

# Unsafe Tract Revelation and Inspection Using Machine Learning

<sup>1</sup>Nikhil Nirgudkar, <sup>2</sup>Ravinder Singh, <sup>3</sup>Abhishek Singh, <sup>4</sup>Margi Patel

<sup>1,2,3</sup>B.Tech. Scholars, <sup>4</sup>Assistant Professor, Department of Computer Science & Engineering  
Indore Institute of Science and Technology, Indore, MP, India.

<sup>1</sup>nikhilkavita19@gmail.com, <sup>2</sup>rskravinder21@gmail.com, <sup>3</sup>avigaharwar07@gmail.com,  
<sup>4</sup>margi.patel22@gmail.com

**ABSTRACT** - Towns of diverse provinces are getting more unsafe day by day. The data following varied crime reports are available for these unsafe towns but, even such reports captioned are not perceived to 80% inhabitants of the town. The intent of this article is to decipher datasets which comprise of two distinct types of the dataset, one stands from real-life established crime dataset taken from police and another one stands from Safety Audit (survey) dataset which is performed by inhabitants of the town and using these pairs, anticipating tracts that may come to be unsafe in the future hinging upon numerous circumstances. In this article, we would be employing the procedure of Machine Learning and Data Science for unsafe tract revelation using the Indore city crime dataset as well as the Safety Audit dataset. The crime data has been obtained from the cyber website portal of Indore city's police. It comprises of criminal parameter evidences like geolocation caption, classification of different criminal activities, duration i.e. date & time. Before the building and training of the processing model, data pre-processing would be mounted. Onto the next step, drafting of useful extracted features and scaling up of them would be performed which on the precision obtained would be increased further. The various distinct algorithms (such as KNN, Linear & Logistic Regression, SVC etc.) would be tested for unsafe tract revelation and an odd one with satisfactory precision would be opted for analysis. Anticipation of the dataset would be performed in terms of graphical depiction of many cases. For example, at what duration the frequency of criminal rates are high or at which sight the criminal undertaking are high. Safety Audit dataset contains information about various circumstances of a locale such as Light, Visibility, Transport, Security, Walk path, People, Time. The sole intent of this project is to give just an idea of how Machine Learning could be used to provide useful information about unsafe tracts to the user. It is not only restricted to Indore City but could also be used in other provinces being sure of upon the availability of the dataset.

**Keywords:** Data Science, Decision Trees, KNN Classification, Linear Regression, Logistic Regression, Machine Learning, Naïve Bayes Classifier, Safety Audit, Sklearn Library, Support Vector Classifier (SVC).

## I. INTRODUCTION

1.1 Criminal movements are a notable threat to humankind. Various transgressions come unflinchingly at a regular rate. Conceivably it is escalating with high frequency rate. The extent of criminal movement could be from a small village, town or big cities too. Criminal activities are catalogued into – Robbery, Murder, Ravishment, Assault, Battery False Imprisonment, abduction, Homicide. Since such criminal activities are expanding, there is a necessity to uncover and aware the inhabitants about these unsafe tracts. The frequency rate of such crimes is very elevated. As a direction towards the support of the police guard system, there is a necessity for

technical assistance through which inhabitants would be attentive and receptive while moving through those tracts.

1.2 The overhead dilemma made us go for scrutiny about how we can provide information about those unsafe tracts. Through various documents augmentation and registered criminal cases, it came out that some newly developing technical concepts like Machine Learning and Data Science would bring about the job easier and faster.

1.3 The intent of this article is to uncover unsafe tracts using the attributes present in the dataset. The datasets are extracted from the authorized cyber sites of police as well as safety audit from inhabitants. With an aid of varied

Machine Learning Algorithm, revelation unsafe locales would more efficient.

1.4 The crux is to design a training model for analysis. The training would be accompanied with training set of data which on further examined using the testing set of data. Assembling up of model will be carried out using a better algorithm depending upon the exactitude. Distinct classification algorithms will be utilized for unsafe track revelation. Anticipation of the different sets of fata is performed to interpret the areas which would be unsafe. This effort would help people to be cautious while moving through those locales.

## II. METHODOLOGY

For optimum inspection and revelation of unsafe tracts, predictive modelling must be single outed. Predictive modelling is a statistical technique to predict future aspects. This process works by analysing historical (raw) & current data (discrete data) and availing it to generate favourable future outcomes. Predictive modelling can be further categorised into Regression Analysis & Patter Recognition (PR) or Classification.

Regression models are based on the analysis of relationships between dependent (target) variables and independent (predictor) variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression analysis. In contrast to Regression models, the task of Pattern Recognition (PR) or Classification is to put the data in a categorical way based on what it learns from historical or indiscrete data. An example of a classification model is - A pattern recognition task in weather forecasting could be the identification of a sunny, rainy, or snowy day based upon different parameters such as coastal, high peak etc.

Pattern classification tasks can be designated into two segments: -Supervised & Unsupervised learning.

A supervised learning algorithm learns from labelled training data, helps you to predict outcomes for unforeseen data.

Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Types of predictive model: -

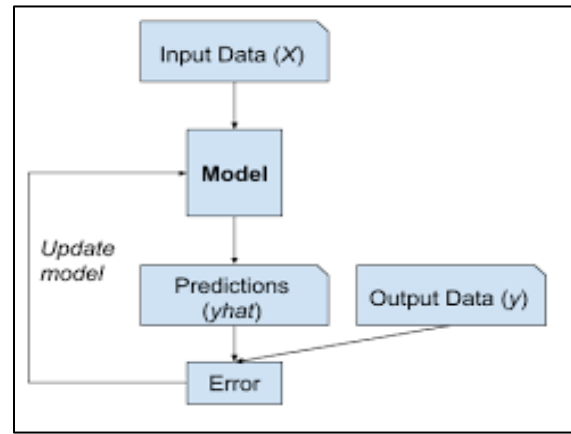


Figure 1: Predictive Model

Classification and Decision Trees: The crux of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from prior data (training data).

Naive Bayes - In machine learning, Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

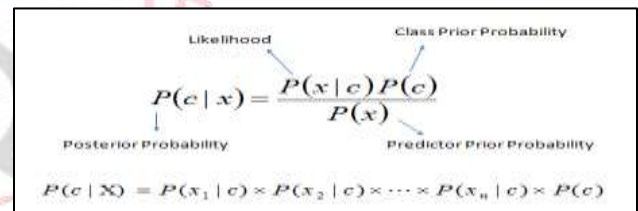


Figure 2: Naïve Bayes Classifier formula

Linear Regression – Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model.

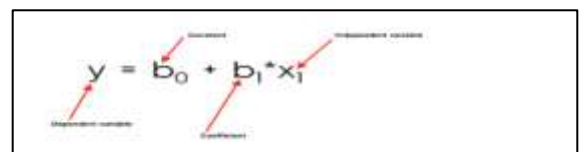


Figure 3: Linear Regression equation

Logistic Regression - Logistic regression is a linear method, but the predictions are transformed using the logistic function.

Logistic Regression Equation:

$$\frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

Figure 4: Logistic Regression equation

## 2.1 Data Collection & Pre-processing

The collected data comprises of two distinct types of the dataset, one stands from real-life established crime dataset taken from Indore city police and another one stands from Safety Audit (survey) dataset which is performed by inhabitants of the town

The data from police officials comprises of criminal parameter evidences like geolocation caption, classification of crime, duration i.e. date & time while audit data comprises of Light, Visibility, Transport, Security, Walk path, People, Time.

The fundamental stages included in data pre-processing are Formatting, Cleaning and Sampling.

**FORMATTING:** The availability of the data was in the form of Hindi language. To carry out the mathematical deduction, the data needs to be transformed into mathematical objects.

**CLEANING:** Cleaning process is performed for dismissal or embedding of some null data. After making a suitable dataset for analysis the further process includes methods to remove any NULL values or infinite values which may affect the accuracy of the system.

**SAMPLING:** Sampling is the process where appropriate data are used which may reduce the running time for the algorithm. Using python, the pre-processing is done.

### Functional Diagram of Proposed Work

Categorisation is mainly into four contiguous parts:

1. Illustrative analysis on the Data
2. Formatting of data (embedding null values)
3. Data Modelling
4. Evaluation of performance

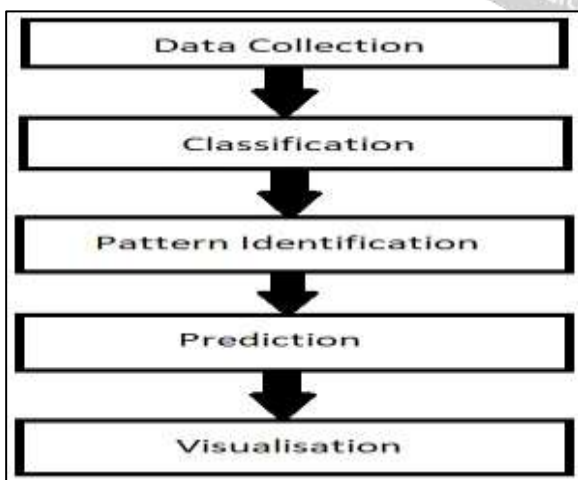


Figure 5: Predictive Modelling Architecture

### 2.1.1 Prepare Data

The data acquired from is transformed into suitable mathematical deductible format for further analysis.

1. Information (data) gathering
2. Formatting
3. Data cleaning.

### 2.1.2 Analyse and Transform Variables:

The formatting of variables is applied by employing one of the approaches:-

- Normalization or standardization.
- Embedding null values.

### 2.1.3 Random Sampling (Set of training data and set of testing data)

- Training Sample: Analytical prototype is developed on this set of data (mainly about 75% of data is used).
- Test Sample: Prototype precision performance is carried out on this set of data (mainly 25% of data is acquired for such task).

### 2.1.4 Prototype Election

Continuing to analysis, there are varied algorithms which are sub-category of supervised and unsupervised. Some of which are: -

- KNN Classification
- Logistic Regression
- Decision Trees Classifier
- Random Forest Classifier
- Support Vector Classifier (SVC)
- Naïve Bayes Classifier

### 2.1.5 Assembling/Developing/Training Prototype

1. Ratify the presumptions of the selected algorithm.
2. Assemble/Train Prototype on set of training data.
3. Check Prototype efficiency in terms of errors.

## 2.2 Anticipation of data

Data anticipation is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations.

Analysis of process model could easily done with help of data anticipation.

### 2.3 Validate/Test Prototype

1. Precision testing set of data.
2. Accuracy estimation.

## III. IMPLEMENTATION

The informative data used in this article is obtained from www.indorepolice.org and survey from people.

The implementation of this project is divided into following steps –



### 3.1 Data collection

The collected data comprises of two distinct types of the dataset, one stands from real-life established crime dataset taken from Indore city police and another one stands from Safety Audit (survey) dataset which is performed by inhabitants of the town

The data from police officials comprises of criminal parameter evidences like geolocation caption, classification of crime, duration i.e. date & time while audit data comprises of Light, Visibility, Transport, Security, Walk path, People, Time.

### 3.2 Data Pre-processing

More than 10k records are filed in the dataset. The null values are embedded by the expression `data = data.dropna()` where data is the "DataFrame". The categorical attributes (location caption, classification of crime, date, time, Light, Visibility, Transport, Security, Walk path, People) are converted into numeric using Label Encoder.

The date attribute has been split into new attributes like month and hour which can be used as feature for the model.

The two figures below shows the raw dataset (Table 1) and pre-processed dataset (Table 2).

क्र. सं.	क्षेत्र	दिनांक	समय	व्यक्तिगत विवरण	व्यक्तिगत विवरण	व्यक्तिगत विवरण	व्यक्तिगत विवरण	व्यक्तिगत विवरण	व्यक्तिगत विवरण
1	High Court	01-04-20	18:30	...	...	...	...	...	...
2	Champhar	04-04-20	04:45	...	...	...	...	...	...
3	W.P. Nagar	04-04-20	01:00	...	...	...	...	...	...
4	W.P. Nagar	04-04-20	18:00	...	...	...	...	...	...
5	W.P. Nagar	01-04-20	18:30	...	...	...	...	...	...
6	W.P. Nagar	04-04-20	18:15	...	...	...	...	...	...

Table 1: Raw Dataset

Area	Occurance	Crime
53 South Kamthipura, Indore	Day	Vehicle theft
Annapurna, Indore	Night	Vehicle theft
Apna Hotel, Indore	Night	Threatening
Banganga, Indore	Day	Vehicle theft
Banganga, Indore	Night	misdemeanour
Bherughat, Indore	Day	Vehicular Rampage
Bicholi Hapsi, Indore	Day	Vehicle theft
Bombay Hospital, Indore	Day	Vehicle theft
Choithram, Indore	Day	Vehicular Rampage
Devguradia, Indore	Night	Vehicle theft
Dewas Bypaas Road, Indore	Day	Vehicular Rampage
Gangwal Bus Stund Square, Indore	Day	misdemeanour

Table 2: Processed Dataset

### 3.3 Feature selection

Features selection is done which can be used to build the model. The attributes used for feature selection are geolocation caption, types of criminal activities, duration i.e. date & time.

### 3.4. Building and training Model

After the above process geolocation with its frequency rate, different types of crime attribute with duration are used for training. The whole set of data is divided into pair of xtrain, ytrain and xtest, y test. The algorithms is imported form sklearn package library. Assembling of prototype is done using expression Fit (xtrain, ytrain).

### 3.5. Prediction

After the prototype is assembled using the above process, inspection is done by expression model.predict(xtest). The accuracy estimation is done by using expression accuracy\_score imported from metrics.accuracy\_score (y\_test, y\_pred).

### 3.6. Visualization

Using matplotlib library from sklearn. Analysis of the crime dataset is done by plotting various graphs.

## IV. RESULTS

The results are obtained after undergoing various processes that comes under machine learning. With reference of the processed data (shown in Table 2) after dividing the data set into training sample and testing sample the prototype is trained using suitable algorithm. The accuracy is calculated using the expression score\_accuracy imported from metric.sklearn. The accuracy is mentioned in the table below.

Generally the ratio of train sample to test sample is 3:1 i.e. out of 100 samples 75 samples would be classified as train

sample & remaining 25 samples would be classified as test sample.

ALGORITHM	ACCURACY by score_accuracy function
KNN Classifier	0.734501375931487
SVC	0.471478227903329
Decision Tree Classifier	0.701348420347235

**Table 3: Accuracy Analysis Table**

As from the above Accuracy Analysis Table (Table 3) the highest accuracy was obtained from KNN algorithm with a precision score of 0.735 approx.

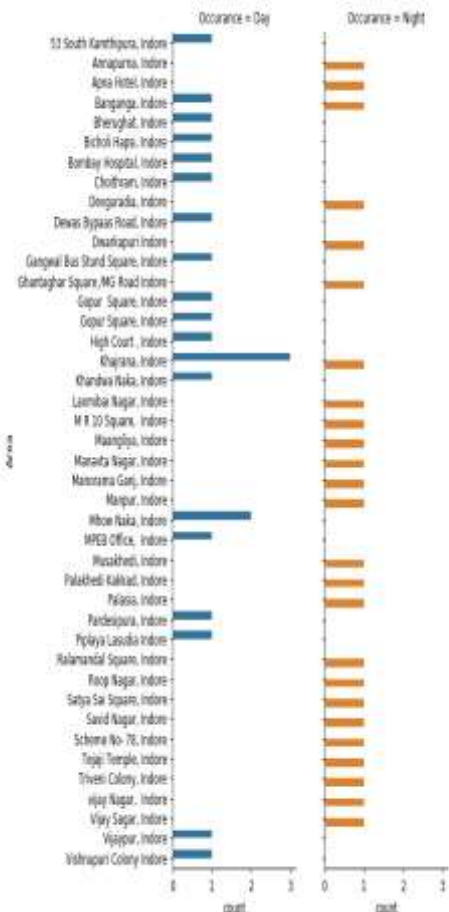
**4.1 Analysis Visualizations**

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

The visualizations are:-

**DATASET 1**

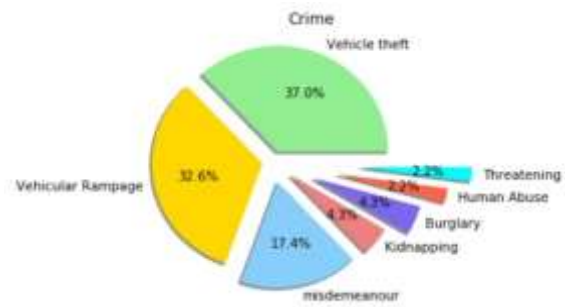
Data obtained from cyber site of police officials.



**Figure 6: Frequency rate of crime in different areas**

The above bar graph is a count-plot of rate of crime in different tracts of Indore city along with a depending parameters i.e. Day & Night, where the blue bars are for Day representation and orange is for Night representation.

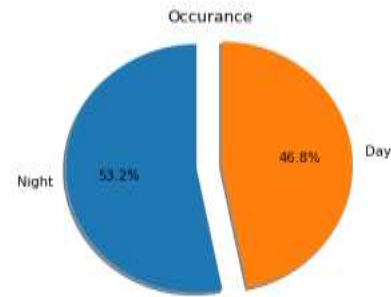
The x coordinate denotes Areas and y coordinate denotes count.



**Figure 7: Percentage of types of crimes**

The above pie-chart represents the various types of crime with their probability of occurrence.

It could be noted that majority cases are with respect to vehicles.



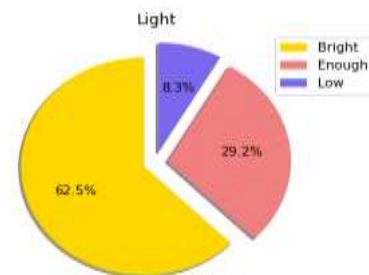
**Figure 8: Percentages of Occurance of crime Day vs Night**

The analysis shows that the occurrence is more at night time when compared with Day time.

**DATASET 2**

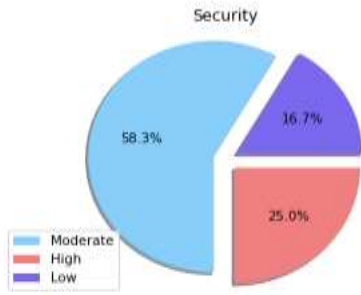
The previous analysis visualizations was on the basis of the data obtained from police officials.

There are notable things given by different inhabitants for different areas. Some of which are visualized as:



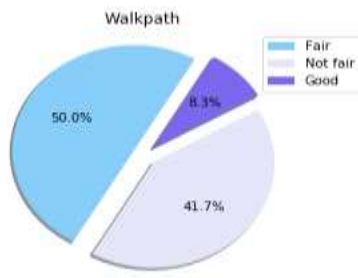
**Figure 9: Percentages of Light visual**

It could be noticed that majority sections lies in bright. But even there is record of 8.3% as low light.



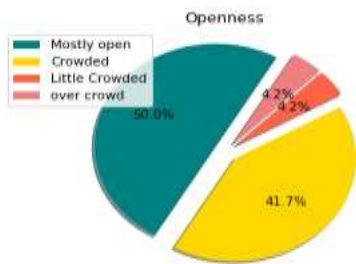
**Figure 10: Percentages of Security Level in different areas**

It was accounted that 16.7% areas there with low security and even only 25.0% accounts for high security.



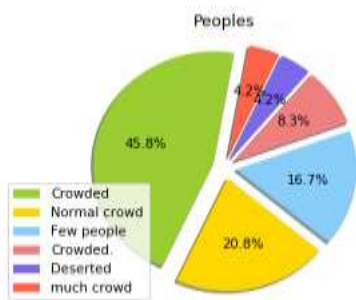
**Figure 11: Percentages of Walk path in different areas**

It was analysed that only half of the percentages accounts for fair walk path.



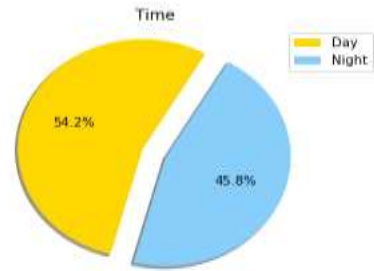
**Figure 12: Percentages of Openness in different areas**

About half of the percentage account for openness in their respective areas.



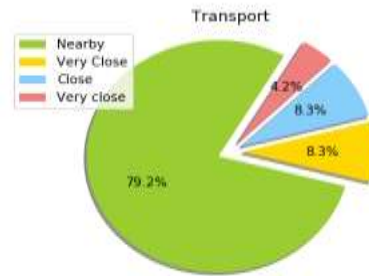
**Figure 13: Percentages of number of surrounding people**

Regarding the no. of peoples to ones surrounding 4.2% accounts for deserted i.e. less no. of surrounding people.



**Figure 14: Percentages of duration Day vs Night (Audit)**

Even according to the audit dataset there is 45.8% occurrence of criminal activities at night.



**Figure 15: Percentages of availability of transport**

All main parameters like Light, Visibility, Transport, Security, Walk path, People, Time was opted for citizen's feedback in their areas.

## V. CONCLUSION

Under the aegis of Machine Learning Technology, it has become quite easier and helpful to extract out the distinct patterns and their relationship other parameters of dataset. The work in this article mainly revolves around revealing the unsafe tracts of the cities and inspecting the locales where different kinds of criminal activities have occurred and its relatable augury.

With the help of Machine Learning we have assembled a prototype using training sample and testing sample which was formatted and cleaned to get mathematical deductible information to the dataset. The prototype predicts the varied criminal activities in respective areas with precision of 73.5%. Using graphical representation, the analysis of the available dataset is more precise but also is user friendly to view. The graphs include bar, piegraphs each having its own characteristics. We provoked many graphs and found interesting statistics that helped in understanding Indore city crimes datasets along with Safety Audit dataset that helped in apprehending the factors that can help in keeping society safe. The above audit dataset was mainly relatable to a respective city.

More such type of initiative like the “Safety Audit” will not only spread awareness about the same to others but also will great contribution for others who at present working the minimization of criminal activities frequency rate.

Distinct classification prototypes were compared during analysis. Machine Learning could help various law enforcing agencies to fight against criminal activities and reducing the frequency rate to a greater extent. Updating the data available is must, as better as precise the data is, the analysis would be strong enough.

## REFERENCES

- [1] Aurelien Geron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems”.
- [2] <https://www.numbeo.com/> [Accessed 25 January 2020]
- [3] Vineet Jain, Yogesh Sharma , Ayush Bhatia, Vaibhav Arora,”Crime Prediction using K-means Algorithm” . GRD journals,Volume 2 , Issue 5 ,April 2017 , ISSN: 2455-5703
- [4] Crime Prediction and Analysis Using Machine Learning Available on <https://www.irjet.net/>e-ISSN: 2395-0056//p-ISSN: 2395-0072
- [5] Tayebi, M. A., Gla, U., &Brantingham, P. L. (2015, May). Learning where to inspect: location learning for crime prediction. In Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on (pp. 25-30).
- [6] Sathyadevan, S., &Gangadharan, S. (2014, August). Crime analysis and prediction using data mining. In Networks & Soft Computing (ICNSC), 2014 First International Conference on (pp. 406-412).
- [7] Sasha Kapoor, AbhineetKalra, “Data Mining for Crime Detection”. International Journal of Computer Engineering and Applications, Volume VII, Issue III, September 14
- [8] Nath, S. V. (2006, December). Crime pattern detection using data mining. In Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 iee/wic/acm international conference on (pp. 41-44). IEEE.
- [9] Manish Gupta, B. Chandra and M. P. Gupta, “Crime Data Mining for Indian Police Information System”, Computer Society of India, Vol. 40, No. 1, pp. 388- 397, 2008
- [10] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal, “Crime Analysis using K-Means Clustering”, International Journal of Computer Applications (0975-8887), Vol. 83, No. 04, 2013