iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38558

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Crime Analysis and Prediction Using Data Mining and Machine Learning Techniques

Jasmeet Kaur[1], Tanmay Malu[2], Simran Gill[3]
[1, 2, 3]*IIIT Allahabad*

Abstract: ***With the increase in crime rates across the world, it has become important for the Government and crime handling agencies to control the situation as it has put every person in distress. This paper is an attempt to systematically analyze and identify the crime trends across the years, the inter-state relations based on crime rates and categories through the data available, which will help in predicting the crime trends in future and will be instrumental for the Government to take informed actions and improve the country's situation. This paper applies various data mining techniques in order to analyze the crime records in India. The results of analysis have been compared for various algorithms in the domain of Association Rule Mining, Clustering, Outlier Analysis, Regression and Classification. The paper also attempts to predict the future occurrences of crimes using classification and regression algorithms which use data mining techniques .***
Keywords: Crime Analysis, Data Mining, Association Rule Mining, Clustering, outlier Analysis, Classification, Regression

I. INTRODUCTION

With the rapid growth and development of the country, the chart of crimes in India has also witnessed a sudden increase. This sudden increase in crimes and malpractices has put everyone in an alarming situation. Repeated thefts, robberies, murders, rapes, suicides have made everyone feel uncomfortable and even worried. Systematic analysis and detection of patterns and trends linked to crime, malpractices and disorder is a law enforcement feature known as the Crime Analysis. Law enforcement agencies can use these results in a productive way to distribute resources more efficiently and to assist detectives in locating and apprehending offenders. Currently many Indian law enforcement agencies have deployed user interactive interfaces and software to analyze and monitor the crime hot-spots by integrating with the National Crime Records Bureau (NCRB) **[31]**. Analysis of crime also plays a part in devising solutions to crime issues and formulating successful methods for reducing crime. With the advent of data mining techniques, it has helped us to query and mine useful information from large databases and form an incisive analysis for future developments **[32]**.

This paper aims at analysing the crimes reported in India with a goal to identify useful patterns and meaningful insights that can help the Indian Government curb this crime menace. The heinousness of a crime can vary largely from civil infraction such as illegal driving to terrorism mass murder such as the 9/11 attacks **[33]**. The inter-relation between the crimes, that is if growth/decline in the crime rate of a particular crime type also increases the growth/decline rate of some other type of crime and vice-versa, then there are chances that if more focus is given to understand and reduce the cases for such crimes, it will also impact the related ones. The political and economic situations of the districts can also be a factor affecting the crime rates. Change in the governing body, enforcement of a major policy or the global recession are also some of the factors that led to increase or decrease of some crimes in a few states. Hence, detecting these situations can help us determine the factors that led to it and if such a situation is likely to be repeated in future then the crime handling agencies can be prepared to handle the cases well in advance. Moreover, some states are also related in their patterns of crimes. Finding a cluster of states that are related to one another in their crime-rate can help the crime-handling agency of one state to look up to other states and the way they handled the situation. This will give them some prior information to handle it in a better way. Many researchers believe that without data mining techniques it would be very inefficient and time consuming to use human intelligence to monitor and control crimes **[34]**.

The analysis of the data will be done using the Data Mining Techniques listed below:

1) Association Rule Mining
2) Clustering
3) Outlier Analysis

Prediction of areas of high, medium and low crime figure out interesting hidden patterns in the data. Chen Hischun [1] worked on establishing the relationships between different data mining techniques and various crime categories. In their work; they used association, prediction, entity extraction and visualization to categorize each crime type.

H. Benjamin Fredrick David and A. Suruliandi [35] surveyed many supervised and unsupervised data mining techniques used for crime identification and detection.

Sathyadevan **[2]** used Naive Bayes to classify various crime categories and to visualize the areas with high probabilities of crime occurrences. Agarwal [3] performed the crime data analysis based on spatial distribution of existing data. He used a rapid miner tool and performed the analysis using K-Means Clustering algorithm.

Tong Wang [36] proposed a pattern detection algorithm called Series Finder. It incorporates both the common characteristics of all patterns and the unique aspects of each specific pattern to automate and detect the crime committed by the rates in the country will be done using the following same individual(s).

Models

a) C4.5 Decision Tree

b) Naive Bayes

The above listed algorithms will also be compared Huang **[4]**worked on detecting criminal activities in urban cities of the San Francisco Bay Area. He used the Haversine formula to calculate the distance between the crime location and venue location using Google Maps API.

Rasoul Kiani [37] proposed a theoretical model on the paradigms of accuracy, recall, precision and based on data mining techniques such as f-score to analyse the algorithm which yields best clustering and classification to real crime dataset results. The crime rates for an area will be recorded by police in England and Wales within predicted on the basis of several attributes using :

- Multi Variable Linear Regression
- Multi Variable Polynomial Regression

Going further, we'll also analyse the correlation between different states in terms of crime rates and which states are more prone towards the occurrence of various categories of crimes.

II. LITERATURE REVIEW

With the increase in crime rates and anti-social elements in the country, crime analysis and prediction has become a field of great importance. Because of the large size of crime data sets, data mining techniques can be used to visualize and 1990 to 2011. He assigned weights to the features in order to improve the quality of the model and remove low value of them. The Genetic Algorithm (GA) was used for optimizing Outlier Detection operator parameters using the RapidMiner tool.

Varvara **[5]** applied linear regression, logistic regression and gradient boosting to predict the crime rates in Saint Petersburg. They used social factors such as the number of bars, schools, churches and population; to determine how these affect the crime rates in the various cities.

III. DATASET AND EXPERIMENTAL SETUP

The dataset used for crime analysis is downloaded from data.gov.in, which is an Open Government Data (OGD) Platform to support the Open Data Initiative by Government of India. This becomes fundamental in providing transparency into the working of Government and also the data can be used for various research purposes to benefit the country.

We have also collected the data from the Ministry of Statistics And Programme Implementation, to gather information about factors such as number of workers, factories, etc which also influence the crime rate trends in the country.

The dataset contains the district-wise crime records in the different states and union territories of the country from 2001- 2016. The crimes include rape, abduction, dowry, murder, theft, female foeticide, sexual harassment, acid attacks, riots. There are a total of 20,772 crime records from 2001-2016 containing 24 attributes including state/union territory, district, year and crime categories. Mostly, the attributes are numerical in nature.

The dataset is divided in the ratio of 80/20, that is 80% of the dataset is used for training purposes and 20% for testing so as to establish the accuracy and precision of the algorithms.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

IV. APPROACH

A. Data Visualization

Graphical representation of data is fundamental in aspects of analysing the trends and patterns in the data. Graphs, pie charts can be used to graphically represent the dataset as:

1) *Year Wise:* Year v/s Number of crimes graph can be plotted which describes the rate of growth or decline in the number of crimes with respect to the year. In the current graph of crime rate of India over 16 years we can see that crime rate is at its all time low during this span in the year 2003. The reason for which can be understood as elections due to which the ruling party in government enforced certain measures in order to curb the crime. After that it came down to 171.35 (per lakh population) in 2006 but after that it has grown steadily to 224.318 in 2015. Crime rate growth has lowered from 2013.

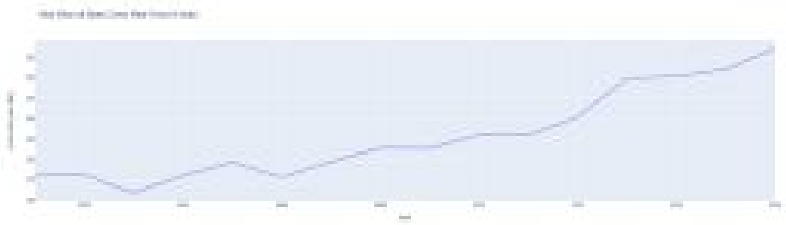


Figure 1. Crime Rate Variation In India from (2001-2016)

2) *Crime Wise:* The graph corresponding to the number of various crimes and years can also be plotted, and the crime categories which show similar growth or decline rates over the years may provide some useful insight into how these crimes are related and curbing one may affect another.

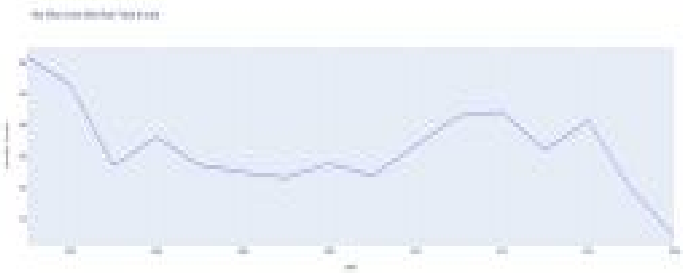


Figure 4. Murder year wise trend in India

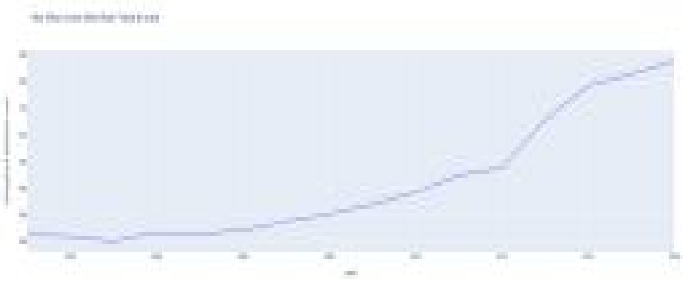


Figure 5. Kidnapping year wise trend in India

The crime trend plot clearly shows that the murder and other crime rates are not consistent. All the the violent crime rates in India show a declining trend except for rape and kidnapping & abduction although the data seems to be inconsistent. We can see an interesting trend that murder count has decreased hugely from 36k to 30k from 2001 to 2016. This has been the case for other violent crime as well except crimes like rape and kidnapping which shows the opposite trend. Thus it can be inferred that there might be some correlation between the latter two crimes. As we can see kidnapping has increased exponentially from 20k to 90k over 16 years hence an immediate measure is needed to curb this crime as this leads to many other violent crimes.

3) State Wise: Graphically representing the contribution of each state towards the crime rate of the country can help the Government take viable actions.

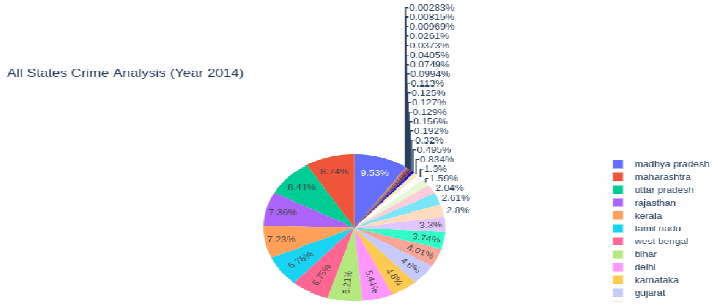


Figure 6. State Wise Contribution In India's Crime Rate (2014)

As seen in the pie chart for the state crime analysis of 2014 where a major portion of India's total crimes are covered by Madhya Pradesh, Uttar Pradesh, Maharashtra and Rajasthan. As they are neighbouring states they surely indicate the inter related crimes or patterns. Hence, they can be categorized under a single cluster. We will be performing clustering to validate the same.

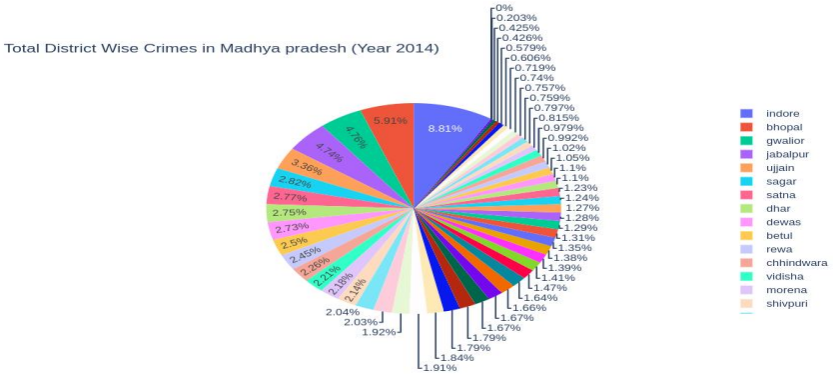


Figure 7. District Wise Contribution In MP's Crime Rate (2014)

Given is the Pie Chart for district wise total crimes in Madhya Pradesh in the year 2014. The 4 major districts of Madhya Pradesh (namely Indore, Bhopal, Gwalior and Jabalpur) constitutes of 24.22% of total crimes in Madhya Pradesh with highest crimes in Indore(8.81%) which can be easily reasoned by the fact that Indore is the most populated and busiest district of MP.
Given is the Pie Chart for district wise total crimes in Punjab in the year 2014. The 4 major districts of Punjab (namely Ludhiana, Patiala, SAS Nagar and Bathinda) constitutes of 29.18% of total crimes in Punjab with highest crimes in Ludhiana(9.54%) which can be easily reasoned by the fact that Ludhiana is the most populated and busiest district of Punjab.

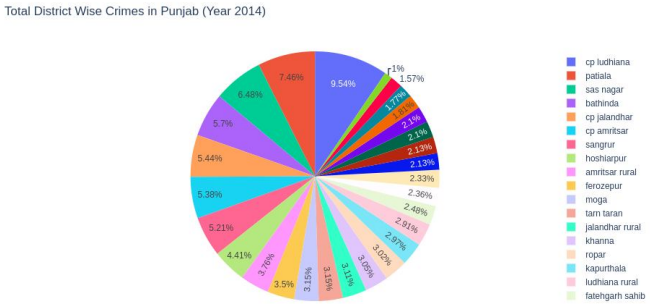


Figure 8. District Wise Contribution In Punjab's Crime Rate (2014)

4) *Correlation Between Crimes:* Graphically representing the correlation of murder and dowry deaths which have the maximum correlation among them i.e is 0.664.

Figure 9. Co-relation Graph between Crimes: Dowry Death And Murder

B. Data Pre-processing

In order to yield the most accurate and best results, the data set needs to be pre-processed before it is fed into any of the data mining algorithms.

1) *Data Cleaning:* Data Cleaning aims to bring consistency in the data set as the real world data set may be incomplete, contain missing values and noisy **[38]**. Since for our data set, the attribute tags and values may differ in cases, that is, some values may be in upper case and some in lower case. We make each of these values case-uniform. There are also certain fields which have NaN values. These must be taken care of in order to prepare a clean data set. There are many strategies taken into account when the data is incomplete:

a) *Removing the Tuple:* This is generally done, when the majority of the values in a tuple are missing. In such a case, the tuple is considered invalid and is removed from the data set.

b) *Replacing The Missing Values With The Attribute Mean:* This is a more realistic and better approach as compared to the first one. If the number of values missing in a tuple are not more than a particular threshold, then that value can be filled with the attribute mean for all samples belonging to the same class as that of the tuple.

2) *Data Aggregation:* Since the data is sparse in nature with most of the values being zeroes, the data can be grouped together in order to produce more consistent results. We can group the data in a state-wise fashion, that is the tuples belonging to the same state can be combined giving a state wise analysis, which is also instrumental for detecting the crime hot-spot areas.

C. Data Analysis

1) *Association Rule Mining:* This technique is used to determine the association rules which are used to establish the relation between different crime c categories. Apriori Algorithm is one of the ways by which association rule mining is performed.

a) *Apriori Algorithm:* It is used for the generation of frequent itemsets. This algorithm is based on Apriori Principle, which states that the subsets of a frequent itemset must also be frequent. Since this can be applied for only discrete dataset we will modify the method of finding support as described in **[6]**. New Association rules are then generated from the frequent itemsets obtained. The association rules should satisfy both minimum support and minimum confidence criteria where:

$$support(A => B) = P(AUB)\ confidence(A => B) = P(B|A)\ support(AUB)/support(A)$$

2) *Clustering:* Clustering can be used to find-out the correlation between different states in terms of crime-rates (collective as well as for a particular category of crime). Techniques used for Clustering:

a) *Agglomerative Hierarchical Clustering:* This technique generates a hierarchical tree from a set of nested clusters, represented in the form of a dendogram, which shows the grouping of tuples **[39].** There are various ways calculating the inter-cluster distance, which give way to its two variations:

- *Single Linkage:* In this, the inter-cluster distance is taken to be the distance between two closest pairs of data points belonging to different clusters.
- *Complete Linkage:* In this, the inter-cluster distance is taken to be the distance between two farthest pairs of data points belonging to different clusters.

b) *K Means Clustering:* This algorithm initially selects k random objects as the cluster centroids, and then iteratively updates these centroids to come up with the best possible clustering. The number of clusters(k) are already provided in the input to this algorithm.

3) *Outlier Analysis:* In order to account for exceptional crime records from 2001-2016, outlier analysis can be applied.

a) *DBSCAN:* DBSCAN stands for Density-Based Spatial Clustering of Applications With Noise. This is used to find clusters with arbitrary shapes. After getting the clusters from this algorithm, the points which do not belong to any of the clusters can be considered as outliers, hence accounting for the exceptional crime-rates.

D. Data Prediction

1) *Classification Algorithms:* Classification Algorithms can be used to predict the crime rates for the subsequent years.

a) *C4.5 Decision Tree:* This is a decision tree algorithm, where each leaf node represents the class (or a decision) and each internal node represents a test with the branches representing the limitation that Information Gain is biased towards the attributes with a large number of the outcomes of the tests. C4.5 is different from ID3 in the sense that it uses Gain Ratio to account for the limitation that Information Gain is biased towards the attributes with a large number of values. It defines another parameter called SplitInfo to normalize the Information Gain.

$$GainRatio(A)=InformationGain(A)/SplitInfo(A)$$

b) *Naïve Bayes:* The naïve Bayes classification technique is based on Bayes Theorem. The naive Bayes classifier calculates the probabilities of a given sample belonging to a particular class. It assumes that all the attributes are conditionally independent. We can then calculate the accuracy, recall, precision and F-score for the classified tuples using the test data.

2) *Regression Algorithms:* In order to predict the crime rates per hundred population, based on given input parameters, we make use of two regression algorithms listed below:

a) *Multi Variable Linear Regression:* This is used to deduce a linear equation for determining the value of y (or dependent variable [Crime Rate (in our case)]) in terms of X (feature vector or independent variables) **[40].**

b) *Multi Variable Polynomial Regression:* In this regression algorithm, instead of determining the best-fit linear equation for y, we find a higher-order equation, which is best suited to represent y in terms of feature vector X.

We evaluate these algorithms on the basis of mean squared error, which is the average squared difference between the actual and the value predicted by the model.

V. EXPERIMENTAL RESULTS

A. Crime Prediction

We divided the crime rate per population of each state into three categories of high, medium and low. And considering this as or target class, we predict which states will have high, medium or low crime rates per hundred in future depending upon the following input parameters :

- Year
- State
- Population
- Sex Ratio
- Literacy Rate
- Number of Factories
- Number of Workers
- Total Persons Engaged
- Gross Fixed Capital
- Net Value Added

To predict this class label (Crime Category), we have made use of two Classification Algorithms, C4.5 Decision Tree and Naive Bayes. We also evaluate both these algorithms in terms of :

- *Accuracy:* It is the measure of closeness between the predicted class label and actual class label. Accuracy can be calculated as : *Accuracy=(True positive + True negative)/Total*
- *Precision:* Precision is the measure of correctly predicted positive instances by total positive predicted instances. *Precision=True positive/(True positive + False positive)*
- *Recall:* Recall is the ratio of correctly predicted positive instances by actual positive instances. *Recall = True positive/(True positive + False negative*
- *F1 Score:* F1 score is used to take into account both recall and precision and is calculated as the weighted average of both. *F1 Score = 2 ∗ (Recall ∗ Precision)/(Recall + Precision)*

1) Classification Algorithms To Predict Crime Category Label

a) C4.5 Decision Tree: We used decision tree algorithm and applied 10 fold cross validation on our dataset which leads to the following decision tree :

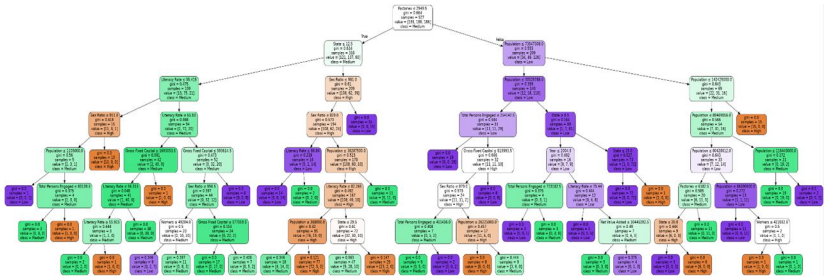


Figure 10. Graph showing Decision Tree

Using 10 fold cross validation, we get an accuracy of 92.22%, 92.62% precision, 92.22% recall and F1 score of 0.92.

b) Naive Bayes: Naive Bayes algorithm classifies the object on the basis of their probabilities of belonging to a particular class. We ran Gaussian NB using both 10 fold and 5X2 cross validation and compared the results of both. For 10 fold cross validation using Gaussian NB, we obtain an accuracy of 38.7% and F1 score of 0.38 For 5X2 cross validation using Gaussian NB, we obtain an accuracy of 36.9% and F1 score of 0.29.

2) Comparative Study of Classification Algorithms: In this section, we compare the results of Decision Tree (10 fold cross validation) and Gaussian Naive Bayes (10 fold and 5X2 cross validation). The results are summarized in the table below :

Metric	DT	NB(10 fold)	NB(5X2)
Accuracy	92.220113	38.707644	36.981132
Precision	92.695735	38.707644	36.981132
Recall	92.220113	38.707644	36.981132
F1 Score	0.924573	0.387046	0.297991

We see that Decision Tree has very good accuracy and all performance metrics whereas Naive Bayes performs relatively poor for predicting the crime category labels. The low metrics value (accuracy, precision, recall and F1 score) for Naive Bayes occur because of the algorithm's Independent Attributes Assumption. Naive Bayes assumes that the attributes are independent and do not affect each other, whereas in reality there is high correlation among various attributes which are used for predicting crime categories.

The graph below shows co-relation between attributes Sex Ratio and Number of Workers: According to Naive Bayes' assumption, Sex Ratio and Number of Workers must be independent of each other, but in reality they have a co-relation of 21%. Same is applicable for other attributes.

Figure 11. Co-relation Graphs for attributes : Sex Ratio and Number of Workers

3) *Regression Algorithms to Predict Crime rate per Hundred Population:* In order to predict the crime rates per hundred population, based on given input parameters, we make use of two regression algorithms listed below:

a) *Multi Variable Linear Regression:* This is used to deduce a linear equation for determining the value of y (or dependent variable [Crime Rate (in our case)]) in terms of X (feature vector or independent variables). The best fit line is used to represent the relation between y and feature vector X.

b) *Multi Variable Polynomial Regression:* In this regression algorithm, instead of determining the best-fit linear equation for y, we find a higher-order equation, which is best suited to represent y in terms of feature vector X.

We evaluate these algorithms on the basis of mean squared error, which is the average squared difference between the actual and the value predicted by the model.

Avoid overfitting of the model, the dataset is i i ed in the ratio of 80/20, where 80% of the t is used for training the models and 20% is to carry out testing.

4) *Comparative Study of Regression Algorithms*

Algorithm	Mean Square Error
Multi Variable Linear Regression	0.0086772795
Multi Variable Polynomial Regression	0.0031219028

We see that the mean squared error is reduced to approximately one-third while using Multi Variable Polynomial Regression and hence this proves to be a better algorithm in our case for crime rate prediction.

B. Association Rule Mining Analysis

We performed the Min Apriori ARM algorithm to deduce the relations between different crime categories for a particular state.
For the state of Andhra Pradesh, with minimum support of 0.80 and minimum confidence of 0.90, the following patterns were identified:
Murder − − −− > Burglary {confidence - 0.93}
Murder − − −− > Kidnapping And Abduction
{confidence - 0.90}
Kidnapping And Abduction − − −− > Theft
{confidence - 0.94}
Murder − − − − − > Dowry Deaths {confidence -
0.94} Rape − − − − − > Women Harassment
{confidence - 0.95} Kidnapping And Abduction, Robbery − − −− > Murder {confidence - 0.97} Murder, Kidnapping And Abduction − − −− > Theft
{confidence - 0.98}
Robbery, Women Harassment − − −− > Rape
{confidence - 0.988}
Rape, Dowry Deaths − − −− > Kidnapping And
Abduction {confidence - 0.99}
Burglary, Women Harassment − − −− > Kidnapping
And Abduction {confidence - 0.99}

We can draw the following insights from the above results :

1) If Kidnapping and abduction is controlled along with either of Rape, Murder or Women Harassment then the rate of crime for all four will decrease as kidnapping and abduction mostly leads to crime against women.
2) Since robberies occur out of force and lead to violent crimes, there are high chances of occurrence of murder when robberies happen. Hence controlling robberies will decrease the rates of murders.
3) Women are most prone to kidnapping. There are more chances of a woman getting kidnapped because of harassment and burglaries. So, special measures can be taken for women safety which will eventually reduce the kidnapping rates.
4) Thefts also generally occur as a result of kidnapping. Hence we can see that Kidnapping is one of the most prevalent crimes and thus controlling it becomes a necessity.

C. Clustering Analysis

Using Hopkins Statistic, the clustering tendency of the dataset was found to be 95.2%. Due to the high clusterability of the dataset, various clustering techniques were used to analyse the data. The states falling in the same cluster are similar to one another on the basis of crime rates and different from the states falling in another cluster for a particular year. We used K-Means and Agglomerative Hierarchical Clustering to determine state clusters. Since we have 22 dimensions because of which the data points will become sparse and hence while performing clustering the distance between will mostly be uniform. Hence we applied Principal Component Analysis (PCA) to reduce the number of dimensions to two. Below is the graph showing the variance ratio for PCA components. We can see from the graph, that by using 2 PCA components, we can retain almost 70% of

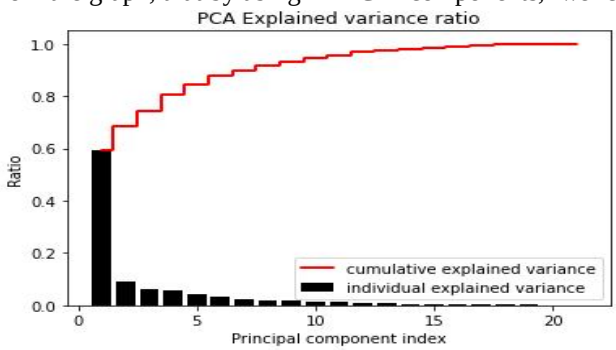


Figure 12. PCA: Explained Variance Ratio the information.

1) *K-Means Clustering:* On the PCA transformed data, we perform year-wise K-Means Clustering for all states, with **k = 5** (number of clusters). Euclidean Distance is calculated between any two rows and is used as the distance measure. All the states are grouped into five clusters on the basis of their similarities in crime rates.

The heat map for the clustered states as per 2014 data is as shown below :

Figure 13. K-Means Clustering of states (2014)

As we can see in the graph below, Madhya Pradesh, Rajasthan and Maharashtra fall in the same cluster, which is also reflected by the NCRB dataset of 2014, where these states were the prime contributor's in the country's crime rate.

2) *Agglomerative Hierarchical Clustering:* This is used to generate hierarchical tree to form a set of nested clusters. The clustering of states has been performed year-wise. Keeping the number of clusters as five, we have used two metrics to compute the inter-cluster distance :

a) *Single Linkage:* In this type of clustering, the distance between two clusters is calculated as the minimum of distance between two closest points in two different clusters. As we can see from the graph, that the states Jammu Kashmir, Punjab, Himachal Pradesh, Uttar Pradesh and all north eastern states fall into the same cluster, which can be validated because all these states have international borders and this can be a reason for them having similar crime trends.

b) *Complete Linkage:* In this type of clustering, the distance between two clusters is calculated as the maximum of distance between two farthest points belonging to two different clusters.

We can see that in Single Linkage Madhya Pradesh and Maharashtra are in different clusters whereas in Complete Linkage they fall into the same clusters which should be the case according to their crime rates.

Figure 14. Agglomerative Hierarchical Clustering : Single Linkage (2014)

D. Outlier Analysis

We used the **DBSCAN** Algorithm for outlier analysis taking **min_points = 50 and $\epsilon = 2$** .

Number Of Kidnapping and abduction cases increased greatly in 2009. Cases doubled which continued for a few years. It became 3 times in 2013 which continued in 2014 as well. If a careful look is drawn towards this, it can be seen that all these 3 years that saw exceptional increase, were years of election (2013- Madhya Pradesh Assembly elections and 2009/14- Lok Sabha Elections). So the police might be too strained during these years in campaign protection duties which gave a way to increased criminal activities.

In 2013, Uttar Pradesh witnessed an increase in its kidnapping and abduction cases to double its previous figures. The state was already having a lot of cases and the situation further worsened in this year. A few possible reasons can be the Muzaffarnagar riots of August 2013 and KumbhFair that was organised in early 2013. After riots, poverty and hatred increases so even small clashes see greater after effects. Due to the fare, crores of tourists and outsiders had entered the state. So the crime might have increased due to more people and the strain on police machinery as well.

From 2013 onwards, the number of kidnapping cases in Maharashtra almost doubled every year. There was a drought in Maharashtra at that time. There was a huge economic loss and unemployment was prevailing. So this could have led to increased participation in criminal activities by the youth. The drought persisted for a few years continuously, which might have led to a further aggravated situation.

In 2013, the number of cases for kidnapping and abduction in Chhattisgarh increased to 5 times. The possible reason for this could be the naxal activities in the state coupled with an election year. Naxals abduct the voters and election officials. Also, there was strain on administration to maintain law and order.

VI. FUTURE WORK

This work can be further extended as :

A. Clubbing the dataset with the age groups of a particular state or district, so as to analyze which age-group people are prone to be the victim of a crime in an area.

B. This project is worked upon by using a dataset of crime records from 2001-2016, the dataset can be further extended until 2020, so as to get more accurate results.

C. More Classification Algorithms for prediction of crime categories can be applied and analyzed such as Support Vector Machines.

VII. CONCLUSION

This paper attempts to analyze the crime records in India and gather meaningful insights which can help the Government in deciding the law enforcement. The data is analyzed using various data mining techniques such as Association Rule Mining, Clustering, Classification and Outlier Analysis. All these are used to analyze the trends and predict the future crime patterns in the country. We have used the decision tree, Naive Bayes algorithm to predict the crime categories. A comparative study of all the applied algorithms has been done by computing their accuracy, recall, precision and F scores. Linear Regression and Polynomial Regression has been applied to predict the future crime rates. They are compared on the basis of their mean square errors.

REFERENCES

[1] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, and M. Chau, “Crime data mining: a general framework and some examples,” IEEE Journals and Magazines, Vol 37, Issue 4, 2004.

[2] S. Sathyadevan, D. M.S, and S. Gangadharan, “Crime Analysis and Prediction Using Data Mining,” In first International Conference on Networks and Soft Computing. ACM, 2014.

[3] J. Agarwal, R. Nagpal, and R. Sehgal, “Crime Analysis using K-Means Clustering”, International Journal of Computer Applications, Vol. 83, No. 4, pp. 1-4, 2013.

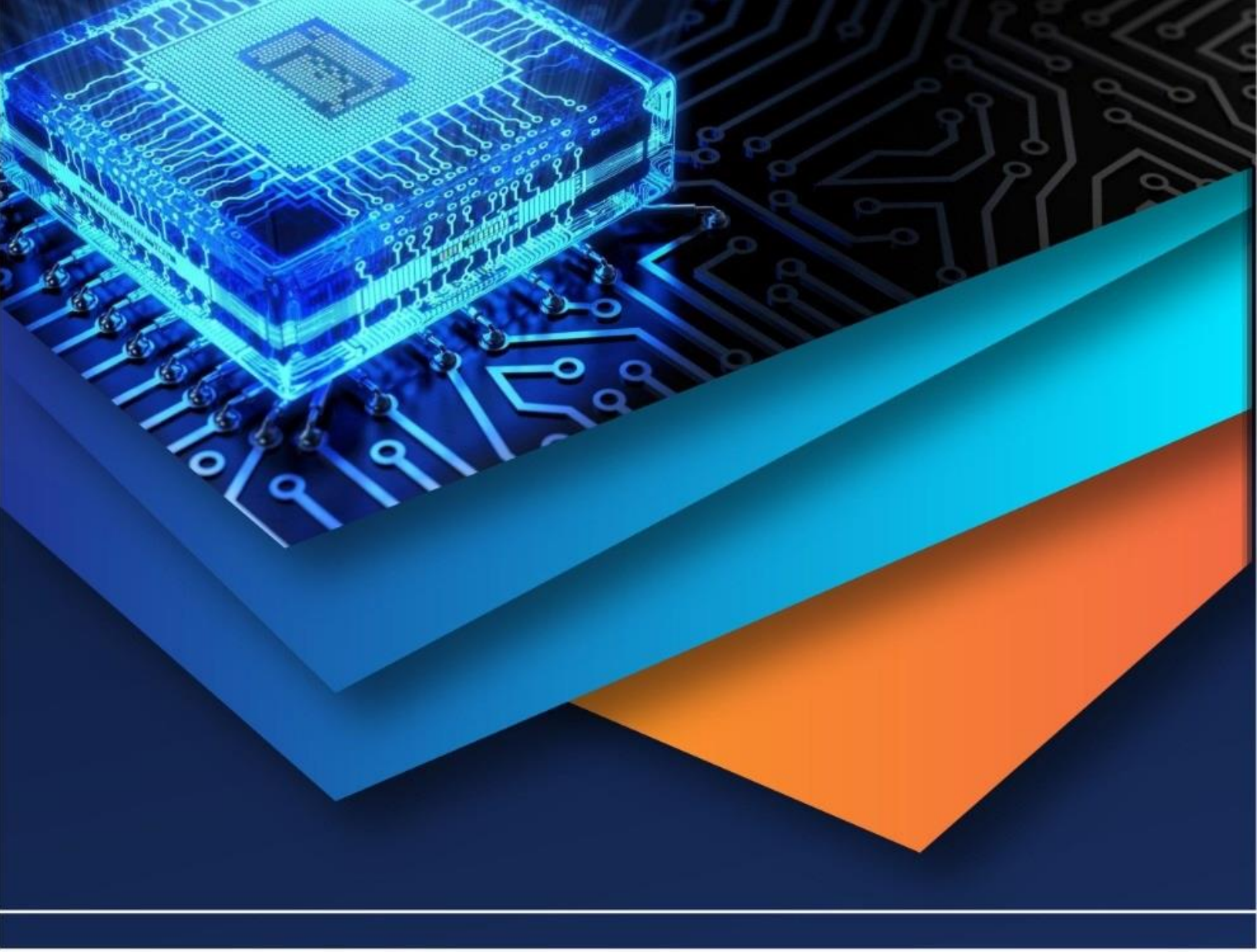
[4] Y. Huang, C. Li and S. Jeng, “Mining Location based Social Networks for Criminal Activity Prediction”, Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp.185-190, 2015.

[5] V. Mogilevich and S. Ivanov, “Crime rate prediction in the urban environment using social factors”, 7th International Young Scientist Conference on Computational Science, 2018.

[6] E. Han, G. Karypis and V. Kumar , “Min-Apriori: An Algorithm for Finding Association Rules in Data with Continuous Attributes”, Department of Computer Science and Engineering/Army HPC Research Center University of Minnesota, 1997.

[7] R. Iqbal, M. Murad, A. Mustapha, P. Panahy, and N.Khanahmadliravi, “An Experimental Study of Classification Algorithms for Crime Prediction”, Indian Journal of Science and Technology, Vol. 6, No. 3, pp. 4219-4225, 2013.

[8] F. Qin, X. Tang and Z. Cheng, “Application and research of multi label Na¨ıve Bayes Classifier”, Proceedings of the 10th World Congress on Intelligent Control and Automation.
[9] S. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 21, Issue 3, May/Jun 1991.
[10] B. Patel and K. Rana, “A Survey on Decision Tree Algorithm For Classification”, International Journal of Engineering Development and Research, Vol. 2, Issue 1, 2014.
[11] G. Uyanik and N. Guler, “A Study on Multiple Linear Regression Analysis”,4th International Conference on New Horizons in Education, Vol. 106, Pages 234-240, 2013.
[12] L. Aiken, S. West, S. Pitts, A. Baraldi, I. Wurpts, “Multiple Linear Regression”, Research Methods in Psychology, Vol.2.
[13] J. Stimson, E. Carmines and R. Zeller “Interpreting Polynomial Regression”, Sociological Methods and Research, 1978.
[14] E. Ostertagia, “Modelling using Polynomial Regression”, Modelling of Mechanical and Mechatronics Systems, Vol. 48, Pages 500-506, 2012.
[15] S. Taheri and M. Mammadov, “Learning the naive Bayes classifier with optimization models”, International Journal of Applied Mathematics and Computer Science, Vol. 23, Issue 4.
[16] E. Schubert, J. Sander, M. Ester, H. Kriegel and X. Xu, “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN”, ACM Transactions on Database Systems, Article No. 19, July 2017.
[17] S. Chakraborty, N. Nagwani, L. Dey, “Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms”, International Journal of Computer Applications (0975 – 8887), Vol. 27, No.11, August 2011.
[18] P. Arora, D. Dr. and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm For Big Data”,1st International Conference on Information Security Privacy, Vol. 78, Pages 507-512, 2016.
[19] W. Bae and S. Roh, “A Study on K -Means Clustering”, Communications for Statistical Applications and Methods, Vol. 12, Issue 2, Pages.497-508, 2005.
[20] K. Arai and A. Barakbah, “Hierarchical K-means: an algorithm for centroids initialization for K-means”, Reports of the Faculty of Science and Engineering,Saga University, Vol. 36, No.1, 2007.
[21] K. Sasirekha and P. Baby, “Agglomerative Hierarchical Clustering Algorithm- A Review”, International Journal of Scientific and Research Publications, Vol. 3, Issue 3, March 2013.
[22] F. Murtagh and P. Contreras, “Methods of Hierarchical Clustering”, May 2011.
[23] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm”, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, pp. 100-108, 1979.
[24] A. Likas, N. Vlassis and J. Verbeek, “The global k-means clustering algorithm”, Pattern Recognition, Vol. 36, Issue 23, Pages 451-461, February 2003.
[25] D. Pham, S. Dimov and C. Nguyen, “Selection of K in K-means clustering”, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, Vol. 219, Issue 1, January 2005.
[26] M. Al-Maolegi and B. Arkok, “An Improved Apriori Algorithm For Association Rules”, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014.
[27] J. Dongre, G. Prajapati and S. V. Tokekar, “The role of Apriori algorithm for finding the association rules in Data mining”, International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014.
[28] A. Bhandari, A. Gupta and D. Das, “Improvised Apriori Algorithm Using Frequent Pattern Tree for Real Time Applications in Data Mining”, Proceedings of the International Conference on Information and Communication Technologies, ICICT, December 2014.
[29] S. Chakraborty and N.K. Nagwani, “Analysis and Study of Incremental DBSCAN Clustering Algorithm”, International Journal of Enterprise Computing and Business Systems, Vol. 1, Issue 2, July 2011.
[30] J.R. Quinlan, “Learning decision tree classifiers”, ACM Computing Surveys, March 1996.
[31] M. Gupta, B. Chandra and M.P. Gupta, “Crime Data Mining for Indian Police Information System”, Journal of Crime, Vol. 2, No. 6, pp. 43-54, 2006.
[32] M. Chen, J. Han and P. Yu, ”Data mining: an overview from a database perspective,” in IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, Dec. 1996.
[33] P. Thongtae and S. Srisuk, ”An Analysis of Data Mining Applications in Crime Domain”, IEEE 8th International Conference on Computer and Information Technology Workshops, Sydney, QLD, pp. 122-126, 2008.
[34] S. Nath, ”Crime Data Mining”, Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp. 405-409.
[35] H. Benjamin Fredrick David and A. Suruliandi,”Survey on Crime Analysis And Prediction Using Data Mining Techniques”, ICTACT Journal on Soft Computing, Vol. 7, Issue 3, April 2017.
[36] T. Wang, C. Rudin, D. Wagner and R. Sevieri,”Detecting patterns of crime with Series Finder”, AAAI Workshop - Technical Report, pp. 140-142, 2013.
[37] R. Kiani, S. Mahdavi and A. Keshavarzi,”Analysis and Prediction of Crimes by Clustering and Classification”, International Journal of Advanced Research in Artificial Intelligence, Vol. 4, Issue 8, 2015.
[38] E. Rahm and H. Do, ”Data Cleaning: Problems and Current Approaches”, IEEE Data Eng., 2000.
[39] L. Billard and E. Diday, ”Agglomerative Hierarchical Clustering”, Clustering Methodology for Symbolic Data, 2019.
[40] G. Mislick and D. Nussbaum, ”Multivariable Linear Regression Analysis”, Cost Estimation: Methods and Tools, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY