



# **Walchand College Of Engineering, Sangli.**

(An Autonomous Institute)

Department Of  
Information Technology

Project Report  
on

## **Police Companion**

*Submitted by*

**Mr. Vinayak Ukkalgaonkar**

**Mr. Tushar Phadatare**

**Mr. Bhushan Deshpande**

**Mr. Rohit Yadav**

**2013BIT065**

**2017BTEIT00002**

**2017BTEIT00042**

**2017BTEIT00038**

Under the Guidance  
Of

**Prof. T. A. Mulla**

Information Technology Department,  
WCE, Sangli

**Year:2020-2021**



## **Walchand College Of Engineering, Sangli.**

(An Autonomous Institute)

### **Department Of Information Technology**

## **Certificate**

This is to certify that the Project Report entitled, "Police Companion" submitted by

**Mr. Vinayak Ukkalgaonkar**

**2013BIT065**

**Mr. Tushar Phadatare**

**2017BTEIT00002**

**Mr. Bhushan Deshpande**

**2017BTEIT00042**

**Mr. Rohit Yadav**

**2017BTEIT00038**

to Walchand College of Engineering, Sangli, India, is a record of bonafide Project work carried out by them under my supervision and guidance and is worthy of consideration for the award of the degree of Bachelor of Technology in Information Technology of the Institute.

**Prof. T. A. Mulla**

Guide

Information Technology Dept.

WCE, Sangli

**Dr. A. J. Umbarkar**

Head Of Department

Information Technology Dept.

WCE, Sangli

## **Acknowledgement**

We feel immense pleasure in submitting this Project report entitled "Police Companion".

We are thankful to our guide Prof. T. A. Mulla for their valuable guidance and kind help during completion of Project and feel great to express our sincere gratitude to other all staff members of IT Department.

We are also thankful to the Head of the 'Department of Information Technology' Dr. A. J. Umbarkar for their valuable guidance during the completion of Project. We would like to thank all faculty members and staff of Department of Information Technology for their generous help in various ways for the completion of this thesis.

We would like to thank all our friends and especially our classmates for all the thoughtful and mind stimulating discussions we had, which prompted us to think beyond the obvious. we have enjoyed their companionship so much during our stay at WCE, Sangli.

## **Declaration**

We hereby declare that work presented in this project report titled "Police Companion" submitted by us in the partial fulfillment of the requirement of the award of the degree of **Bachelor of Technology (B.Tech)** Submitted in the **Department of Information Technology, Walchand College of Engineering, Sangli**, is an authentic record of my project work carried out under the guidance of Prof. T. A. Mulla.

**Mr. Vinayak Ukkalgaonkar**

**2013BIT065**

**Mr. Tushar Phadatare**

**2017BTEIT00002**

**Mr. Bhushan Deshpande**

**2017BTEIT00042**

**Mr. Rohit Yadav**

**2017BTEIT00038**

Date: 15/06/2021

Place: Sangli

Signature

## **ABSTRACT**

'Police Companion' is an AI/ML driven project where crime analysis and prediction are done by identifying and analyzing the patterns or trends of crime happenings present in the Police department record. The approach is to make modern technology and administration go hand-in-hand. The methodological and systematic approach will make the task of the enforcement agencies much easier. For prevention from any danger, be it human induced or nature driven, it is always important to get prepared beforehand. And for prevention, the trend and pattern analysis of available data is crucial thing. In overview approach, we are going to use different machine learning algorithms to train and classify the dataset. Following the process of analysis using various algorithms, our aim is to predict two things:

- Location of crime and timeline of crime happenings as an input and Class of crime enlisted in the dataset as an output.
- Class of crime and timeline of crime happenings as an input and location of crime as an output.

To sum up, 'Police Companion' is about applying AI/ML based technology to analyze the current crime data and based on it to predict (to prevent) the future possibility of crime that can happen and its relative position.

## Contents

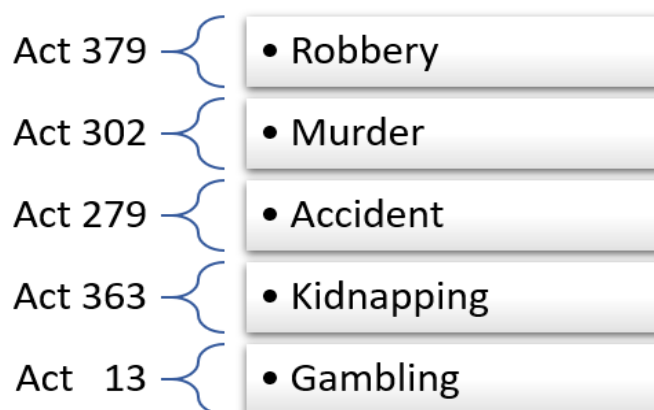
<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>2</b>	<b>Objectives</b>	<b>9</b>
<b>3</b>	<b>Methodology and Implementation</b>	<b>10</b>
3.1	Flow Diagram.....	10
3.2	Data Collection .....	10
3.2.1	Data Pre-processing .....	11
3.3	Analytical Visualization.....	14
3.4	Prediction: Application of algorithms.....	20
3.4.1	Prediction of type of crime.....	20
3.4.2	Prediction of location of crime.....	20
3.5	Accuracy: .....	26
3.5.1	Crime Prediction: .....	26
3.5.2	Location Prediction: .....	26
<b>4</b>	<b>Case Diagrams</b>	<b>28</b>
4.1	Activity Diagram.....	28
4.2	Sequence Diagram .....	29
<b>5</b>	<b>Tools</b>	<b>30</b>
5.1	Software Details: .....	30
5.2	Hardware Details: .....	30
<b>6</b>	<b>Screenshots</b>	<b>31</b>
<b>7</b>	<b>Future Scope</b>	<b>35</b>
<b>8</b>	<b>References</b>	<b>36</b>

# 1 Introduction

Crime is nowadays one of the most dominating and biggest threats to society. Daily there are 'n' number of cases of crimes and the main task of enforcement agencies is to prevent their happening. This requires keeping a broad dataset which can be used for further reference by these agencies. Due to the advent of various technologies, the data has been shifted from register books to spread sheets, so it is quite of transferable nature now. The question here is, how various agencies can work upon it and use it for future reference i.e., for prediction. Keeping data in a systematic and classified form is very hectic task.

Though we cannot predict the exact suspect of crime, we can identify the location at which it is likely to happen. In a similar sense, we can also predict the type of crime that may happen in future. The another aspect of this is, we cannot predict such kind of things with 100 percent accuracy, but here is when machine learning algorithms come in. By studying various algorithms separately and doing their comparative analysis, we can calculate the maximum possibility with which a crime can happen and the relative location at which it can happen.

For reference, we have obtained the dataset from online sources. The types of crimes mentioned in the dataset are:



Also, the algorithm which we are using in our project are:

- KNN
- Random forest
- Decision trees
- Multioutput linear regression

So, the aim of our project is to apply these ML algorithms on the dataset that we have got from online resources to identify the pattern and trends of various attributes present therein and to develop such a platform- **"POLICE COMPANION"** which will help law enforcement agencies to mitigate the problems regarding crime prevention.



## **2 Objectives**

The main objectives of our project are:

- Applying machine learning regression algorithms like KNN, logistic regression, decision trees, random forest, and doing comparative analysis among these for maximum efficiency.
- Crime detection and prediction by analyzing features in standard dataset to help police dept. for maximum resource allocation.

## 3 Methodology and Implementation

### 3.1 Flow Diagram

This flow diagram depicts overall methodological approach of the project:

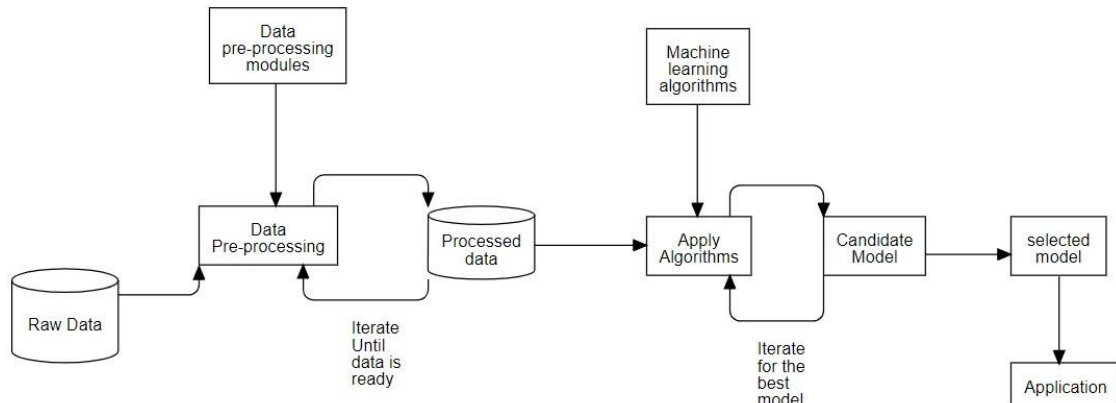


Figure 1:

### 3.2 Data Collection

As project will be working on any real-time data, as of now, we have taken the data set from online resources. It is a process in which we collect and measure information for various sources, but in order to make it sense to the machine learning process for which we are using it, it needs to be collected and stored in a proper way.

Collecting data gives you the record of past events and data analysis can be performed on it for predicting repetitive behavior. we build predictive models using machine learning algorithms which will help us to predict future trends. The goodness of predictive models is mostly dependent on the form in which we collect and store data. So, the usable data must be errorless i.e., there should be garbage value omissions. Hence, we perform pre-processing on the gathered dataset in order to make it free from unnecessary entities, attributes and to make suitable for the machine learning project on which work is to be done.

### **3.2.1 Data Pre-processing**

When we talk about data, we often see it is an aggregated interception of rows and columns, but actually it can be in variety of formats like- structured data, video, audio, etc. The point is machine does understand it in its original form. They are designed to operate in 0s and 1s. So, in machine learning process, data pre-processing means the process of transforming, encoding the available data in such a format which can be easily recognizable and operable by machine.

The steps that are involved in data pre-processing are:

1. Data Quality Assessment
2. Feature Aggregation
3. Feature Sampling
4. Dimensionality Reduction
5. Feature Encoding

1. **Data Quality Assessment:** Since the data is usually taken from multiple sources, it would be completely unrealistic to assume that it would be perfect. It has may have human errors or limitations in the data collection process. The prior problems and solutions are:
  - Missing values: this is very usual problem. It can happen during the data collection process. The possible solutions may be:
    - eliminate rows with missing data.
    - Fill them in with the mean, mode and median value of respective feature.
  - Inconsistent values: It is more often that data may have inconsistent values. Like address field combined with phone number. The possible solution is to performing data assessment like what type of the data of features should be and whether it is consistent or not.

- Duplicate values: The data may also contain duplicacy. it could be due to the reason that the same person submits the same form of data more than once. The remedy can be using deduplication method to avoid biasness.

2. **Feature Aggregation:** Feature aggregation is done in order to make data look in a better perspective. Say, daily transaction of sales is aggregated into monthly transactions or yearly transactions.

The advantages are:

- Less memory consumption and processing time.
- It provides high level view of the data is aggregated data is more stable than the individual ones.

3. **Feature Sampling:** Feature sampling means selecting a subset of a dataset we are analyzing. This will help us to overcome time and memory constraints. The condition for this is the sample should be representative of the original dataset. Means it should have the same properties of the original dataset. It requires choosing the correct sample size and sampling strategy.

This can be done in two ways:

- Sampling without Replacement: For each selected item, it is removed from the set of objects forming the original dataset.
- Sampling with Replacement: Items are not removed from the dataset as a whole i.e., they can get selected for more than once.

4. **Dimensionality Reduction:** Usually large datasets come with large number of features. These features are called as dimensions. These dimensions are in the form of geometric planes. More the features, more the dimensions, more the geometric planes. Hence it becomes difficult to visualize and model the data-set. Dimensionality reduction maps the data on the lower dimension space.

Some benefits of this technique are:

- Data analysis algorithm work better if the dimensionality is low because the irrelevant and useless noisy data has now been removed.

- Models which are designed over the top of the low dimensional space are more understandable.
5. **Feature Encoding:** The whole purpose of this is to make the data readable to the machine. Feature encoding is a process in which the data is transformed into such a form which is easily readable by the machine still holding its basic features.

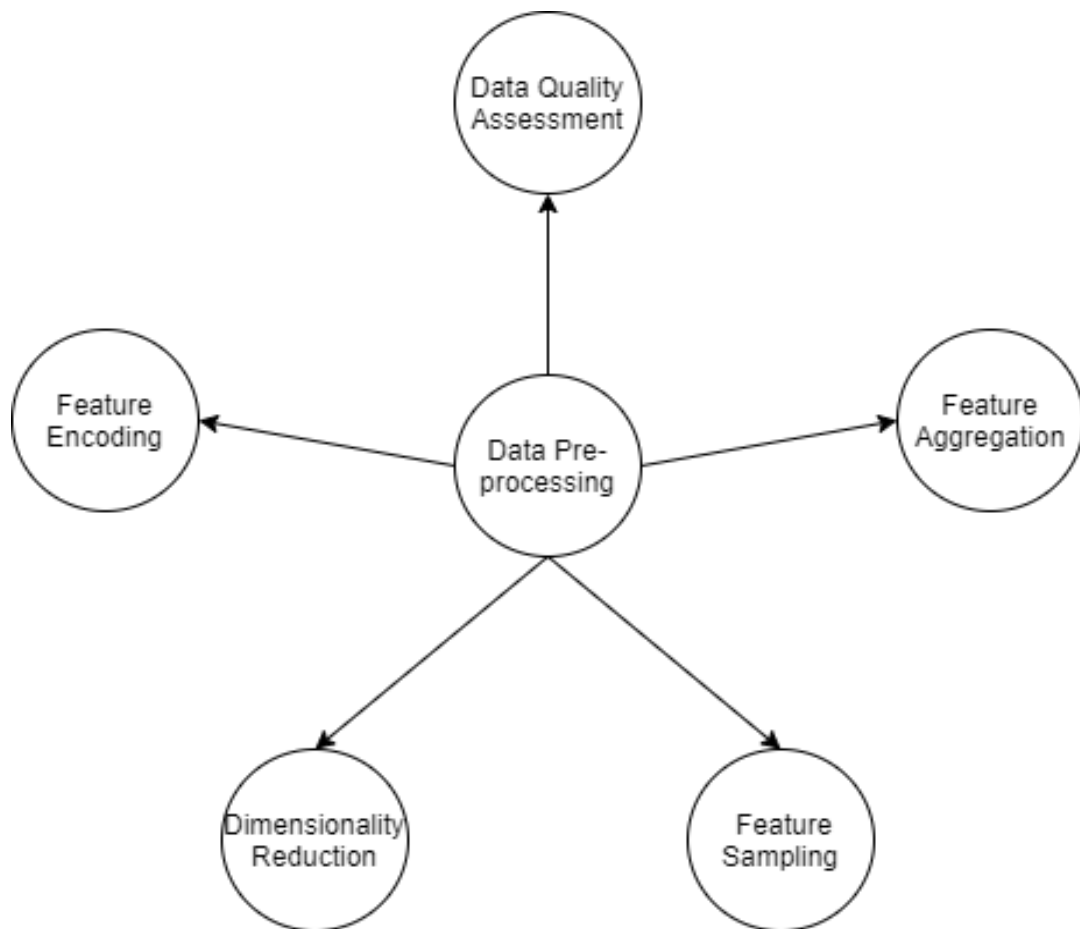


Figure 2:

Some of the techniques are: Nominal, Ordinal, Interval, Ratio.

These are the universal steps that are used in the pre-processing of data

depending upon the type of classification we want to perform. In our project, we have undertaken following steps to remove irrelevant entities.

- police station, Complainant name address, station number, accused name address have been removed.
- Resolution, Address and Description dropping: Description and Resolution of crime are only recorded after the commitment of crime. Hence, if we go onto practical approach to predict the crime, this data is irrelevant. Again, as we are using latitude and longitude as location ordinates, the address portion also becomes irrelevant.
- The timestamp originally in the format Year, date, time was decomposed into year, month, date, hour, minute.

After accomplishing pre-processing looks like this.

```
[ ] dataset.head()
```

	timestamp	Robbery	Gambling	Accident	Violence	Kidnapping	Murder	latitude	longitude
0	28-02-2018 21:00	1	0	0	0	0	0	22.737260	75.875987
1	28-02-2018 21:15	1	0	0	0	0	0	22.720992	75.876083
2	28-02-2018 10:15	0	0	1	0	0	0	22.736676	75.883168
3	28-02-2018 10:15	0	0	1	0	0	0	22.746527	75.887139
4	28-02-2018 10:30	0	0	1	0	0	0	22.769531	75.888772

**Fig 3**

### 3.3 Analytical Visualization

The content in the pre-processed data can be seen in various forms such as whisker plot, box-plots, pie charts, bar diagrams, curves, etc. with the help of various packages available in python.

**The libraries/packages used for this are:**

- **matplotlib.pyplot:** matplotlib is used to visualize data into structured, animated and interactive format. pyplot is a collection of functions that makes matplotlib work like MATLAB. Every pyplot feature makes changes in the figure such as creating the figure, plotting points into the figure and like that. The various plots we can visualize using pyplot are- histogram, scatter, polar, 3D plot, contour, image.
- **seaborn:** seaborn is a data visualization library which is closely integrated with pandas and served on the top of matplotlib. It is used to explore and understand the data in better form.

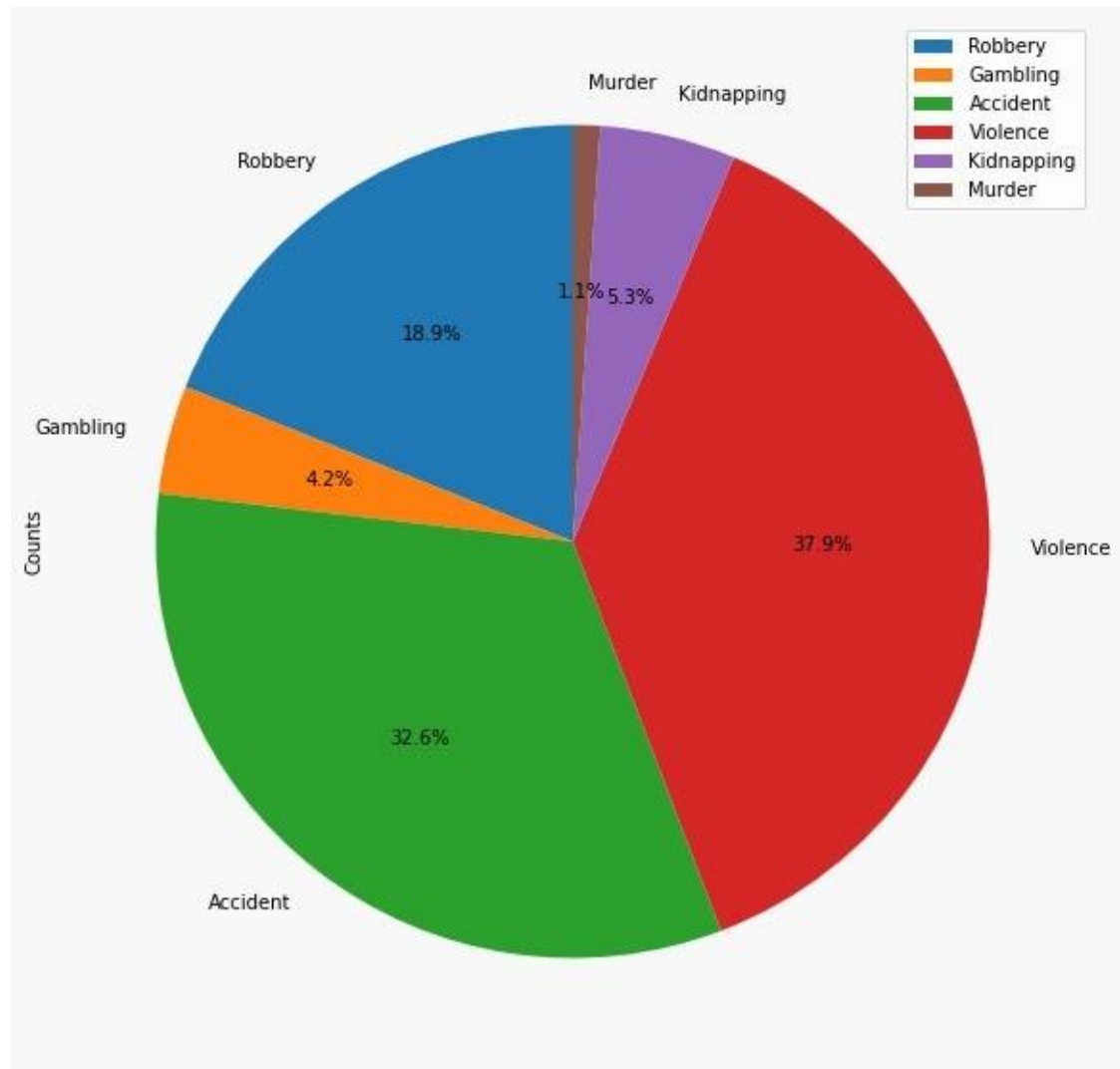
It offers the following functionalities:

- It helps to determine the relationship between different variables by providing dataset-oriented API.
- Linear regression plots are automatically estimated and plotted.
- High level abstraction for multiplot grids has been provided.

Using seaborn we can plot ranges of plots such as:

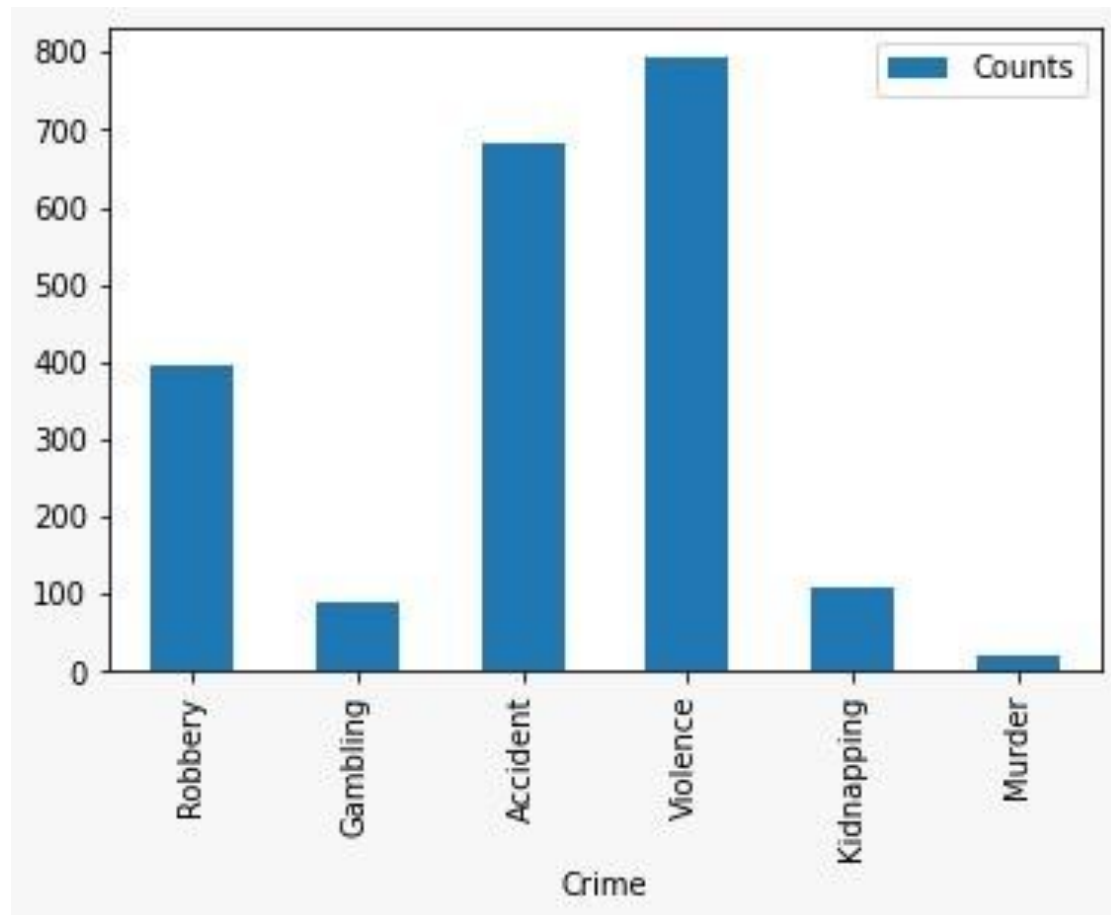
- Distribution Plots
- Scatter Plots
- Heat maps
- Pair Plots
- Pie Chart Bar Chart

- **Whisker Plot:** It is used for visualizing the result in the form of box-plots. It uses properties such as First quartile, third quartile, minimum, maximum, median. The box is created that ranges from first quartile to third quartile and the median passes from their middle. X axis denotes the data to be plotted and Y axis denotes the frequency of happening.

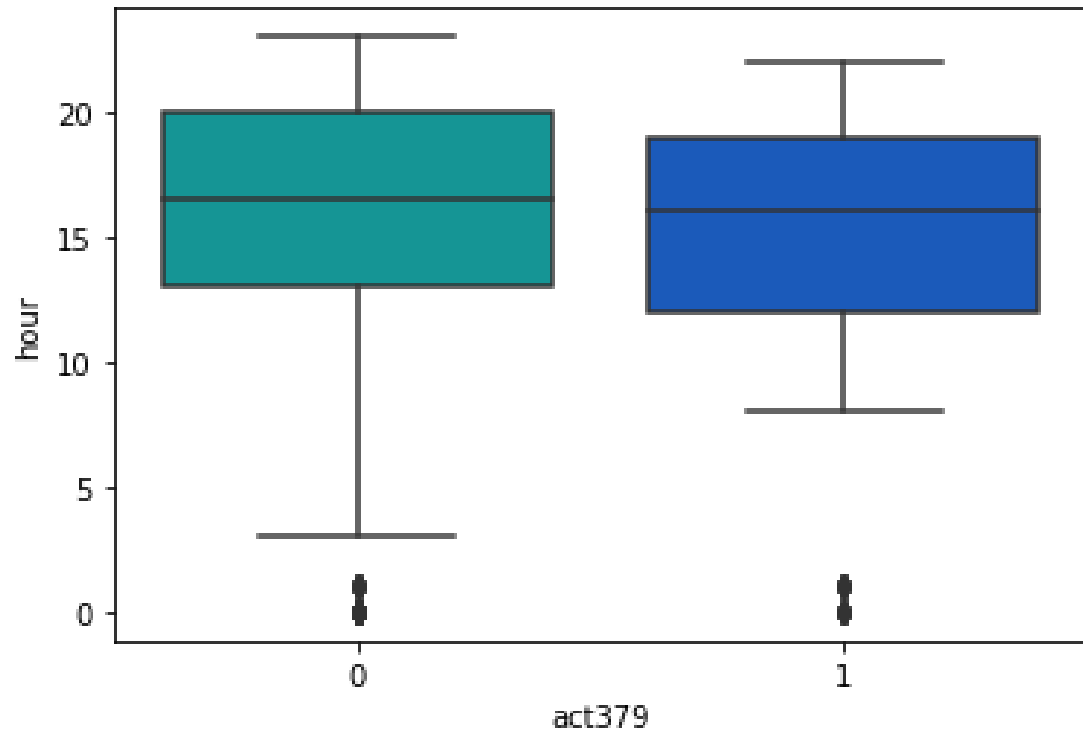


**Fig 4**

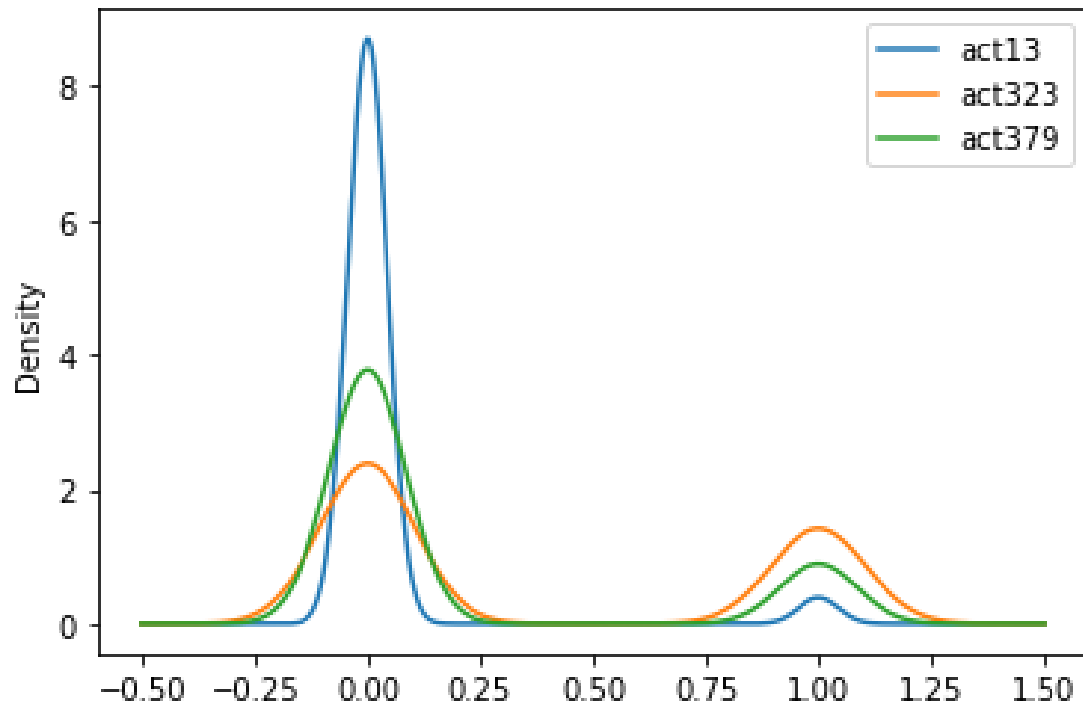
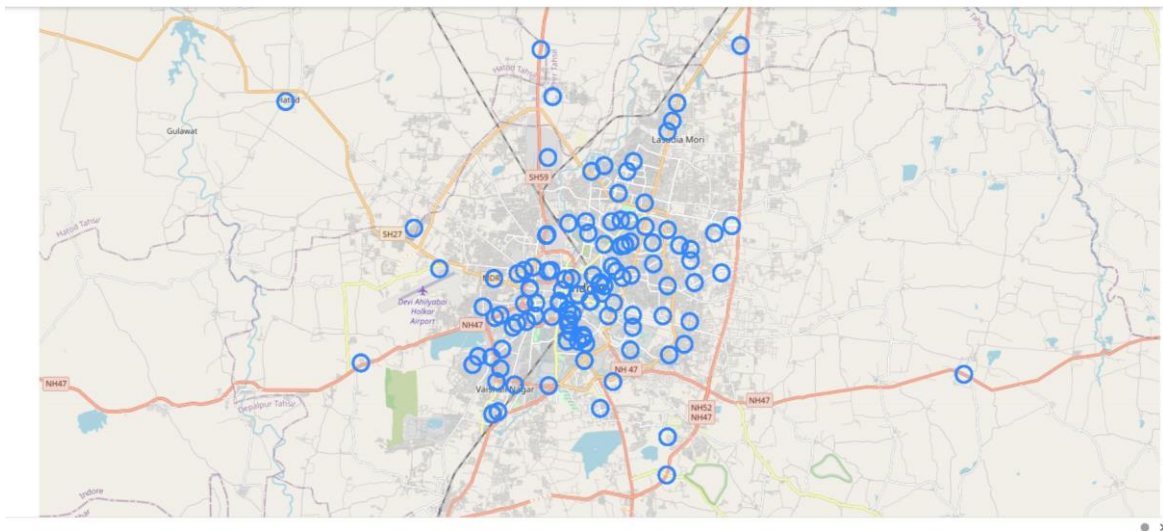




**Fig 5:** Crime vs Counts



**Fig 6:** Act379(Robbery vs Hour)

**Fig 7:****Fig 8:**

### **3.4 Prediction: Application of algorithms**

In this project we are performing prediction process for two parts:

- Class of crime and timestamp of crime happenings as an input and location of crime as an output.
- Location of crime and timestamp of crime happenings as an input and Class of crime enlisted in the dataset as an output.

#### **3.4.1 Prediction of type of crime**

This is the first part of our prediction process. Taking location and timestamp as inputs, our aim is to predict the future probability of the type of crime to happen. For carrying out such prediction process, we need to implement various algorithms on our data-set. Their comparative efficiency analysis will give us the best-fit algorithm with which we can predict the future possibility of any crime to happen. So, we carry out the comparative study of enlisted algorithms one-by-one.

#### **3.4.2 Prediction of location of crime**

This is the second part of our prediction process. type of crime and timestamp as inputs, our aim is to predict the location where the crime is likely to happen in future. For carrying out such prediction process, we need to implement various algorithms on our data-set. Their comparative efficiency analysis will give us the best-fit algorithm with which we can predict the location of crime that is likely to happen in future. So, we carry out the comparative study of enlisted algorithms one-by-one.

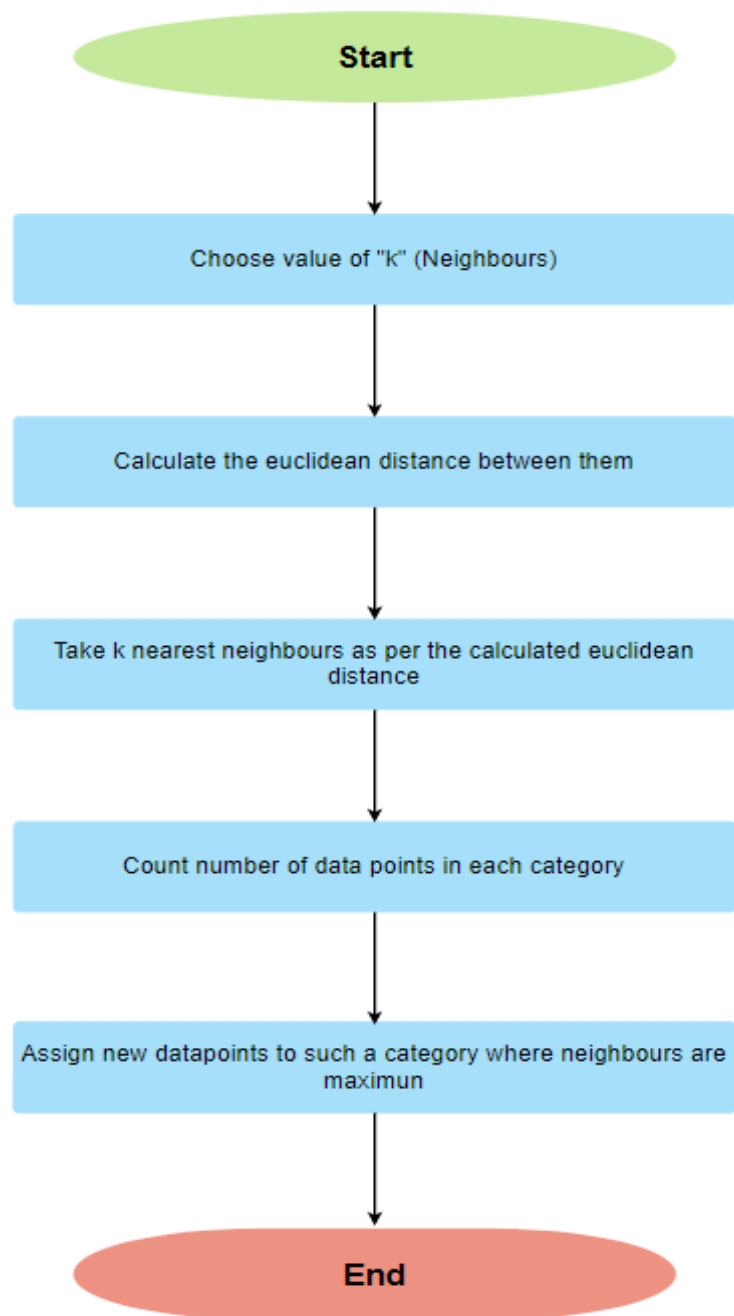
The algorithms that are applied for both prediction processes are:

- KNN
- Multioutput linear regression
- decision trees
- random forest

- **KNN:**

It is one of the simplest machine learning algorithms. It is based on supervised learning. The base of its working is to find the similarity between the new data entered and the old data available. It primarily stores the dataset without immediately working on it and when new data is entered, it starts its classification process where it classifies the current data which is much similar to new data. It does not make any assumptions on underlying data.

The steps that are involved in working of KNN are:

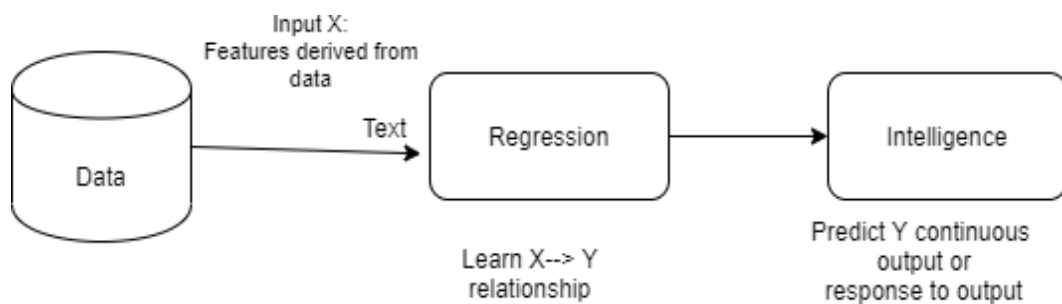


**Fig 9: KNN**

- **Multioutput linear regression:**

It is easier and one of the most popular machine learning algorithms. It carries out prediction in statistical form. It makes predictions for continuous as well as numeric attributes such as sales, age, salary, etc.

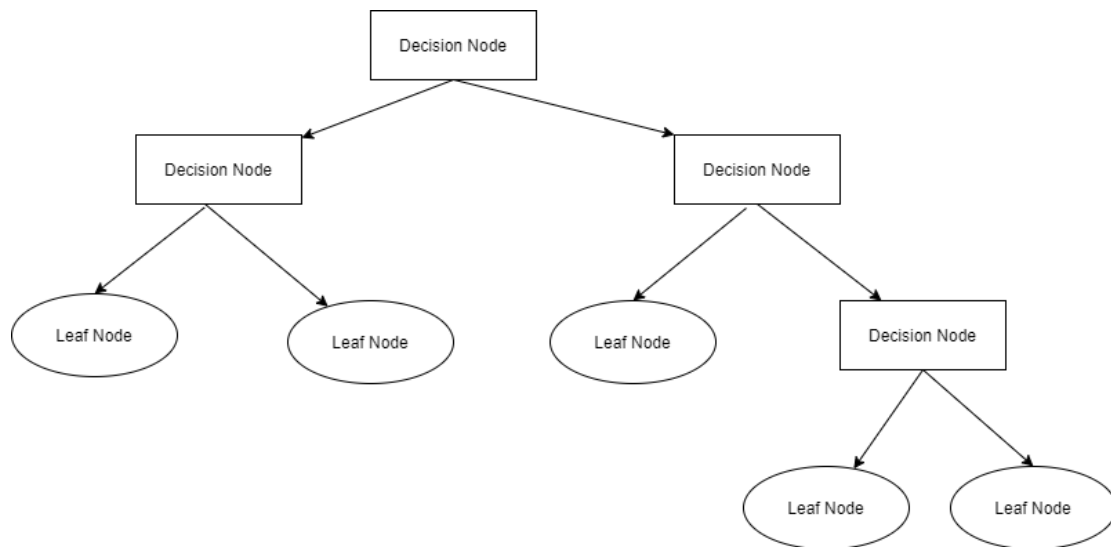
Linear regression model shows us the relationship between one dependent and one or more independent variables; hence it is called as linear regression model. Since it establishes linear relationship, it depicts how the value of dependent variable is changing with respect to independent variable or variables. It provides sloped line representing relationship between dependent and independent variables. Multioutput regression model is used for predicting two or more numerical values given an example as input.



**Fig 10**

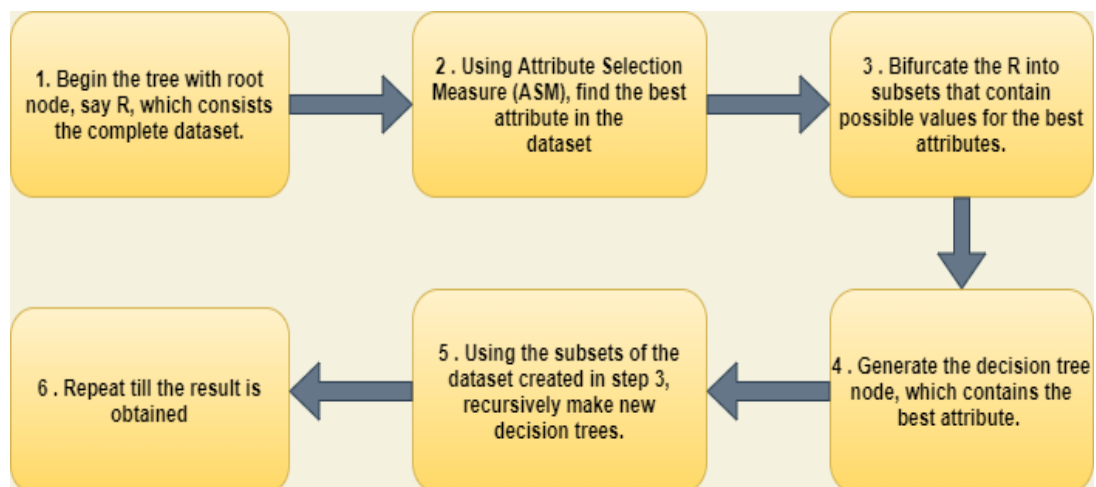
- **Decision trees:**

It is a supervised learning algorithm. It is used for both classification and regression. But mostly it is used for classification. It has mainly three parts. Decision Node, branches and leaf node. Internal nodes represent features of dataset, branches represent decision rules and leaf nodes represent output. The tests or decisions are performed based on the features of the dataset. It gives graphical representation with all possible solution for given problem based on the conditions in the dataset. Based on Yes/No process, each node splits into further sub trees.



**Fig 11:** Decision Tree

Steps involved in the decision tree prediction process are: prediction processes are:

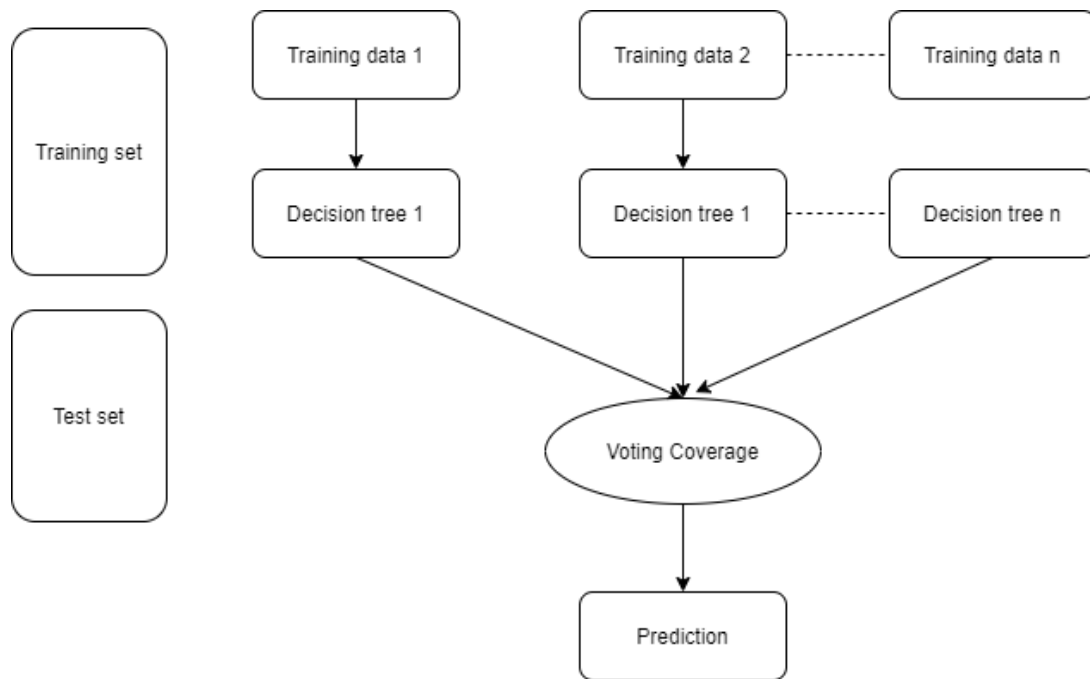


- **Random forest:**

Random forest is a popular supervised learning algorithm. It can be used for classification as well as regression. It combines the output of multiple classifiers and their average output helps us to improve the efficiency of



prediction i.e., ensemble learning. It is a classifier which has number of subsets on a dataset and takes the average to improve predictive efficiency of dataset. It takes output from number of trees instead of relying on a single entity and based on the majority votes of prediction, it gives the final output.



**Fig 12**

**Steps performed during operation are:**

1. Select random K datapoints from the available dataset
2. build the decision trees which are associated with the selected datapoints. (Subset formation.).
3. choose number N for decision trees that we want to build.
4. Iterate steps 1 and 2.
5. for the new data point, find the prediction of each decision tree and the datapoint will be assigned to the category that wins the majority.

### 3.5 Accuracy:

#### 3.5.1 Crime Prediction:

Out of above mentioned 4 algorithms, we have applied KNN, random forest and decision tree for prediction of crime. After performing their comparative analysis to get one most efficient algorithm out of them, we came to a conclusion that KNN gives the accuracy up to 93.23 percent, random forest gives accuracy up to 98.06 percent and decision tree gives the accuracy up to 98.06 percent. For our convenience, we have selected random forest algorithm for predicting class of crime.

Accuracy Measurement method	KNN	DT	Random Forest
Accuracy in %	93.23671497584542	98.06763285024155	98.0686328502415
R2 Score	0.8409464479819618	0.9123849813900212	0.9123849813900212
Explained Variance	0.8432201825668598	0.9124907966298896	0.9124907966298896
Mean Absolute Error	0.161	0.005	0.005
Mean squared Error	0.01328502415458937	0.00644122383252818	0.00644122383252818

**Fig 13**

#### 3.5.2 Location Prediction:

Out of above mentioned 4 algorithms, we have applied KNN, multioutput linear regression and decision tree for prediction of location of crime. After performing their comparative analysis to get one most efficient algorithm out of them, we came to a conclusion that r2-score and explained-variance-score are the two parameters with which we can measure the accuracy of algorithm to predict location of crime. The algorithm which will have the value of these two

parameters near to 1 will be considered as the most efficient one. So, the respective values of r2-score and explained-variance-score for KNN, multioutput linear regression and decision tree are 0.6497 and 0.6520, 0.0876 and 0.0881, 0.7058 and 0.7076 respectively. from this it is clear that decision tree model gives the highest accuracy, hence it is implemented.

Accuracy Measurement method	Linear regression (Multi Output)	KNN	DT
Accuracy in %	8.045679479166365	71.99873190754299	76.12991966466133
R2 Score	0.08762420572487117	0.6497734409427314	0.7058899578571258
Explained Variance	0.08819459071103586	0.652048447393606	0.7076468030798932
Mean Absolute Error	0.033	0.011	0.010
Mean squared Error	0.00270609137294599	0.0010503209051527654	0.000884954577809402

**Fig 14**

## 4 Case Diagrams

### 4.1 Activity Diagram

This is the activity diagram for "Police Companion". Left side represents prediction process for crime and right side represents prediction process for location.

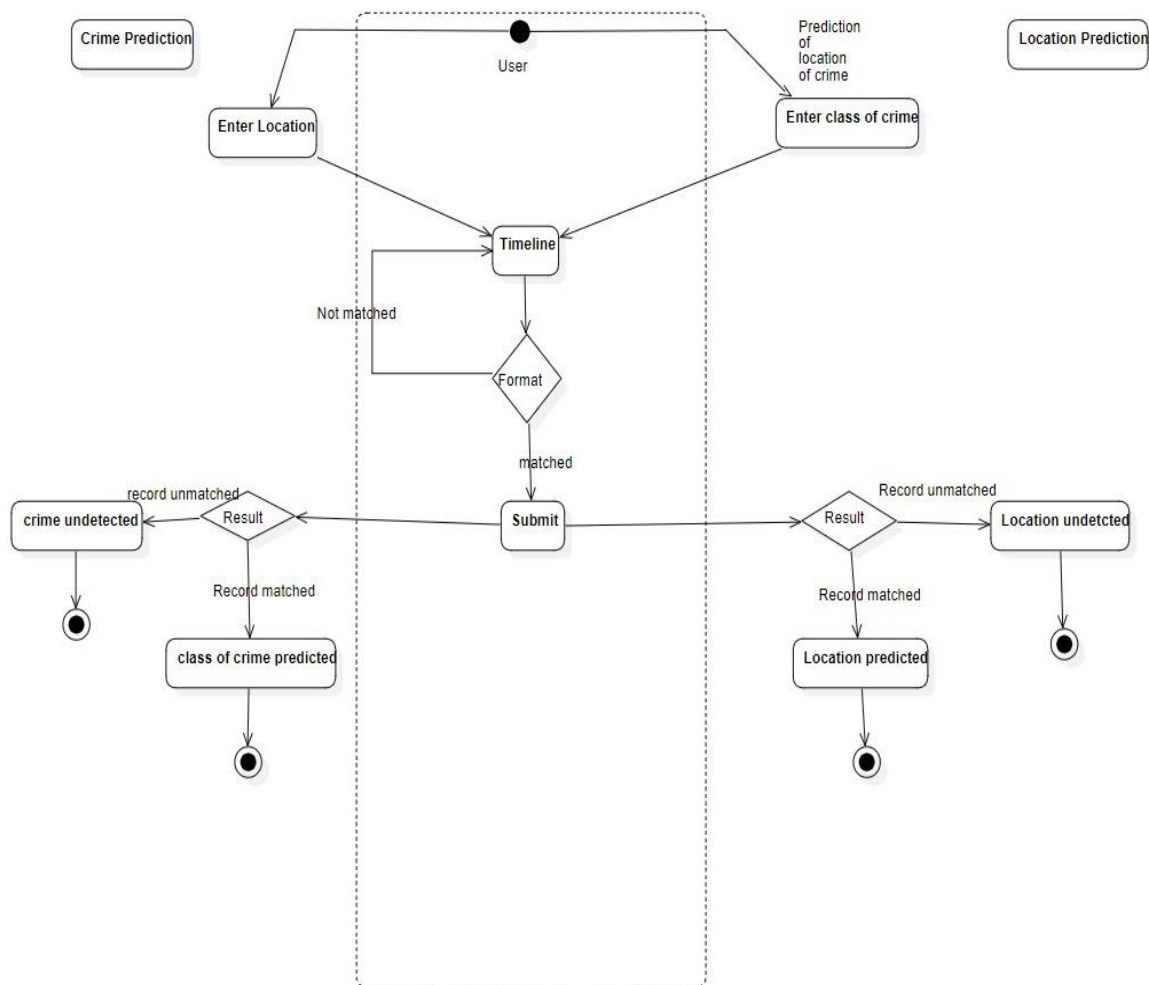
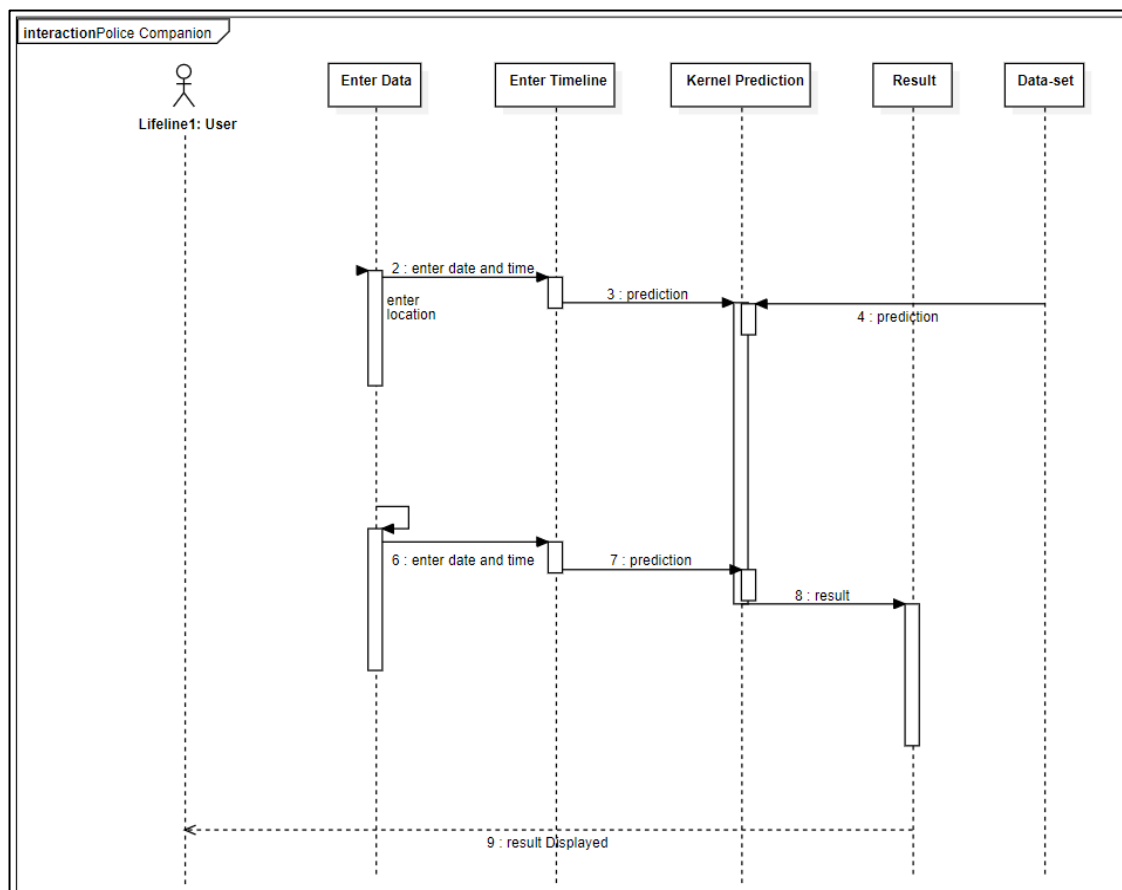


Fig 15

## 4.2 Sequence Diagram

This is the sequence diagram for "Police Companion". The kernel predictive model accomplishes the machine learning operations performed on data-set.

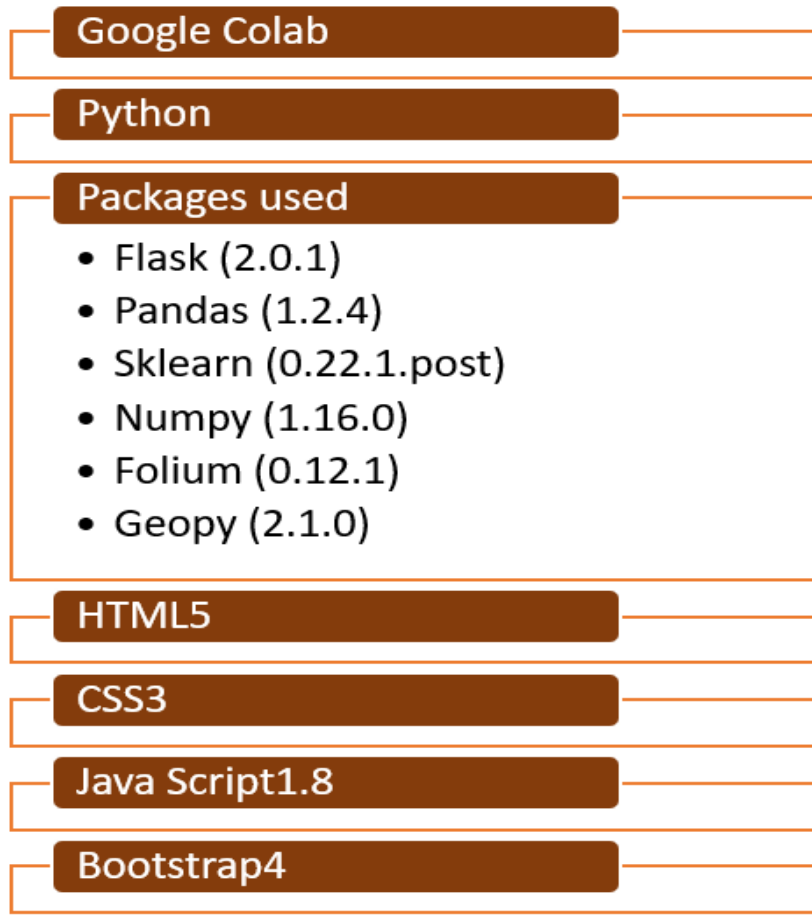


**Fig 16**

## 5 Tools

### 5.1 Software Details:

Software requirements are:



### 5.2 Hardware Details:

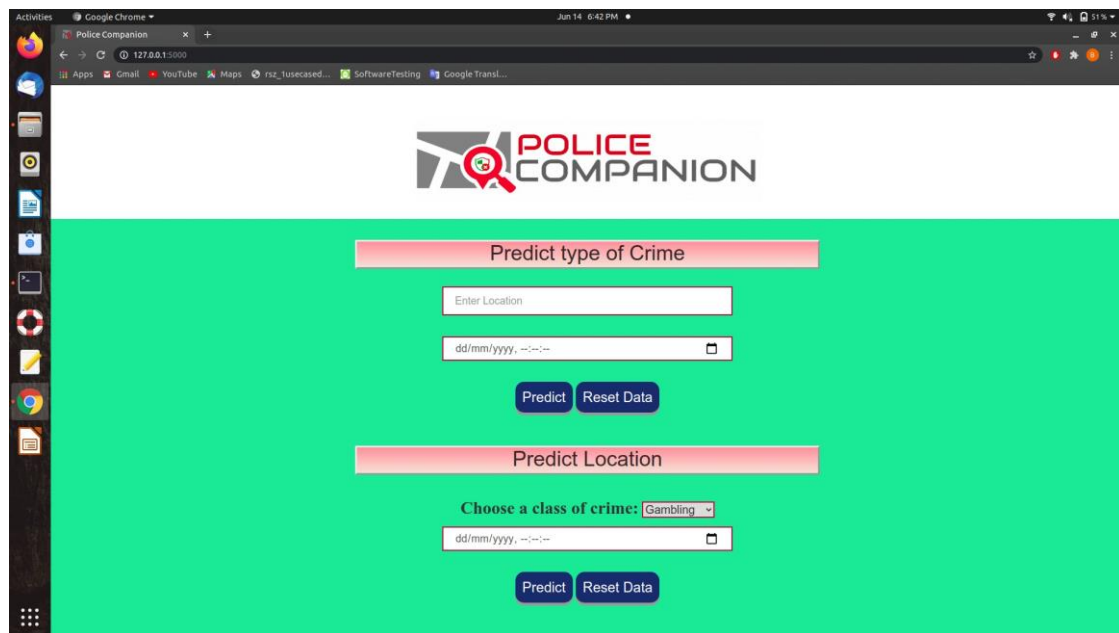
Hardware requirements are:

- OS: Windows 7 or newer, Linux, or 64-bit macOS 10.9+.
- System architecture: Windows or Linux with 64-bit x86, 32-bit x86.
- RAM: 4 GB or greater.

## 6 Screenshots

This screenshot section consists of the screenshots of the GUI that we have framed.

This is how GUI looks like at opening glance:



**Fig 17**

Giving Location and timestamp as inputs, the predicted class of crime is seen as follows:



  <b>Predict type of Crime</b> <input type="text" value="palasiya indore"/> <input type="text" value="23/06/2021, 03:42:38"/> <input type="button" value="Predict"/> <input type="button" value="Reset Data"/>  <b>Predict Location</b> <input type="text" value="Choose a class of crime: Gambling"/> <input type="text" value="dd/mm/yyyy, --:--:--"/> <input type="button" value="Predict"/> <input type="button" value="Reset Data"/>	<b>Output:</b>   <b>Predicted type of crime : Violence-Act 323</b>
--	---

Fig 18



  <b>Predict type of Crime</b> <input type="text" value="palasiya indore"/> <input type="text" value="23/06/2021, 06:51:48"/> <input type="button" value="Predict"/> <input type="button" value="Reset Data"/>  <b>Predict Location</b> <input type="text" value="Choose a class of crime: Gambling"/> <input type="text" value="dd/mm/yyyy, --:--:--"/> <input type="button" value="Predict"/> <input type="button" value="Reset Data"/>	<b>Output:</b>   <b>Predicted type of crime : Accident-Act 279</b>
--	---

Fig 19



If no record matches, then it will give the result like this:



**POLICE COMPANION**

**Predict type of Crime**

A.B.Road Indore

23/06/2021, 06:42:36

Predict Reset Data

**Predict Location**

Choose a class of crime: Gambling

dd/mm/yyyy, --:--:--

Predict Reset Data

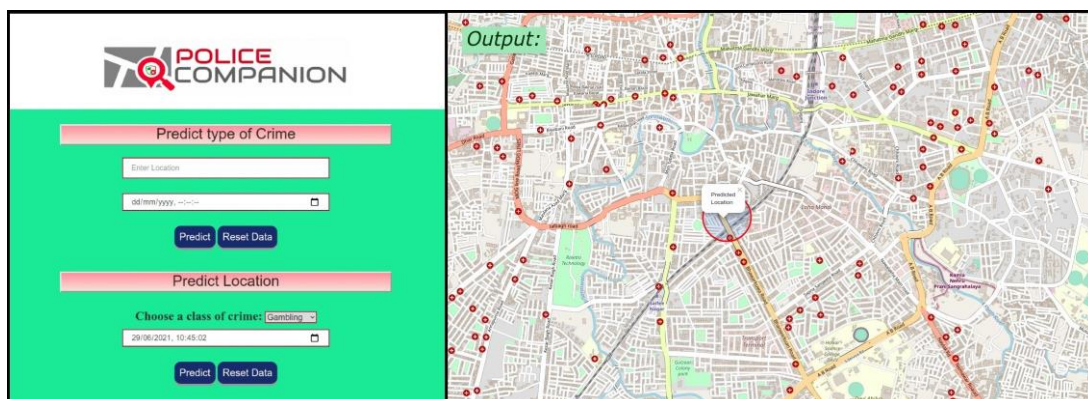
**Output:**

**POLICE COMPANION**

Place is safe no crime expected at that timestamp.

**Fig 20**

Giving Class of crime and timestamp as inputs, the location of crime is predicted as follows:



**POLICE COMPANION**

**Predict type of Crime**

Enter Location

dd/mm/yyyy, --:--:--

Predict Reset Data

**Predict Location**

Choose a class of crime: Gambling

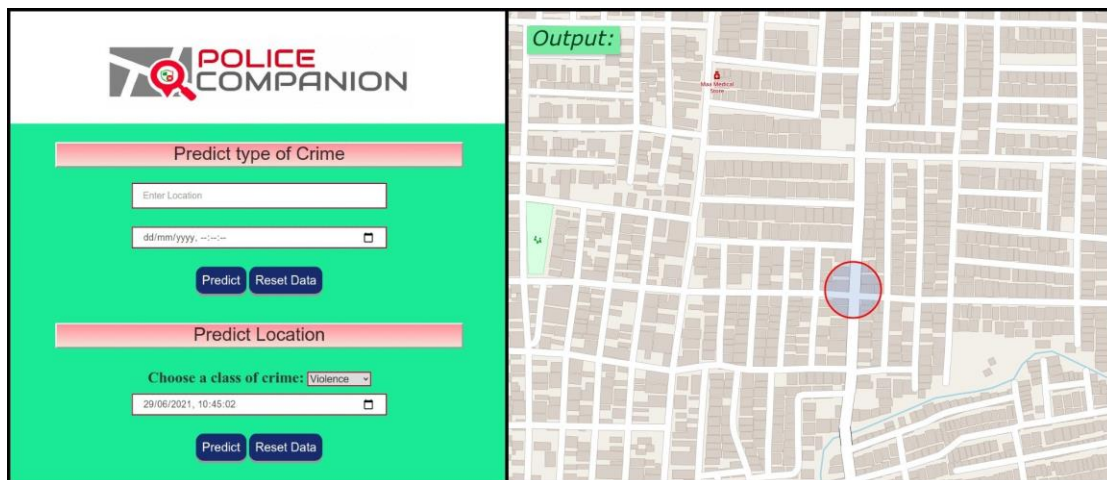
29/06/2021, 10:45:02

Predict Reset Data

**Output:**

Predicted Location

**Fig 21**



**Fig 22**

## **7 Future Scope**

The future scope of this project is to apply more machine learning techniques in the current project so as to help police department along present measures. One of them can be predicting the suspect of the crime using face recognition techniques. The system will check if there are any suspicious movements or changes in particular area. For example-suspects moving in and around certain place over and over again.

The second scope of improvement can be collecting the stored crime data from all police stations in city and combining them into one single dataset which would contain crime data of all the station at single place. This would increase the overall accuracy of prediction as more data would be available. Also, one station will have the look over the crime activities of area coming under another station's jurisdiction. So, the access of information will become much easier. Collecting data from various sources would result in re-modification of the available data.

In this way, the current project can be extended to various limits depending upon the relevance and efficiency of various machine learning techniques that are suitable for work related to enforcement department.

## **8 References**

- [1] Alkesh Bharati, Dr Sarvanaguru RA.K , "Crime Prediction and Analysis Using Machine Learning" , "International Research Journal of Engineering and Technology (IRJET)",Volume: 05 , Sep 2018
- [2] Suhong Kim, Param Joshi, Parminder Singh Kalsi, Pooya Taheri, "Crime Analysis Through Machine Learning", "IEEE" , 17 January 2019
- [3] Ying-Lung Lin, Tenge-Yang Chen,Liang-Chih Yu , "Using Machine Learning to Assist Crime Prevention" , "IEEE" , 16 November 2017
- [4] <https://jupyternotebook.readthedocs.io/en/stable/notebook.html>
- [5] NumPy community, "NumPy User Guide" , January 31, 2021
- [6] [https://pandas.pydata.org/docs/user\\_guide/index.html](https://pandas.pydata.org/docs/user_guide/index.html)