International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

# Crime Analysis and Investigation (CANI)

**Himank Budhiraja[1] , Deepak Vats[2], Nishant Tiwari[3] , Aman Kumar Agarwal[4] , Sharanya Chandran[5]**

*Department of Computer Science, HMR Institute of Technology and Management, Delhi*

## Abstract

Crimes are the significant threat faced by the society. Large number of crimes are committed on the daily basis across the country. In order to prevent the criminal activities, firstly we need to keep a track of the criminal activities by storing them into a database and thus using it for further analyzation and prediction. The major problem faced is to maintain a dataset with all the updated criminal activities and use them for analyzing and predicting. In this project, the efforts have been made to develop an algorithm that can analyze the dataset and predict about the type of crime which may happen in future even more accurately. In this project, the criminal data is extracted from the official portal of Government of India (Data.gov.in), on which we will be using machine learning algorithm that learns certain properties from a training dataset in order to make predictions. This predictive modelling consists of two areas – regression and pattern classification. The data from the dataset is extracted and mapped into a statistical graph using regression. The slope is calculated and polynomial degree is used to map the curve and get the expected prediction of the number of crime cases that may occur in future.

## Introduction

The crime rate is increasing very fast and so is the need to solve the cases at a faster rate. Due to availability of modern gadgets, tools and technology, it is quite easy for a criminal to meet his misdeeds within no time. The crime rates are not only rising in cities, but nowadays, a large number of crime cases are being reported in small towns and villages as well. According to the crime records, the crimes such as robbery, arson have been decreased while the crimes like sex abuse, murders, gang rape have been increased. All criminals have played a lot with the emotions of innocents and even after staying in the headlines for several week, so many cases are still unsolved, making people losing their faith from judiciary system resulting in injustice to the victim and its family. With the help of the Machine learning algorithm used in this project, we can also keep a track of all the pending cases that needs to be solved as soon as possible. The crime cannot be predicted accurately as it is not a systematic thing or done at random. We cannot predict about the victims or the criminals that are going to commit crime. But instead, with the help of the machine learning algorithms, we can identify and analyze patterns and trends in crime and can predict about the area and the type of crime that can occur in future.

The prediction cannot assure 100% accuracy, but the results have been proven to help in decreasing crime rates by providing police security and alert to crime sensitive areas or crime hot spots. The very first step is crime analysis. Crime analysis includes exploring, inter relating and detecting relationship between the various crimes and

characteristics of the crime. This analysis helps in preparing statistics, queries and maps on demand. It also helps to see if a crime in a certain known pattern or a new pattern necessary [1]. Second step is to use regression on the data of the crime records and to establish a statistical. Third step includes pattern classification. Pattern classification consists of two sub parts, Supervised and unsupervised learning. In supervised learning, the training dataset containing particular output is known which will be used to train and predict for unseen data.

## Background or related work

Crime is a big problem for everyone in the world. Thus, it is necessary to find a way to minimize the crime rates and catch the culprits. There are many already existing approaches for the efficient prediction of the crime and number of researches are currently going on to find the optimum solution/ best approach.

Earlier, techniques like Crime Detection and Criminal Identification (CDCI) were used in Indian cities. In this technique, identification of the criminals was done on the basis of the parameters like sex, name, facial features, weapon used etc. It consisted of 6 steps – data extraction followed by data pre-preprocessing, clustering, map representation, classification and WEKA tool. It used K-means algorithm and provides a group of similar crime as its output. The KNN was used to spot the criminal [2].

The major drawback was the main causes of the crime were being ignored and daily factored were focused more.

The paper concentrates on analyzing approaches from both the ends. Ie, Computer Science and Police Department. The author implemented pattern detection technique and also provides the suggestion for the future prediction. Clustering is done with the help of K-means algorithm and thus, patterns of crime can be identified resulting in solving the crimes at faster rate. Semi supervised technique is implemented in order to improve its efficiency and accuracy. The crimes are represented using Geo-spatial spots. There is graphical representation of the crime type and the geographical region as the output based on the selection of time range [3].

In this paper, the key features are focused which led to the increase in crime rate. The final predictions are made based on the rankings of the features. The Random Forest Classifier gives the best accuracy is given by The Random Forest Classifier [4].

In this paper, data mining techniques like clustering and classification are used. The developed system is responsible for analyzing the crime information which further helps in brutalize the investigation in India. This tool is very helpful for the identification of the criminals thereby boosting up the speed of investigation [5].

In this paper, the author developed a forecasting model in partnership with the police division of the US. The first step to this approach is extraction of the crime data and form a dataset from the available crime records. The dataset consists of all the crime information like crime location, time of crime and other crime related aspects. This is also known as Classification method. The drawback to this paper was it is best for analyzing but does not predict about crimes that can happen in future [6].

In this paper, the author used various algorithms like Naïve Bayesian, KNN, and Neural Networks. The results came out were proven to be better than Decision Tree and Support Vector Machine. The dataset is tested with two different

feature selection methods. The area under the curve (AUC) was compared in different algorithms. In addition to this, Chi-Square feature selection technique improves the efficiency of data mining results. KNN is proved to be more efficient and gives better results by using Chi-Square feature selection technique.[7].

Hotspot Mapping technique is very common for analyzing crime characteristics. Socio-economic and other crime factors play an important role for distribution of spatial crime. A tool to identify crime hotspots was invented- Hotspot Optimization Tool (HOT). Geospatial Discriminative Patterns is the main component of HOT. GDP have the ability to differentiate between two classes in a dataset that comprises of spatial information. The accuracy of HOT for mapping crime hotspots is nearly 100% and found to be very efficient in searching patterns and utilizing it in geospatial space. Spatial distributions are represented as output using Grid thematic mapping.[8].

## Proposed system

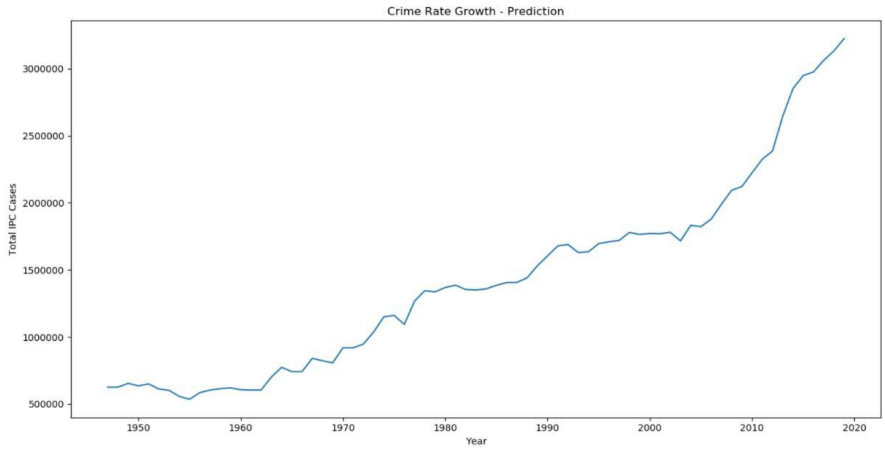
### Data collection and Pre-Processing

In this step, we retrieve the criminal records and crime rates from the Government of India's official website. The data retrieved is very inefficient to work on. Therefore, before starting with the analysis, data preprocessing is done. The extracted raw data is converted into data frames/ dataset.

In the data frames, each crime record is associated with its details in a proper format. Each record consists of every possible information about the crime. Different data frames/ data sets are made for each state of the country. Each data set consists of crime details for every location of the state it is associated with.

### Plotting of Data

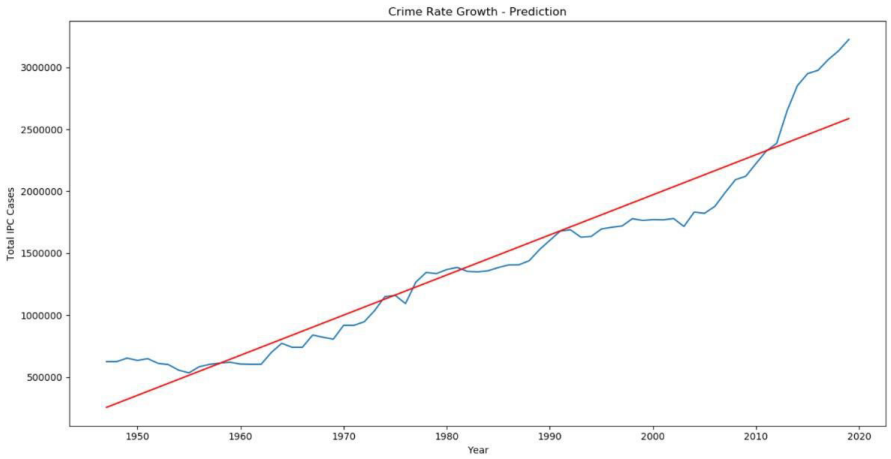
The data frames are then plotted on the x-y axis using matplotlib. Matplotlib consists of various functions and can acts like MATLAB. The plotted data is then checked weather it is scattered or not. Scattering of data is prevented.



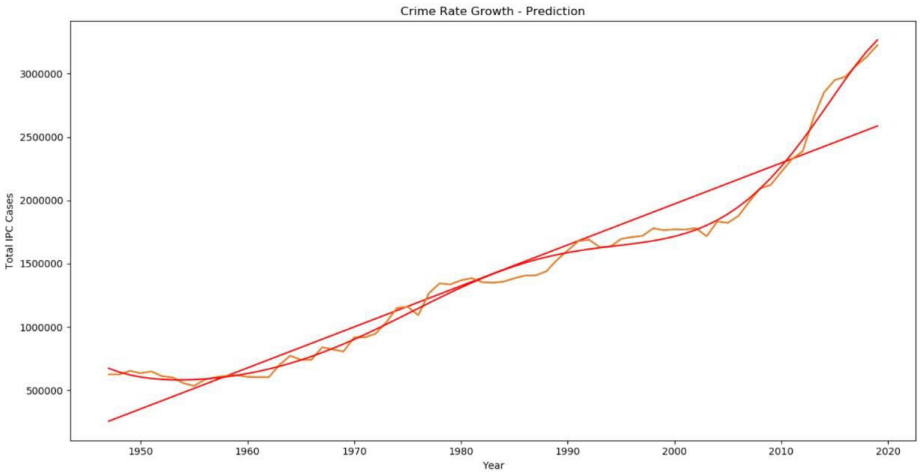
## Analysis:

### Linear Regression

Linear regression is a very powerful algorithm used for statistical analysis in predictive modelling. Using linear regression, firstly we calculate the slope of the plotted curve.




In this approach, we use polynomial regression for analysis. We simply keep on increasing the polynomial degree of regression and check whether the curve is distorted or not. In our research, we are able to reach $17^{Th}$ polynomial degree. The polynomial regression with $17^{th}$ degree is done on the given data frames. After $17^{th}$ degree, the curve gets distorted.



**Coefficient of determination and Prediction**

Coefficient of prediction (P value) refers. It is to the statistically significant test results. P values represents the probability of observing a sample statistic taking assumption as null hypothesis is true.

The null hypothesis is a hypothesis of "no difference".

High P value represents that sample results are consistent with a null hypothesis that is true.

Low P value refers to the inconsistent sample results with a null hypothesis that is true.

Coefficient of Determination (R-squared value) represents the goodness-of-fit measure for regression models. It measures the strength of relationship between the dependent variable and the model on a scale 0-100%.

The P value and R-Squared value of the curve is calculated.

Coefficient of Determination (R-squared value): 0.9933766740070443

Coefficient of Prediction (P value): 3267758

**Prediction**

With the help of the $17^{th}$ polynomial degree, the curve is mapped and, in the output, we can see the line is Ponting to nearly exact value of the number of criminal cases that may occur in future. This model helps us to predict the approximate value of the number of cases that may occur in future.

## Experimentation and result

With the help of the various machine learning algorithms, nowadays, it is easy and efficient to keep the track of the criminal cases and to predict the upcoming number of cases. The prediction cannot be exact because the crime cases are uncertain and we cannot predict about when and who will be committing next crime. The predictive model in this project is capable of predicting with an efficiency over 95%.

The Coefficient of Determination (R-Squared value) comes out to be:0.9933766740070443

The Coefficient of Prediction (P value) comes out to be: 3267758

## Conclusion

The factors like corruption, unemployment, drug abuse, poverty etc. are the main reason for the increasing of number of crime cases in India. This approach is very useful for detecting and analyze the rate of crime and thus we can predict the area and the number of crimes may happen in the future. With the help of prediction, we can take several steps like – deployment of special forces, making police alert, high area alert, barricading etc. Which would lead to decrease the number of crime rates and thus solving the cases even faster. The model used in this project is very useful for police as well as investigating agencies. This model can be applied to other country's data set also. By identifying the

crime zones, the general public can also be alerted about the crime in different states of the country. Future work on this project emphasizes on training bots to automatically predict the areas that are crime prone. Reliability, accuracy and data privacy and the availability of crime data can be improved for enhanced prediction.

## References

[1] Deepika K.K, Smitha Vinod," Crime analysis in India using data mining techniques", International Journal of Engineering & Technology

[2] Devendra Tayal, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, Nikhil Tyagi, "Crime detection and Criminal identification in India using data mining technique", Ai & Society, 2014, https://doi.org/10.1007/s00146-014-0539-6

[3] Shyam Nath, "Crime Pattern Detection Using Data Mining", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006, DOI: 10.1109/WI-IATW.2006.55

[4] Prajakta Yerpude and Vaishnavi Gudur, "Predictive Modelling of Crime Dataset Using Data Mining", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.7, No.4, July 2017, DOI: 10.5121/ijdkp.2017.7404

[5] Arunima Kumar, Raju Gopal, "Data mining-based crime investigation systems: Taxonomy and relevance", 2015 Global Conference on Communication Technologies (GCCT) – 2015, DOI: 10.1109/GCCT.2015.7342782

[6] Chung-Hsien Yu, Max Ward, Melissa Morabito, Wei Ding, "Crime Forecasting Using Data Mining Techniques", 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, DOI: 10.1109/ICDMW.2011.56

[7] Somayeh Shojaee, Aida Mustafa, Fatimah Sidi, Marzanah Jabar, "A Study on Classification Learning Algorithms to Predict Crime Status", International Journal of Digital Content Technology and its Applications (JDCTA), Volume 7, Number 9, 1-3, 2013, DOI: 10.4156/jdcta.vol7.issue9.43

[8] Dawei Wang, Wei Ding, Henry Lo, Tomasz Stepinski, Josue Salazar, Melissa Morabito, "Crime hotspot mapping using the crime related factors—a spatial data mining approach", Applied Intelligence, 2012, https://doi.org/10.1007/s10489-012-0400-x