MAJOR PROJECT REPORT

at

Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of

Bachelor of Engineering Degree in Computer Science and Engineering

By

Busupalli Harinath Reddy(Reg.No.38110063)

Avala Pavan Kumar (Reg. No.38110058)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI – 600119, TAMILNADU

MARCH 2022

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119

www.sathyabamauniversity.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Avala Pavan Kumar(38110058), Busupalli Harinath Reddy(38110063)** who carried out the project entitled "**A SYSTEMATIC APPROACH TOWARDS DESCRIPTION AND CLASSIFICATION OF CRIME INCIDENTS**" under my supervision from January 2022 to April 2022.

Internal Guide

Dr. R. AROUL CANESSANE M.E., Ph.D.,

Head of the Department

Dr. S.VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner **External Examiner**

DECLARATION

We, **Avala Pavan Kumar (38110058**), **Busupalli Harinath Reddy(Reg.No.38110063)** hereby declare that the Project Report entitled done by me under the guidance of **Dr. R. AROUL CANESSANE M.E., Ph.D.,** at Sathyabama institute of science andtechnology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing , **Dr.S.Vigneshwari M.E., Ph.D., and Dr.L.Lakshmanan M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R. AROUL CANESSANE M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TABLE OF CONTENT

INDEX NO	TITLE	PAGE NO
1.	ABSTRACT	6
2.	INTRODUCTION	7
3.	AIM	13
4.	SCOPE	13
5.	MODULES AND MODULE DESCRIPTION	14
6.	SYSTEM ANALYSIS	20
7.	SYSTEM ARCHITECHTURE	22
8.	CONCLUSION	28
9.	SCREENSHOTS	29
10.	REFERENCES	35

ABSTRACT

Crime analysis and prediction is a systematic approach for identifying the crime. This system can predict region which have high probability for crime occurrences and visualize crime prone area. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. The extraction of new information is predicted using the existing datasets. Crimes are treacherous and common social problem faced worldwide. Crimes affect the quality of life, economic growth and reputation of nation. With the aim of securing the society from crimes, there is a need for advanced systems and new approaches for improving the crime analytics for protecting their communities. We propose a system which can analysis, detect, and predict various crime probability in given region. This paper explains various types of criminal analysis and crime prediction using several data mining techniques.

INTRODUCTION

What is Machine Learning?

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to pedict an output which can be used to make actionable insights.

The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.

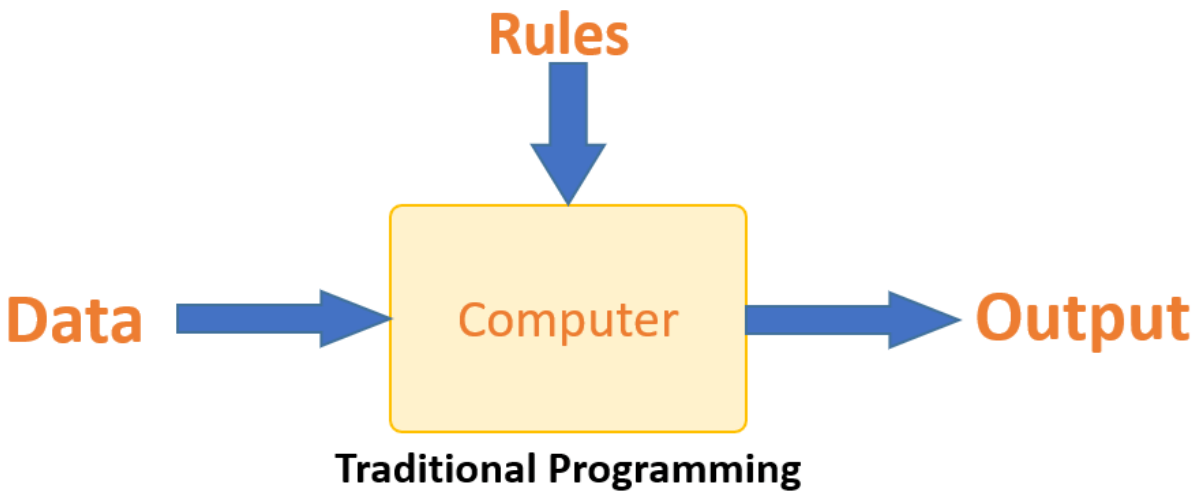
A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

Machine Learning vs. Traditional Programming

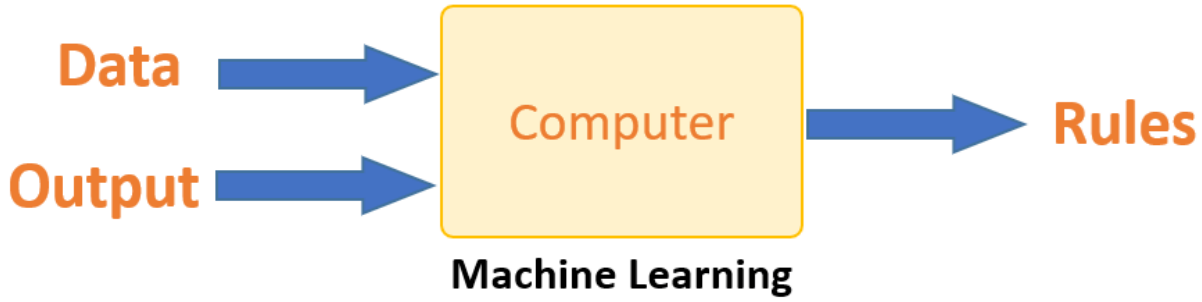
Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.



Traditional Programming

Machine learning is supposed to overcome this issue. The machine learns how the input and output data are correlated and it writes a rule. The programmers do not need to write new rules each time there is new data. The algorithms adapt in response to new data and experiences to improve efficacy over time.



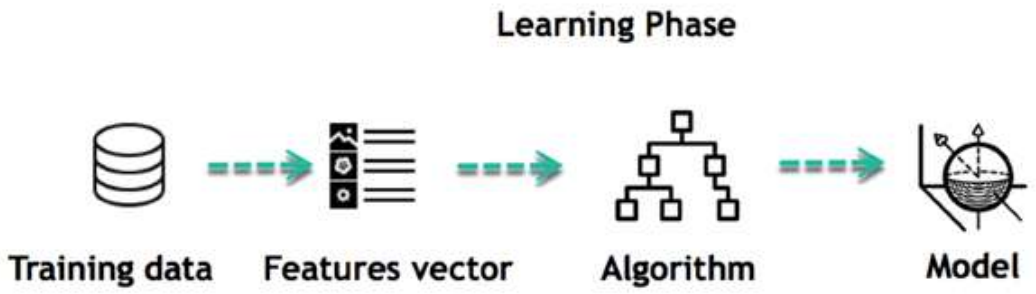
Machine Learning

How does Machine Learning Work?

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the **learning** and **inference**. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the **data**. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a **feature vector.** You can think of a feature vector as a subset of data that is used to tackle a problem.

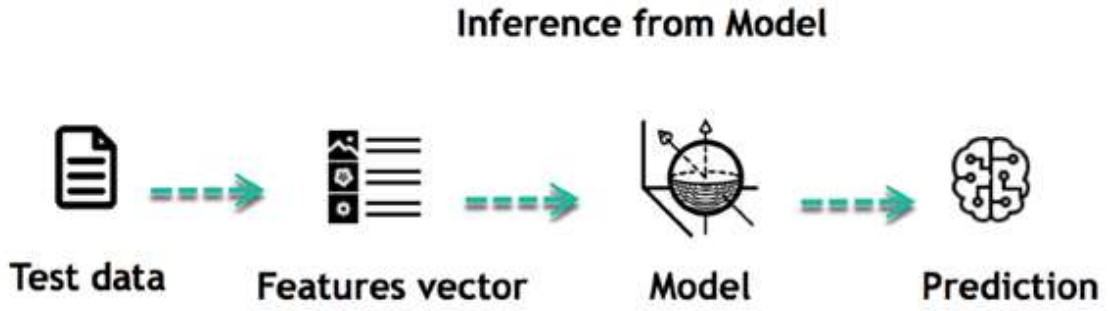
The machine uses some fancy algorithms to simplify the reality and transform this discovery into a **model**. Therefore, the learning stage is used to describe the data and summarize it into a model.



For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model

Inferring

When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

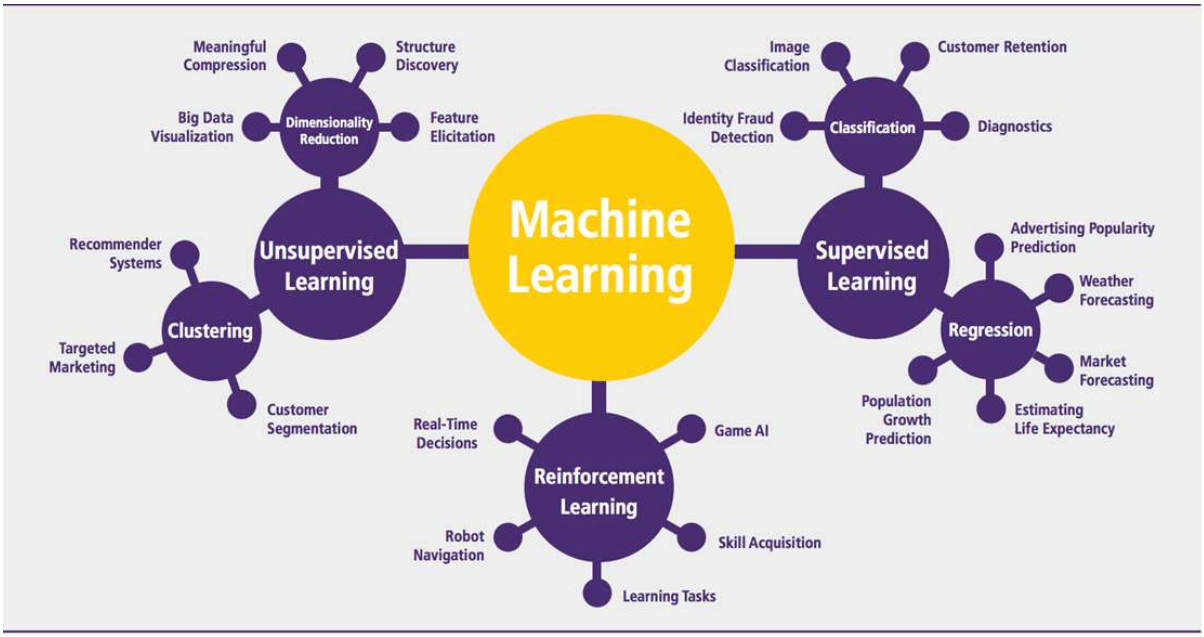


The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

Machine Learning Algorithms and Where they are Used?



Machine learning Algorithms

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms

Supervised learning

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.

You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

- Classification task
- Regression task

Classification

Imagine you want to predict the gender of a customer for a commercial. You will start gathering data on the height, weight, job, salary, purchasing basket, etc. from your customer database. You know the gender of each of your customer, it can only be male or female. The objective of the classifier will be to assign a probability of being a male or a female (i.e., the label) based on the information (i.e., features you have collected). When the model learned how to recognize male or female, you can use new data to make a prediction. For instance, you just got new information from an unknown customer, and you want to know if it is a male or female. If the classifier predicts male = 70%, it means the algorithm is sure at 70% that this

customer is a male, and 30% it is a female.

The label can be of two or more classes. The above Machine learning example has only two classes, but if a classifier needs to predict object, it has dozens of classes (e.g., glass, table, shoes, etc. each object represents a class)

Regression

When the output is a continuous value, the task is a regression. For instance, a financial analyst may need to forecast the value of a stock based on a range of feature like equity, previous stock performances, macroeconomics index. The system will be trained to estimate the price of the stocks with the lowest possible error.

Algorithm Name	Description	Type
Linear regression	Finds a way to correlate each feature to the output to help predict future values.	Regression
Logistic regression	Extension of linear regression that's used for classification tasks. Th output variable 3is binary (e.g., only black or white) rather tha continuous (e.g., an infinite list of potential colors)	Classification
Decision tree	Highly interpretable classification or regression model that splits data feature values into branches at decision nodes (e.g., if a feature is color, each possible color becomes a new branch) until a fina decision output is made	Regression Classification
Naive Bayes	The Bayesian method is a classification method that makes use o the Bayesian theorem. The theorem updates the prior knowledge o an event with the independent probability of each feature that ca affect the event.	Regression Classification
Support vector machine	Support Vector Machine, or SVM, is typically used for th classification task. SVM algorithm finds a hyperplane that optimall divided the classes. It is best used with a non-linear solver.	Regression (not ver common) Classification

Algorithm Name	Description	Type
Random forest	The algorithm is built upon a decision tree to improve the accurac drastically. Random forest generates many times simple decisio trees and uses the 'majority vote' method to decide on which label t return. For the classification task, the final prediction will be the on with the most vote; while for the regression task, the averag prediction of all the trees is the final prediction.	Regression Classification
AdaBoost	Classification or regression technique that uses a multitude of model to come up with a decision but weighs them based on their accurac in predicting the outcome	Regression Classification
Gradient-boosting trees	Gradient-boosting trees is a state-of-the-art classification/regressio technique. It is focusing on the error committed by the previous tree and tries to correct it.	Regression Classification

Unsupervised learning

In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns)

You can use it when you do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you

Algorithm	Description	Type
K-means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters)	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommender system	Help to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension Reduction

AIM AND SCOPE OF THE PRESENT INVESTIGATION

AIM :

OUR AIM TOWARDS THIS PROJECT IS TO PREDICT THE CRIME INCIDENTS THAT HAPPENS IN FUTURE. 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.

SCOPE :

A SYSTEMATIC APPROACH TOWARDS DESCRIPTION AND CLASSIFICATION OF CRIME INCIDENTS

EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHM USED

MODULES:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

MODULES DESCSRIPTION:

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Dataset:

The dataset consists of 520 individual data. There are 23 columns in the dataset, which are described below.

1. **ID**: Unique identifier for the record.
2. **Case Number**: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
3. **Date**: Date when the incident occurred.
4. **Block**: address where the incident occurred
5. **IUCR**: The Illinois Unifrom Crime Reporting code.
6. **Primary Type**: The primary description of the IUCR code.
7. **Description**: The secondary description of the IUCR code, a subcategory of the primary description.
8. **Location Description**: Description of the location where the incident occurred.
9. **Arrest**: Indicates whether an arrest was made.

10. **Domestic**: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
11. **Beat**: Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car.
12. **District**: Indicates the police district where the incident occurred.
13. **Ward**: The ward (City Council district) where the incident occurred.
14. **Community Area: Indicates** the community area where the incident occurred. Chicago has 77 community areas.
15. **FBI Code**: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
16. **X Coordinate**: The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
17. **Y Coordinate**: The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
18. **Year**: Year the incident occurred.
19. **Updated On**: Date and time the record was last updated.
20. **Latitude**: The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
21. **Longitude**: The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
22. **Location**: The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Data Preparation:

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
Split into training and evaluation sets

Model Selection:

We used Random Forest Classifier machine learning algorithm , We got a accuracy of 80.7% on test set so

we implemented this algorithm.

The Random Forests Algorithm

Let's understand the algorithm in layman's terms. Suppose you want to go on a trip and you would like to travel to a place which you will enjoy.

So what do you do to find a place that you will like? You can search online, read reviews on travel blogs and portals, or you can also ask your friends.
Let's suppose you have decided to ask your friends, and talked with them about their past travel experience to various places. You will get some recommendations from every friend. Now you have to make a list of those recommended places. Then, you ask them to vote (or select one best place for the trip) from the list of recommended places you made. The place with the highest number of votes will be your final choice for the trip.

In the above decision process, there are two parts. First, asking your friends about their individual travel experience and getting one recommendation out of multiple places they have visited. This part is like using the decision tree algorithm. Here, each friend makes a selection of the places he or she has visited so far.
The second part, after collecting all the recommendations, is the voting procedure for selecting the best place in the list of recommendations. This whole process of getting recommendations from friends and voting on them to find the best place is known as the random forests algorithm.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

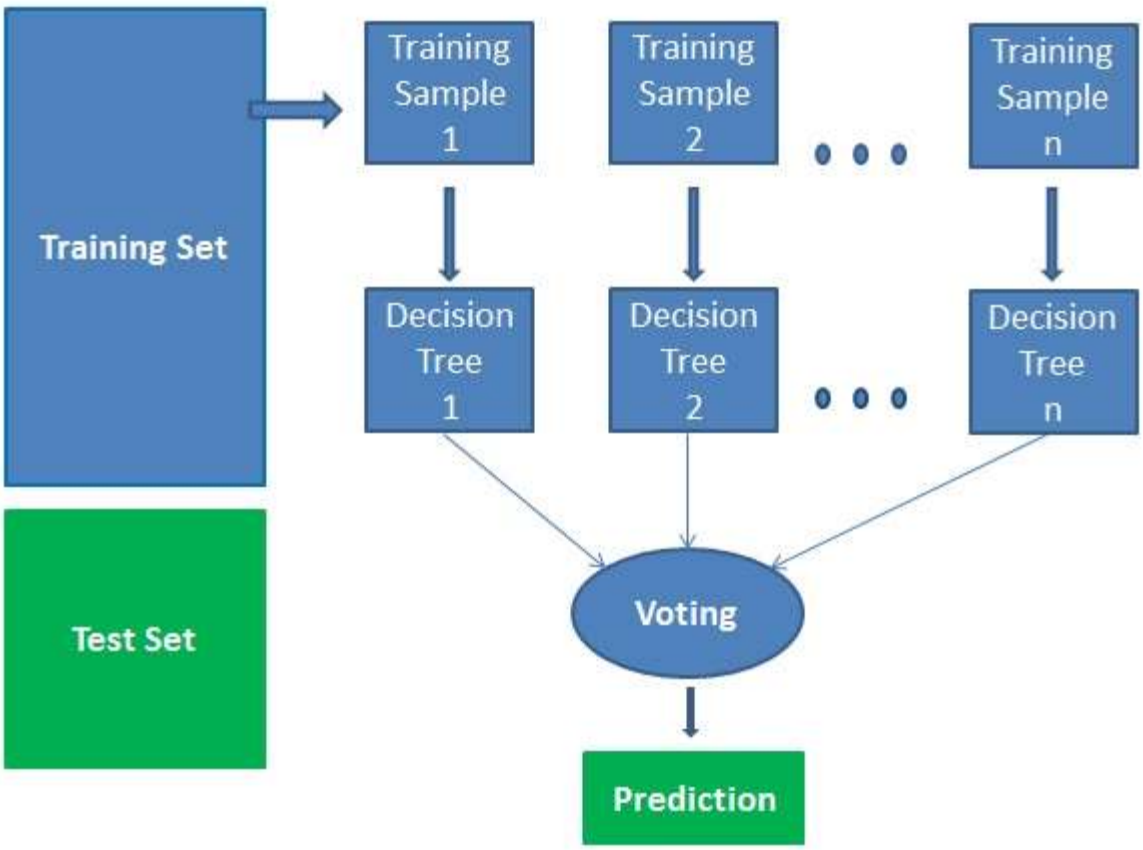
How does the algorithm work?

It works in four steps:
Select random samples from a given dataset.
Construct a decision tree for each sample and get a prediction result from each decision tree.
Perform a vote for each predicted result.

Select the prediction result with the most votes as the final prediction.

Advantages:

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- The algorithm can be used in both classification and regression problems.
- Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.



Disadvantages:

- Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
- The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

Finding important features

Random forests also offers a good feature selection indicator. Scikit-learn provides an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction. It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.

This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.

Random Forests vs Decision Trees

- Random forests is a set of multiple decision trees.
- Deep decision trees may suffer from overfitting, but random forests prevents overfitting by creating trees on random subsets.
- Decision trees are computationally faster.
- Random forests is difficult to interpret, while a decision tree is easily interpretable and can be converted to rules.

Analyze and Prediction:

In the actual dataset, we chose only 8 features :

1. **Year** : Year when the incident occurred.
2. **Month**: Month when the incident occurred.
3. **Day**: Day when the incident occurred.
4. **Day Of Week**: Day Of Week when the incident occurred.
5. **Minute**: Minute when the incident occurred.
6. **Second**: second when the incident occurred.
7. **Latitude**: The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
8. **Longitude**: The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block

Accuracy on test set:

We got an accuracy of 80% on test set.

Saving the Trained Model:

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .
Make sure you have pickle installed in your environment.
Next, let's import the module and dump the model into .pkl file

SYSTEM ANALYSIS

EXISTING SYSTEM:

- In pre-work, the dataset obtained from the open source are first pre-processed to remove the duplicated values and features.
- Decision tree has been used in the factor of finding crime patterns and also extracting the features from large amount of data is inclusive. It provides a primary structure for further classification process.
- The classified crime patterns are feature extracted using Deep Neural network. Based on the prediction, the performance is calculated for both trained and test values. The crime prediction helps in forecasting the future happening of any type of criminal activities and help the officials to resolve them at the earliest.

DISADVANTAGES OF EXISTING SYSTEM:

- The pre-existing works account for low accuracy since the classifier uses a categorical values which produces a biased outcome for the nominal attributes with greater value.
- The classification techniques does not suited for regions with inappropriate data and real valued attributes.
- The value of the classifier must be tuned and hence there is a need of assigning an optimal value.

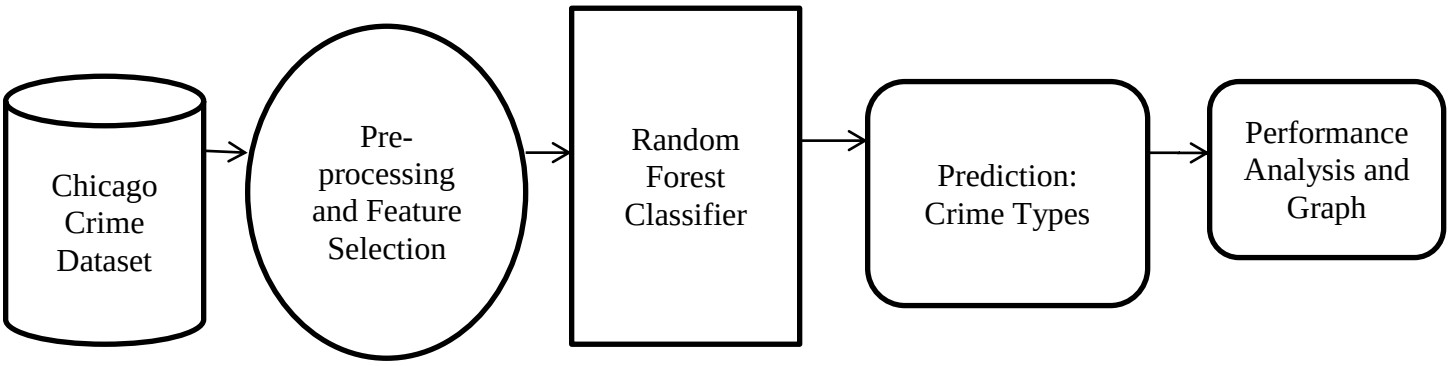
PROPOSED SYSTEM:

- The data obtained is first pre-processed using machine learning technique filter and wrapper in order to remove irrelevant and repeated data values. It also reduces the dimensionality thus the data has been cleaned. The data is then further undergoes a splitting process. It is classified into test and trained data set.
- The model is trained by dataset both training and testing .It is then followed by mapping. The crime type, year, month, time, date, place are mapped to an integer for ensuring classification easier. The independent effect between the attributes are analysed initially by using Random Forest Classifier.
- The crime features are labelled that allows to analyse the occurrence of crime at a particular time and location. Finally, the crime which occur the most along with spatial and temporal information is gained. The performance of the prediction model is find out by calculating accuracy rate. The language used in designing the prediction model is python and run on data analysis and machine learning model.

ADVANTAGES OF PROPOSED SYSTEM:

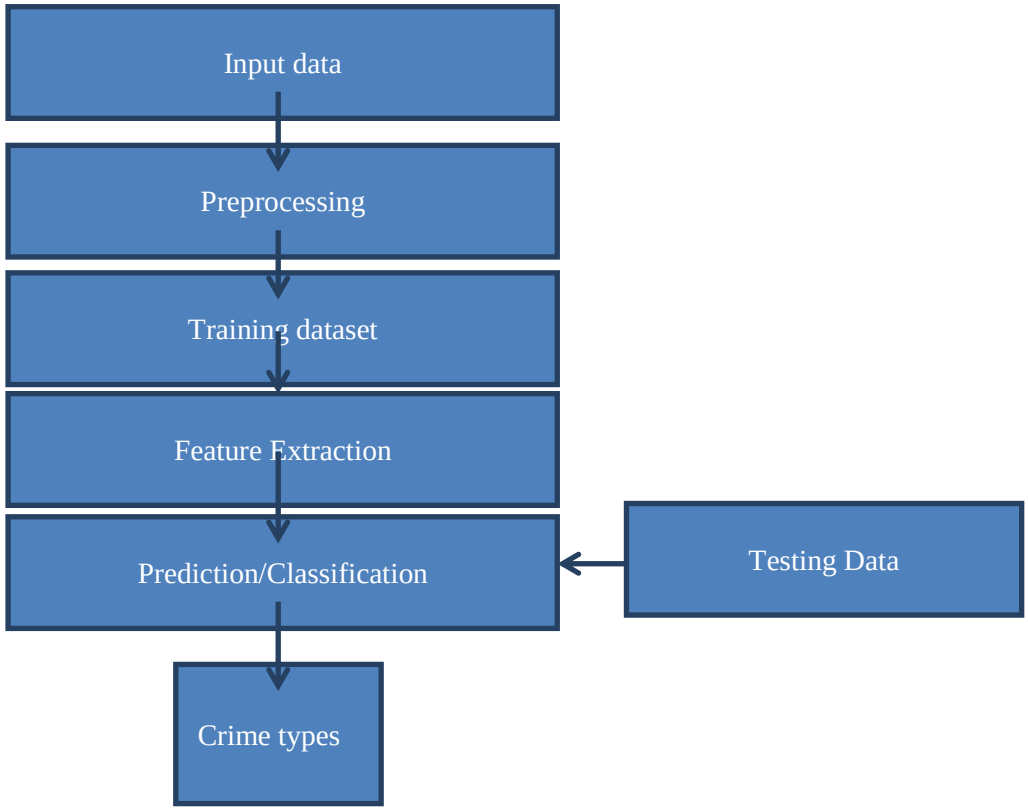
- The proposed algorithm is well suited for the crime pattern detection since most of the featured attributes depends on the time and location.
- It also overcomes the problem of analyzing independent effect of the attributes.
- The initialization of optimal value is not required since it accounts for real valued, nominal value and also concern the region with insufficient information.
- The accuracy has been relatively high when compared to other machine learning prediction model.

SYSTEM ARCHITECTURE



DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

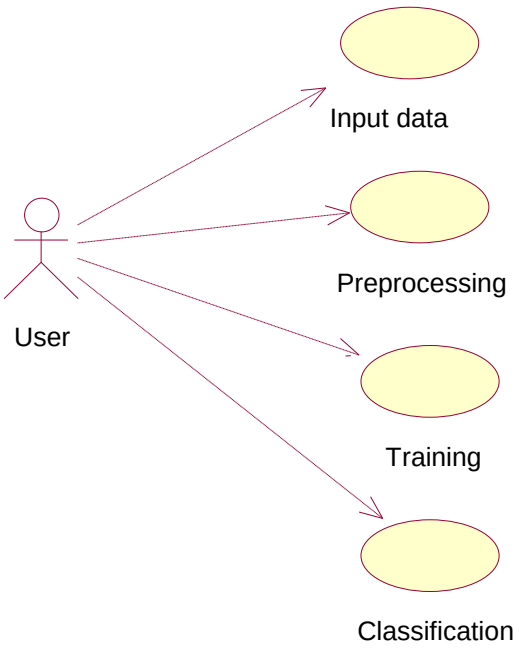
GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

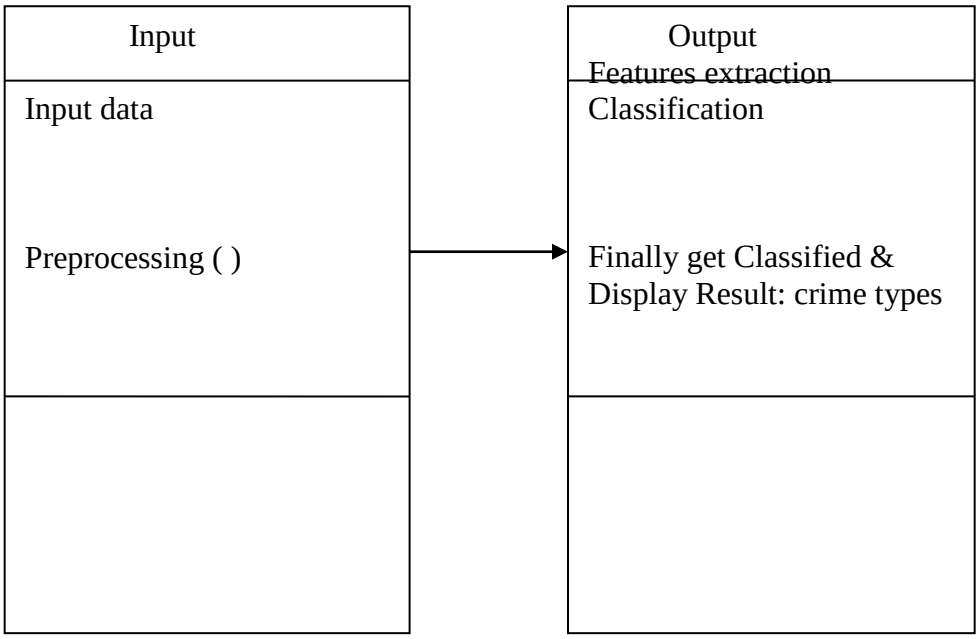
USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



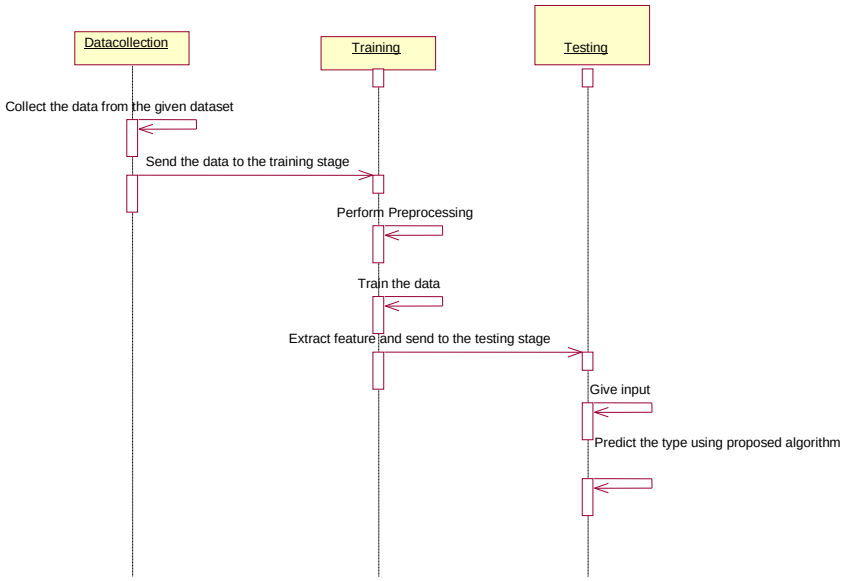
CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



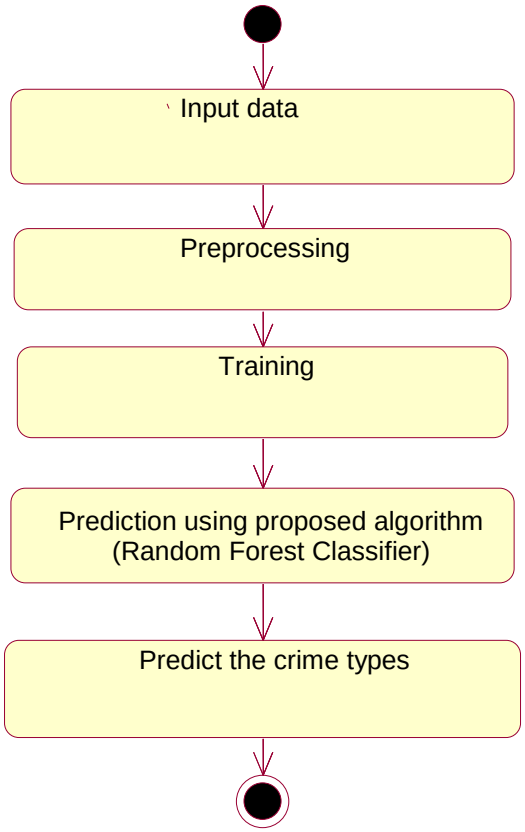
SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.


Input data
Preprocessing
Training
Prediction using proposed algorithm
(Random Forest Classifier)
Predict the crime types

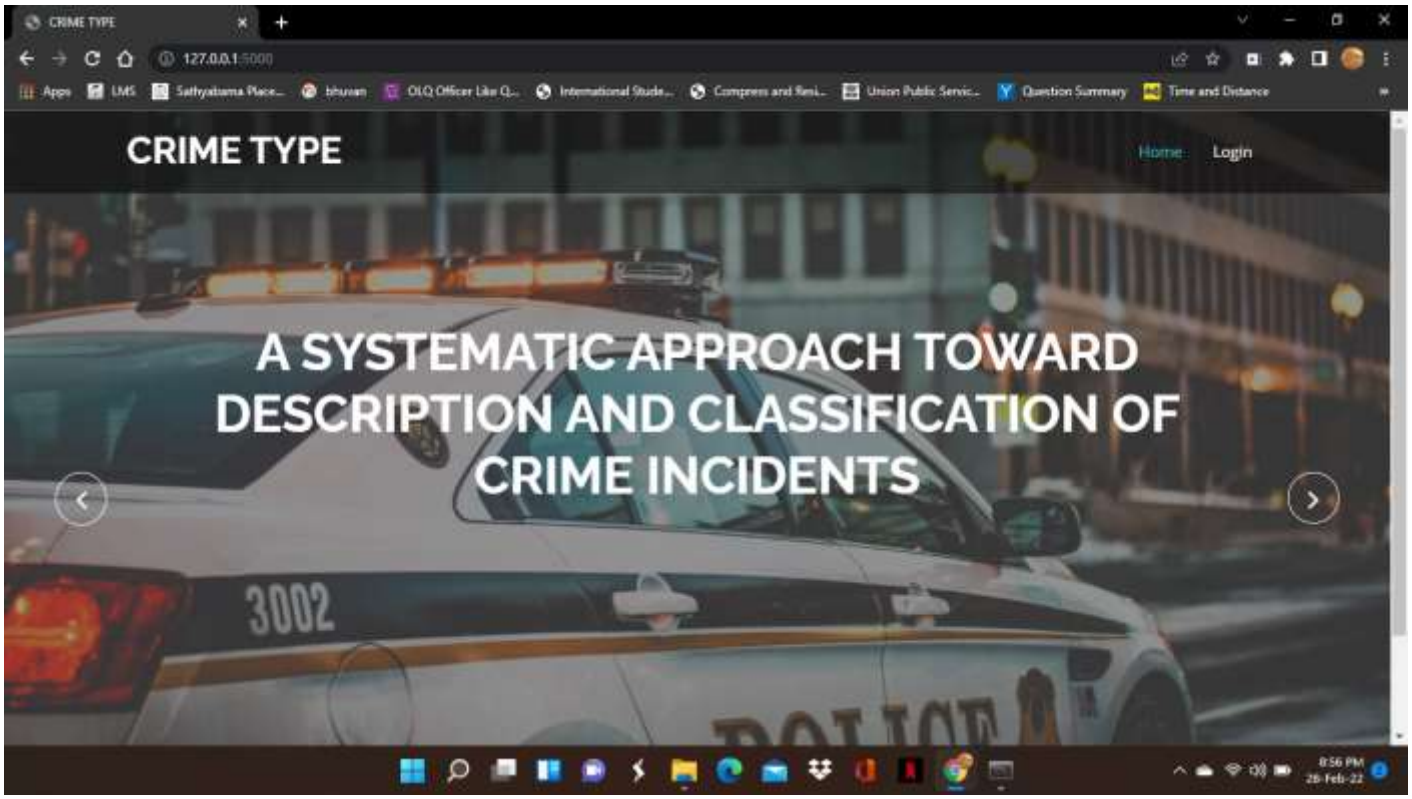
CONCLUSION

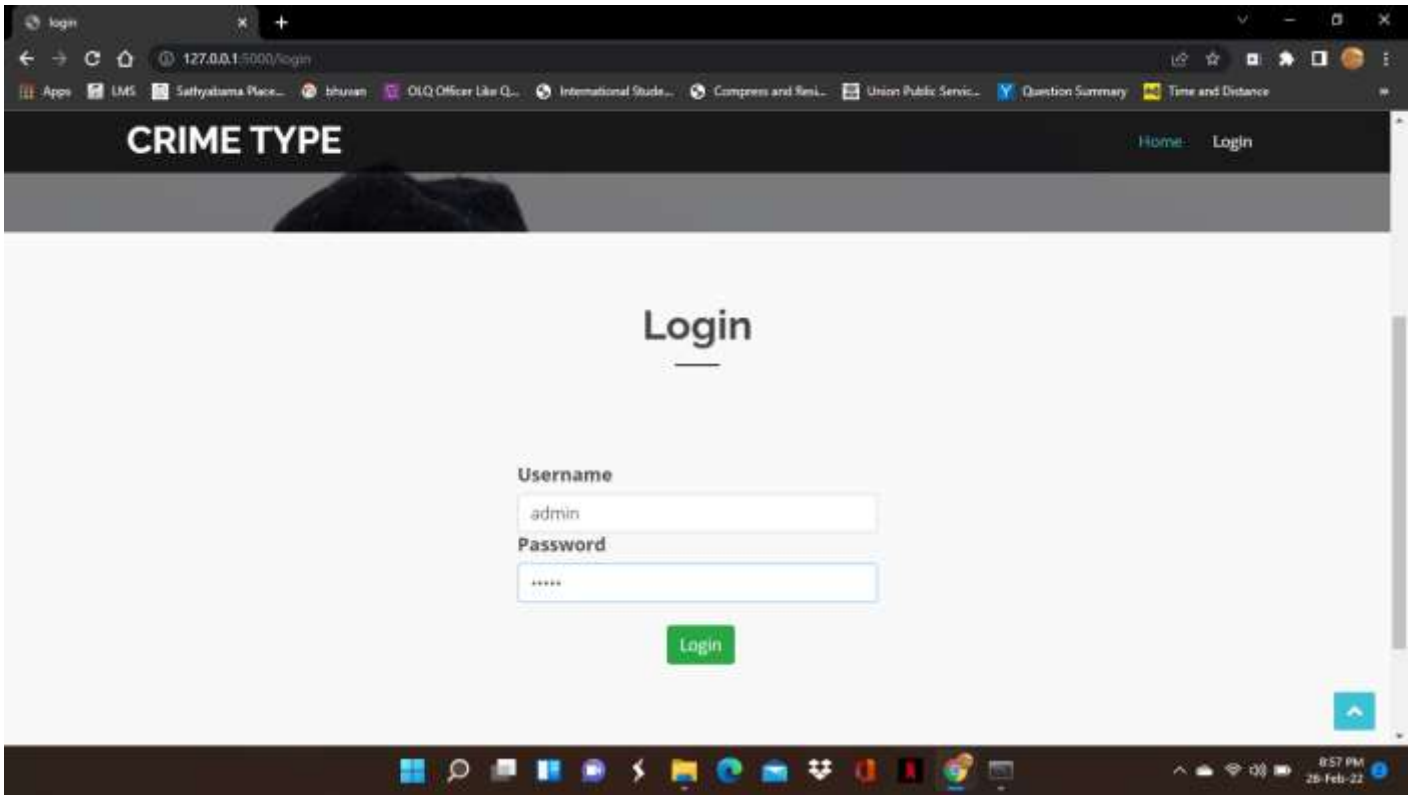
In this paper, the difficulty in dealing with the nominal distribution and real valued attributes is overcome by using two classifiers such as Multinominal NB and Gaussian NB. Much training time is not required and serves to be the best suited for realtime predictions. It also overcomes the problem of working with continuous target set of variables where the existing work refused to fit with. Thus the crime that occur the most could be predicted and spotted using Random Forest Classification. The performance of the algorithm is also calculated by using some standard metrics. The metrics include average precision, recall, F1 score and accuracy are mainly concerned in the algorithm evaluation. The accuracy value could be increased much better by implementing machine learning algorithms.

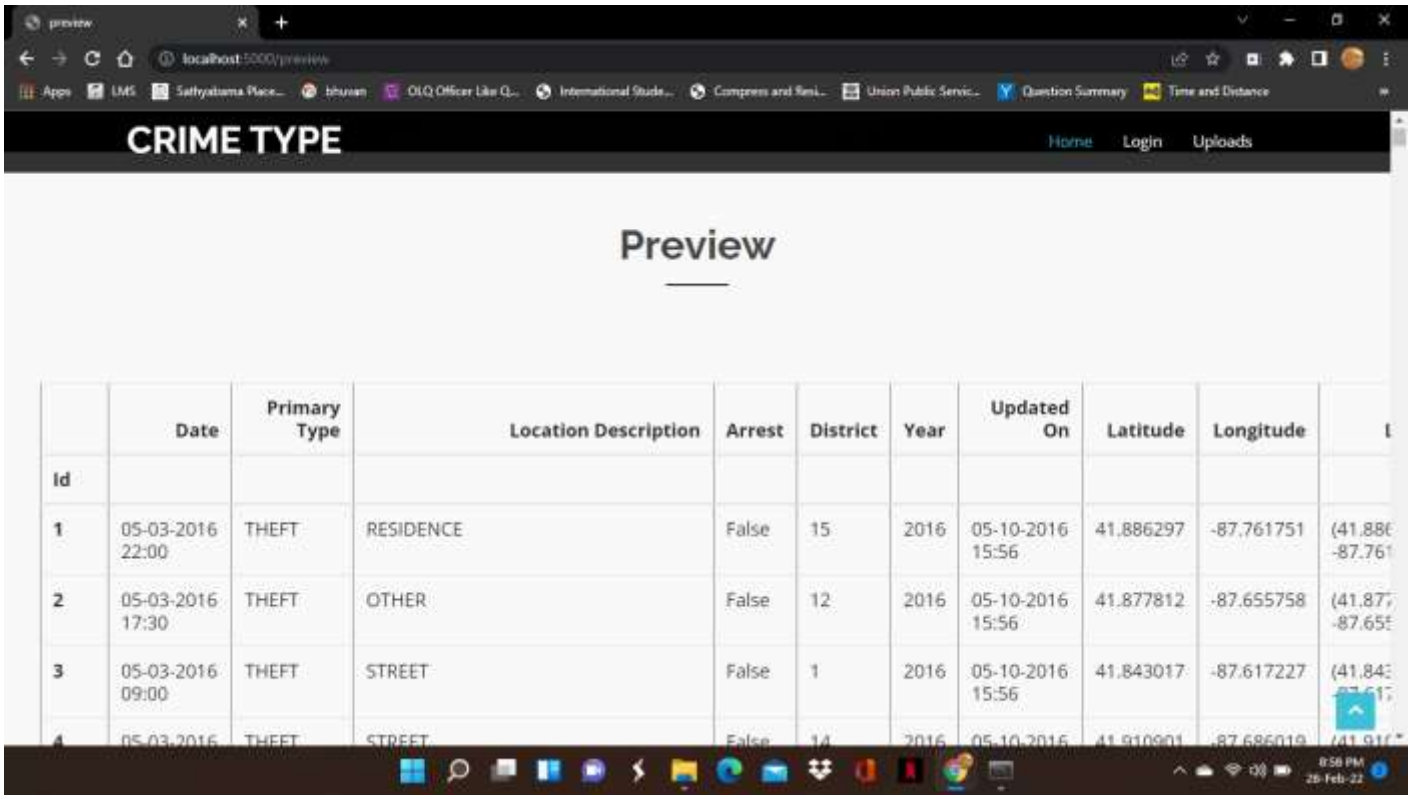
Future Work

Though it overcomes the problem of the existing work, it has some limitations. In the situation of absence of class labels, then the probability of the estimation will be zero. As a future extension of the proposed work, the application of more machine learning classification models proves to increase accuracy in crime prediction and will enhance the overall performance. It helps in providing a better study for the future improvement by taking the income information into consideration for neighborhoods places in order to foresee if any relationship between the income levels of a particular in the neighborhood places and their crime rates.

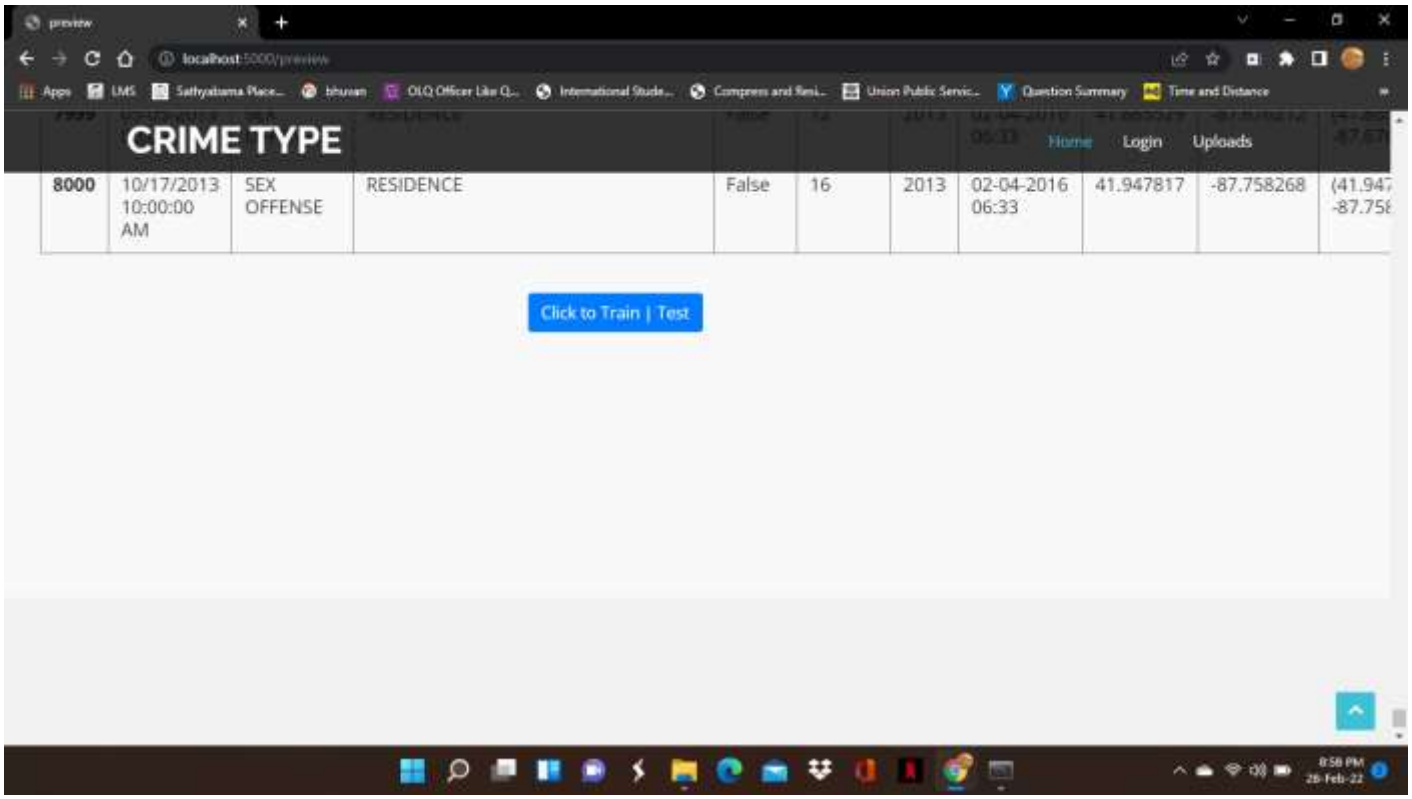
SCREENSHOTS

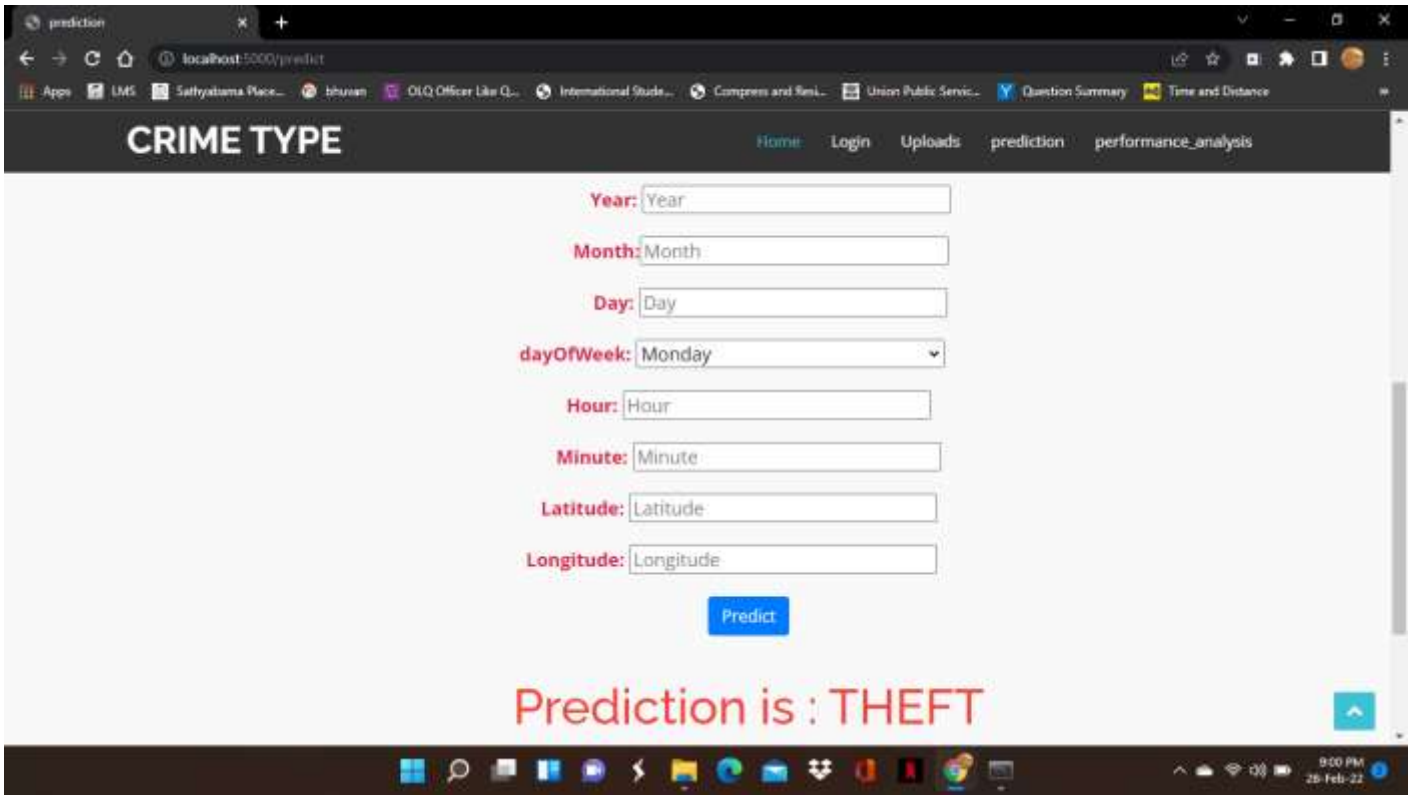






Id	Date	Primary Type	Location Description	Arrest	District	Year	Updated On	Latitude	Longitude
1	05-03-2016 22:00	THEFT	RESIDENCE	False	15	2016	05-10-2016 15:56	41.886297	-87.761751
2	05-03-2016 17:30	THEFT	OTHER	False	12	2016	05-10-2016 15:56	41.877812	-87.655758
3	05-03-2016 09:00	THEFT	STREET	False	1	2016	05-10-2016 15:56	41.843017	-87.617227


CRIME TYPE
Home
Login
Uploads
8000
10/17/2013 10:00:00 AM
SEX OFFENSE
RESIDENCE
False
16
2013
02-04-2016 06:33
41.947817
-87.758268
Click to Train | Test


CRIME TYPE
Home
Login
Uploads
prediction
performance_analysis
Year:
Year
Month:
Month
Day:
Day
dayOfWeek:
Monday
Hour:
Hour
Minute:
Minute
Latitude:
Latitude
Longitude:
Longitude
Predict
Prediction is : THEFT

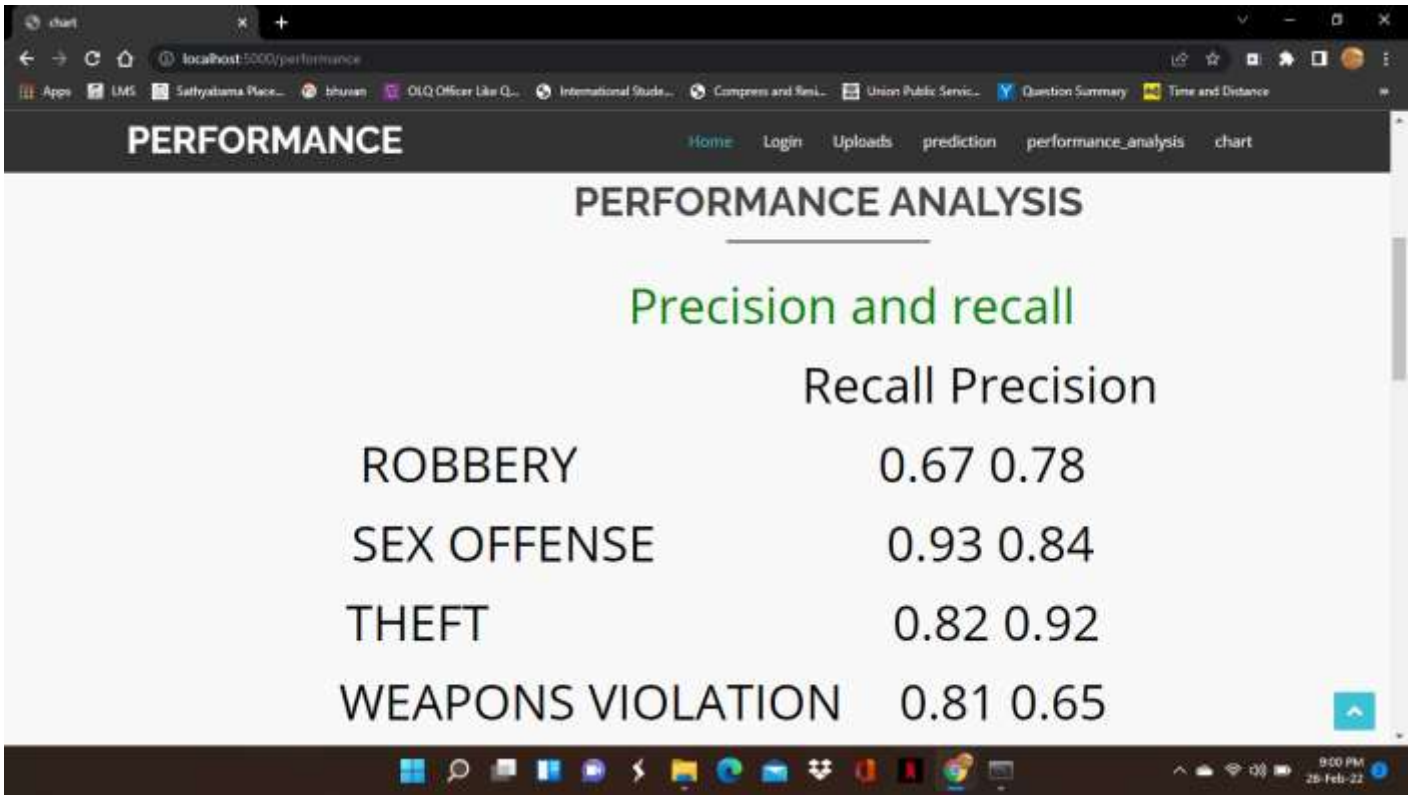


chart
localhost:5000/performance
PERFORMANCE
Home
Login
Uploads
prediction
performance_analysis
chart
PERFORMANCE ANALYSIS
Precision and recall
Recall Precision
ROBBERY 0.67 0.78
SEX OFFENSE 0.93 0.84
THEFT 0.82 0.92
WEAPONS VIOLATION 0.81 0.65

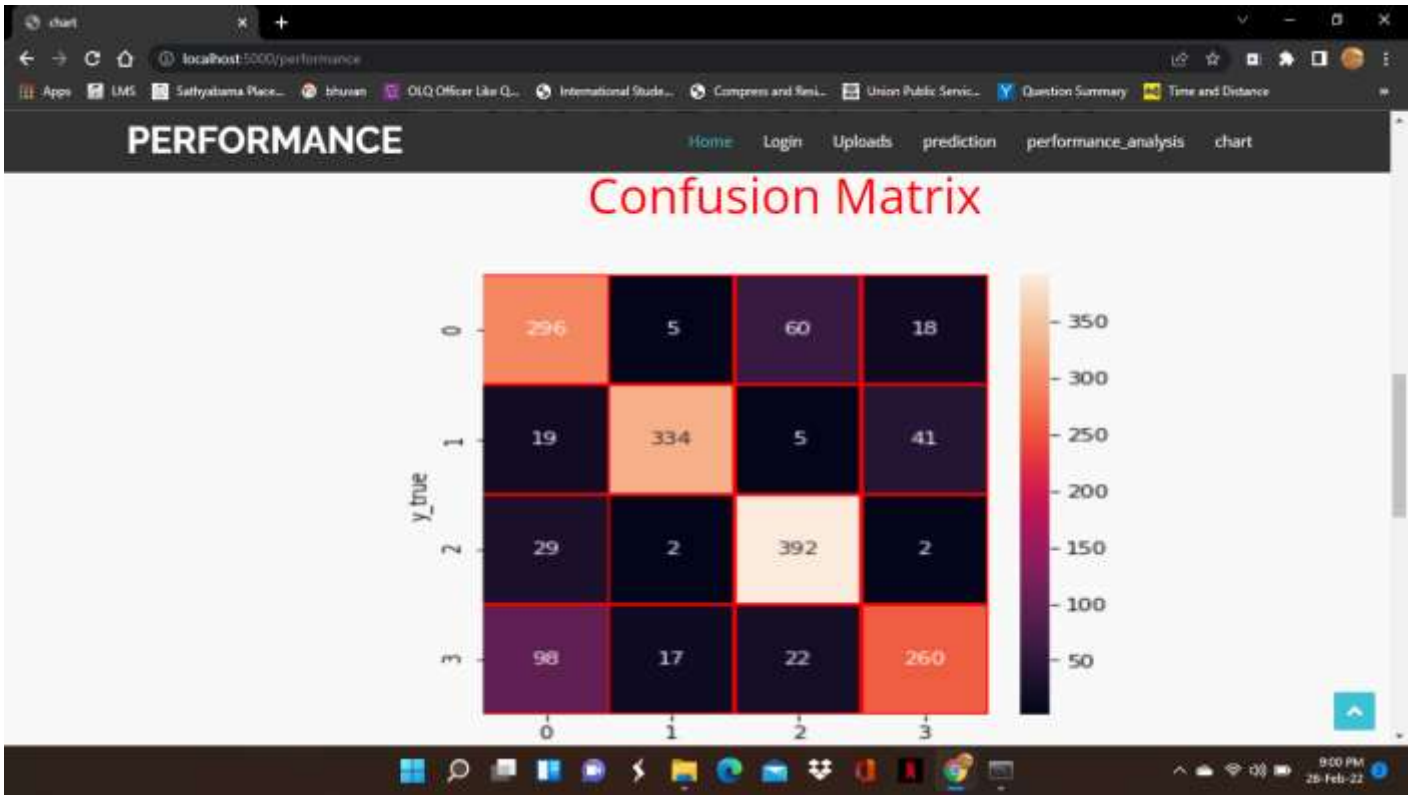


chart
localhost:5000/performance
PERFORMANCE
Home
Login
Uploads
prediction
performance_analysis
chart
Confusion Matrix
y_true
296
5
60
18
19
334
5
41
29
2
392
2
98
17
22
260
0
1
2
3
350
300
250
200
150
100
50

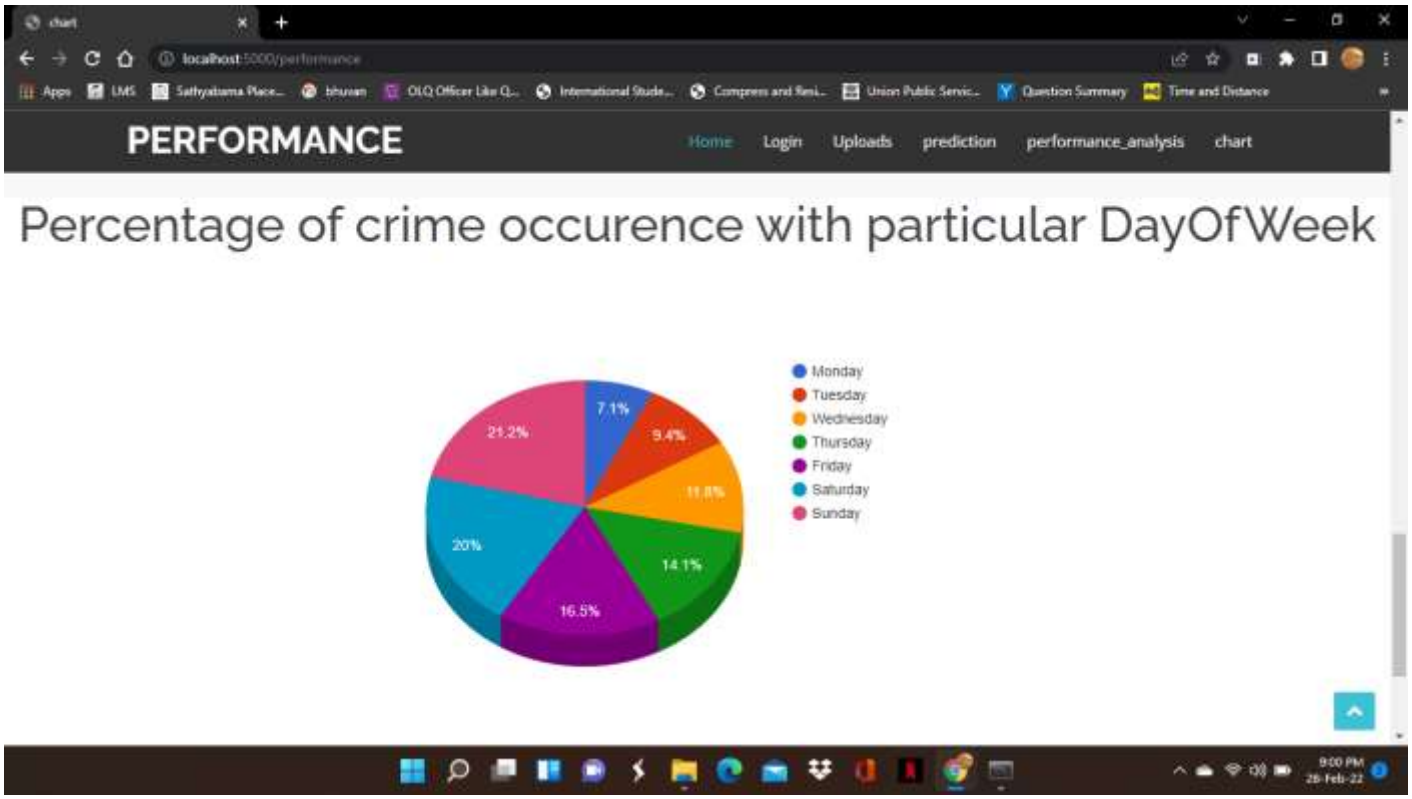


chart
localhost:5000/performance
PERFORMANCE
Home
Login
Uploads
prediction
performance_analysis
chart
Percentage of crime occurence with particular DayOfWeek
7.1%
9.4%
11.8%
14.1%
16.5%
20%
21.2%
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday


chart
localhost:5000/chart
CRIME TYPE
Home
Login
Uploads
prediction
performance_analysis
chart
Logout
Chart
Highest crime type
100
90
80
70
60
50
99
85
70
60
ROBBERY
THEFT
SEX OFFENSE
WEAPONS VIOLATION


chart
localhost:5000/chart
CRIME TYPE
Home
Login
Uploads
prediction
performance_analysis
chart
Logout
Percentage of crime occurence with particular DayOfWeek
7.1%
9.4%
11.8%
14.1%
16.5%
20%
21.2%
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

REFERENCES

[1] Ginger Saltos and Mihaela Coacea, An Exploration of Crime prediction Using Data Mining on Open Data, International journal of Information technology & Decision Making,2017.

[2] Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.

[3] Khushabu A.Bokde, Tisksha P.Kakade, Dnyaneshwari S. Tumasare, Chetan G.Wadhai B.E Student, Crime Detection Techniques Using Data Mining and K-Means, International Journal of Engineering Research & technology (IJERT) ,2018.

[4] H.Benjamin Fredrick David and A.Suruliandi,Survey on crime analysis and prediction using data mining techniques, ICTACT Journal on Soft computing, 2017.

[5] Tushar Sonawanev, Shirin Shaikh, rahul Shinde, Asif Sayyad, Crime Pattern Analysis, Visualization And prediction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE), 2015.

[6] RajKumar.S, Sakkarai Pandi.M, Crime Analysis and prediction using data mining techniques, International Journal of recent trends in engineering & research,2019.

[7] Sarpreet kaur, Dr. Williamjeet Singh, Systematic review of crime data mining, International Journal of Advanced Research in computer science , 2015.

[8] Ayisheshim Almaw, Kalyani Kadam, Survey Paper on Crime Prediction using Ensemble Approach, International journal of Pure and Applied Mathematics,2018.

[9] Dr .M.Sreedevi, A.Harha Vardhan Reddy, ch.Venkata Sai Krishna Reddy, Review on crime Analysis and prediction Using Data Mining Techniques, International Journal of Innovative Research in Science Engineering and technology ,2018.

[10] K.S.N .Murthy, A.V.S.Pavan kumar, Gangu Dharmaraju, international journal of engineering, Science and mathematics, 2017.

[11] Deepiika k.K, Smitha Vinod, Crime analysis in india using data minig techniques , International journal of Enginnering and technology, 2018.

[12] Hitesh Kumar Reddy ToppyiReddy, Bhavana Saini, Ginika mahajan, Crime Prediction &Monitoring Framework Based on Spatial Analysis, International Conference on Computational Intelligence Data Science (ICCIDS 2018).