IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Crime Pattern Detection, Analysis and Prediction

[1]Aishwarya Retawade, [2]Shivraj Yelave, [3]Rutika Salunkhe,Anutej Ware,Jinesh Melvin
[1]Information Technology,
[1]Pillai College of Engineering, New Mumbai ,India

Abstract : Crime is one of the dominant aspect of our society. Everyday lots numbers of crimes are committed, these frequent crimes have made the lives of citizens restless. So, preventing the crime from occurring is a vital task. In the recent time, it is seen that Machine Learning and Artificial Intelligence has shown its importance in almost all the field and crime prediction and analysis is one of them. However, it is needed to maintain a proper database of the crime that has occurred as this information can be used for future reference. Prediction of crime is a systematized method that classifies and examines the crime patterns. There exist various algorithms for crime analysis and pattern prediction but they do not reveal all the requirements. Predicting the crime accurately is a challenging task because crimes are increasing at an alarming rate. Thus, the crime prediction and analysis methods are very important to detect crimes and to reduce them.

I. INTRODUCTION

”1. Introduction

With the increase in the number of crimes being recorded crime analysis is no longer a choice but has become a necessity. Crime analysis can be performed by collecting crime-related data from various sources such as newspapers, online sources, etc. Then the collected data can be analyzed with the help of efficient algorithms thus drawing useful conclusions regarding the safety of a particular area, crime trends, and other such useful information. According to Crime Records Bureau crimes like burglary, arson etc have been decreased while crimes like murder, sex abuse, gang rape etc 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 . Since the availability of criminal data or records is limited we are collecting crime data from various sources like web sites, news sites, blogs, social media, RSS feeds etc. This huge data is used as a record for creating a crime record database. So the main challenge in front of us is developing a better, efficient crime pattern detection tool to identify crime patterns effectively.

2. Literature Survey

Various techniques have been used by authors such as Random forest, Prophet, Linear Regression, etc. for crime analysis. Some of the approaches were used for crime prediction while some aimed crime detection. Many researches have been done which address this problem of reducing crime and many crime-predictions algorithms has been proposed. The prediction accuracy depends upon on type of data used, type of attributes selected for prediction.

2.1 Technique Category One

A Machine Learning Technique

To forecast crime using location and time data. The dataset does not use neighborhood cleaning rules. The author has used San Francisco's crime records from 2003 – 2015 to classify a crime by depending on its time and location. The author used different classification models such as Prophet, Random Forest and K-nearest-neighbour to depict the crime.

1. Technique One

Although there are several drawbacks, the model can be used to generate crime location on a geographical map which will be easier for the police to stop the casualties. Imbalanced classes are one of the major obstacles to reach a better result. The author was unable to find the major socioeconomic cause of the crime

2. Technique Two

As a result of using under sampled data, accuracy of 81.93% is achieved using Adaboost decision which is defined as the best classification model by the author as it brought the highest accuracy as compared to other algorithms of machine learning.

1.1 Technique Category Two

1. **K means clustering Technique**

K means clustering algorithm that can be used to predict the location of crime so that preventive measures can be taken by the police to prevent it, depending on the location. K mean partitions the data into clusters or groups based on the means and when mean value becomes constant in next iteration, algorithm finishes. A participant shared his approach on Kaggle in which the clustering technique is used to convert the dataset into groups related to crime, a group of dense clusters depicts large crime-prone areas and vice versa.

1. **Proposed Technique**

In the suggested approach, the data can be analyzed to depict the most frequent time a crime occurs, so that it may help the Police department of the nation to take proper prevention measures. Although, K means algorithm is fast but it has some major drawbacks like choosing the initial value of K manually makes this algorithm dependent on initial values. If the clusters are showing diversity in size and densities, it needs to be generalized.

2. **Hybrid Approach On**

In the proposed model, all the factors that affect the crime rate are found. Working on these factors, authorities can reduce the crime rate at their location. Correlated factors are found in the dataset used and also the work done includes designing the model which helps in predicting the crime rate using the aforementioned factors.

1. **Implemented System**

The proposed system has been divided into four major steps as follows,

A. Data Collection:

In the data collection step, data is gathered from various sources like websites, official police reports, social media, etc. The gathered data is stored into a database for further processing. During the process of data collection, various types of data were given different priority under the umbrella of crime data such as primary data like crime type, location, time, day, date, and secondary data like location type. Primary data is the most valuable data as it provides valuable insight and has a major contribution in the analysis of data here crime type can signify whether the recorded crime comes under the category of theft, murder, sex offense, etc. Location and time provide details regarding the place where the crime occurred the day month year etc. Not all the recorded crime is always a success thus the status of crime provides valuable insight into whether the recorded crime was an attempt or a success.
Secondary data is the data that can provide valuable insight in some cases however this data can contain anomalies or can be inaccurate. Such types of data in this case are the data related to Location Type which can be residential, commercial, etc. Literacy levels, etc.

B. Data Pre-processing:

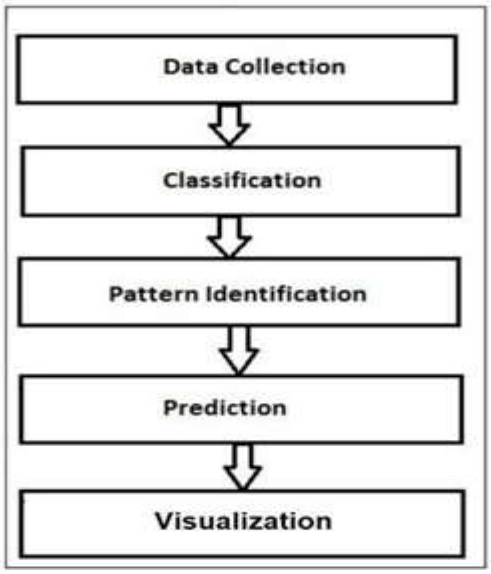
In this step all the null values are removed. The categorical attributes are converted into numeric using OneHotEncoder which can be understood by the models. OneHotEncoder Encodes categorical integer features as a one-hot numeric array. Its Transform method returns a sparse matrix if sparse=True, otherwise it returns a 2-d array. There exist some samples which are considered to be outliners, those samples have been removed after checking location of each point and if not in San Francisco range then it was removed. There also exist two features Descript and Resolution are considered redundant as they do not exist in testing values so they were removed.

C. Pattern Identification:

This step is the example of recognizable growth of crime where patterns and examples in collected crime data need to be distinguished. Climatic conditions, region affectability, outstanding occasions, nearness of criminal gatherings zones, and properties are taken relating to every area. In order to find the criminal hotspots of a particular place, when a new crime will occur, and when a similar kind of event will take place at the same place then that place can be stated as the zone of the possibility of that crime. The police can maintain appropriate distance from events by keeping a watch in that zone, CCTV provisions, settling Robert alerts, patrolling, and so on. If a certain crime happened and again a similar type of crime is expected in the same area then that place happens to be the likelihood for crime in that place.

D. Prediction

For prediction we are using the decision tree concept. A decision tree is similar to a graph in which an internal node represents a test on an attribute, and each branch represents the outcome of a test. The main advantage of using a decision tree is that it is simple to understand and interpret. The other advantages include its robust nature and also it works well with large data sets. This feature helps the algorithms to make better decisions about variables . Corresponding to each place we build a model. So for getting the crime prone areas we pass current date and current attributes into the prediction software. The result is shown using some visualization mechanisms.



00

Figure3.1 Proposed System

Existing System Architecture

Finding the patterns and trends in crime is a challenging factor. To identify a pattern, crime analysts takes a lot of time, scanning through data to find whether a particular crime fits into a known pattern. If it does not fit into an existing pattern then the data must be classified as a new pattern. After detecting a pattern, it can be used to predict, anticipate and prevent crime. The reason for choosing this method is that we have only data about the known crimes we will get the crime pattern for a particular place. Therefore, classification techniques that rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Also the nature of crimes change over time, so in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

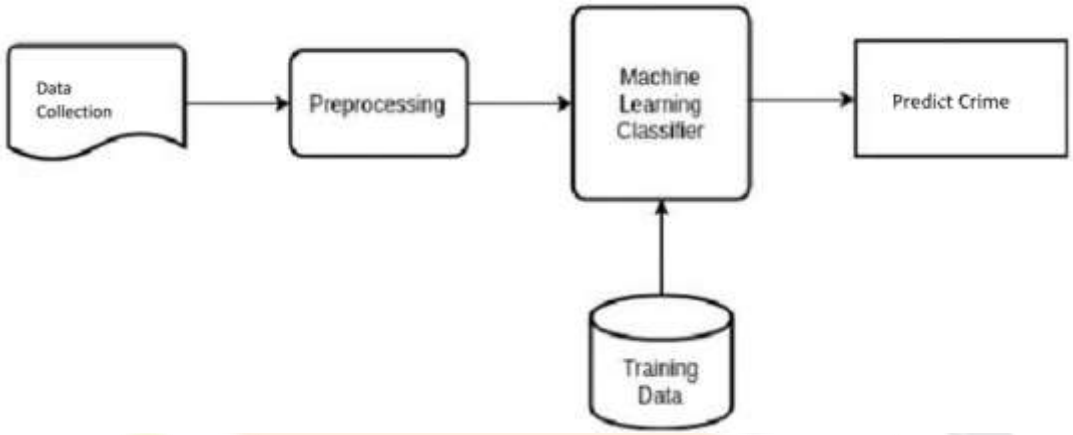


Fig. 3.2 Existing system architecture used for Content -based Systems

3.1.1 Proposed System Architecture

The previous sections discussed the strengths and weaknesses of existing system. In order to achieve better domain results, researchers combined both techniques to build Hybrid domain systems, which seek to inherit vantages and eliminate disadvantages.

In general, hybrid recommenders are systems that combine multiple recommendation techniques together to achieve a synergy between them. Although there exist a number of recommendation approaches that are practical to merge, our work will mainly focus on the combination of CF and CBF techniques. The proposed architecture is shown in Figure 3.3

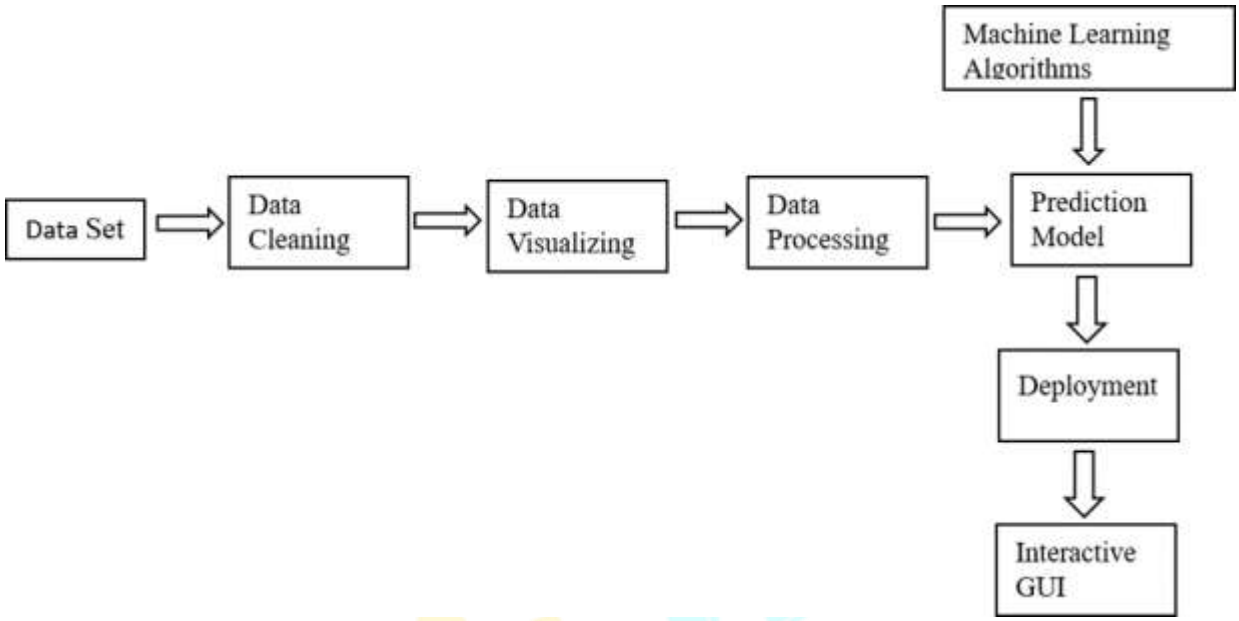


Fig. 3.3 Proposed system architecture

3.1.2 Implementation Details

Data is the backbone of any system and data collection is the means to obtain this data. Here we are collecting crime related data in order to draw useful conclusions from well established facts. Starting with the collection the world wide web enables us to access crime data easily. However it is important to identify the reliable sources of crime data as well as the various secondary inputs that need to be collected. Here for data collection we have considered three processes : data collection from online newspaper sources, crowdsourcing and Official FIRs. Below figures 3 illustrate the process of collecting data from Mumbai Mirror's online website.

3.2.1 Algorithms:

Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Random Forest:

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

K-Means:

Unsupervised Machine Learning learning is the process of teaching a computer to use unlabeled, unclassified data and enabling the algorithm to operate on that data without supervision. Without any previous data training, the machine's job in this case is to organize unsorted data according to parallels, patterns, and variations. The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another.

Naive Bayes:

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.

3.2.2 Use Case Diagram / Activity 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 factors, 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.

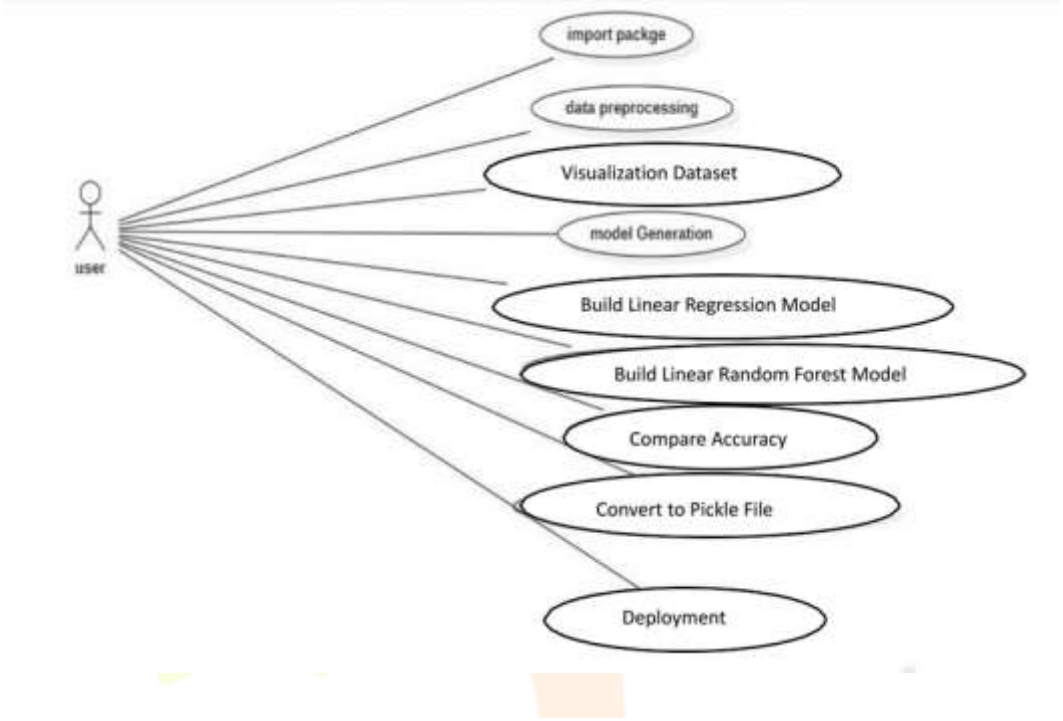


Fig: 3.4 Use Case Diagram

Activity Diagram

The process flows in the system are captured in the activity diagram. Similarity a state diagram, an activity diagram also consists of activities, actions, transitions,initial and final states, and guard conditions.

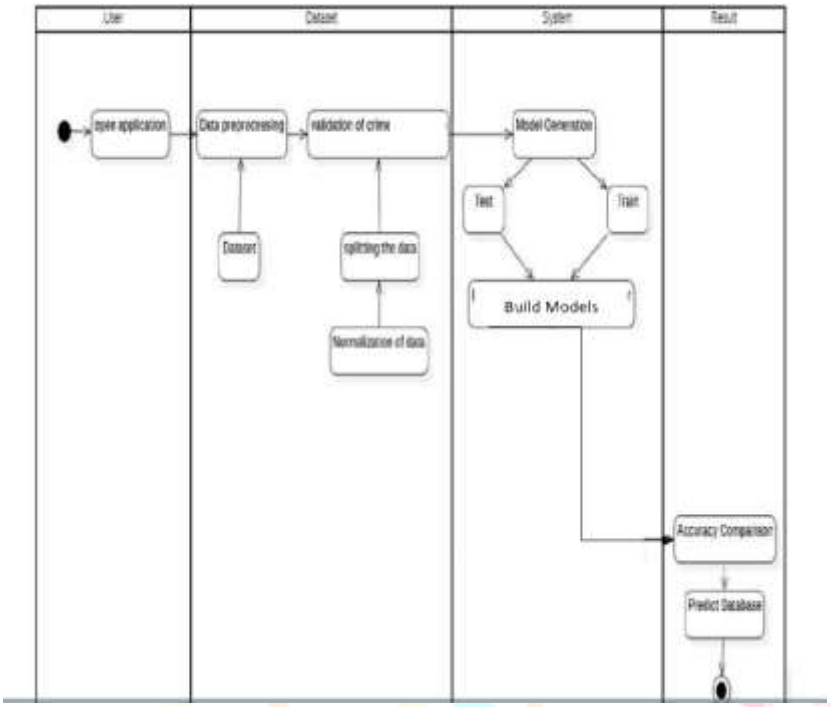


Fig: 3.5 Activity Diagram

3.2.3 Sample Dataset Used

The crime dataset is extracted from primary data collection based on field work. This dataset consists of about 200000 in 10 rows details. The key features such as Years, Months, Crime Type, Crime Areas, and Months are selected from the dataset as the system input features.

.

Table 3.1 Sample Dataset Used for Experiment

offence	reportedy	reportedmc	reportedd	reportedh	occurrenc	occurrenc	occurrenc	occurrenc	occurrenc	occurrenc	MCI
Assault	2015	December	Friday	3	2015	December	18	352	Friday	3	Assault
Assault	2015	August	Monday	22	2015	August	15	227	Saturday	21	Assault
B&E	2015	August	Tuesday	14	2015	August	16	228	Sunday	16	Break and
B&E	2015	December	Friday	13	2015	Novembe	26	330	Thursday	13	Break and
Assault	2015	December	Friday	19	2015	December	18	352	Friday	19	Assault
B&E	2015	December	Friday	14	2015	April	10	100	Friday	10	Break and
Assault	2015	December	Wednesd	18	2015	December	14	348	Monday	21	Assault
Assault	2015	December	Wednesd	18	2015	December	14	348	Monday	21	Assault
Assault	2015	December	Wednesd	18	2015	December	14	348	Monday	21	Assault
Robbery - Taxi	2015	November	Tuesday	2	2015	Novembe	17	321	Tuesday	1	Robbery
Assault - Resist	2015	November	Wednesd	5	2015	Novembe	18	322	Wednesd	5	Assault
Assault	2015	December	Tuesday	15	2015	December	1	335	Tuesday	15	Assault
Assault	2015	December	Tuesday	22	2015	December	1	335	Tuesday	22	Assault
B&E	2015	December	Tuesday	6	2015	December	8	342	Tuesday	6	Break and
Assault	2015	December	Tuesday	9	2015	December	8	342	Tuesday	9	Assault
Assault	2015	December	Friday	13	2015	December	11	345	Friday	13	Assault
Assault	2015	December	Wednesd	16	2015	December	16	350	Wednesd	16	Assault
Assault	2015	December	Wednesd	19	2015	December	16	350	Wednesd	17	Assault

3.2.4 Hardware and Software Specifications

Processor	2 GHz Intel
HDD	180 GB
RAM	2 GB

Table 3.2 Hardware details

3.1 Evaluation Metrics

For evaluating classification models that were implemented for the purpose of classification and prediction. The metrics used are accuracy, f beta-score. Precision is a measure which identifies positive cases from all the predicted cases.

Next is recall it measure which correctly identifies positive cases from all the actual positive cases.

$$rc = \frac{tv}{(tv + fnv)} \; q = \frac{tv}{(tv + fv)}$$

Accuracy is one of the most commonly used metric which measure all the correctly identified value without caring about the wrongly identified values. So, instead of using accuracy the measure that is used to check the performance is F-beta score.

$$accuracy = \frac{tv + tnv}{(tv + fv + tnv + fnv)}$$

F-beta Score is the harmonic mean of Recall and precision which gives a better measure of incorrectly classified cases than that of Accuracy Metric.

$$F - betascore = \frac{(\hat{} * (rc * q))}{(rc + q)}$$

Here tv stands for true positive, fv stands for false positive, fnv stands for false negative, tnv stands for true negative, rc stands for recall and q stands for precision.

4. Result and Applications

4.1 Output:

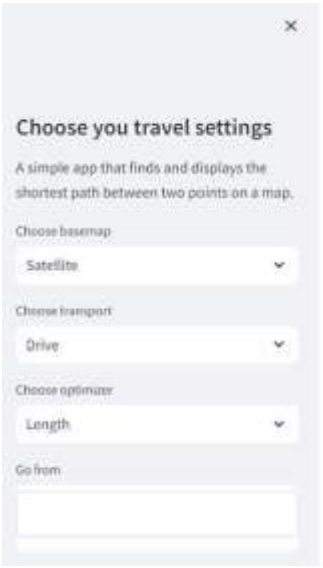
4.1.1 HomePage

4.1.2 About Page

4.1.3 Expert Analysis Page

4.1.4 Crime Prediction Page

4.1.5 Map/Navigation page with Police Station Marked along with phone number



4.2 Applications:

There are various applications of this domain system. The application is listed here. Social:

Police Analytics

Police departments are increasingly using predictive algorithms to determine "hot spot" potential crime areas. PredPol utilizes commonly-understood patterns for when, where, and how crimes occur, and formalizes those patterns using an algorithm that predicts locations where a crime is likely to occur in the near future. Studies show that that crime tends to be geographically concentrated, but that these “hot spots” are often dispersed throughout a city. Using a machine learning model originally built to predict earthquakes, PredPol uses location, timing, and type of crime as inputs

Technical:

Crime forecasting

Crime forecasting refers to the basic process of predicting crimes before they occur. Tools are needed to predict a crime before it occurs. Currently, there are tools used by police to assist in specific tasks such as listening in on a suspect’s phone call or using a body cam to record some unusual illegal activity. Below we list some such tools to better understand where they might stand with additional technological assistance. One good way of tracking phones is through the use of a stingray, which is a new frontier in police surveillance and can be used to pinpoint a cell phone location by mimicking cell phone towers and broadcasting the signals to trick cellphones within the vicinity to transmit their location and other information.

5. Summary

In this report, The implementation has been done in Python language. Here, we find out which state has more or fewer criminals based on each cluster’s values. It is really helpful for the authorities to beware of criminal incidents. The result of the optimized k-means algorithm is efficient and provides improved accuracy of the final cluster reduced the number of iterations. In the future, the result of crime analysis can be used to make various strategies for crime control and the optimal deployment of resources in crime avoidance. 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.

6. References

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pent-land, “Once upon a crime: towards crime prediction from demographicsand mobile data”,IEEE,Proceedings of the 16th international conferenceon multimodal interaction, 2014, pp. 427-434.

[2] Ubon Thansatapornwatana,”A Survey of Data Mining Techniques for-Analyzing Crime Patterns”, Second Asian Conference on DefenseTechnology ACDT, IEEE,Jan 2016, pp. 123–128.

[3] H. Adel, M. Salheen, and R. Mahmoud, ”Crime in relation to urbandesign. Case study: the greater Cairo region,” Ain Shams Eng. J., vol.7, no. 3, pp. 925-938, 2016.

[4] J. L. LeBeau, ”The Methods and Measure of Centrography and thespatial Dynamics of Rape” Journal of Quantitative Criminology, Vol.3,No.2, pp.125-141, 1987.

[5] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, FabioPianesi, Alex Pentland.”Once Upon a Crime: Towards Crime Predictionfrom Demographics and Mobile Data”.

[6] Sai Tarlekar: Geogriphical Crime rate prediction,sep-2021

[7] Amrita Vishwa,Devan M.S:Crime analysis and prediction using data mining,March 2020

[8] Krishnendu S.G, Lakshmi P.P:Crime Analysis and Prediction using Optimized K-Means Algorithm, sep-2020

[9] Shiju sathyadevan m.s, surya gangadharan: crime analysis and prediction using data mining,in networkssoft computing(icnsc),(sep 2020) first international conference.