govern.in that we are using in our current project. This is information is maintained that is updated with every change by Chicago police department.

working on the project is followed by several steps they are

*A. Collection of Information*

We collected the dataset from Data.govern.in in .csv extension.

Our dataset name is Crimes_-_One_year_prior_to_present dataset.

*B. Data preprocessing*

Our dataset consists of --- entries.First we have converted all attributes into numerical datatype using label encoder and then replaced all Null Values using median Strategy.

*C. Feature selection*

Generally feature selection plays a crustial role that is it used to be build the model. The attributes which are used for feature selection are Block,Location,case,X coordinate , Y coordinate, Latitude , Longitude.

*D. Building and Training the model*

Location and --- attributes are used for the training after feature selection and then the dataset is classified into xtrain, ytrain and xtest, ytest. Sklearn is used to import the model alogorithms[fit(xtrain, ytrain)].

*E. Prediction*

model.predict(xtest) is used to done the prediction after done the all previous process and build the model. accuracy_score is used to calculate the accuracy which is acutally imported from the metrics.accuracy_score(ytest, predicted). This is the process we generally used in prediction process.

*F. Visualization*

We used sklearn to import the matplotlib library for visualization. We represented the crime analysis in many ways like plotting graphs and represented in pie charts.

*G. Results and Discussion*

We can obtain the results after undergoing into many processes with many functions that are through machine learning.

## V. DATA VISUALIZATION

Crime visualization totally allowed to show the dataset analysis in a visualized format like in way that a normal user can easily go through it. In detail way to plot the graphs or make them to understand by bar graphs etc… The analysis can be done through.

- In a period of time the number of crimes is might committed.
- By taking a city to observe the number of crimes overall crime types
- Ratio of taking them in custody that is the ratio of arrest in police records.
- By considering the different locations observing the committed crimes.
- Details of crimes that are happen majorly in city.

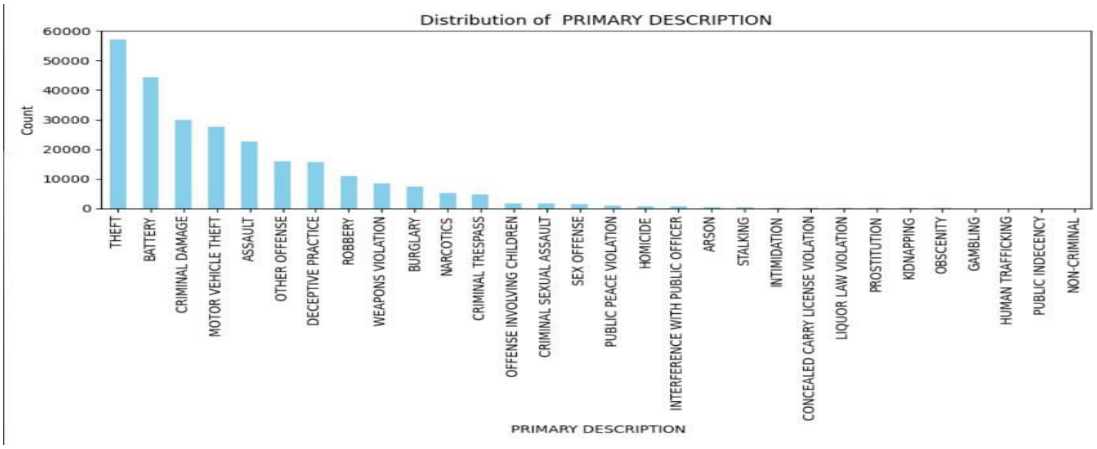


Fig.5 Major Crime Indicator

By analyzing the previous crimes we representing in the form of a bar graph which crime indicates mostly. Here Theft is the major crime indicator, secondary is the Battery Indicator etc… Here we represented x-axis as crime names and y-axis as crimes count. The bar graph represents the major crimes to minor crimes from left to right.

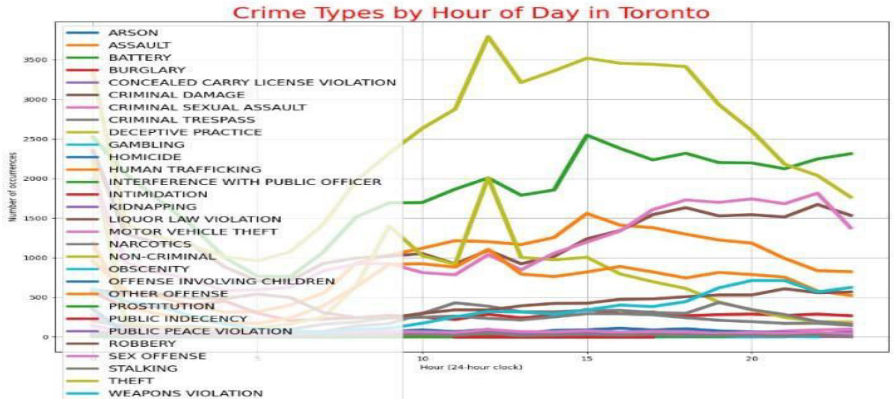


Fig.6 Crime Types by Hours

Here we have used line graph, In one graph we combined all lines into one graph. Here each line represents one type of crime, each line represents the crime and it represents the time on that day of occurrence which is in 24-hr format.

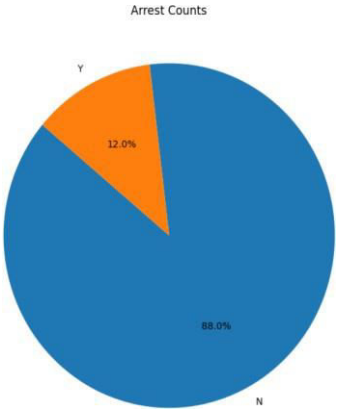


Fig.7 Pie Chart representing the Arrest counts

By using Pie chart we have represented the arrest counts.If we see in past 2023 Jan-2024 Jan only 12% cases accused has been arrest remaining 88% cases are still in pending.

## VI. SAMPLE SCREENSHOTS

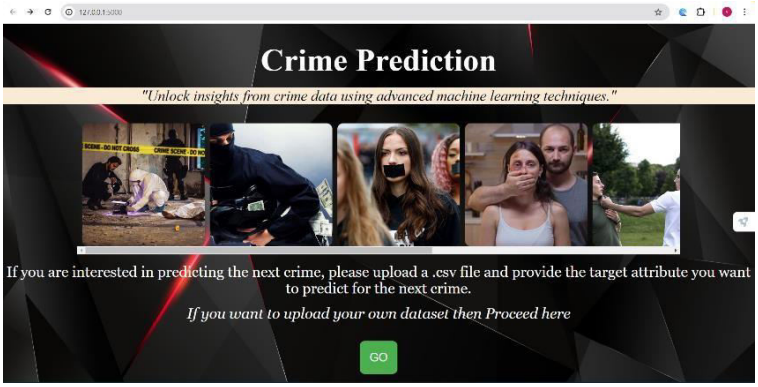


Fig.8 Home Page

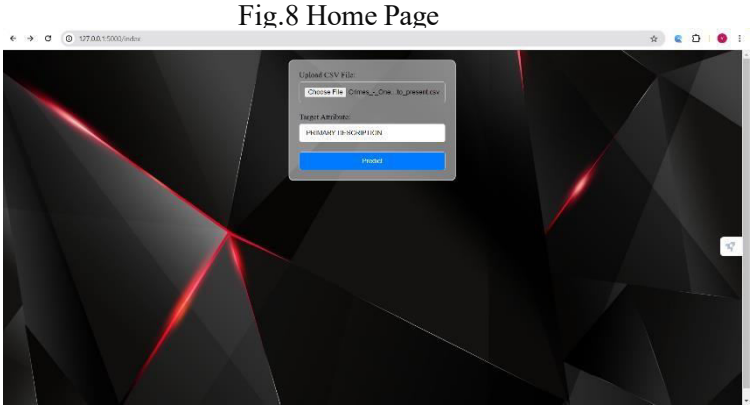

Fig.9 Index Page

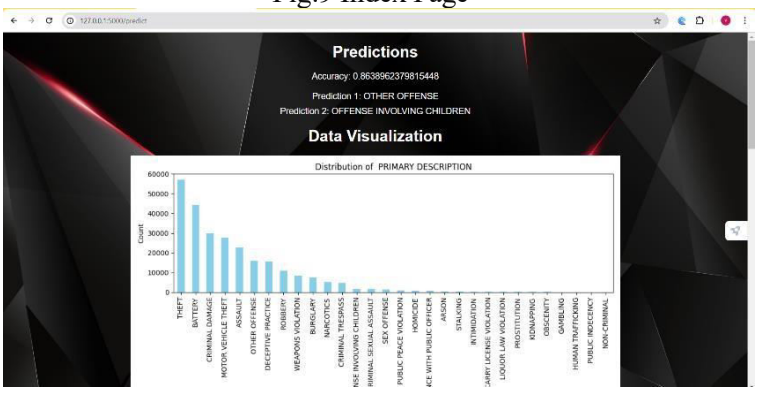


Fig.10(A) Output Page

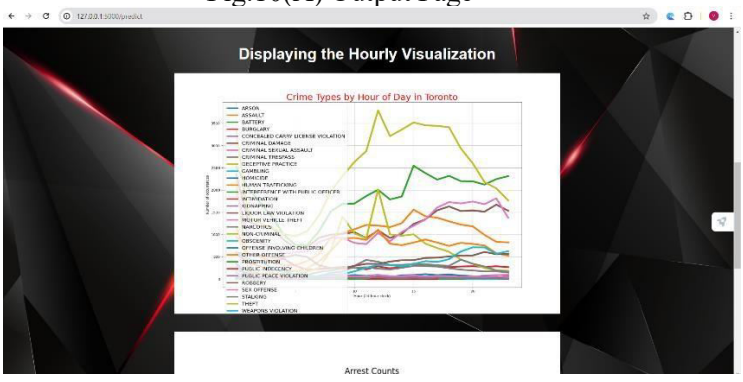


Fig.10(B) Output Page

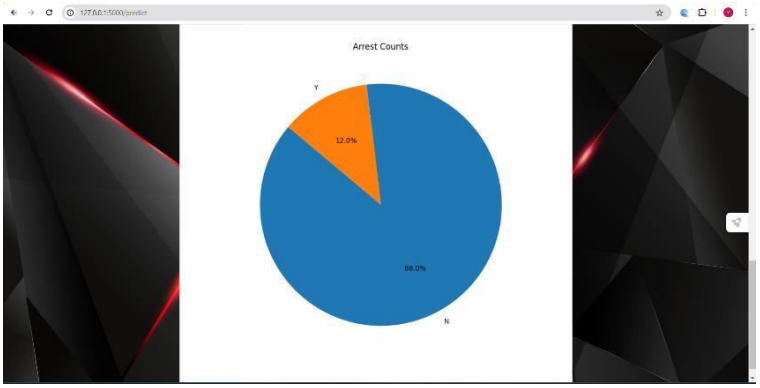


Fig.10(C) Output Page

## VII. CONCLUSION

From the entire work we can conclude that our main agenda is to predict the occurrence of crime, it will be possible by know the location of where it occur. The entire process can easily build by the help of Machine Learning techniques. Through the data we have it make us easy to find the patterns among the relation of occurrence of crime. With the data we have, we undergo with preprocessing like removing null values and delete unwanted data. We got accuracy 86%. We use bar graphs, pie charts etc… for easily understand about the concept. This research and work would help to society effectively.

## VIII. FUTURE SCOPE

Future research directions could focus on enhancing the predictive capabilities of crime prediction models by incorporating additional data sources, such as social media activity, weather patterns, and urban infrastructure. Moreover, exploring advanced machine learning techniques, such as deep learning and ensemble methods, may further improve the accuracy and robustness of predictive models in this domain.

## REFERENCES

1) Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In Electronics, Communication and AerospaceTechnology (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. In Student 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) Chen, Hsinchun, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau.

7) "Crime data mining: a general framework and some examples." computer 37, no. 4 (2004): 50-56.