IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 9, September 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

Analyzing the Indian Crime Dataset Using Machine Learning Techniques with the Help of R Software

Sarthak Mishra[1], Jyoti Badge[2]

Undergraduate Student, Department of Computer Science and Engineering, VIT Bhopal University, Ashta, Sehore, India[1]

Assistant Professor, School of Advanced Sciences and Languages, VIT Bhopal University, Ashta, Sehore, India[2]

ABSTRACT: Analyzing the Indian crime data set to produce some useful inferences: 01_District_wise_crimes_committed_IPC_2001_2012.csv (available at data.gov.in). In our research paper, we have used the above-mentioned data set and other transformed and processed forms of the data set to produce useful data visualization and also have applied different machine learning algorithms namely, Linear regression, Multiple linear regression, Polynomial regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression to build some useful regression models. We have also implemented clustering algorithms k means and hierarchical clustering with the given data set and data sets derived from it. We have also performed time series analysis and association rule mining algorithms, namely, eclat and apriori algorithms on the transformed form of the data set. We have implemented all the data visualization, machine learning algorithms, time series analysis, and association rule mining in R, and for deriving the needed data set and data set transformations we have used SQL.

KEYWORDS: Linear Regression, Multiple Regression, Polynomial Regression, Support Vector Machine, Decision Tree, Random Forest

I. INTRODUCTION

The system can predict regions that have a high probability for crime occurrence and can visualize crime-prone areas. Using the concept of Machine Learning in R we can extract previously unknown, useful information from unstructured data [1]. The aim is to investigate machine learning-based techniques for crime rate by prediction results in best accuracy and the analysis of dataset by supervised machine learning technique(SMLT) to capture several pieces of information like variable identification, univariate analysis, bivariate and multivariate analysis, missing value treatments and analyze the data validation, data cleaning/preparing, and data visualization will be done on the entire given dataset [3].The model has implementation of Linear Regression, Additive Regression, and Decision Stump algorithms, on the Crime Dataset. The scope of this research paper is to prove how effective and accurate the machine learning algorithms used in data mining analysis can be at predicting violent crime patterns [4]. This study unfolds major aspects. Like the impact of data mining and machine learning approaches, the utility of time series analysis techniques and deep learning techniques in crime trend prediction, and the inclusion of spatial and temporal information in crime datasets making the crime prediction systems more accurate and reliable [2].

With the chosen data set we have to build a predictive model which can predict the total number of specific crimes and a total number of total IPC crimes for a given year, state, district/area or any combination of these three provided as input. Other than this, time series analysis and possible set of crimes committed frequently in India are based on given data. And to group different districts (treated as instances within a year). We have used different regression, association rule mining, and clustering algorithms on the chosen data set to produce solutions to the given problem.

II. METHODOLOGY

Rstudio, Notepad, and MySql DBMS R packages used: caTools, e1071, rpart, randomForest, arules, ggplot2, dummies, and cluster. We have used different regression, association rule mining, and clustering algorithms on the chosen data set to produce solutions to the given problem. Needed data pre-processing (according to the algorithms to be applied or according to the specific objective), data transformation/deriving new data sets from the chosen data set using R and

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

SQL. Handling missing values in R. Dummy coding the categorical variables: STATE/UT and DISTRICT before applying regression algorithms on the data set. Implementing association rule mining on the chosen dataset. Implementing apriori algorithm on the dataset. Producing a transaction data set from the chosen dataset. We calculated the median of each numerical data column (represents the median of the count of specific crimes and total ipc crime in the time period of 2001-2012) i.e. median for all the 30 variables. Then we used SQL queries to replace the cell value with less than the median value with 0 and greater one with. Then we changed the data type of each column to text from int using MySQL Workbench data import wizard and then replaced the 1s with the corresponding crime name and zeroes with no value.

Data visualization: bar plots: Specific crime count/Total crime count (State level/ National Level) / and line plots: Specific crime/Total Crimes vs. Year (State level/National Level). We have implemented all the data visualization, machine learning algorithms, time series analysis, and association rule mining in R, and for deriving the needed data set and data set transformations we have used SQL. Fitting our data sets to different algorithms according to our objectives, visualizing the results, measuring and comparing the models.

III. EXPERIMENTAL RESULTS

- **Plot of Crime vs its frequency**:

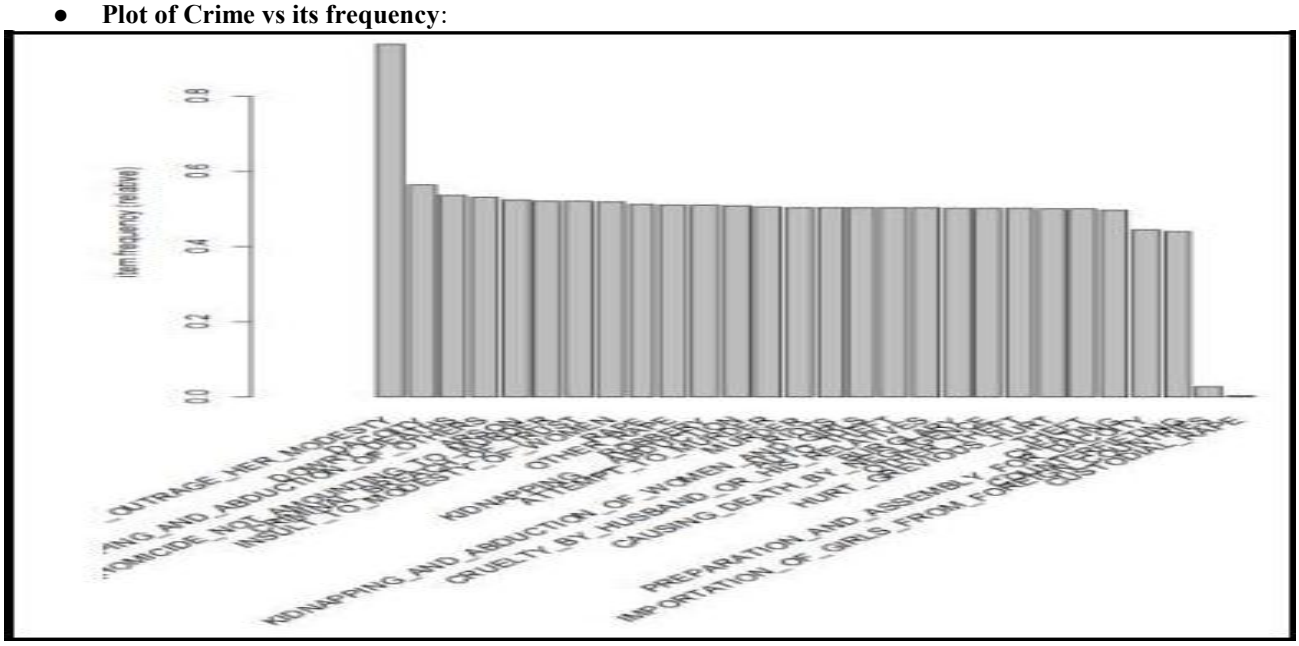

Figure 1: Crime vs its frequency

- **Analysis of Crime vs its frequency**:

The crime that occurred on high scale in most of the part in the country is: ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY. Followed by crimes: DOWRY_DEATHS, ARSON, DACOITY, KIDNAPPING AND ABDUCTION OF OTHERS and so on. With 242 representing that in a specific year no crime committed greater than the median value of that crime count in 12 years in different areas all over the country and 7 records representing just the opposite extreme.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

- **Plot showing evolution of crime in India**

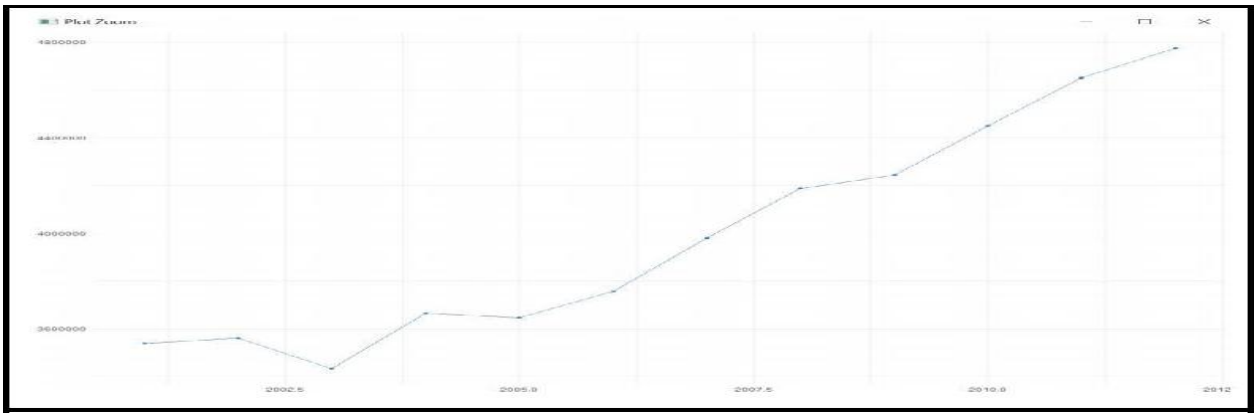

Figure 2 : Evolution of crime in India

How has Crime Evolved over time in India ? To answer the question we plot the number of crimes per year from 2001 to 2012 . The graph shows that Crime in India has increased year after year with continuous Incline. To understand the complexity in numbers, trend in the data and to observe the entire Murder data through a single visualization . Murder crime count in Year over the period of 11 years for the India is represented as map in given figures. Darker color shades of blue represents lower crime count due to Murder in each Year and lighter color shades represents higher crime count due to Murder. X axis contains Year 2001-2012 representing Year from 2001 to 2012. These maps give clear understanding to Police about the crime history of India in a single glance.

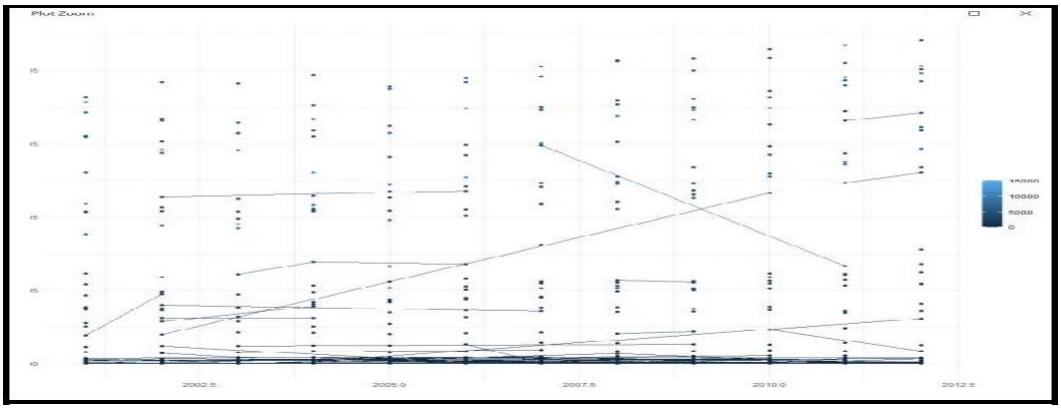

Figure 3: Graph for Murder

Like wise Murder this is graph for Rape . Higher crime rate due to rape in Year 2001 – 2012. We can see different crime rates due to different crimes in the year like Dowry, kidnapping , etc.

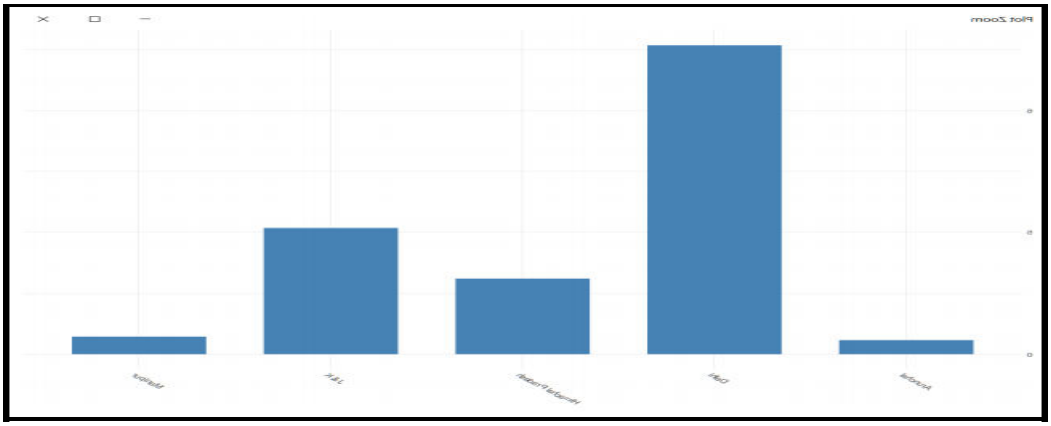

Figure 4 Graph for Rape

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

A decision tree is a machine learning algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. Various branches of variable length are formed. By applying modern technology Decision techniques to these cities' crime data, Most Crime States can be predicted in different categories . This research analyzes crime data and gives visualizations for easy understanding of the results. It also uses past 12 years' crime data from data.gov.in to Crime States can be predicted in different categories .

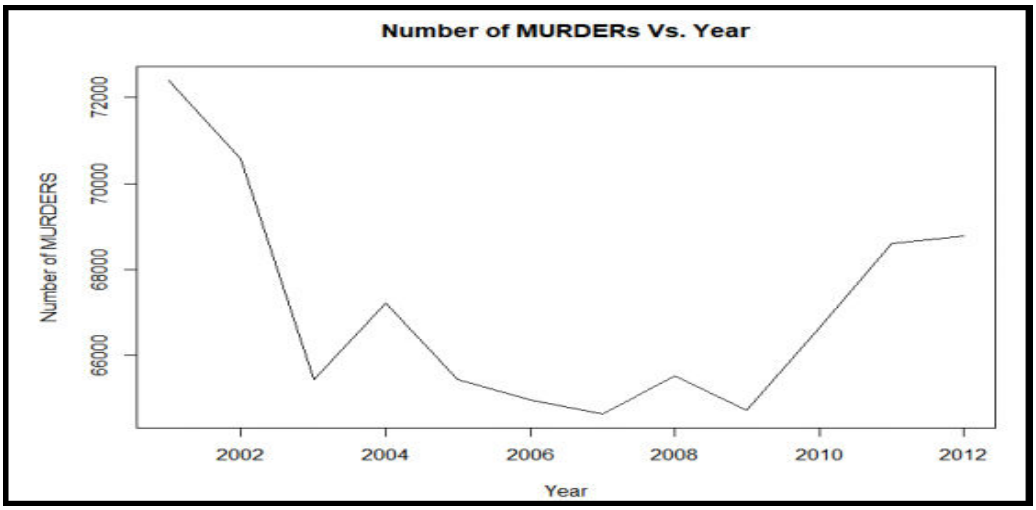

Figure 5 Number of Murders vs Year

- **Analysis of Figure 5:**

No Trend is observed. From 2001 to 2012 count of murder cases are lower with a difference of around 4700. In 2001 to 2012, Lowest number of murder cases was recorded in 2007 with around 64,400 registered cases. In 2001 to 2012, Highest number of murder cases was recorded in year 2001 with around 72,400 registered cases.

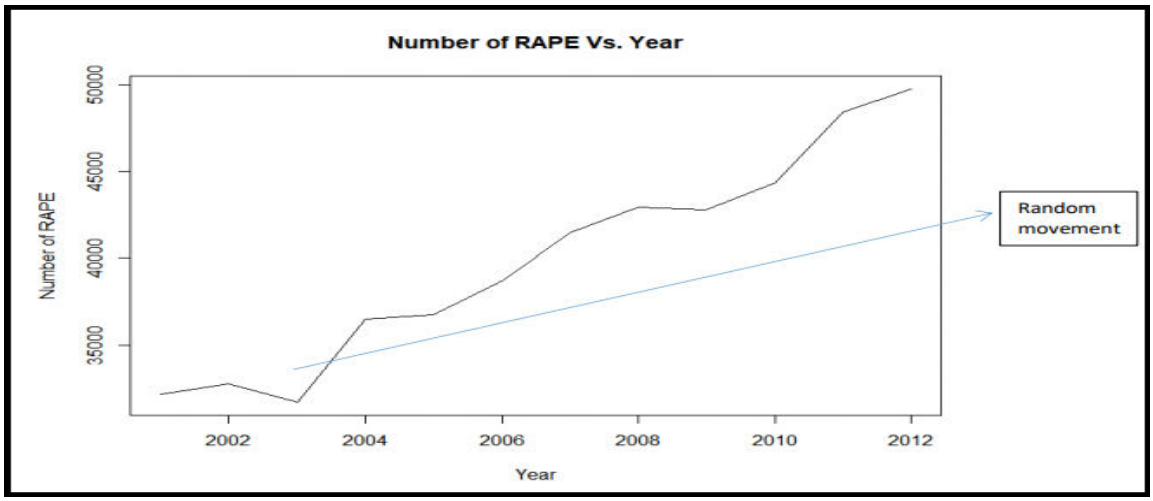

Figure 6: Number of Rape vs Year

Analysis of Figure 6:

A Trend can be observed from the above plot, number of registered rape cases was increasing in the period of 2001 to 2012 in India. But we have a random movement due to decrease of number of registered rape cases in between year 2002 and 2004 due to lowest number of registered rape cases around 36,000 cases registered in the year 2003. Lowest number of registered rape cases, around 36,000 registered cases in 2003 and highest around 50,000 registered cases in 2012. One of the important observation is that the plot between total ipc crimes vs. Year and Number of rape cases vs. Year are identical:

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

- With the same trend.
- Random movement, highest y value and lowest y value at same corresponding years and that too with very less difference in ratio, following ratios:

    - Highest total ipc crime cases registered / Highest number of registered rape cases = 4768564/49752 = 95.8

    - Lowest rape cases registered / Lowest number of registered rape cases =3432240/31694=108.2930523127406

From this we can conclude that number of total ipc crime cases registered and number of rape cases registered in 2001 to 2012 years are positively correlated and hence, as the total crime increased the rape cases contributing to it also increased.

- **State wise Crime**

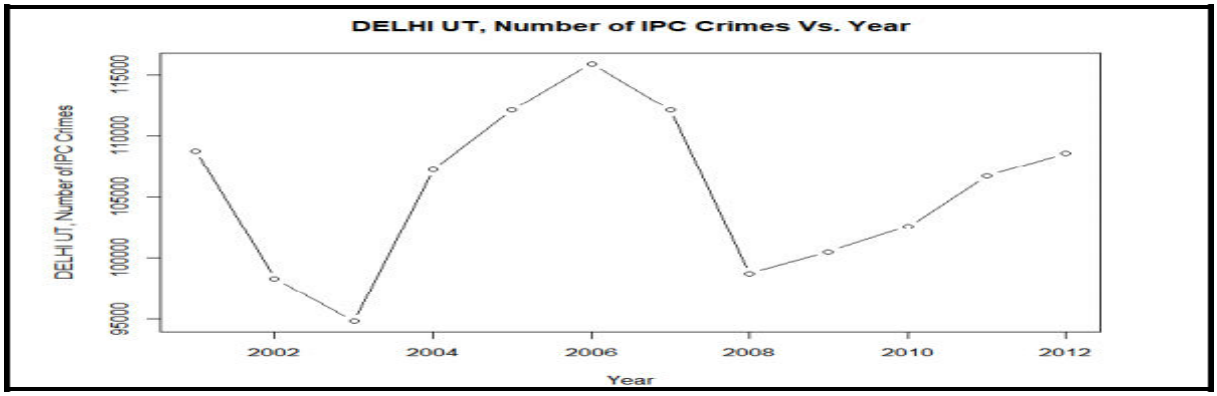

Figure 7:Crime in Delhi

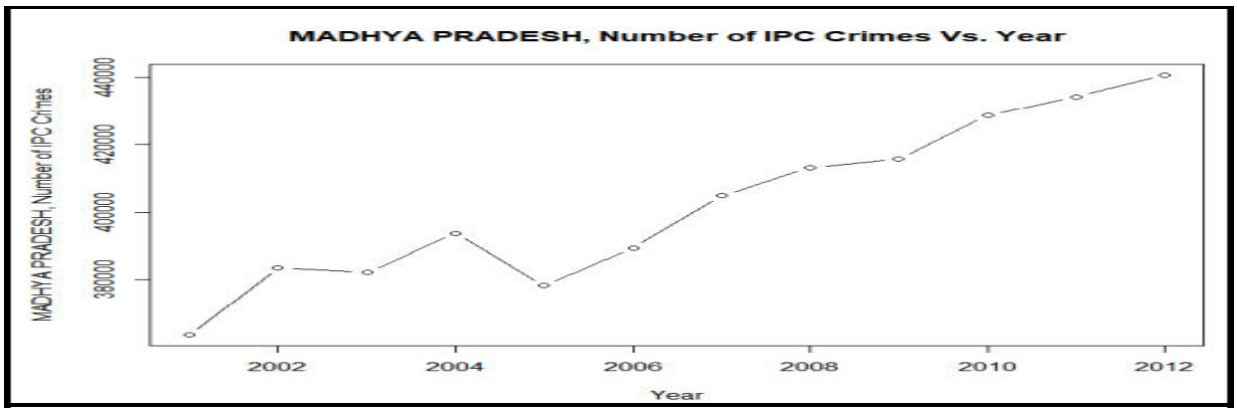

Figure 9:Crime in Madhya Pradesh

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

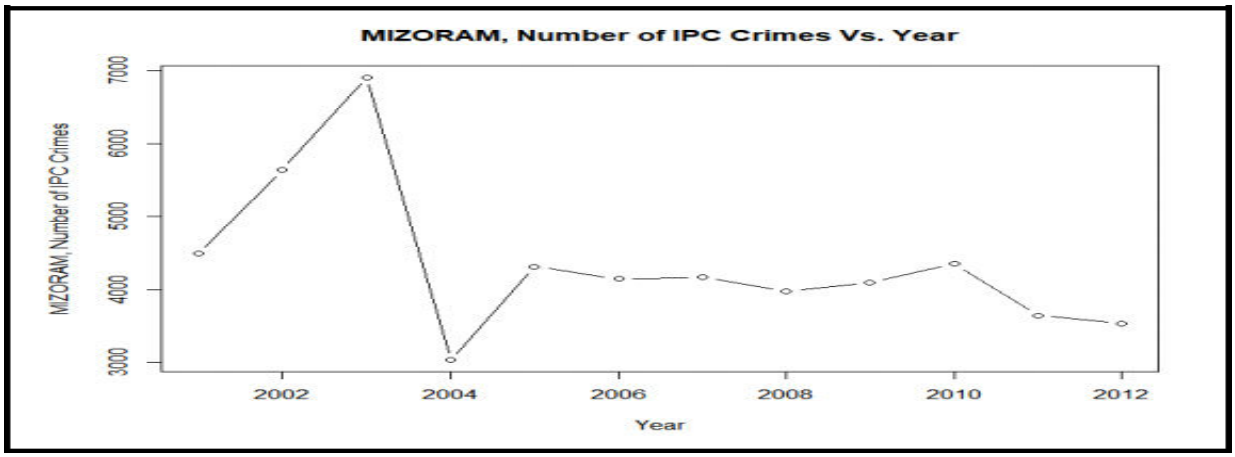

Figure 10:Crime in Mizoram

- **Bar Plots**

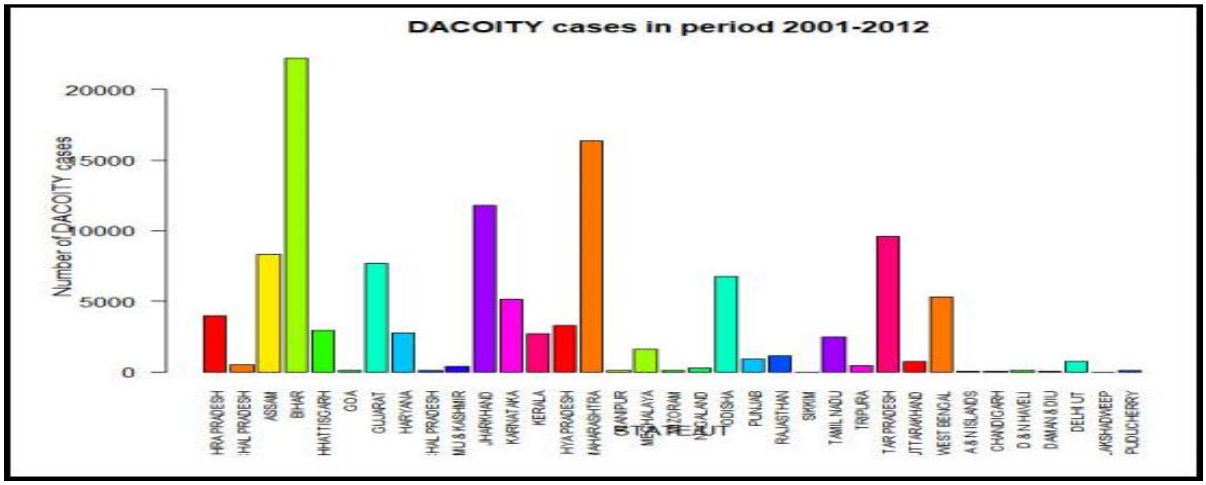

Figure 12:Bar Plot of Dacoity (2001-2012)

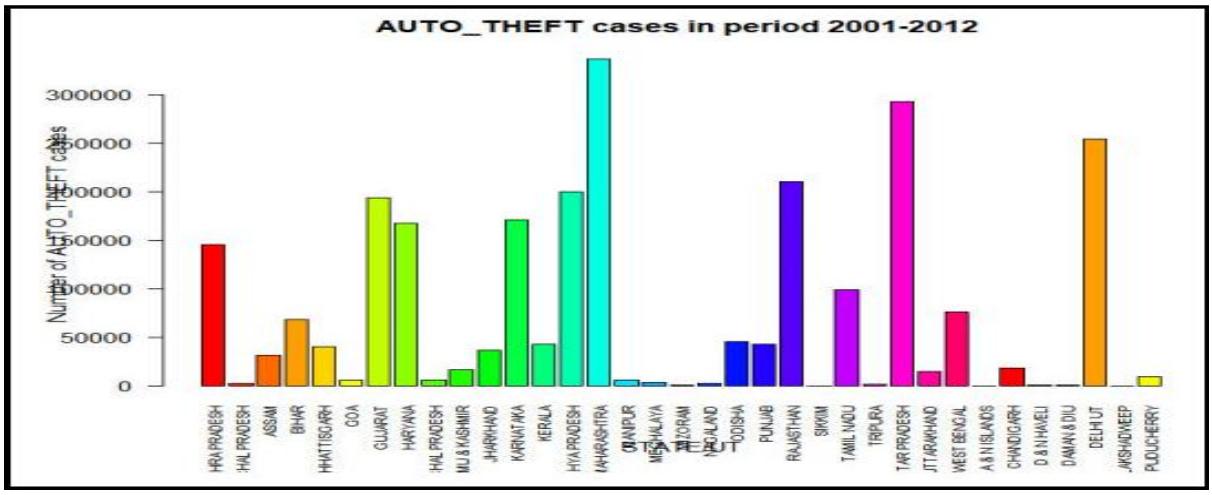

Figure 13:Bar Plot of Auto Theft (2001-2012)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

- **Clustering**

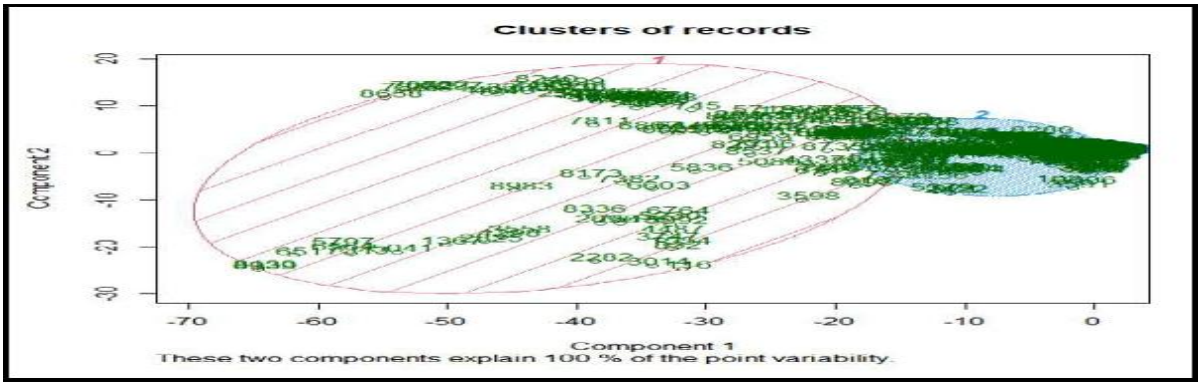

Figure 14:Clusters of records

- **Clustering Analysis:**

We can observe applying pca was successful as the two components explain 100% point variability. We can observe that on the basis of similarity in the number of cases registered against different crimes corresponding to 9012 district, year pairs can be segmented into two groups. We can observe that cluster 2 has the majority of records belonging to its region : PC1 ranging from -20 to 0 and PC2 ranging from -10 to 10. We can observe that cluster 1 has a minority of records belonging to region PC1 ranging from -70 to around -20 and PC2 ranging from -30 to around +20. The relation between principal components and original variables can be analyzed using ggbiplot. If we consider region, year pairs as an instance of a region in the past then having the data of regions or instances of regions sharing some similarity can add a lot to crime management.

Table 1: Comparing the four regression we built on the chosen dataset

Regression Model Used	R Squared Value	Adjusted R Squared Value
Multiple Linear Regression	0.0323	-0.06447
Polynomial Regression	0.0323	- 0.06447
Support Vector Regression	0.8026651	0.782916
Decision Tree Regression (min. Split = 2)	0.9646263	0.9610889
Random Forest	0.8484668	0.8333135

We can infer from the above table that

- We got the best fit with the Decision Tree Regression model.
- We got the worst fit with MLR and PLR.
- The order of best fitting models is as follows:

MLR = PLR < SVR < Random Forest < Decision tree regression

IV. CONCLUSIONS

We expect that from our research paper we are able to draw useful inferences that will have potential to become useful for other researchers or crime investigation departments to use it and to learn from the past trend and nature of crime rates, similarities among areas and set of crimes that occur frequently. Can be used to improve crime management. The models can be used to apply to the similar data but after 2012 whenever it is available in bulk from the data.gov.in.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 9, September 2021 ||

| DOI:10.15680/IJIRSET.2021.1009009|

V. ACKNOWLEDGEMENTS

This paper and the exploration behind it would not have been conceivable without the remarkable help of my administrator, Dr. Jyoti Badge. Her excitement, information and demanding scrupulousness have been a motivation and kept my work on target from my first experience with the log books of My associates at VIT University, have likewise investigated my records and replied with unfailing persistence various inquiries concerning the language, I am additionally thankful for the astute remarks offered by the unknown companion audits. The liberality and ability of the whole gang have improved this investigation innumerably and saved me from numerous blunders; those that definitely remain are altogether my own duty.

REFERENCES

[1] Suryansh Singh Raghuvanshi, Deepak Kumar. Crime Analysis and Prediction System (CAPS) India. *Walailak J. Sci. & Tech.* 2012.

[2] Umair Muneer Butt, Sukumar Letchmunan,Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir, Hafiz Husnain Raza. Spatio-Temporal Crime HotSpot Detection and Prediction. Pakistan, 2020, p. 166553-166574.

[3] Alkesh Bharati1, Dr Saravanaguru RA.K. *Crime Prediction and Analysis Using Machine Learning*, VIT University Vellore, Tamil Nadu, India.

[4] Lawrence McClendon and Natarajan Meghanathan (2015, March), Using Machine Learning Algorithms to Analyze Crime Data, Jackson State University, 1400 Lynch St, Jackson, MS, USA.

[5] Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime prediction methods. In Student Project Conference (ICT-ISPC), 2017 6th ICT International (pp. 1-5). IEEE.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details