CRIME RATE PREDICTION & ANALYSIS USING K_MEANS CLUSTERING ALGORITHIM

YANDAMURI RAMASATISH[1], K·RAMBABU[2]
1. PG STUDENT, D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202.
Email id: - ramasatish5445@gmail.com
2. Assistant Professor in DEPARTMENT OF MASTER OF COMPUTER SCIENCE, BHIMAVARAM-534202.
Email id: - kattaramababudnr@gmail.com

ABSTRACT

In India, the crime rate is increasing each day. In the current situation, recent technological influence, effects of social media and modern approaches help the offenders to achieve their crimes. Both analysis and prediction of crime is a systematized method that classifies and examines the crime patterns. There exist various clustering algorithms for crime analysis and pattern prediction but they do not reveal all the requirements. Among these, K means algorithm provides a better way for predicting the results. The proposed research work mainly focused on predicting the region with higher crime rates and age groups with more or less criminal tendencies. We propose an optimized K means algorithm to lower the time complexity and improve efficiency in the result

KEYWORDS: Clustering, k-means Algorithm, Crime

1. INTRODUCTION

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

2. LITERATURE SURVEY AND RELATED WORK

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

Paper-1 Summary: Proposed by Sutapat Thirprungsri

The purpose of this study is to examine the possibility of using clustering technology for continuous auditing. Automating fraud filtering can be of great value to preventive continuous audits. In this paper, cluster-based outliers help auditors focus their efforts when evaluating group life insurance claims. Claims with similar characteristics have been grouped together and those clusters with small population have been flagged for further investigations. Some dominant characteristics of those

clusters are, for example, having large beneficiary payment, having huge interest amount and having been submitted long time before getting paid. This study examines the application of cluster analysis in accounting domain. The results provide a guideline and evidence for the potential application of this technique in the field of audit.

Paper-2 Summary: Proposed by K. Zakhir Hussain

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

3. EXISTING SYSTEM

Crime analysis tool is developed using various distinct data mining methods. It supports the police officers for investigating crimes. Implementing a clustering algorithm on crime datasets enables analysis of crimes . It makes identification and analysis of various criminality trends over the years through their conclusion. The random initial starting points produced by K-means which gives results in the form of cluster that helps in reaching the local optima [8]. So to overcome this problem, the partitioned data along with the data axis with the highest variance for assigning the initial centroid for K- Means clustering was applied. So it is observed that the proposed technique uses a lesser number of iteration thereby reducing the clustering time. Using merge sort, K- means algorithm can be improved for clustering the Hidden Markov Model (HMM).

EXISTING SYSTEM DISADVANTAGES:

1.LESS ACCURANCY

2. LOW EFFICIENY

4. PROPOSED SYSTEM

We are working on Spyder for implementation. Here we use a Spyder 3.7 version. Spyder is an integrated development environment for systematic programming in Python. Here we implemented different packages like matplotlib, numpy, sklearn, pandas, etc. Which helps to plot elbow graph and data frame table using a K-means clustering algorithm? Dataset is collected from Kaggle datasets and import datasets into Spyder in CSV format. We perform normalization for finding the accurate number of clusters (k) using the elbow method. The elbow method performs k- means clustering on the obtained dataset for a range of values of k (2-15) and calculates the SSE. A line chart of the SSE is plotted for each value of k

PROPOSED SYSTEM ADVANTAGES:

1.**HIGH ACCURACY**

2.HIGH EFFICIENC

SYSTEM ARCHITECTURE

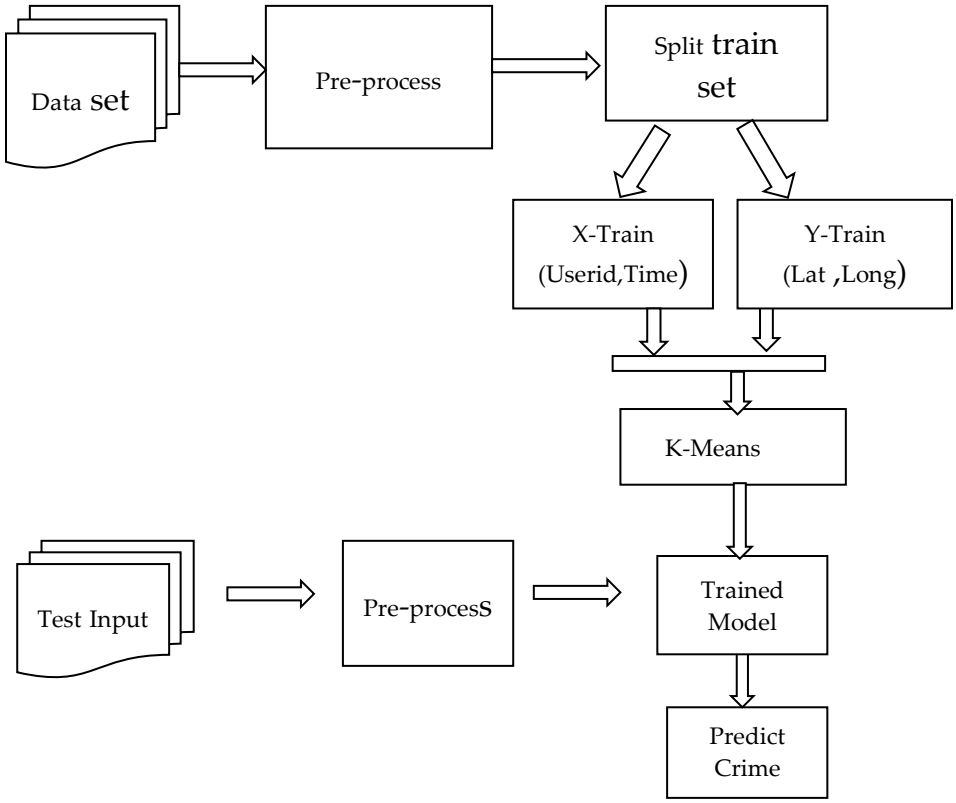


Fig 1: SYSTEM ARCHITECTURE

3 METHODOLOGIES

4.1Modules

1. Load Dataset:

Load data set using pandas read_csv() method. Here we will read the excel sheet data and store into a variable.

2. Split Data Set:

Split the data set to two types. One is train data test and another one is test data set. Here we will remove missing values from the dataset.

3.Train data set:

Train data set will train our data set using fit method. 80% of data from dataset we use for training the algorithm.

4.Test data set:

Test data set will test the data set using algorithm. 20% of data from dataset we use for testing the algorithm.

5.Predict data set:

Predict () method will predict the results. In this step we will predict the ranking of the google play store app.

5. RESULTS AND DISCUSSION SCREEN SHOTS

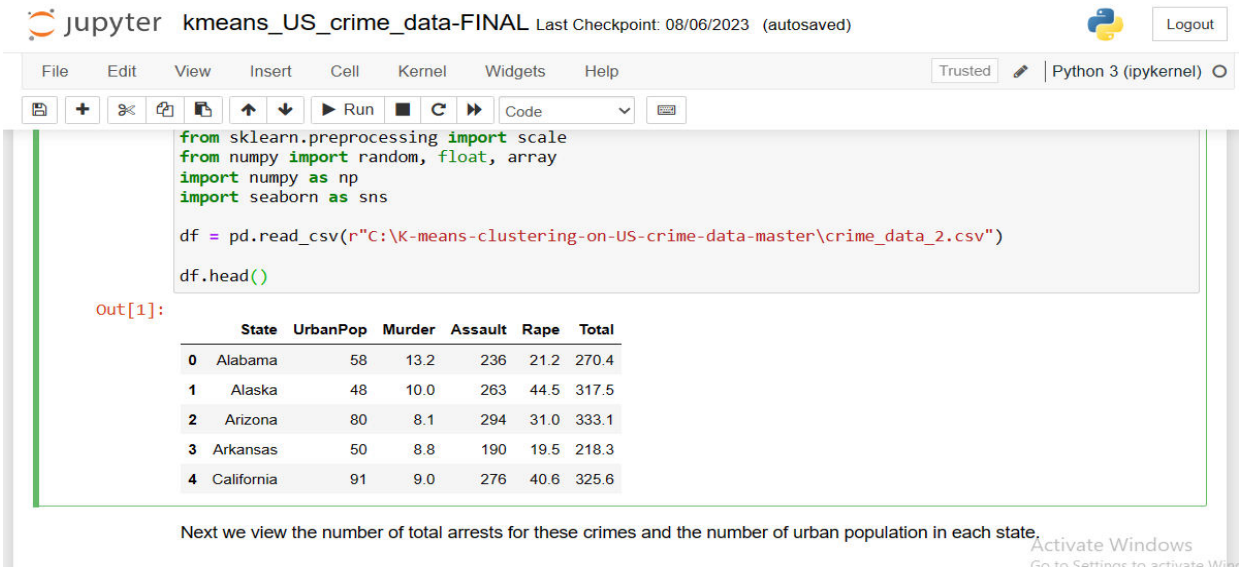


FIG 2:- Next we view the number of total arrests for these crimes and the number of urban population in each state.

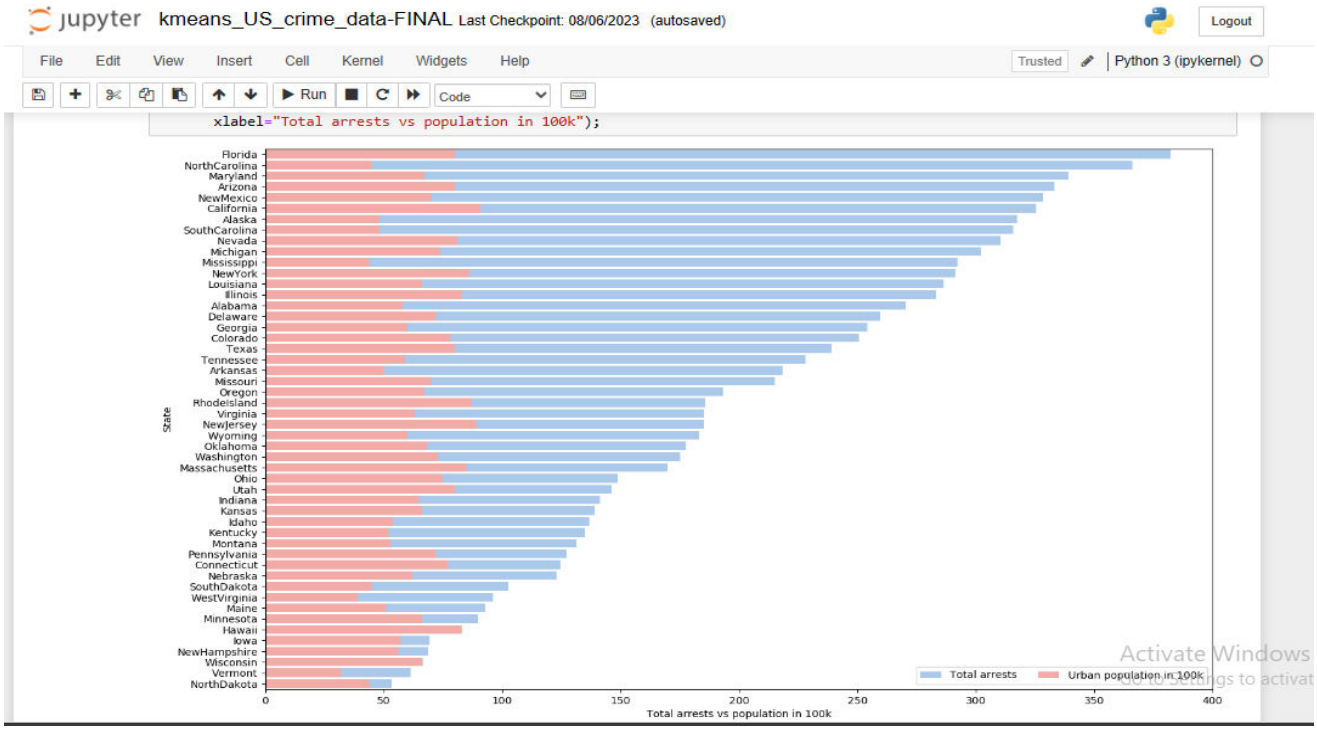


FIG 3 :-Next, we choose the optimal number of clusters using the elbow method by plotting the above table:

The kinks appear to be smoothening out after four clusters indicating that the optimal number of clusters is 4. Next, we divide the data into the chosen number of clusters.

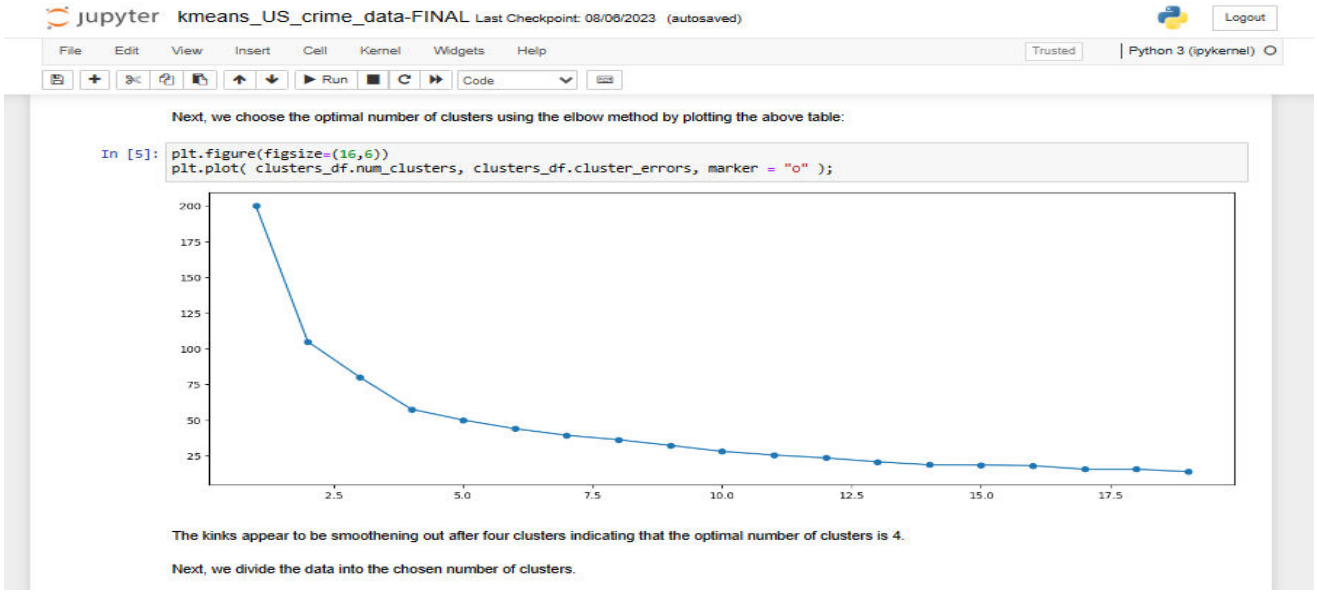


FIG 4 :- Analyzing the data pairwise - UrbanPop & Total We start by looking at the two main variables until digging into separate crime types.

FIG 5 :- Analyzing the data pairwise - Murder & Assault Next up, these two:

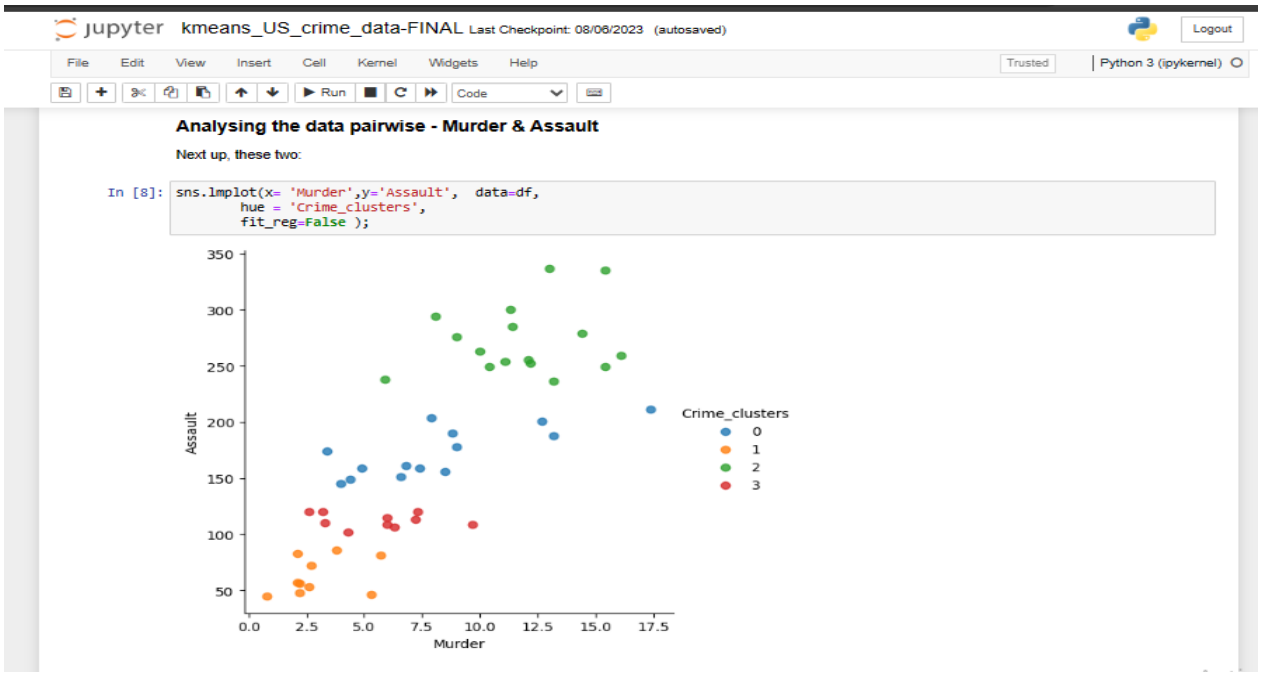


FIG 6 :- ontrary to murders vs. assaults, there is much more spread among the clusters when comparing murders vs. rapes. Some correlation is visible, though; low murder rates in a state seem to indicate lower number of rapes, as well. For the higher rate states, the differences are more scattered

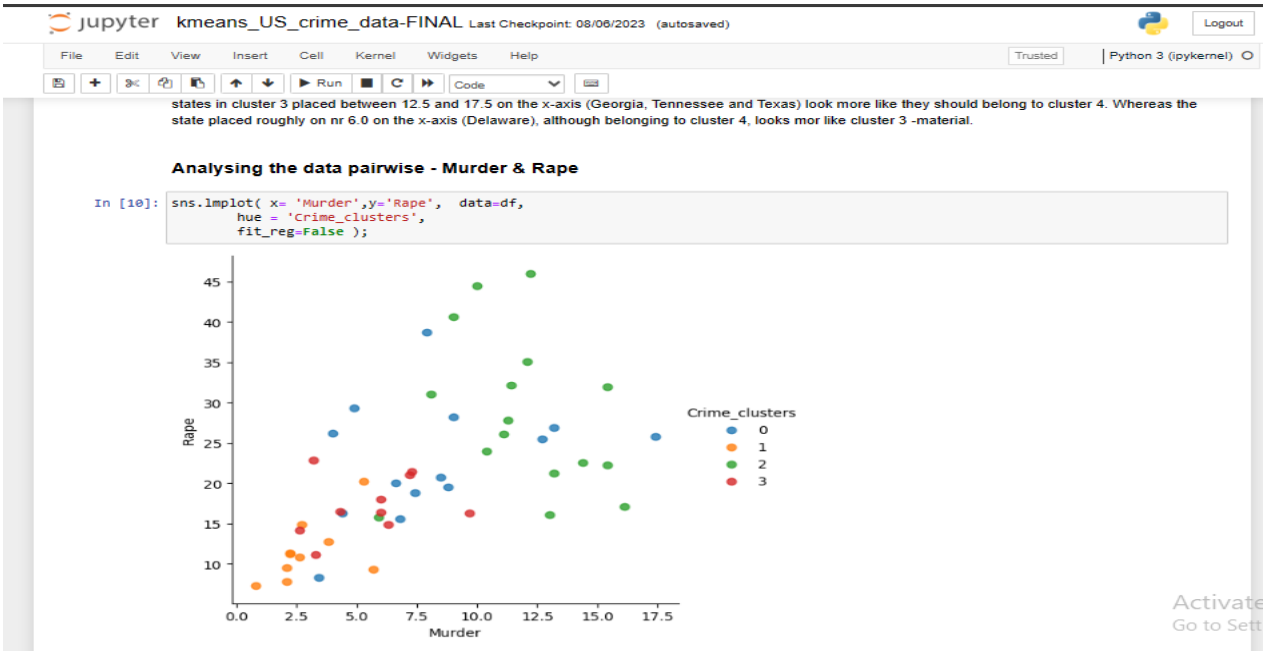
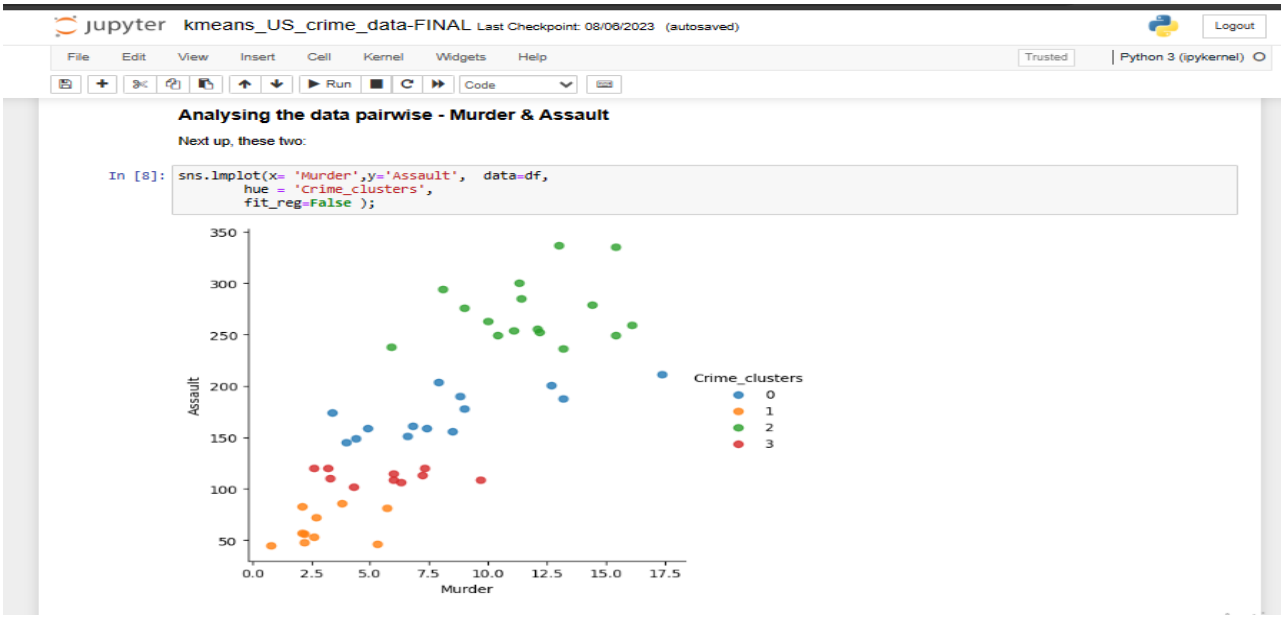


FIG 7:- This is interesting! The table quite well confirms the assumptions regarding variable correlations indicated also by the graphs. For example, murder and assault have the highest correlations, whereas the size of urban



6. CONCLUSION AND FUTURE SCOPE

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

Experimental results prove that application is effective in terms of analysis speed, identifying common crime patterns and crime prone areas for future prediction. From the encouraging results, we believe that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern. As we have applied clustering technique of data mining for crime analysis we can also perform other techniques of data mining such as classification. Also we can perform analysis on various dataset such as enterprise survey dataset, poverty dataset, aid effectiveness dataset, etc.

7. REFERENCES

1. De Bruin ,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J. and Kok,J.N(2006) Data miningapproaches to criminal carrer analysis ,"in Proceedings of the Sixth International Conference on Data Mining (ICDM"06) ,Pp. 171-177.
2. Manish and M. P. GuptaGupta1, B.Chandra1 1,200Information System.
3. Nazlena Mohamad Ali1, Masnizah Mohd2, Hyowon Lee3, Alan F. Smeaton3, Fabio Crestani4 and Shahrul Azman Mohd Noah2 ,2010 Visual Interactive Malaysia Crime News Retrieval System 7 Crime Data Mining for Indian Police.
4. Chung-Hsien Yu, Max W.Ward, Melissa Morabito and Wei Ding,"Crime Forecasting Using Data Mining Techniques", 2011 11th IEEE International Conference on Data Mining Workshops.
5. Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Detecting patterns of crime with series finder. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), 2013.
6. Li Zhang, Yue Pan, and Tong Zhang. Focused named entity recognition using machine learning. In Proceedings of the 27th Annual International.
7. Malathi. A and Dr. S. Santhosh Baboo. Article:an enhanced algorithm to predict a future crime using data mining. International Journal of Computer Applications, 21(1):1–6, May 2011. Published by Foundation of Computer Science.
8. Eibe Frank and Remco R. Bouckaert. Naive bayes for text classification with unbalanced classes. In Proceedings of the 10th European

Conference on Principle and Practice of Knowledge Discovery in Databases, PKDD'06, pages 503–510, Berlin, Heidelberg, 2006. Springer-Verlag.

9. Wikipedia contributors.(9 July 2013), Stanford NLP. [Online].Available :http://www- nlp.stanford.edu/software/dcoref.shtml. Last accessed: 24-Feb-2014, 10:00 AM.
10. Wikipedia contributors.(12 May 2014 at 19:05.), Series Finder. [Online].Available:http://en.wikipedia.org/wiki/Crime_analysis, Last accessed: 12- Feb- 2014, 12:00 PM.