ISSN: 2366-1313

CRIME RATE PREDICTION & ANALYSIS USING K-MEANS CLUSTERING ALGORITHM

[1]Mrs. Rojaramani. Adapa, [2]K.Shiva shankar, [3]K.Douglas enosh, [4]V.Gopinath

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

rojatkrec.cse@gmail.com

[2]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
shivashankar2572@gmail.com

[3]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
dkorada508@gmail.com

[4]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
gopinathvaddagani@gmail.com

Abstract: *Crimes have a negative effect on any society both socially and economically. Law enforcement bodies face numerous challenges while trying to prevent crimes. We propose a Crime Data Analytic Platform (CDAP) to assist law enforcement bodies to perform descriptive, predictive, and prescriptive analysis on crime data. CDAP has a modular architecture where each component is built separately from each other. CDAP also supports plugins enabling future feature expansions. The platform can ingest any crime dataset which has the required attributes to map dataset to attributes required by the platform. It can then analyze them, train models, and then visualize data. We demonstrate the utility of the platform by visualizing spatial and temporal relationships in a set of real-world crime datasets. Predictive capabilities of the platform are demonstrated by predicting crime categories, for which a machine learning approach is used. To construct a model Nave Bayesian, Random Forest Classifier, and Multi-layer Perceptron Network classification algorithms are provided. Identification of optimized police district boundaries and allocating patrol beats are used to demonstrate the prescriptive analytics capabilities of the tool. Heuristic-based clustering approach was taken to define police district boundaries in a way that the identified districts have equitable population distribution with compact shape. The resulting districts are then evaluated on inequality of population and the compactness using Gini Coefficient and Isoperimetric Quotient. Another heuristic-based approach was taken to define new police patrol beats to be optimized on equitable workload distribution, compactness, and minimizing response time for new police patrol beats.*

ISSN: 2366-1313

Keywords: *Cluster, Crime Analysis and Rapid miner, Crime Data Analytic Platform.*

I. INTRODUCTION

In present scenario criminals are becoming technologically sophisticated in committing crime and one challenge faced by intelligence and law enforcement agencies is difficulty in analyzing large volume of data involved in crime and terrorist activities therefore agencies need to know technique to catch criminal and remain ahead in the eternal race between the criminals and the law enforcement. So appropriate field need to choose to perform crime analysis and as data mining refers to extracting or mining knowledge from large amounts of data, data mining is used here on high volume crime dataset and knowledge gained from data mining approaches is useful and support police forces. To perform crime analysis appropriate data mining approach, need to be chosen and as clustering is an approach of data mining which groups a set of objects in such a way that object in the same group is more similar than those in other groups and involved various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. In this paper k means clustering technique of data mining used to extract useful information from the high-volume crime dataset and to interpret the data which assist police in identify and analyse crime patterns to reduce further occurrences of similar incidence and provide information to reduce the crime. In this paper k mean clustering is implemented using open-source data mining tool which are analytical tools used for analysing data. Among the available open source data mining suite such as R, Tanagra ,WEKA ,KNIME ,ORANGE ,Rap id miner. k means clustering is done with the help of rapid miner tool which is an open source statistical and data mining package written in Java with flexible data mining support options. Also, for crime analysis dataset used is Crime dataset an offences recorded by the police in England and Wales by offence and police force area from 1990 to 2011-12. In this paper homicide which is crime committed by human by killing another human is being analysed.

Crimes are a social nuisance and it has a direct effect on a society. Governments spend lots of money through law enforcement agencies to try and stop crimes from taking place. Today, many law enforcement bodies have large volumes of data related to crimes, which need to be processed to turn into useful information. Crime data are complex because they have many dimensions and in different formats, e.g. Most of them contain string records and narrative

ISSN: 2366-1313

records. Due to this diversity, it is difficult to mine them using off the shelf, statistical and machine learning data analytics tools. It is the primary reason for the lack of a general platform for crime data mining. While there are some propitiatory platforms to predict and analyze crime data, they are focused only on certain areas of crimes, not extensible, and do not provide an API to integrate with other tools . Moreover, the same tool cannot be used for the analysis and well as planning such as patrol beads and district boundaries.

MOTIVATION

To solve a case based upon a particular data there should be a thorough investigation and analysis that is to be done internally. With the amount of crime data that is present in India currently the analysis and decision making of these criminal cases is too difficult for the officials. Identifying this a major problem this paper concentrates on creating a solution for the decision making of crime that is committed. the vehicle starts driving on its own. An autonomous driving vehicle performs various actions to arrive at its destination, repeating the steps of recognition, judgment and control on its own. High or increased crime-levels make communities decline, as crimes reduce house prices, neighbourhood satisfaction, and the desire to move in a negative manner. To reduce and prevent crimes it is important to identify the reasons behind crimes, predict crimes, and prescribe solutions. Due to large volumes of data and the number of algorithms needed to be applied on crime data, it is unrealistic to do a manual analysis. Therefore, it is necessary to have a platform which is capable of applying any algorithm required to do a descriptive, predictive, and prescriptive analysis on large volumes of crime data. Through those three methodologies law-enforcement authorities will be able to take suitable actions to prevent the crimes. Moreover, by predicting the highly likely targets to be attacked, during a specific period of time and specific geographical location, police will be able to identify better ways to deploy the limited resources and also to find and fix the problems leading to crimes. Several applications are already developed for crime analysis. Most of these tools are developed to help the police to identify different crime patterns and even to predict criminal activities.

They are complex software which needs a lot of training before use. Designing a tool which is easy to use with minimal training would help law-enforcing bodies all around the world to reduce crimes.

II. LITERATURE SURVEY

Data mining and machine learning have become a vital part of crime detection and

ISSN: 2366-1313

prevention. In this research, we use WEKA, an open source data mining software, to conduct a comparative study between the violent crime patterns from the Communities and Crime Un normalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com. 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.

Crimes and its Effect on the Society

A crime can be defined as any action or omission that violates a law, which results in a punishment. Usually what constitutes a crime depends on the government bodies and laws that are in existence in those places. To understand the nature of crimes, one has to understand not only its spatio-temporal dimensions, but also the nature of the crime, the victim-offender relationship, role of guardians, and the history of similar incidents. Regardless of the reasons why crimes take place, they put a strain on the communities, towns, and cities. Usual monetary costs associated with them include cost of policing crime and prosecuting those who commit crimes. Non-monetary costs consist of social costs, where they affect the quality of life, mental health, and physical security of people living in those areas.

Criminology Theories

According to John and David, theories of crimes can be divided into two categories namely, those that seek to explain the development of criminal offenders and those that seek to explain the development of criminal events.

Criminology has been mainly developed through theories and research on offenders. Only recently it has begun to explain the crimes rather than the criminality of people involved in it. Criminology consists of many theories that explain how and why some offenders act in the way they do. Following are some of theories that explain how places are associated with crimes.

1. Rational Choice Rational Choice suggests that offenders will select targets and define means to achieve their goals in a manner that can be explained. Further it can be explained as that human actions are based on rational decisions, that is they are informed by probable consequences of that action.

ISSN: 2366-1313

2. Routine Activity Theory This theory explains the occurrence of crimes as the result of several circumstances. Namely, a motivated offender, a desirable target, target and offender must be at the same place at the same time, and lastly absent of other types of controllers, intimate handlers, guardians, and place managers.

3. Crime Pattern Theory This theory combines the above two theories and goes on to say that how targets come to the attention of offenders is influenced by distribution of crime events over time, space, and among targets. An offender will come to know of criminal opportunities while engaging in their day-to-day legitimate work. So, a given offender will only know about a subset of available targets. The concept of place is essential to crime pattern theory.

CRIME ANALYSIS

Crime analysis is a difficult task, as it requires both collection and analysis of large volumes of data. For example, Brown states that Richmond city in the USA has approximately 100,000 criminal records per year. Given the data volume and need to apply different algorithmic techniques forbids manual analysis. Whereas an automated analysis of such a rich data set could identify complex crime patterns and assist in solving crimes faster. Data mining techniques can be used in law and enforcement for crime data analysis, criminal career analysis, bank fraud analysis, and analysis of other critical problems Some of the traditional data mining techniques are association analysis, classification and prediction, cluster analysis, and outlier analysis, which identify patterns in structured data. Using criminology theories along with modern technology would help to identify crime patterns quickly and efficiently. To simplify the workload a crime data analytic platform could be used which would help in simplifying the process.

Using machine learning algorithms to analyze crime data.

Data mining and machine learning have become a vital part of crime detection and prevention. In this research, we use WEKA, an open-source data mining software, to conduct a comparative study between the violent crime patterns from the Communities and Crime Un normalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com. 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

ISSN: 2366-1313

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,

III. PROPOSED WORK

The proposed system is made on the basis of the research work that is done by going through various such documentations. Nearly all of the crimes are predicting based on the location and the types of crimes that are occurring in those areas. On surveying previous works, Linear Regression, Decision Tree and Random Forest tend to give good accuracy so these models are used in this paper to predict crimes. The dataset used in this paper is fromdata.world.com. The data set contains different types of crimes that being committed in India according to the state and year respectively.

This paper takes types of crimes as input and gives the area in which crimes are committed as output. The data pre-processing involves data cleaning, feature selection, dropping null values, data scaling by normalizing and standardizing. After data preprocessing the data is free of null values which may alter the accuracy of the model significantly and feature selection is used to select only the required features that won't affect the accuracy of the model.

After data pre-processing the models chosen i.e., Logistic Regression, Decision Tree and Random Forest are trained by splitting the data into train and test data. As the output required is a categorical value classification models are used here. Python language is used for data prediction. After literature review there is need to use an open-source data mining tool which can be implemented easily and analysis can be done easily. So here crime analysis is done on crime dataset by applying k means clustering algorithm using rapid miner tool

The procedure is given below:

1. First we take crime dataset

2. Filter dataset according to requirement and create new dataset which has attribute according to analysis to be done

3. Open rapid miner tool and read excel file of crime dataset and apply “Replace Missing value operator” on it and execute operation

4. Perform “Normalize operator” on resultant dataset and execute operation

5. Perform k means clustering on resultant dataset formed after normalization and execute operation

6. From plot view of result plot data between crimes and get required cluster

7. Analysis can be done on cluster formed

Learning to detect patterns of crime

We introduce a novel, robust data-driven regularization strategy called Adaptive

ISSN: 2366-1313

Regularized Boosting (AR-Boost), motivated by a desire to reduce overfitting. We replace AdaBoost's hard margin with a regularized soft margin that trades-off between a larger margin, at the expense of misclassification errors. Minimizing this regularized exponential loss results in a boosting algorithm that relaxes the weak learning assumption further: it can use classifiers with error greater than 1 2 . This enables a natural extension to multiclass boosting, and further reduces overfitting in both the binary and multiclass cases. We derive bounds for training and generalization errors, and relate them to AdaBoost. Finally, we show empirical results on benchmark data that establish the robustness of our approach and improved performance overall.

Introduction Boosting is a popular method for improving the accuracy of a classifier. In particular, AdaBoost is considered the most popular form of boosting and it has been shown to improve the performance of base learners both theoretically and empirically. The key idea behind AdaBoost is that it constructs a strong classifier using a set of weak classifiers.

Crime Analysis using K-Means Clustering

In today's world security is an aspect which is given higher priority by all political and government worldwide and aims to reduce crime incidence. As data mining is the appropriate field to apply on high volume crime dataset and knowledge gained from data mining approaches will be useful and support police force. So In this paper crime analysis is done by performing k-means clustering on a crime dataset using a rapid miner tool.

Crime data mining: a general framework and some examples

A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. Detecting cybercrime can likewise be difficult because busy network traffic and frequent online transactions generate large amounts of data, only a small portion of which relates to illegal activities. Data mining is a powerful tool that enables criminal investigators who may lack extensive training as data analysts to explore large databases quickly and efficiently. We present a general framework for crime data mining that draws on experience gained with the Coplink project, which researchers at the University of Arizona have been conducting in collaboration with the Tucson and Phoenix police departments since 1997.

ISSN: 2366-1313

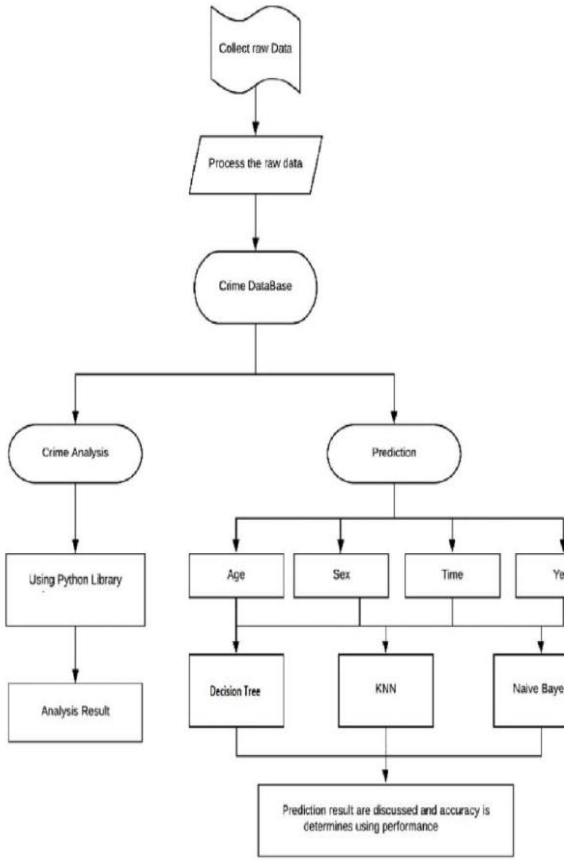


Fig.1 System architecture

IV. RESULTS

Crime Rate Prediction & Analysis using K-Means Clustering Algorithm In this project we are using clustering and regression algorithms to predict crime rate. Clustering algorithm is used to predict HIGH or LOW crime area and regression algorithms are used to forecast future crime rates.

This application consists of two users

1) Admin: admin can login to application using username and password as ‘admin’ and ‘admin’ and after login admin can upload dataset and then train that dataset using machine learning algorithms

2) User: users can enter area name and then enter crime details and then system will predict future crime rate.

ISSN: 2366-1313

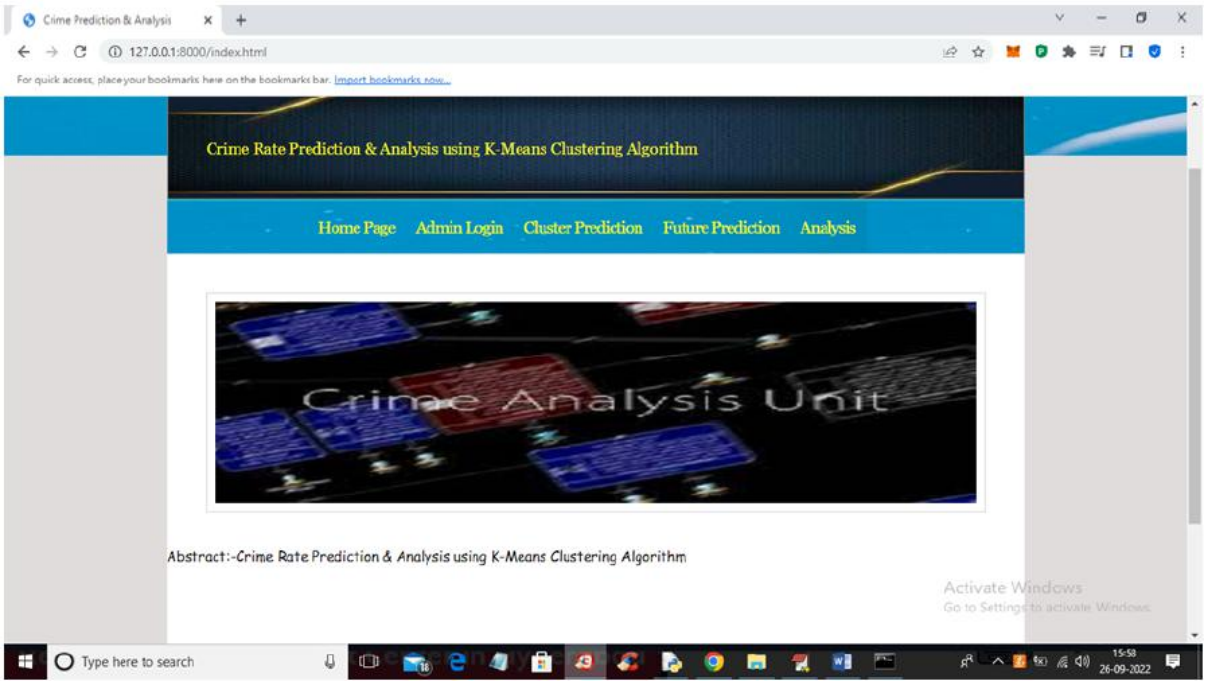


Fig.2 Admin login

In above screen click on 'Admin Login' link to get below login page

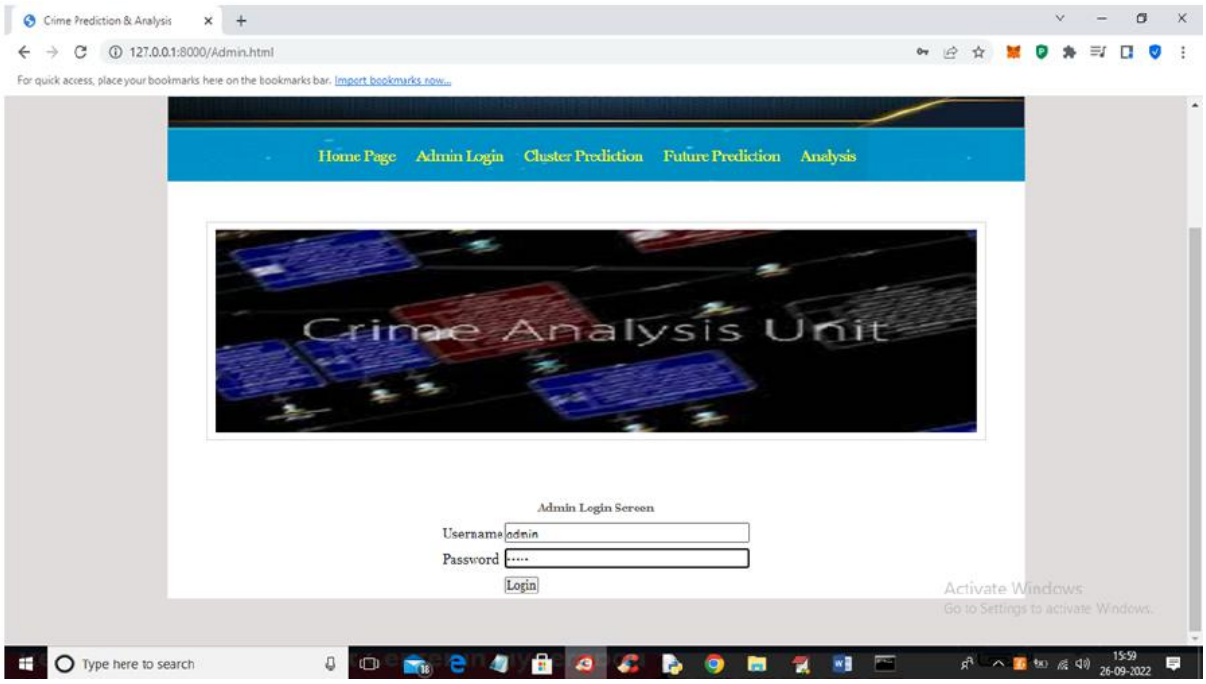


Fig.3 Admin login with his credentials

In above screen admin is login and after login will get below screen In above screen admin is login and after login will get below screen

ISSN: 2366-1313

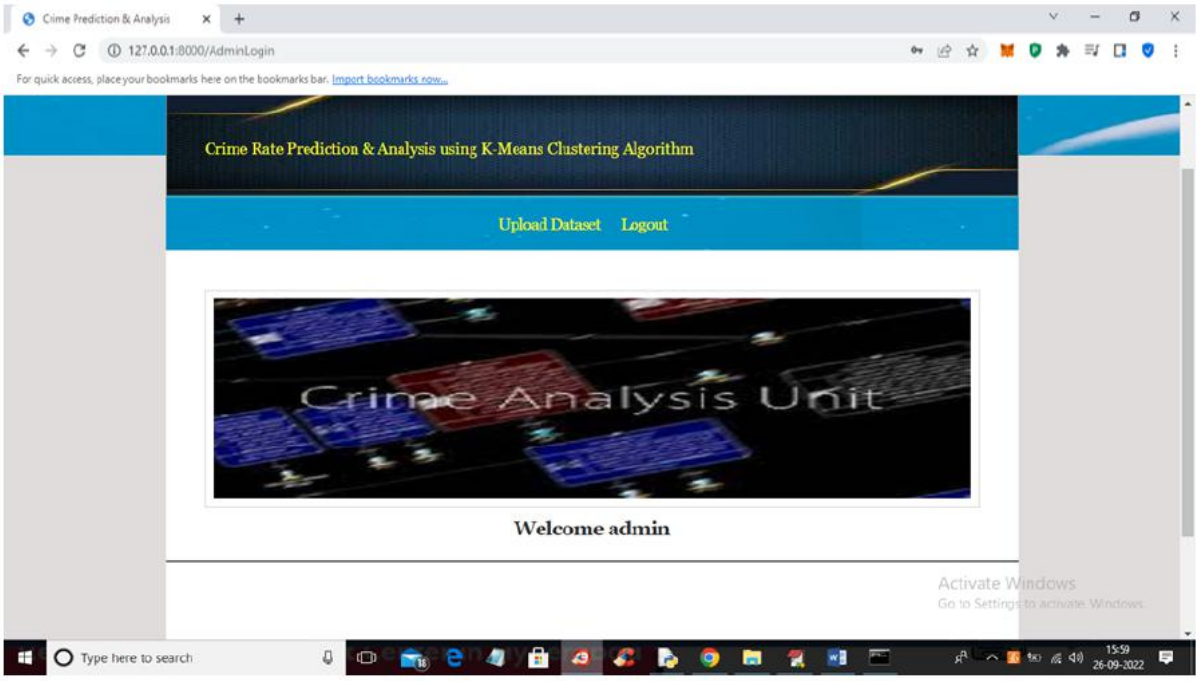


Fig.4 uploading dataset page

In above screen admin can click on 'Upload Dataset' link to upload dataset and then click submit button to load dataset and then train it with machine learning algorithms.

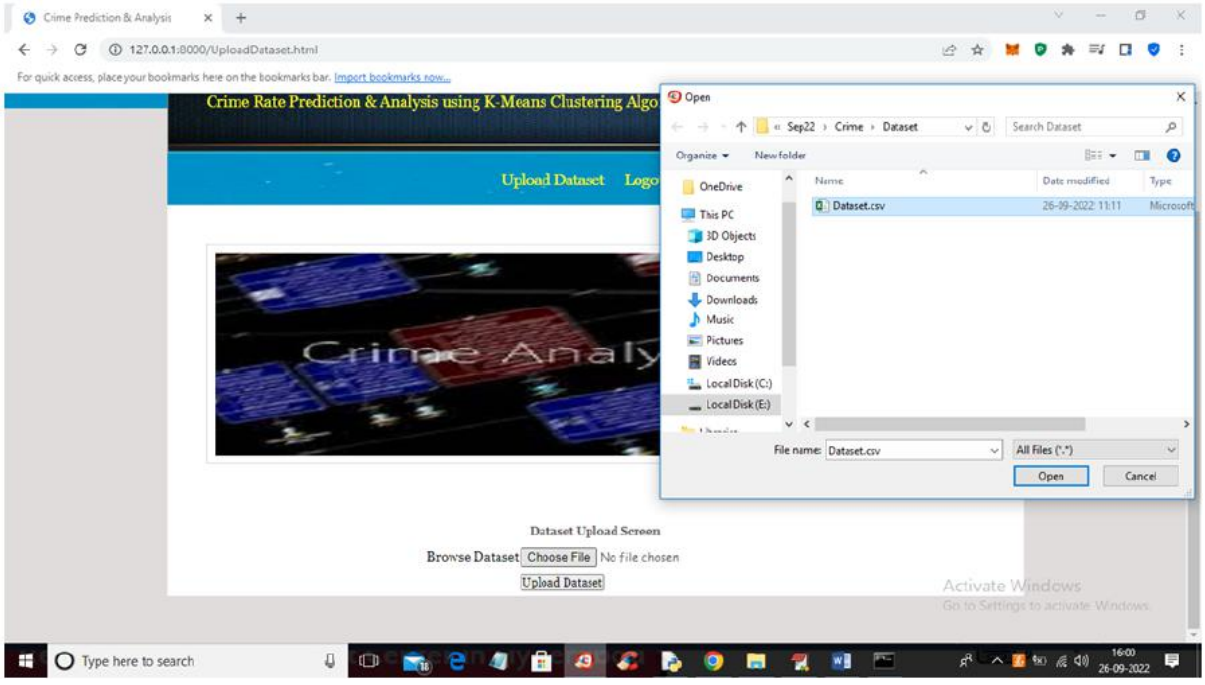
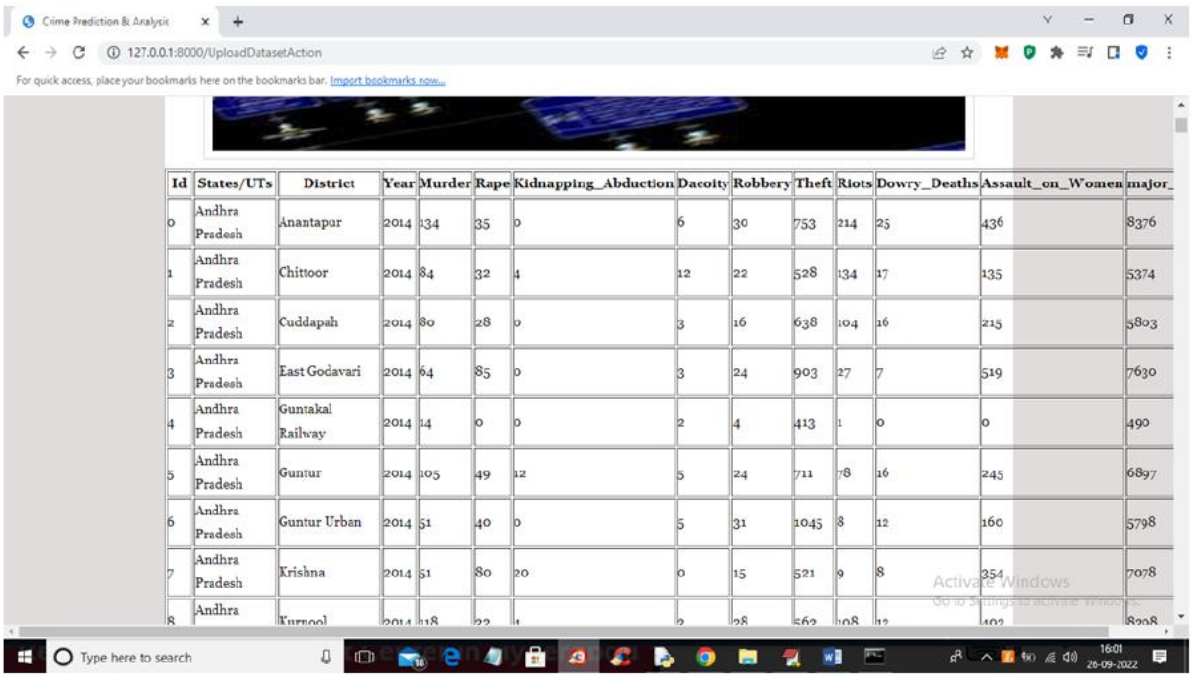


Fig.5 Dataset selected page

In above screen selecting and upload dataset and then click on 'Open' and 'Upload Dataset' button to load and complete training process and get below output

ISSN: 2366-1313



Id	States/UTs	District	Year	Murder	Rape	Kidnapping_Abduction	Dacoity	Robbery	Theft	Riots	Dowry_Deaths	Assault_on_Women	major_
0	Andhra Pradesh	Anantapur	2014	134	35	0	6	30	753	214	25	436	8376
1	Andhra Pradesh	Chittoor	2014	84	32	4	12	22	528	134	17	135	5374
2	Andhra Pradesh	Cuddapah	2014	80	28	0	3	16	638	104	16	215	5803
3	Andhra Pradesh	East Godavari	2014	64	85	0	3	24	903	27	7	519	7630
4	Andhra Pradesh	Guntakal Railway	2014	14	0	0	2	4	413	1	0	0	490
5	Andhra Pradesh	Guntur	2014	105	49	12	5	24	711	78	16	245	6897
6	Andhra Pradesh	Guntur Urban	2014	51	40	0	5	31	1045	8	12	160	5798
7	Andhra Pradesh	Krishna	2014	51	80	20	0	15	521	9	8	354	7078
8	Andhra	Kurnool	2014	118	22	1	2	28	562	108	12	402	8208

Fig.6 In above screen training is completed and then we got all dataset details and now click on 'Logout' link to get below screen

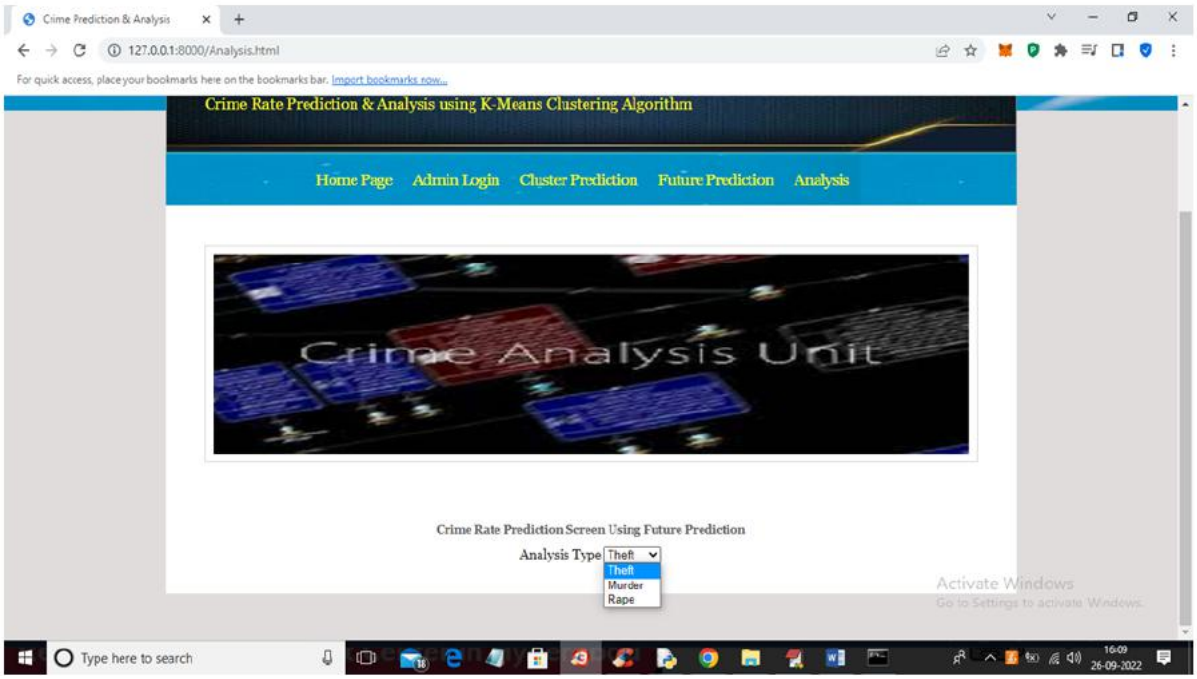


Fig.7 In above screen select the type of analysis and press button to get below graphs

ISSN: 2366-1313

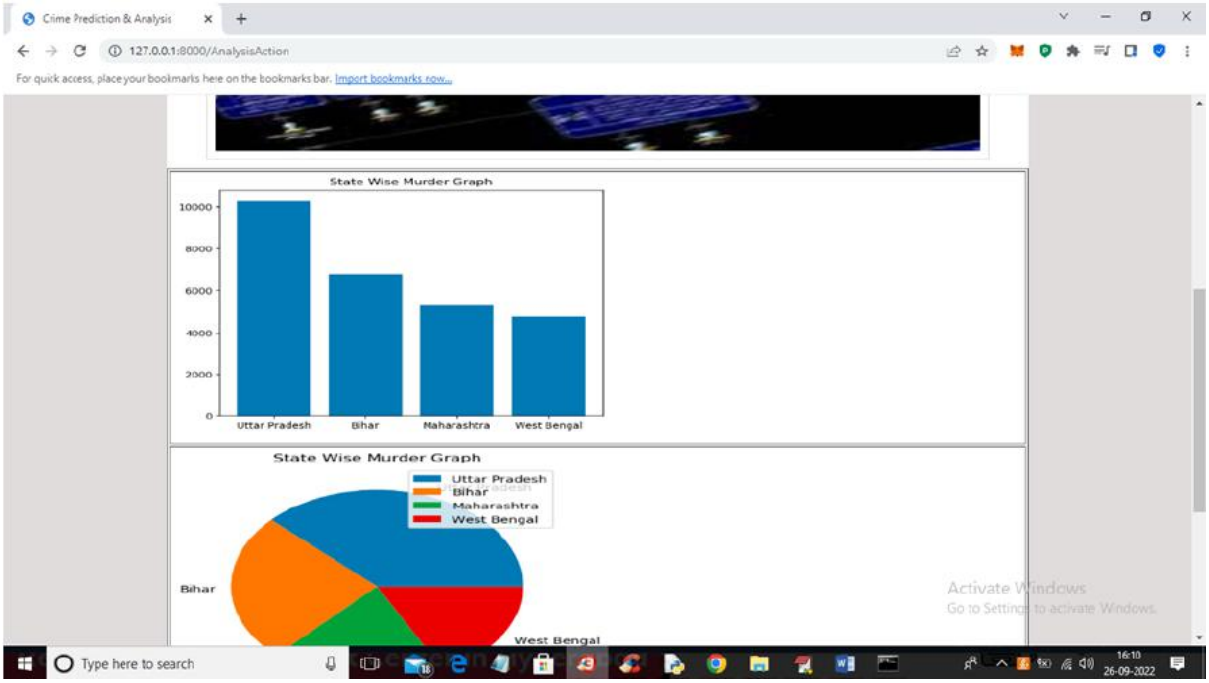


Fig.8 Type of analysis between various states

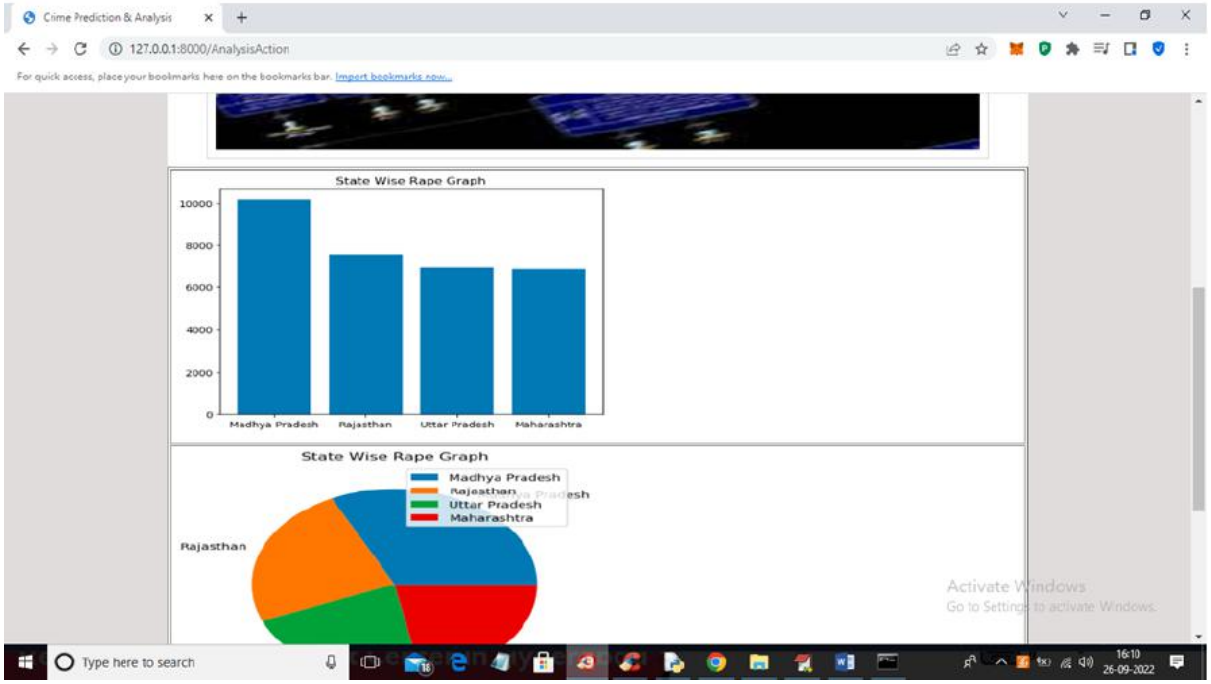


Fig.9 State wise rape graph

V. CONCLUSION

It is clear that basic details of criminal activities in a neighbourhood contain indicators that will be employed by machine learning agents to classify a criminal activity given a location and date. The training agent suffers from imbalanced categories of the dataset, it had been ready to overcome the problem by oversampling and under-sampling the

ISSN: 2366-1313

dataset. This paper presents a crime data prediction by taking the types of crimes as input and giving are in which these crimes are committed as output using Jupyter notebook having python as a core language and python provide inbuilt libraries such as Pandas and Numpy through which the work will be completed faster and Scikit provides all the processes of how to use different libraries providing by the python. Results of prediction are different for different algorithms and the accuracy of Boosted Decision Tree Classifier found to be good with the accuracy of 95.122%.

REFERENCES

[1] S.K.Lodha and A.K.Verma, "Spatio-temporal visualization of urban crimes on a gis grid," Sth ACM Intl. symposium on Advances in Geographic Information Systems, pp. 174-179, 2000.

[2] J. Forgeat. (2015) Data processing architectures lambda and kappa. [Online]. Available: https://www.ericsson.com/research-blog/data-knowledge/ data-processing-architectures-lambda-and-kappa/

[3] C. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," I Ith IEEE Intl. Conf. on Data Mining Workshops, pp. 779-786, 2011.

[4] A. T. Murray, I. McGuffog, J. S. Western, , and P. Mullins, "Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment," British Journal of criminology, vol. 41, no. 2, pp. 309-329, 2001.

[5] R. Krishnamurthy and J. S. Kumar, "Survey of data mining techniques on crime data analysis," International Journal of Data Mining Techniques and Applications, vol. 1, no. 2, pp. 117-120, 2012.

[6] D. E. Brown, *The regional crime analysis program (recap): a framework for mining data to catch criminals," IEEE Intl. Conf. on Systems, Man, and Cybernetics, vol. 3, pp. 2848-2853, 1998.

[7] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder, "Coplink: managing law enforcement data and knowledge," Communications of the ACM, vol. 46, no. 1, pp. 28 34, 2003.

[8] A. Verma, R. Ramyaa, S. Marru, Y. Fan, and R. Singh, "Rationalizing police patrol beats using voronoi tessellations," IEEE Intl. Conf. on Intelligence and Security Informatics (ISI), pp. 165-167, 2010.

[9] J. E. Eck and D. L. Weisburd, "Crime places in crime theory," Crime and place: Crime prevention studies, vol. 4, 2015.

[10] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining:a general framework and some examples," Computer, vol. 37, no. 4, pp.

ISSN: 2366-1313

50-56, 2004.

[11] P. S. Mitchell, "Optimal selection of police patrol beats," Criminal Law and Criminology, vol. 63, p. 577, 1972. 42

[12] Prasadu Peddi (2019), "Data Pull out and facts unearthing in biological Databases", International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.