P.D.Patil *et al*, International Journal of Computer Science and Mobile Applications,
Vol.7 Issue. 4, April- 2019, pg. 9-16 **ISSN: 2321-8363**
Impact Factor: 5.515

An Approach for Crime Rate Prediction Using Data Mining

P.D.Patil[1]; P.H.Patil[2]; S.N.Isani[3]; R.R.Jagtap[4]; L.M.Koli[5]; D.B.Shukla[6]

[1,2,3,4,5,6] Department of Computer Engineering
[1]P.S.G.V.P. Mandal's D.N.Patel College of Engineering, Shahada, Maharashtra, India, pllvpatil1905@gmail.com
[2]P.S.G.V.P. Mandal's D.N.Patel College of Engineering, Shahada, Maharashtra, India, patil.prg2998@gmail.com
[3]P.S.G.V.P. Mandal's D.N.Patel College of Engineering, Shahada, Maharashtra, India, safiyaisani@gmail.com
[4]P.S.G.V.P. Mandal's D.N.Patel College of Engineering, Shahada, Maharashtra, India, ritikajagtap14@gmail.com
[5]P.S.G.V.P. Mandal's D.N.Patel College of Engineering, Shahada, Maharashtra, India, kolileena1710@gmail.com
[6]P.S.G.V.P. Mandal's D.N.Patel College of Engineering, Shahada, Maharashtra, India, dheerajbshukla@gmail.com

Abstract: Crime rate is increasing now-a-days in many countries. There may be several types of crimes that occurs like murder, sexual assault, traffic violence, burglary etc. This higher crime rate must be reduced. For reducing the rate of crime we propose a system in which earlier or existing data is imported into system or admin will enter whole dataset and data mining algorithm such as k-means algorithm is used for clustering to predict the crime information. Algorithm will cluster collaboration and dissolution of organized crime groups, identifying various relevant crime patterns and analysis of crime data. Clustering will be done on the basis of location, gang or time. To determine position, time, location or gang related with all type of crime. The system will help to prevent crime occurring in society.

Keywords: Data mining, k-means, Cluster, Clustering, Crime Analysis.

1. 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 analysing 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 be chosen 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 are 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

analysis for Crime against women, analysis for Victims of murder will be done on the basis of dataset being used and result (prediction) is being generated.

2. Related Work

Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. J. Agarwal, R. Nagpal and R. Sehgal in [5] have analyzed crime and considered homicide crime taking into account the corresponding year and that the trend is descending from 1990 to 2011.They have used the k-means clustering technique for extracting useful information from the crime dataset using RapidMiner tool because it is solid and complete package with flexible support options. In [5] an integrated system called PerpSearch have proposed by L. Ding et al. It has been combined using two separate categories of visualization tools: providing the geographic view of crimes and visualization ability for social networks. It will take a given description of a crime, including its location, type, and the physical description of suspects (personal characteristics) as input. To detect suspects, the system will process these inputs through four integrated components: geographic profiling, social network analysis, crime patterns and physical matching. D. E. Brown (1998) in [13] constructed a software framework called ReCAP (Regional Crime Analysis Program) for mining data in order to catch professional criminals using data mining and data fusion techniques. Data fusion was used to manage, fuse and interprets information from multiple sources. The main purpose was to overcome confusion from conflicting reports and cluttered or noisy backgrounds. Data mining was used to automatically discover patterns and relationships in large databases. Anshu sharma, et al., proposed k means clustering algorithm which was used for constructing patterns of data. Data were collected and distributed, two third of true data and misrepresentation history information were utilized for preparing and remaining information were utilized for forecast and web crime discovery. The precision of the proposed work was 94.75 % and it productively recognized the false rate of 5.28%.K. K. Sindhu et al., explained scientific investigation ventures in the capacity media and hidden data investigation in the record framework, network forensic and cyber-crime mining. Device was proposed by combining digital forensic investigation and mining of crime data intended for discovering motive and pattern of attacks and hacks of assaults sorts occurred in that time period.

3. Proposed Methodology

Crime rate is increasing in many states in Country under consideration Maharashtra, Madhya Pradesh, Andhra Pradesh, Tamil Nadu etc. In today's world with such higher crime rate and brutal crime happening, there must be some protection against this crime. Here we introduced a system by which crime rate can be reduced. The proposed system that takes the existing or past data, analyses it and produce result. The system will predict the time when crime will happen, on which location and gang or persons involved in particular crime. The system will give what kind of chances or percentage we have to occur kidnap, Robbery, Arson, Kidnapping, Murder type of crime, in a specific location, in a particular time span by a particular gang. It is crucial to identify where the crime will happen, when it will happen and may be the person involve in that crime. Crime data must be imported into the system.

This system is implemented using unsupervised data mining algorithm (K-means Clustering algorithm). It plays an important role in analysing and predicting crimes. K-means algorithm will cluster co-offenders, collaboration and dissolution of organized crime groups, identifying various relevant crime patterns, Hidden links, link prediction and statistical analysis of crime data. This system will prevent crime occurring in society. Crime data is analysed which is stored in the database. Data mining algorithm will extract information and patterns from database (CSV format). System will group crime into kidnap, Robbery, Arson, Kidnapping, Murder type of crime. Clustering will be done based on places where crime occurred, culprit who involved in crime and the timing crime took place. This will help to predict crime which will occur in future. Admin will enter crime details into the system which is required for prediction. Admin can view criminal historical data. Crime incident prediction mainly dependent on the historical crime record and various geospatial and demographic information. The system will look at how to convert crime information into a data-mining problem, so that it will help detectives in solving crimes faster. In terms of crime a cluster is a group of crimes in a geographical region or a hot spot of crime. Whereas, in terms of data mining a cluster is the group of a

particular set of objects based on their characteristics of possible crime pattern. Thus relevant clusters or a sub set of the cluster will have a one-to-one correspondence to crime patterns. The Proposed system focuses on:

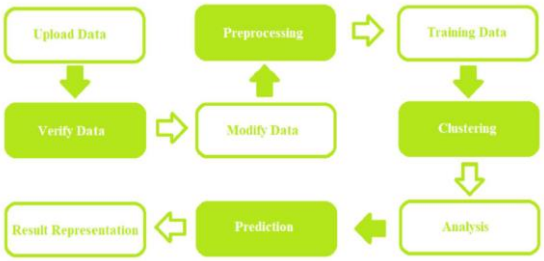

Figure 1: Framework of Proposed System

***3.1 Upload Data*:**
Upload dataset of crime information, new data into the database, and filter dataset according to requirement: Since there may be data in the read dataset that would not be used according to our method, the unnecessary data have to be filtered.

3.2 Verify Data:
Data verification is a process in which different types of data are checked for accuracy and inconsistencies after data migration is done.

It helps to determine whether data was accurately translated when data is transferred from one source to another, is complete and supports processes in the new system. During verification there may be a need for a parallel run of both systems to identify areas of disparity and forestall erroneous data loss.

3.3 Modify Data:
In Modify data, replaces the dataset missing values with a new value, and adds it to our previous dataset. This can be done by one of the Minimum, Maximum, Average and None functions which is determined by the Default parameter. If none were selected, it will be leaded to no replacement.

3.4 Preprocessing
We performed the following pre-processing steps on the dataset:

3.4.1 Data Cleaning: There are some missing values in some attributes such as last occurrence date and incident address in dataset. However, we found that all attributes containing missing values are not of our key attributes. Therefore, we did not need to clean them. All key attributes completed with cleaned values in dataset.

3.4.2 Data Reduction: For crime dataset, we needed to apply data reduction. We implemented dimensionality reduction using attribute subset selection. For example, among the available 19 attributes in crimes dataset, we just selected four of them. The selected attributes are the related ones or the key attributes for our mining purpose. We removed all the other irrelevant attributes from the dataset.

3.4.3 Data Integration: We performed several steps of data integration for our datasets. First, to avoid different attribute naming, we unified the key attribute names for crime datasets as follow: Crime Type, Crime Date, and Crime Location, represents the neighbourhood attribute for dataset whereas the Area attribute for dataset. Our mining study requires analysing the date and time info on different granularities.

3.4.4 Data Transformation and Discretization: We finished our data integration process by having 24 different distinct values for the Crime Time attribute and 14 types for the Crime Type attribute. We realized

that it is necessary to reduce the diversity of these two attribute values. Thus, we applied data transformation to both attributes by mapping their values to fall within smaller groups. Our goal was to get more frequent patterns and to increase the model accuracy. For the crime types feature, we minimize the type list by grouping them into six new types.

3.5 Training:

In this phase production of training data is done using Sample (Stratified) and Set Minus operators for increasing confidence in the response without replacement.

3.6 Clustering:

Division of a set of data or objects to a number of clusters is called clustering. Thereby, a cluster is composed of a set of similar data which behave same as a group. It can be said that the clustering is equal to the classification, with only difference that the classes are not defined and determined in advance, and grouping of the data is done without supervision.

3.7 Analysis:

Today, collection and analysis of crime-related data are imperative to security agencies. The use of a coherent method to classify these data based on the rate and location of occurrence, detection of the hidden pattern among the committed crimes at different times, and prediction of their future relationship are the most important aspects that have to be addressed.

3.8 Prediction:

The actual core part comes in this module. System gives prediction to user according his/her basic information and areas related crime. The prediction of safety is varies according to area and user information.

3.9 Result representation:

Results will be represented in the form of correlation between various crime and location of crime i.e. state/city. Crime can also be correlated on the basis of age group, location of crime & type of crime. Prediction of the crime will be displayed using various diagrams pie charts, heat maps, spikes and graphs.

4. Algorithm

K-means is the simplest and most commonly used partitioning algorithm among the clustering algorithms in scientific and industrial software. Acceptance of the k-means is mainly due to its being simple. This algorithm is also suitable for clustering of the large datasets since it has much less computational complexity, though this complexity grows linearly by increasing of the data points. Beside simplicity of this technique, it however suffers from some disadvantages such as determination of the number of clusters by user, affectability from outlier data, high dimensional data, and sensitivity toward centers for initial clusters and thus possibility of being trapped into local minimum may reduce efficiency of the k-means algorithm.

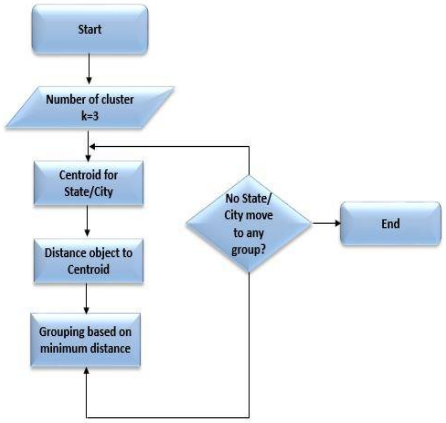

Figure 2 : k-means Algorithm

K-means algorithm mainly used to partition the clusters based on their means. Initially number of objects are grouped and specified as 'k' clusters. The mean value is calculated as the mean distance between the objects. The relocation iterative technique which is used to improve the partitions by moving objects from one group to other. Then number of iterations is done until the convergence occurs. K-means algorithm steps are given as:

Input : Number of clusters k.
Step 1: Arbitrarily choose k objects from a dataset D of N objects as the initial cluster centers.
Step 2: Reassign each object which distributed to a cluster based on a cluster center which it is the most similar or the similar.
Step 3: Update the cluster means, i.e. calculate the mean value of the object for each cluster.
Output: A set of k clusters. K-means algorithm is a base for all other clustering algorithms to find the mean values.

5. Result and Discussion

5.1 Dataset Used

This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7 belong to test set, week 2,4,6,8 belong to training set.

Table 1: Crime Dataset

2003-01-07 07:52:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	5TH ST / SHIPLEY ST	-122.402843	37.779829
2003-01-07 04:49:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Tuesday	TENDERLOIN	ARREST, BOOKED	CYRIL MAGNIN STORTH ST / EDDY ST	-122.408495	37.784452
2003-01-07 03:52:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	OFARRELL ST / LARKIN ST	-122.417904	37.785167
2003-01-07 03:34:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	DIVISADERO ST / LOMBARD ST	-122.442650	37.798999
2003-01-07 01:22:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	900 Block of MARKET ST	-122.409537	37.782691
2003-01-06 23:30:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	BAYVIEW	ARREST, BOOKED	REVERE AV / INGALLS ST	-122.384557	37.728487
2003-01-06 23:14:00	WARRANTS	WARRANT ARREST	Monday	CENTRAL	ARREST, BOOKED	BUSH ST / HYDE ST	-122.417019	37.789110
2003-01-06 22:45:00	WARRANTS	WARRANT ARREST	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:45:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:19:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	GEARY ST / POLK ST	-122.419740	37.785893
2003-01-06 21:54:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	SUTTER ST / POLK ST	-122.420120	37.787757

5.2 Result Analysis by Year

1) According to the dataset being used the Analysis graph result shows Comparison of crime Against Women with respect to States of India happened in the year 2001. This also involves tracking crime rate for the subgroups (States) by clicking on each state.

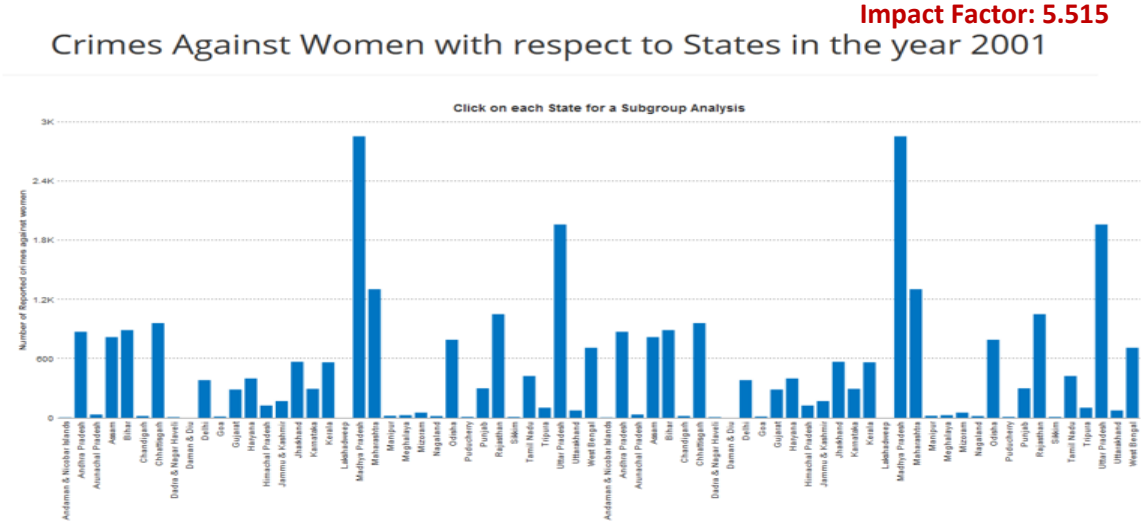

Figure 3 : Crimes against women with respect to States in year 2001

2) Analysis of crime Against Women with respect to States of India happened in the year 2010.

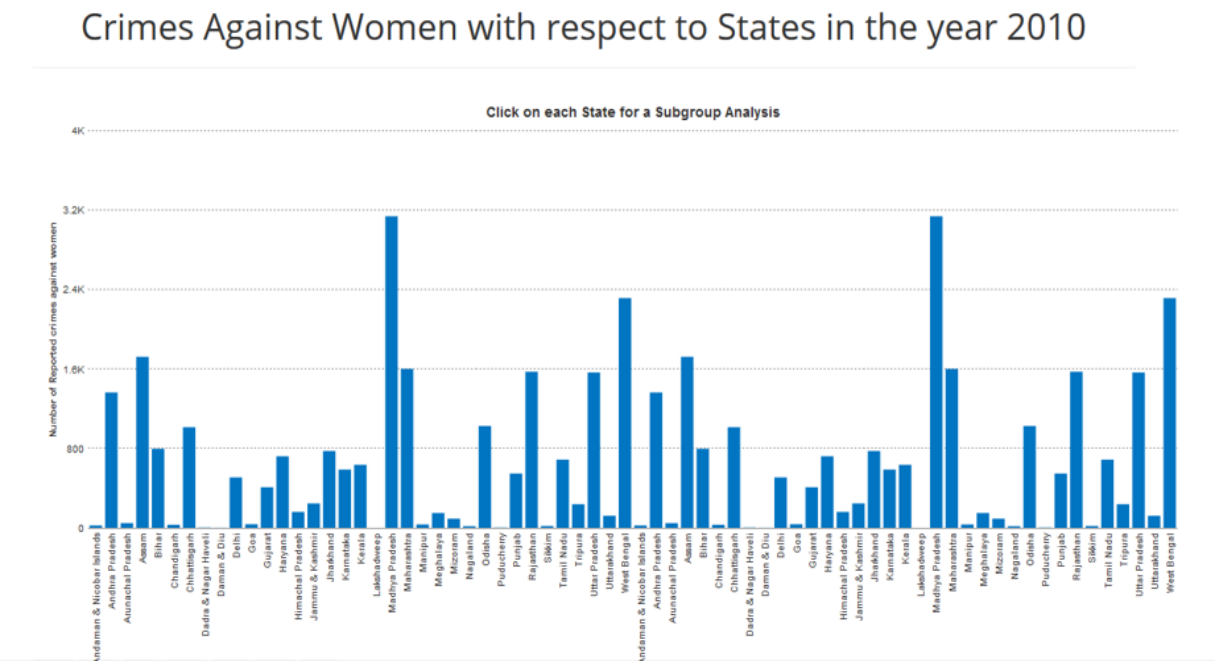

Figure 4 : Crimes against women with respect to States in year 2010

3) Comparison of Murder victims corresponding to the States in the year 2010. This also involves tracking crime rate for a gender by clicking on each state.

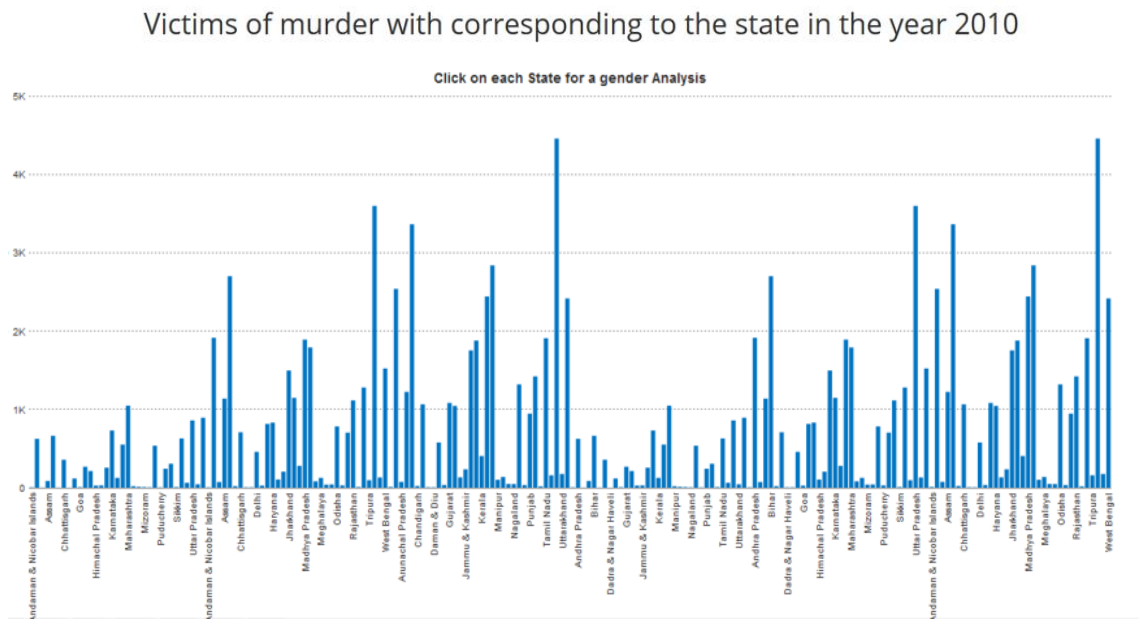

Figure 5 : Victims of murder with corresponding to the states in the year 2010

Result Analysis Comparison

This graph shows Number of student killed vs got Injured in school with respect to different cities. Blue colour shows number killed and Orange shows number got injured.

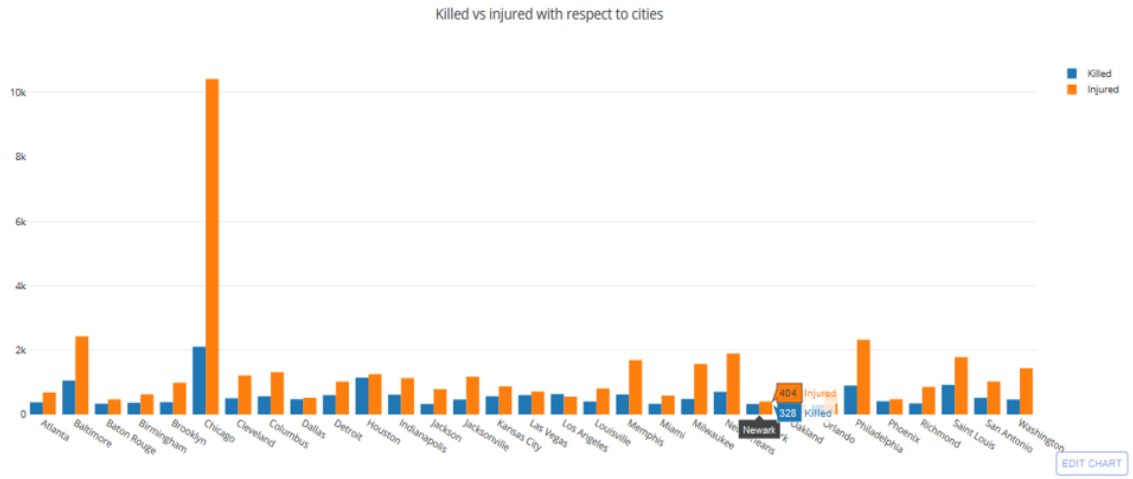

Figure 5 : Killed vs Injured with respect to Cities

6. Conclusion

This paper presents a new framework for analysing and predicting crimes based on real data. Our aim is to provide a most effective platform to the citizen for their safety. We came to the conclusion that by using data mining technique with k-means clustering algorithm, we can predict crime that might happen in future. This system will be able to predict crime, the location where crime may happen and time. We have seen the literature survey, evolution from initial proposal for perfect understanding and proposed a new system called

"Crime Rate Prediction System", which is our aim. This is going to give us more secure life than ever before, will give helping hand to Law Enforcement Agencies in solving cases. As per the result obtained, we finally concluded that the Approach to Crime Rate Prediction can be used for forecasting crime and also for identifying crime trend over years.

References

[1] Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING Concepts and Techniques", Publisher: Morgan Kaufmann, Third Edition 2012, ISBN 978- 0-12-381479-1

[2] R. Kiani, S. Mahdavi, A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No.8, 2015, PP. 11-17.

[3] V. Jain, Y. Sharma, A. Bhatia, V. Arora, "Crime Prediction using K-means Algorithm", *Global Research And Development Journal for Engineering*, Volume 2 Issue 5, April 2017, PP. 206–209.

[4] T. Sonawanev, S. Shaikh, S. Shaikh, R. Shinde, A. Sayyad, "Crime Pattern Analysis, Visualization And Prediction Using Data Mining", *IJARIIE*, Vol-1 Issue-4 2015, PP. 681-686.

[5] J. Agarwal, R. Nagpal, R. Sehgal, "Crime Analysis using K-means Clustering", *International Journal of Computer Applications*, Volume 83 – No4, December 2013.

[6] Malathi A., Dr. S. Santhosh Baboo, "An Enhanced Algorithm to Predict a Future Crime using Data Mining", *International Journal of Computer Applications*, Volume 21-No.1, May 2011, PP. 1-6.

[7] M. Gupta, B. Chandra and M. P. Gupta, "Crime Data Mining for Indian Police Information System", *Computer Society of India*, 2008, PP. 388-397.

[8] Roger S. Pressman, "Software Engineering: A Practitioner's Approach", Fifth Ed., MGH, ISBN 0-07-365578-3

[9] Silberschatz, Korth, Sudarshan, "Database System Concepts", Fourth Edition, The McGraw−Hill Companies, 2001, ISBN 0-07-255481-9

[10] Grady Booch, James Rumbaugh, Ivar Jacobson, "The Unified Modeling Language User Guide", Publisher: Addison Wesley, First Edition October 20, 1998, ISBN 0-201-57168-4

[11] *https://www.kaggle.com/c/sf-crime/data*

[12] L. Ding et al., "PerpSearch: an integrated crime detection system", 2009 *IEEE* 161-163 ISI 2009, June 8-11, 2009, Richardson, TX, USA.

[13] D.E. Brown 1998, "The regional crime analysis program (RECAP): A frame work for mining data to catch criminals", In Proceedings of *the IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3, Pp. 2848-2853

