# CRIME DATA ANALYSIS

**A MINI PROJECT REPORT**

***Submitted by***


**DEENADHAYALAN.V** **420419205003**

**VIKRAM.P** **420419205020**


***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

***in***

**INFORMATION TECHNOLOGY**



**ADHIPARASAKTHI ENGINEERING COLLEGE**

**MELMARUVATHUR**

**ANNA UNIVERSITY: CHENNAI 600025**


**JUNE 2022**

# BONAFIDE CERTIFICATE

Certified that this report titled "**CRIME DATA ANALYSIS"**, is a bonafide work of **DEENADHAYALAN.V (420419205003),VIKRAM(420419205020),** who carried out the work under my supervision.

**HEAD OF THE DEPARTMENT**

**Dr.A.BHUVANESWARI M.E.Ph.D.,**
PROFESSOR,
Dept. of Information Technology,
Adhiparasakthi Engineering College,
Melmaruvathur-603 319.

**SUPERVISOR**

**Dr.A.BHUVANESWARI M.E.,Ph.D.,**
PROFESSOR,
Dept. of Information Technology
Adhiparasakthi Engineering College,
Melmaruvathur-603 319.

# CERTIFICATION OF EVALUATION

College Code/Name: 4204/Adhiparasakthi Engineering College

Branch / Semester : Information Technology / VI

| S.No | Name of the Student & Register Number | Title of the project | Name of the Supervisor with Designation |
|---|---|---|---|
| 1. | DEENADHAYALAN. V (420419205003) | CRIME DATA ANALYSIS | A.BHUVANESWARI ME.,Ph.D., PROFESSOR OF THE DEPARTMENT, Department of IT, Adhiparasakthi Engineering College, Melmaruvathur-603 319 |
| 2. | VIKRAM. P (420419205020) | | |

The report of the project works submitted by the above students in the partial fulfillment for the award of Bachelor of Technology degree in Information Technology of Anna University were evaluated and confirmed to be reports of work done by the above students and then evaluated

Submitted for the Project Work and Viva -voice examination held on _____

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

## ACKNOWLEDGEMENT

It is indeed a great pleasure and proud privilege to acknowledge the help and support we received from the positive minds around us in making this endeavor a successful one. The spiritual blessings of His Holiness **ARULTHIRU AMMA** and the devout guidance **THIRUMATHI AMMA** have undoubtedly taken us to the path of victory in completing this project.

The infrastructural support with all kinds of lab facilities have been a motivating factor in our completion of project work, all because of our **Correspondent Sakthi Dr. G. B. SENTHIL KUMAR** with great pleasure we take this opportunity to thank him.

From the academic side the support from **Dr. J. RAJA M.E., Ph.D., Principal** has encouraged us to work hard and attain this goal of completing the project.

We thank our **Head of the Department Dr.A.BHUVANESHWARI M.E., Ph.D.,** who has given us both moral and technical support adding experience to the job we have under taken.

With great pleasure we thank our **Supervisor Dr.A.BHUVANESHWARI M.E.,PhD.,** and **Project Coordinator, Ms.A.JAYANTHI M.E.,** Department of Information Technology who guided and helped us in crossing obstacles in the path to our glory.

We take immense pleasure in thanking our parents who gave us moral support to complete this project.

We also thank other staff members, and non-teaching staff members of computer Center and library, who have given their constant support and motivation in our endeavors.

# ABSTRACT

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. Our system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. Here we have an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc we are focusing mainly on crime factors of each day.

Data mining and machine learning have become a vital part of crime detection and prevention. We implemented the Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features, on the communities and crime dataset. Overall, the linear regression algorithm performed the best among the three selected algorithms. The scope of this project is to prove how effective and accurate the machine learning algorithms used in data mining analysis can be at predicting violent crime patterns.

# TABLE OF CONTENTS

# LIST OF FIGURES



.

# CHAPTER 1

# INTRODUCTION

## 1.1 GENERAL

Day by day the crime rate is increasing considerably. Crime cannot be predicted since it is neither systematic nor random. Also the modern technologies and hi-tech methods help criminals in achieving their misdeeds. According to rime Records Bureau crimes like burglary, arson etc have been decreased while crimes like murder have been increased. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence. The predicted results cannot be assured of 100% accuracy but the results shows that our application helps in reducing crime rate to a certain extent by providing security in crime sensitive areas. So for building such a powerful crime analytics tool we have to collect crime records and evaluate it.

In this research, we implemented the Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features, on the communities and crime un Normalized dataset to conduct a comparative study between the violent crime patterns from this particular dataset and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com. Some of the statistical data that was provided by neighborhoodscout.com such as the population of Mississippi, population distribution by age, number of violent crimes committed, and the rate of those crimes Per 100K people in the population are also features that have been incorporated into the test data to conduct analysis.

## 1.2 PROBLEM STATEMENT

Criminals are nuisance for the society in all corners of world for a long time now and measures are required to eradicate crimes from our world. Our mission is to offer crime prevention application to keep public safe. Current policing strategies work towards finding the criminals, basically after the crime has occurred. But, with the help of technological advancement, we can use historic crime data to recognize crime patterns and use these patterns to predict crimes beforehand. We are using clustering algorithms to predict crime prone areas.

Crime Pattern detection, analysis and prediction system is mainly based on Data Mining concepts and also implements various machine learning algorithms. The focus of the system is to analyse the existing dataset related to crime records in various areas and to predict the possible type of crime that may happen at various locations.

## 1.3 OBJECTIVE OF THE PROJECT

- The main objective of the project is analyze the crime rate to be happened in future. Based on this Information the officials can take charge and try to reduce the crime rate.

- The concept of Multi Linear Regression is used for predicting the graph between the Types of Crimes (Independent Variable) and the Year (Dependent Variable).

- The system will look at how to convert crime information into a regression problem, so that it will help detectives in solving crimes faster.

- Crime analysis based on available information to extract crime patterns. Using various multi linear regression techniques, frequency of occurring crime can be predicted based on territorial distribution of existing data and Crime recognition.

# CHAPTER 2

# LITRATURE SURVEY

There are various papers which contributed to the study of sentimental classification of citations. Based on the study of these papers, this project was proposed.

## CLUSTER ANALYSIS FOR ANAMOL DETECTION IN DATA

**Summary: Proposed by Sutapat Thirprungsri**

The purpose of this study is to examine the possibility of using clustering technology for continuous auditing. Automating fraud filtering can be of great value to preventive continuous audits. In this paper, cluster-based outliers help auditors focus their efforts when evaluating group life insurance claims. Claims with similar characteristics have been grouped together and those clusters with small population have been flagged for further investigations. Some dominant characteristics of those clusters are, for example, having large beneficiary payment, having huge interest amount and having been submitted long time before getting paid. This study examines the application of cluster analysis in accounting domain. The results provide a guideline and evidence for the potential application of this technique in the field of audit.

## ANALYZING CRIMINAL BEHAVIOUR BY SIMULATION MODEL

**Summary: Proposed by K. Zakhir Hussain**

Crime analysis, a part of criminology, is a task that includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Identifying crime characteristics is the first step for developing further analysis. The knowledge that is gained from data mining approaches is a very useful tool which can help and support in identifying violent criminal behavior. The idea here is to try to capture years of human experience into computer models via data mining and by designing a simulation model.

## AN INTELLIGENT ANALYSIS OF A CITY CRIME DATA

**Summary:**

Crime had been an enormous increase in the crime in the recent past. The concern about national security has increased significantly since the 26/11 attacks at Mumbai, India. However, information and technology overload hinders the effective analysis of criminal and terrorist activities. Crime deterrence has become an upheaval task. The cops in their role to catch criminals are required to remain convincingly ahead in the eternal race between law breakers and law enforcers. Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating such problem. In this paper we use a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyse the city crime data from Police Department.

## DATA CRIME APPROACHES TO CRIMINAL CAREER ANALYSIS

Summary:

Narrative reports and criminal records are stored digitally across individual police departments, enabling the collection of this data to compile a nation-wide database of criminals and the crimes they committed. The compilation of this data through the last years presents new possibilities of analysing criminal activity through time. Augmenting the traditional, more socially oriented, approach of behavioural study of these criminals and traditional statistics, data mining methods like clustering and prediction enable police forces to get a clearer picture of criminal careers.. Four important factors play a role in the analysis of criminal careers: crime nature, frequency, duration and severity. This method yields a visual clustering of these criminal careers and enables the identification of classes of criminals. The proposed method allows for several user-denied parameters.

# CHAPTER 3

# METHODOLOGY

## 3.1 Machine learning Algorithm

Machine learning is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

### Supervised learning:

In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

### Unsupervised learning:

This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

### Semi-supervised learning:

This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

**Reinforcement learning**:

Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

## 3.2 ALGORITHM ILLUSTRATION PROCESS

- Collecting the data because of machines initially learn from the data you give them.
- After you have your data, you have to prepare it.
- A machine learning model determines the output you get after running a machine learning algorithm on the collected data.
- Training the data to the machines, it is the most important in machine learning.
- After training the model, you have to check to see how it's performing.
- In the end, you can see your model on unseen data to make predictions accurately.

## 3.3 Proposed System

In the proposed system, we done crime data analysis of with many parameters and factors including daily arrests, monthly arrests, number of domestic violence, top 5 monthly, weekly and daily crime are visualized.

Crime pattern analysis system is mainly based on data mining concepts and also implements various machine learning algorithms. The focus of the system is to analyse the existing dataset related to crime records in various areas and to predict the possible type of crime that may happen at various location.

## 3.4 System Architecture

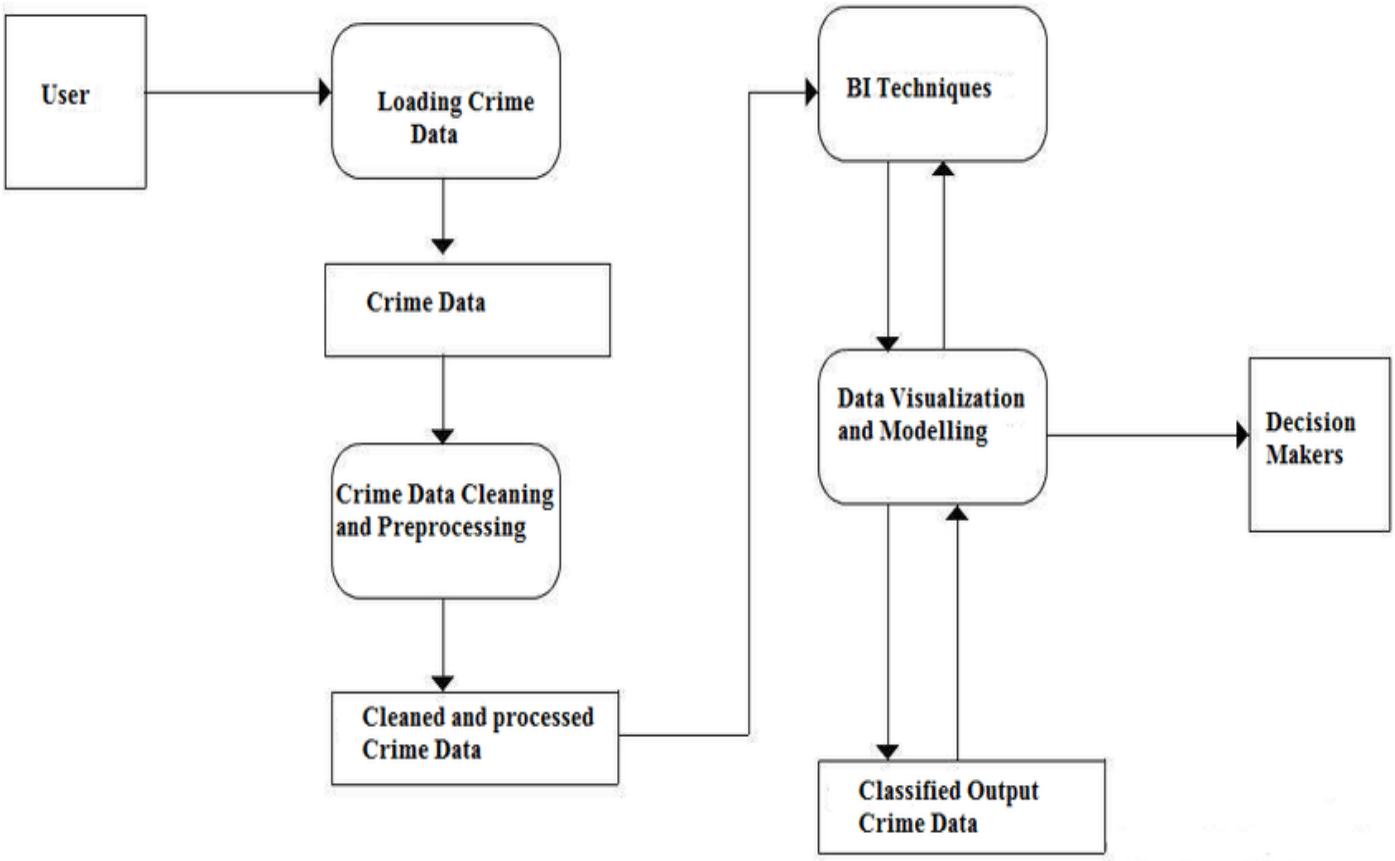


**Figure 3.4.1: Data Flow Diagram**

Intialy the user should upold the crime data set. Then the data set was cleaned and preprocessed. Lastly it gives the vishual representation by using the classification algorithm, it will usefull for decision makers.

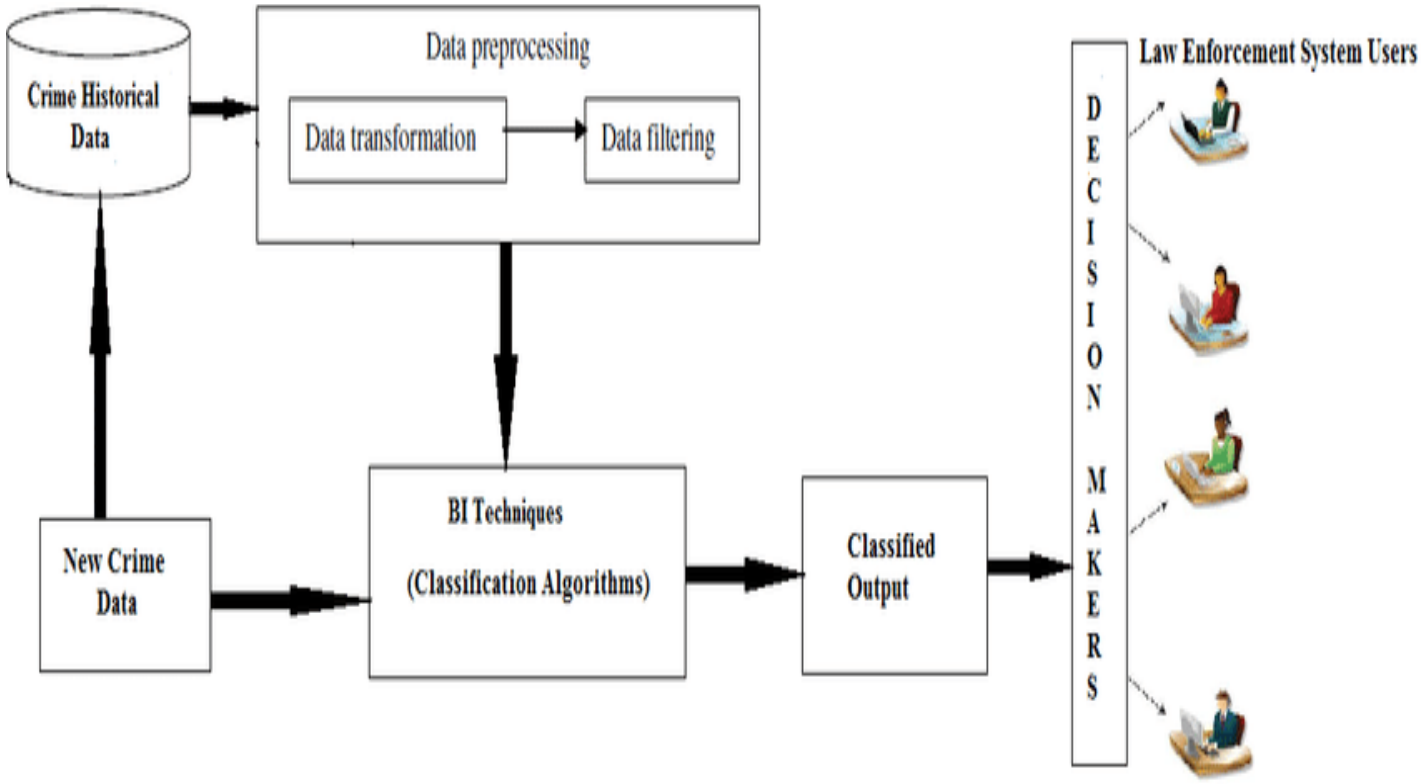


**Figure 3.4.2: Block Diagram**

New crime data will goes to crime data base and it goes classification algorithm. Historical data will preprocessed and it will also goes to classification algorithm. Then it will give the classified output.

# CHAPTER 4

# MODULES USED IN PROJECT

## 4.1 Seaborn

Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

## 4.2 NumPy

NumPy is one of Python's most essential libraries, and it's also one of the most helpful. NumPy is capable of handling big datasets with ease. I can almost see your eyes glinting with excitement at the possibility of mastering NumPy. As data scientists or aspirant data scientists, we must have a strong understanding of NumPy and how it works in Python.

## 4.3 Pandas

Pandas is an open-source library designed to make it simple and natural to work with relational or labelled data. It includes a number of data structures and methods for working with numerical data and time series. This library is based on the NumPy Python library. Pandas is quick and has a high level of performance and productivity for its users .

## 4.4 Matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to matlab.

# CHAPTER 5

# SYSTEM SPECIFICATION

## 5.1 TOOLS

- Python IDLE.
- Tools for implementing as GUI.
- Matplotlib, Pandas, Numpy and Seaborn module.

## 5.2 SUPPORTED OPERTING SYSTEM

- Windows XP (32 bit or 64 bit) or Windows 7(32 bit or 64 bit).
- Linux (Tested on unbound Linux, lucid Linux).
- MAC as x10.5.8 later(x86 only).

## 5.3 DEVELOPING ENVIRONMENT

- Python 3.6/3.7.
- Google Colab.

## 5.4 SYSTEM CONFIGURATION:

**Software Requirements:**

- Operating System : Windows 7 or higher
- Programming : Python 3.6 and related libraries

**Hardware Requirements:**

- Processor : Any Processor above 500 MHz.
- Ram : 4 GB
- Hard Disk : 4 GB
- Input device : Standard Keyboard and Mouse.
- Output device : VGA and High Resolution Monitor.

# CHAPTER 6

# IMPLEMENTSTION

## 6.1 Crime Data Analysis

During crime analysis, the input data is important to be used In training and testing process. The training of the process is used to conduct the crime model and the testing of the process is used to validate the algorithm. Input data can be obtained from the criminal records with the help of the Government, agencies, etc. As a consequence, the collected data is large volumes of data and it is also in the unstructured Data formats. The collected data is stored into different Databases. The many researches are concerning to solve the problem of handling such a huge data. Hence to overcome such a difficulty, of accessing the knowledge from large volumes of data, several other methods may be useful to integrate data.

## 6.2 Data Collection

In data collection step we are collecting data from different websites like blogs, news sites, social media, etc . The data information collected comprises of name, address etc. And this collected data is stored into database for further process. Crime data is an structured data therefore the no of field, content, and size of the document will be same due to which it will greatly improve the efficiency .First we collect the crime data's and verify the features of attributes. For each arrival of crime incidents in the datasets, the upcoming featured attributes are included:

- Category- Type of the crime.
- Descript – A short note describing details of the crime;
- Dates – Date and time of the crime incident;
- X – Longitude of the location of a crime;
- Y – Latitude of the location of a crime;
- Year – which year the case happened;

- Coordinate – Pairs of Longitude and Latitude.

## 6.3 Data Manipulation

After Data Collection we are performing Data manipulation where we are going to add the information of the criminal as the new entry which will then automatically will be stored in the database and will be used for future purpose as well as we can delete the data stored in it . And Foremost we can sort & search the criminal data according the parameters or in the alphabetical order.

## 6.4 Data Pre-processing

It is the cleaning process of tangled data sets for doing Operations and analysis. Before initiating any algorithms and Operations on our datasets, a series of preprocessing steps are Performed for data conditioning as presented below:

- For some missing coordinate attributes in Indian Datasets, assigned uncertain values representative from The non-missing values, computed their mean, and then Replaced the missing ones.
- We also avoid few features that needless.

## 6.5 Clustering of Dataset

Clustering is type of unsupervised learning. It groups the similar data items into clusters. The purpose of crime data clustering is to find a set of similar crime incidents based on an offender's behavioural traits or to find a geographical area with a high concentration of certain types of crimes. The system uses k-means algorithm to group the data into various clusters that will be useful for prediction of crime.

## 6.6 Narrative visualization

After the Data collection and Featured Attributes we Perform data preprocessing and negotiating. Take brief summary of the data frame; It Helps obtain a quick overview of the data set. The Summary includes a list of all columns with their data Types and the number of non null values in each column. And since our values are not null we don't have to fill the Missing values.

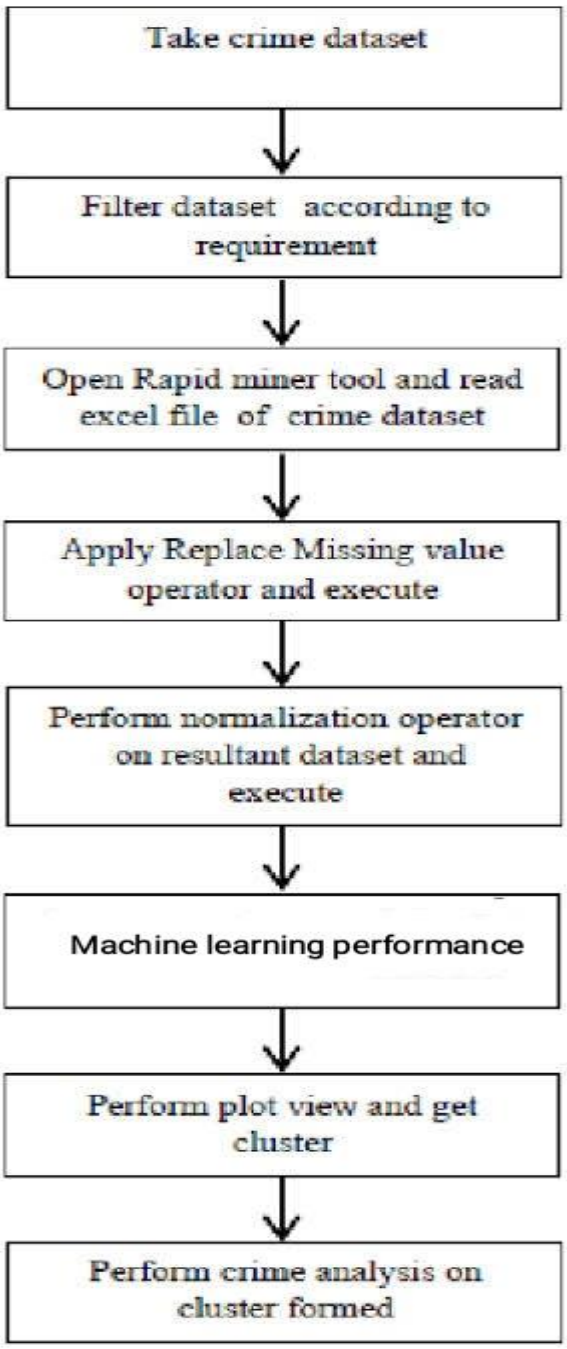


**Figure 6.6.1: Crime Data Analysis**

The flow chart defines the each level of the process. The given data set in evaluate with thease process.

# CHAPTER 7

# DESIGN

## 7.1 UML DIAGRAMS

The design is a plan or drawing produced to show the look and function or workings of an object before it is made. Unified Modeling language (UML) is a standardized modeling language enabling developers to specify, visualize, construct and document artifacts of a software system. Thus, UML makes these artifacts scalable, secure and robust in execution. UML is an important aspect involved in object-oriented software development. It uses graphic notation to create visual models of software systems.

- Use Case Diagram
- Activity Diagram
- Sequence Diagram
- Collaboration Diagram
- Deployment Diagram
- Class Diagram

### 7.1.1 Use case Diagram

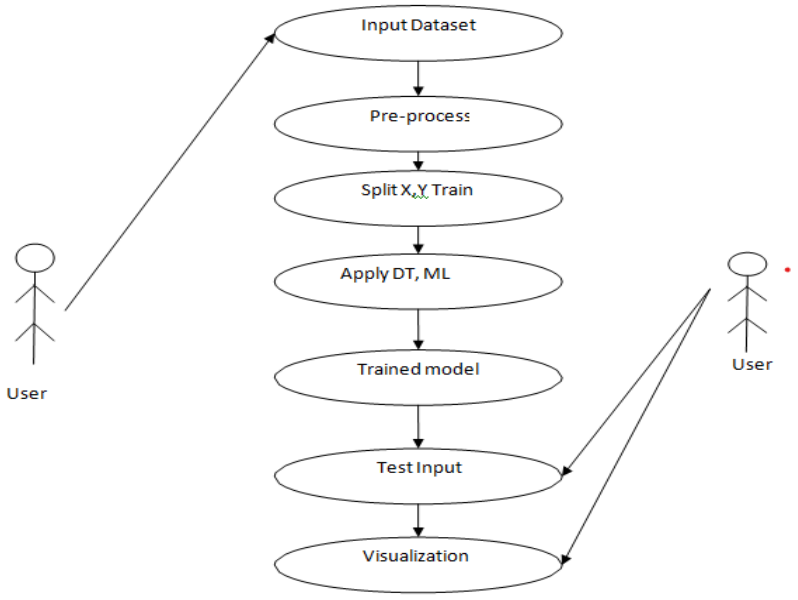


**Figure7.1.1:Usecase Diagram**

The above figure represent use case diagram of proposed system, where user inputs dataset, we pre-process dataset, the algorithm Decision Tree and K-means clustering to generate the trained model to predict the crime type. The actor and use case is represented. An eclipse shape represents the use case namely input image, pre-process, Split features, prediction and output

## 7.1.2 Activity Diagram

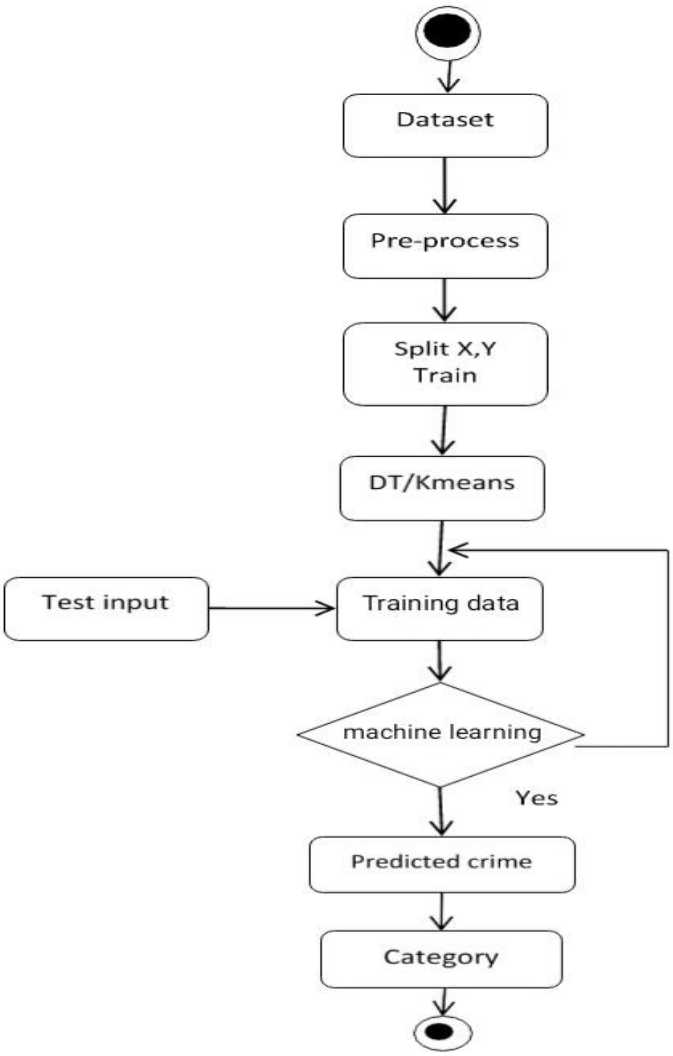


**Figure 7.1.2: Activity Diagram**

The above figure show the activity diagram of the proposed system, where we represented theidentified activities and its functional flow.

## 7.1.3 Sequence Diagram

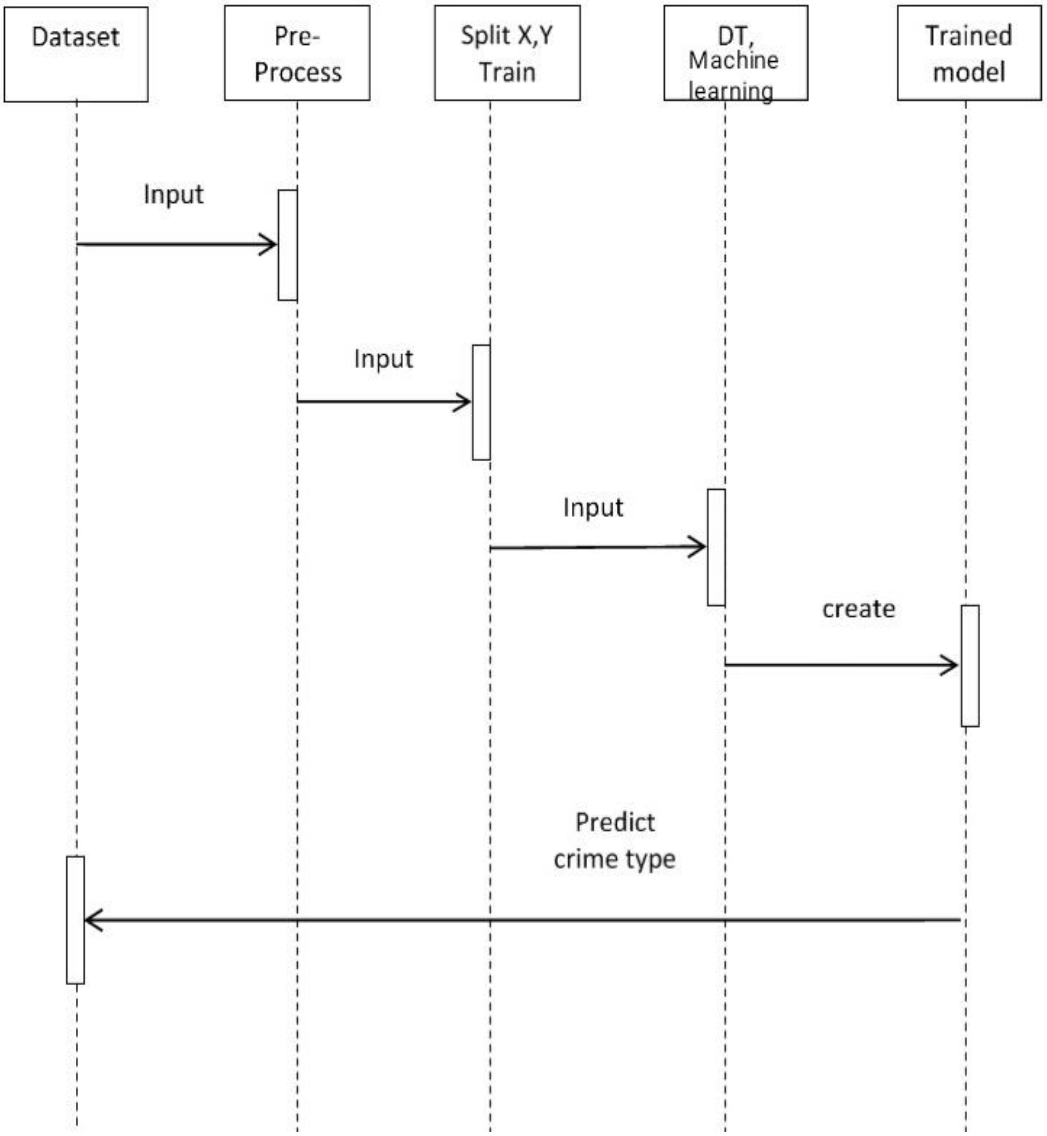


**Figure 7.1.3: Sequence Diagram**

A sequence diagram shows a parallel vertical lines, different processes or objects that live simultaneously, and as horizontal arrows, the messages exchanged between them, in order in which they occur. The above figure represents sequence diagram, the proposed system's sequence of data flow is represented.

## 7.1.4 Collaboration Diagram

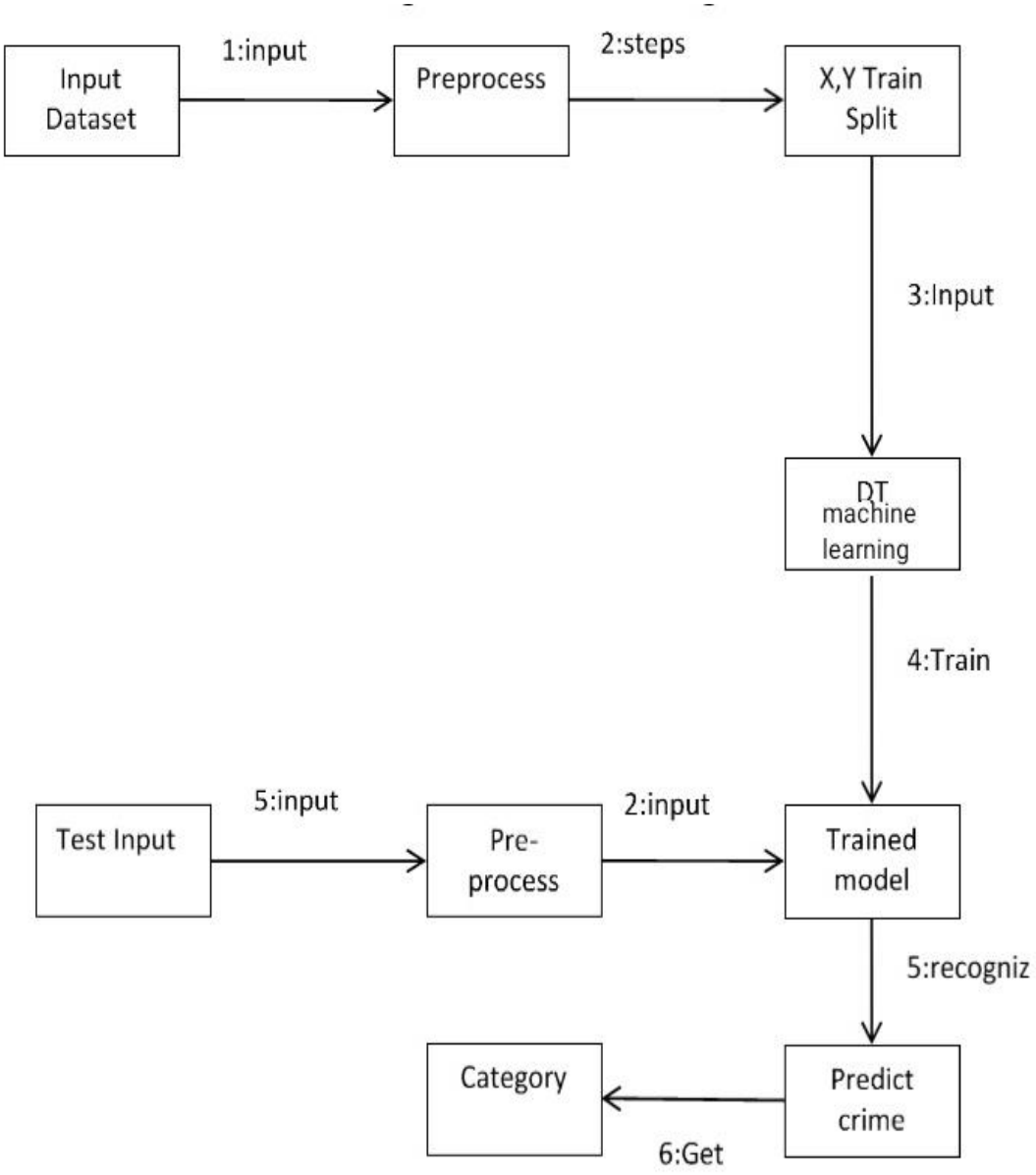


**Figure 7.1.4: Collaboration Diagram**

The above figure shows the collaboration diagram of the proposed system, where we represented thecollaboration between the actor and function modules with sequence number.

## 7.1.5 Deployment Diagram

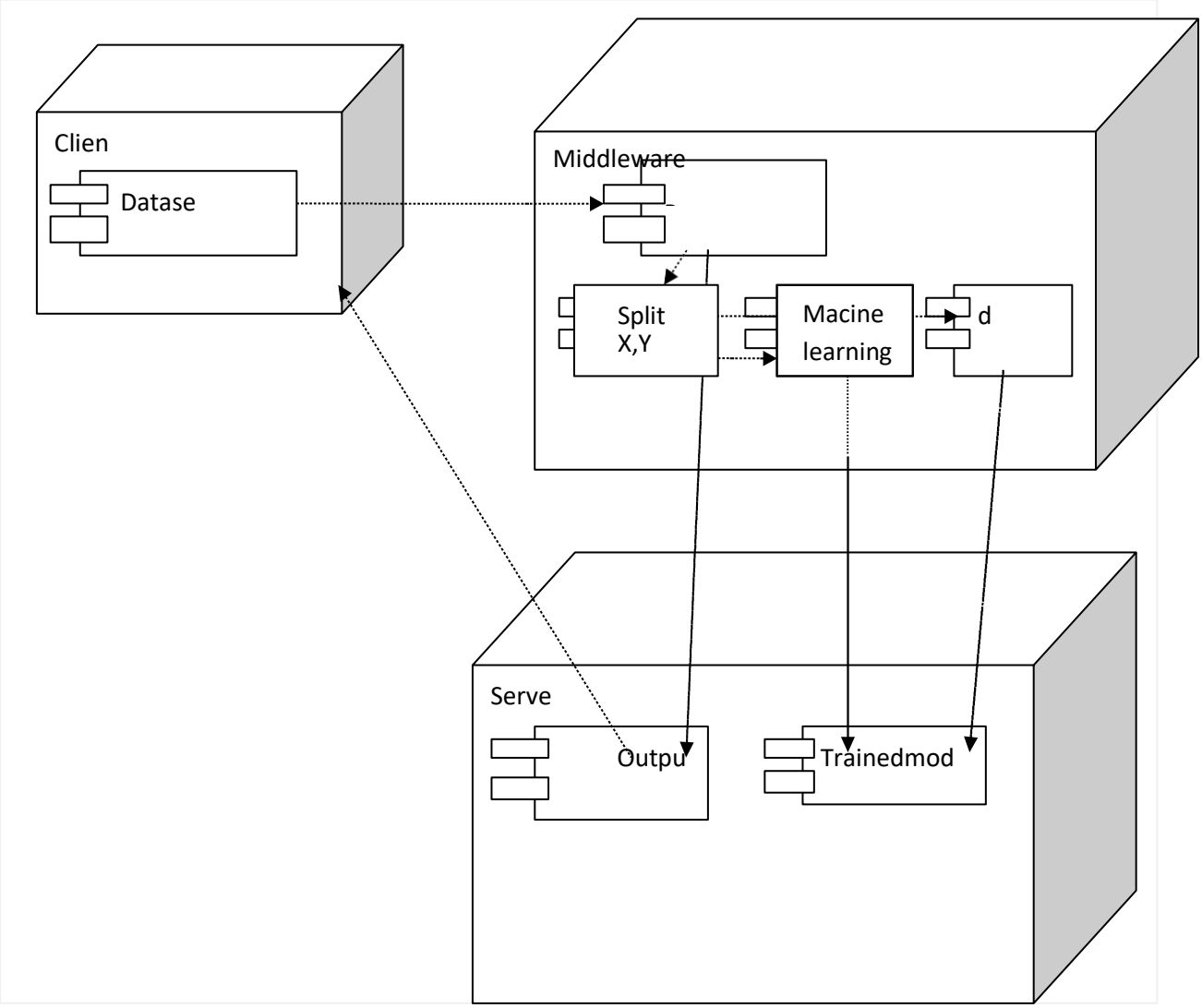


**Figure: Deployment Diagram**

In the deployment diagram the UML models the physical deployment of artifacts on nodes. The nodes appear as boxes, and the artifacts allocated to each node appear as rectangles within the boxes. Nodes may have sub nodes, which appear as nested boxes. A si7ngle node in a deployment diagram may conceptually represent multiple physical nodes, such as a cluster of database servers.

## 7.1.6 Class Diagram

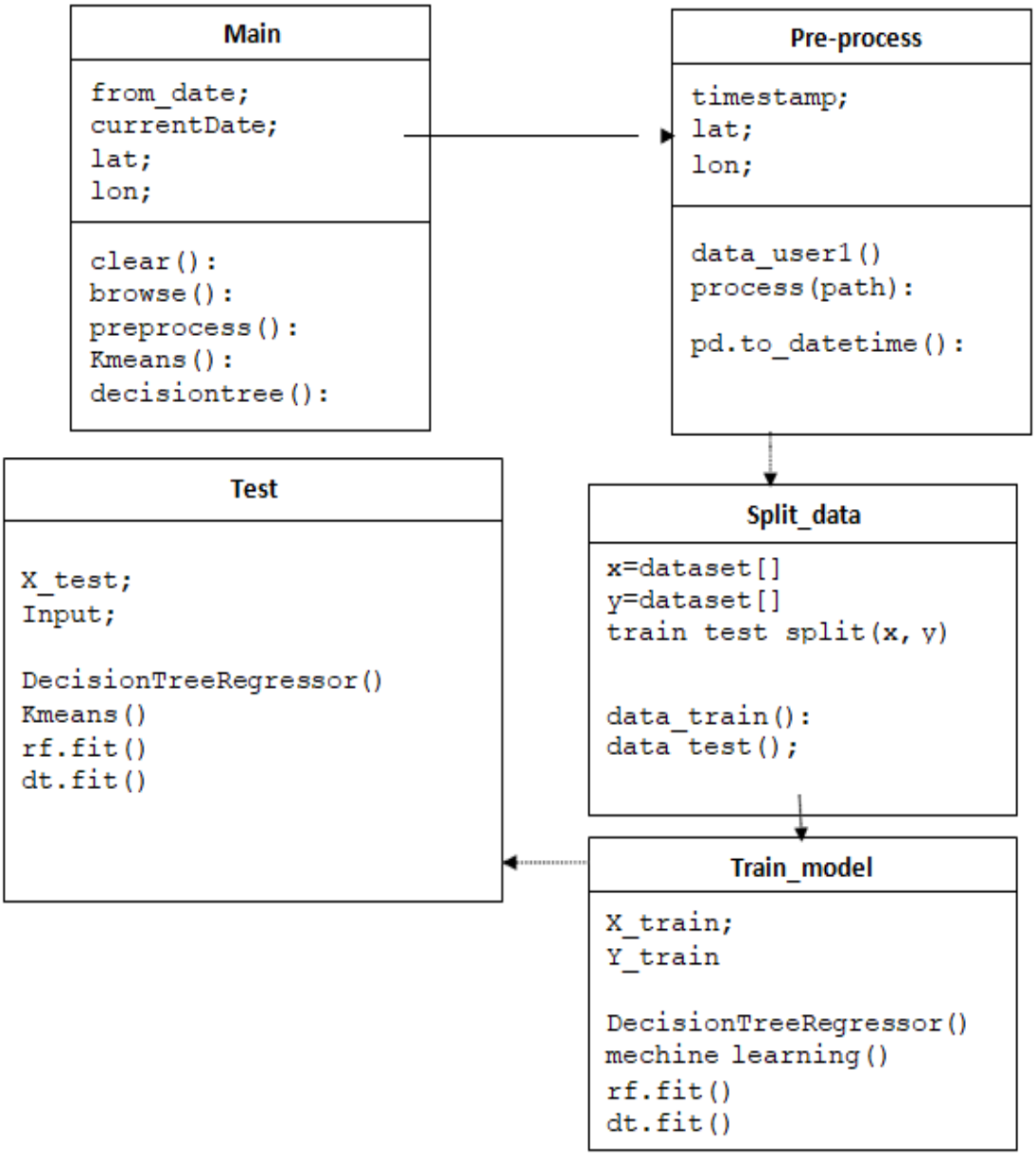


**Figure 7.1.6: Class Diagram**

The class diagram explains about the properties and functions of each class. The classesare Main, pre-process, split data, train and test. In the above diagram, every class is represented with attributes and operations.

# CHAPTER 8

# SOURCE CODE AND OUTPUTS

## 8.1 Source Code

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sb


#Read CSV file

data_frame=pd.DataFrame(pd.read_csv('/content/area_classification.csv'))


#Count of null value in each column

data_frame.isnull().sum()


#average of crime count

data_frame['NoOfCrimes'].mean()


# Rekation between each columns

data_frame.corr()
```

```
#Visualization of levels and counts
sb.countplot(data = data_frame, x = 'Class', color = 'lightblue')
sorted_counts = data_frame['Level'].value_counts()
plt.pie(sorted_counts, labels = sorted_counts.index, wedgeprops = {'width' : 0.2});
plt.axis('square');


#Date wise crimes
base_color = sb.color_palette()[5]
sb.countplot(data = data_frame, y = 'Level', color = base_color)
date=data_frame.loc[:,"Date"]
NoOfCrimes=data_frame.loc[:,"NoOfCrimes"]
plt.plot(date,NoOfCrimes)
plt.xlabel('date')
plt.ylabel('No Of Crimes')


#Area wise Classification
NoOfCrimes=data_frame.loc[:,"NoOfCrimes"]
Area=data_frame.loc[:,"Area/Cluster"]
NoOfCrimes=NoOfCrimes[:10]
Area=Area[:10]
```

```
plt.plot(Area,NoOfCrimes)

plt.xlabel('Area')

plt.ylabel('No Of Crimes')
```

## 8.2 SCREENSHOT

```
[ ] data_frame.isnull().sum()

Area/Cluster    0
NoOfCrimes      0
Class           0
Level           0
Date            0
dtype: int64
```

**Figure 8.2.1: Null Value Check**

The abouve figure shows with “data_frame.isnull” is used to find the null values in the given dataset. It it checks the each column wise and show the result in table formation.

```
[ ] data_frame['NoOfCrimes'].mean()

28392.62
```

```
data_frame.corr()
```

| | Area/Cluster | NoOfCrimes | Class |
|---|---|---|---|
| **Area/Cluster** | 1.000000 | -0.061727 | -0.119370 |
| **NoOfCrimes** | -0.061727 | 1.000000 | 0.903475 |
| **Class** | -0.119370 | 0.903475 | 1.000000 |

**Figure 8.2.2: Data Frame Mean and Corrletion**

data_frame .mean is used to find the average value of No of Crimes data_frame collrlation is used

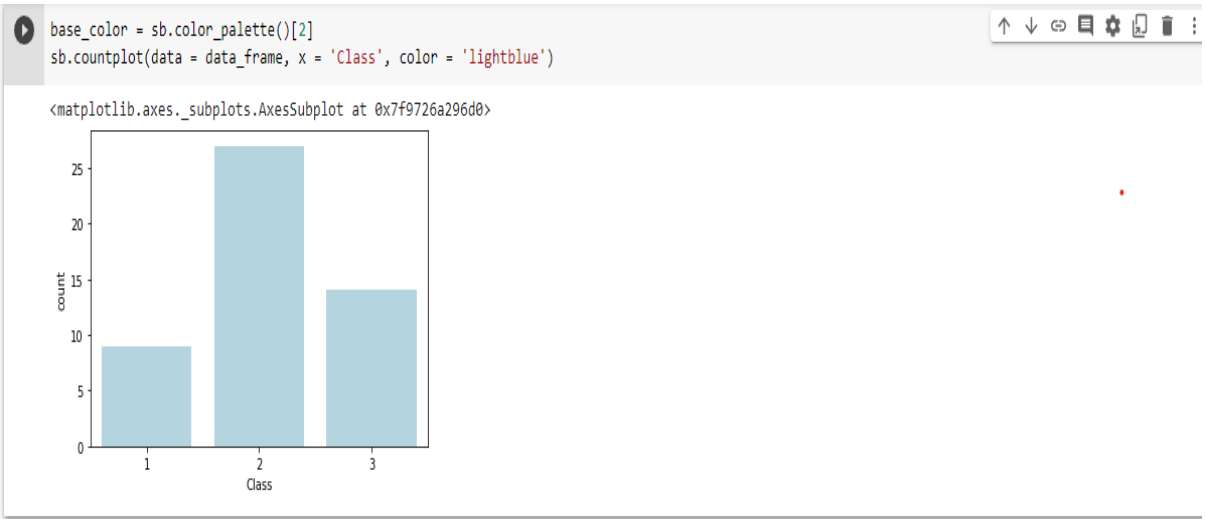


**Figure8.2.3:Graph Of Class and Crime Count**

The graphical representation of class of the crime and crime count denoted by this graph with use of "sb.countplot".


**Figure 8.2.4: Circle Visualization of Crime Level**

The circle representation of crime leve . It is easy to vishualise the crime data.


**Figure 8.2.5:Relation of Crime Level and Count**

The relation between crime level and crime count will be showed in green colour grapgical representation.

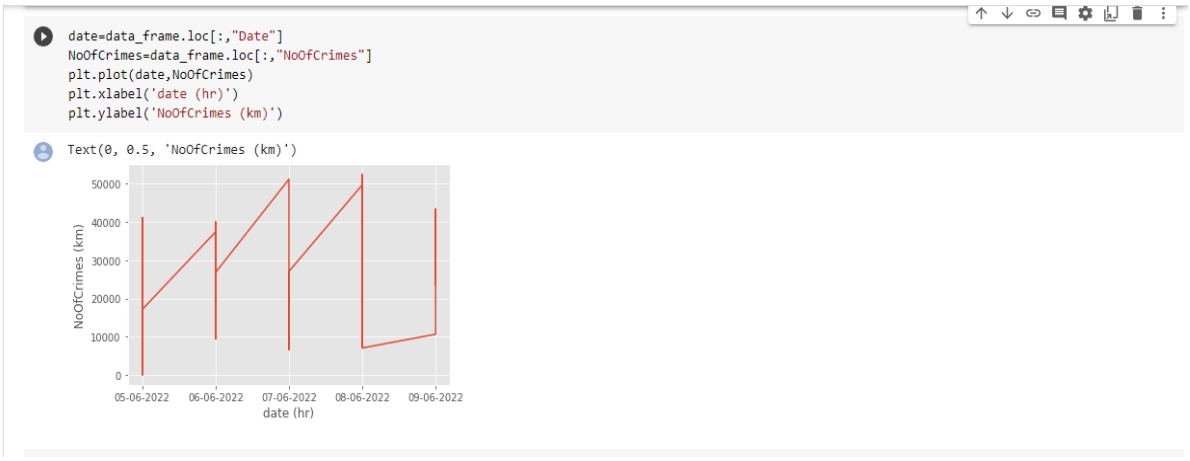


**Figure 8.2.6: Date waise Representation of Crimes**

Date wise crime rate will analysied and showed the above figure. With the use of given data set.

# CHAPTER 9

# CONCLUSION

We have examined the accuracy of class and prediction based totally on different check sets. Classification is done based on the Bayes theorem which showed more than 90% accuracy. Using this algorithm we trained numerous news articles and build a model. For testing we are inputting some test data into the model which shows better results. Our system takes elements attributes of an area and preprocessing offers the frequent patterns of that place. The pattern is used for constructing a model for decision tree. Corresponding to each place we build a model by training on these frequent patterns. Crime patterns cannot be static since patterns change over time. By training means we are teaching the system based on some particular inputs. So the machine automatically learns the converting patterns in crime through examining the crime patterns. Also the crime elements trade over time. By sifting through the crime data we have to identify new factors that lead to crime. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction we have to find more crime attributes of places instead of fixing certain attributes. Till now we trained our system using certain attributes but we are planning to include more factors to improve accuracy. Our software predicts crime prone regions in India on a particular day. It will be more accurate if we consider a particular state/region. Also another problem is that we are not predicting the time in which the crime is happening. Since time is an important factor in crime we have to predict not only the crime prone regions but also the proper time.

# REFERENCE


- De Bruin ,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J. and Kok,J.N(2006) Data mining approaches to criminal carrer analysis ,"in Proceedings of the Sixth International Conference on Data Mining (ICDM"06) ,Pp. 171-177.

- Manish and M. P. GuptaGupta1, B.Chandra1 1,200Information System.

- Nazlena Mohamad Ali1, Masnizah Mohd2, Hyowon Lee3, Alan F. Smeaton3, Fabio Crestani4 and Shahrul Azman Mohd Noah2 ,2010 Visual Interactive Malaysia Crime News Retrieval System 7 Crime Data Mining for Indian Police.

- Chung-Hsien Yu, Max W.Ward, Melissa Morabito and Wei Ding,"Crime Forecasting Using Data Mining Techniques", 2011 11th IEEE International Conference on Data Mining Workshops.

- Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Detecting patterns of crime with series finder. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), 2013.