IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94103

CRIME ANALYSIS USING PREDICTIVE MODELING

Divya Chopra, Deepanshu Kaushik, Mr. Varun Goel, Mr. Sachin Garg

Maharaja Agrasen Institute of Technology (IT)

Abstract: Preventive measures are always better than curative ones. The same is true for crimes as well. The Crime Analysis uses mathematics, predictive modeling, and predictive analysis to help law enforcement in targeting potential criminal and antisocial activities. Studying, observing, and analyzing the patterns formed in crimes are used in various countries and organizations. Crime is dynamic in nature but still, we can find patterns in a crime which will help the authoritative and concerned organizations to find areas that are less affected by a crime and those with a high rate of a particular crime. This research aims to provide people with the idea of how crime patterns can be analyzed and help create a crime-free neighborhood. We have used a clustering method (K-Mean Clustering) to find vulnerable locations to the classified crime. As of now, we have taken six classes of crime (i.e. Robbery, Accident, Gambling, Violence, Kidnapping, and Murder). More classes will be added to extend the model's usefulness. For this research, the datasets were selected from government websites, which were pre-processed and used to find patterns in the various classes of crime occurring in different states according to the jurisdiction of the country

I. INTRODUCTION

The increasing crime rate is alarming in major cities of the country. The traditional methods to prevent crime are not sufficient. Incorporation of technology into the law enforcement system is a must. The punishment should be strict enough to create a fear to follow the laws seriously so that nobody thinks twice before breaking the law but it would be much better if we can stop the crime even before it takes place.

The task is to predict which category of crime is more likely to take place at a given time and place. The idea is to take data with the help of AI-based cameras and sound devices and use that data as an input in our trained model for the next predictions. Since crime rates and places change dynamically, we have to train our models regularly to get accurate results.

We have used various algorithms and different tuning parameters, to get the best possible accuracy. The dataset is trained differently for different algorithms. We used regression algorithms like SVM, Decision tree regression, and Random Forest to get the best predictions possible.

II. PROPOSED ARCHITECTURE

A supervised learning problem for predicting future rates is time series forecasting. We develop a time series model to best capture or describe an observed time series in order to understand the underlying causes. Here we want to know the reason behind the time series dataset. The method by which predictions about the future are made is called extrapolation and refers to it as time series forecasting.

Supervised learning is a form of learning in which we have to enter the input variable(X) and an output variable(y) and an appropriate algorithm is used in order to map the function from the input to the output. Y=f(x)

The supervised learning basically comprises of: -

Classification where we classify the output variable into a particular type for example summer or winter

Regression problems are the ones in which the output variable has a specific real value.

Our problem is a supervised regression problem. We have primarily used three regression algorithms and compared the result to find out which algorithm is giving us the maximum accuracy.

KNN (K nearest neighbor): K nearest neighbor falls under a supervised learning algorithm that can implement the regression and classification problems.[1-3]

Decision Tree Regression: A regression or classification model is made in the form of a tree. It divides the dataset into many small subsets, simultaneously incrementally developing a decision tree. The final tree has two unique node types, leaf nodes, and decision nodes. The decision nodes have two or more child nodes, each representing values for the attribute tested. Leaf nodes represent the target numeric value to be predicted.

Regression Decision Trees are used when the target values take continuous values, which is exactly our case.[1-3]

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94103

Random Forest Regression: A supervised learning method for both Regression and Classification. It contains multiple decision trees and the output is the mean of those individual trees. Random forest trains these trees on different parts of the dataset which helps in reducing the variance. This increases the performance greatly but also increases some bias and loss of interpretability. Violence, Gambling, Murder, and Kidnapping for **acts 379, 279, 323, 13, 302, and 363** respectively.[1-3]

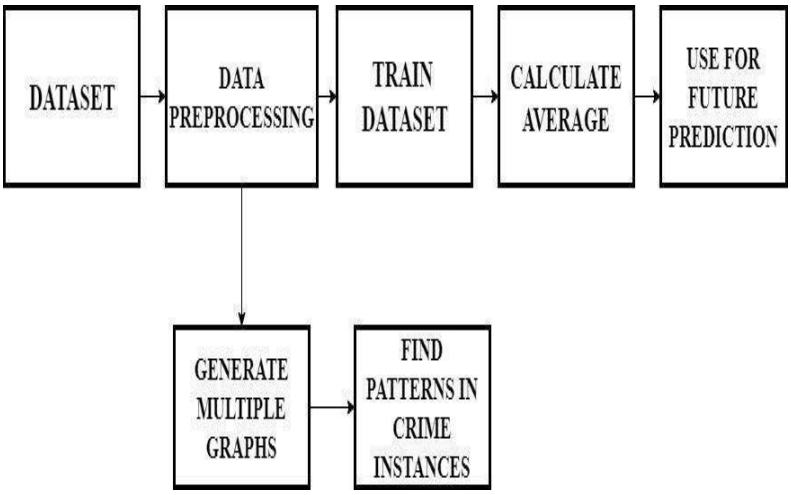


Fig 1. Data Processing Flow

The aim of this paper is to find patterns in criminal
activities and find the future increase or decrease of a particular criminal activity in the state. This is done so that necessary actions can be taken to curb such activities in that state. In order to achieve the goals, we have used the architecture/workflow diagram shown above.

थाना	थाना अपराध /मर्ग क्रमांक	धारा	फरियादी का नाम एवं पता	आरोपी का नाम एवं पता	घटना स्थल	घटना दिनांक व समय
थाना पलासिया	53/18	279, 337 भादवि	मोहिल पिता/पति पिता महेन्द्र कुमार ठाकुर निवासी 22 शक्तिनगर इन्दौर	मोटर सायकिल क्र.MP09VA4249 का चालक------,	पत्रकार चौराहा इन्दौर	10-02-18 10:45 के 10:45 बीच
थाना पलासिया	54/18	379, भादवि	पायस पिता/पति अनिल प्रकाश सेम्युल उम्र 18 साल निवासी 431 स्कीम नं.114 पार्ट 1 विजय नगर इन्दौर	अज्ञात-,	सेन्ट पाल कालेज के पास गार्ड के सामने से लालाराम नगर इन्दौर	09-02-18 12:40 के 13:40 बीच
थाना एमआईजी	117/18	379,	अनिल पिता/पति स्व. रामसेवक चौरसिया निवासी ३९ हीरा बाग कालोनी थाना लसुडिया इन्दौर	अज्ञात चोर-,	सी एच एल अस्पाकताल के पीछ के पार्किंग से इन्दौर	08-02-18 19:30 के 21:30 बीच

Fig 2. data before processing in Hindi

This data is taken from government records[4]

timestamp	act379	act13	act279	act323	act363	act302	latitude	longitude
28-02-2018 21:00	1	0	0	0	0	0	22.73726	75.87599
28-02-2018 21:15	1	0	0	0	0	0	22.72099	75.87608
28-02-2018 10:15	0	0	1	0	0	0	22.73668	75.88317
28-02-2018 10:15	0	0	1	0	0	0	22.74653	75.88714

Fig 3. Data post-processing

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

DOI: 10.17148/IARJSET.2022.94103

The crime which occurred is assigned a value of 1. other columns are assigned a value of 0. The processed data CSV file is available at Kaggle[5]

III. DATA ANAYLYSIS

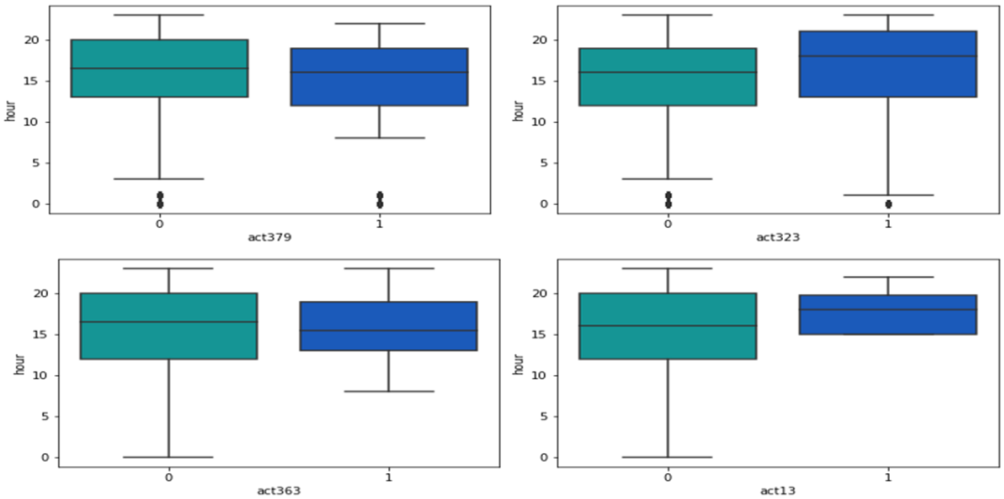


Fig 4. Box Plot for act 13

Individual time series graphs were also generated for each crime and the rate of increase/decrease of crime rates for every hour could also be analysed.

Outlier detection was performed on the data set for which box plots were plotted (as shown above) and hex bin plots show the hourly distribution of crime.

In this way, the government can plan more suitable techniques to handle criminal activities in the area and focus on one or more crime than the other. The analysis are shown below using hexbin plot and data distribution.

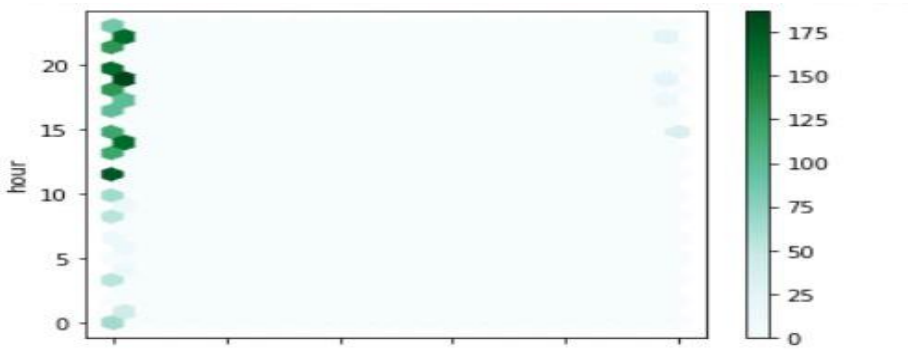


Fig 5. Hexbin plot

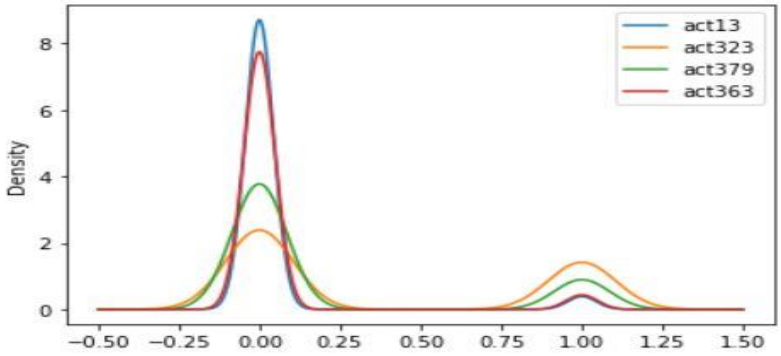


Fig 6. Data Distribution

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 4, April 2022

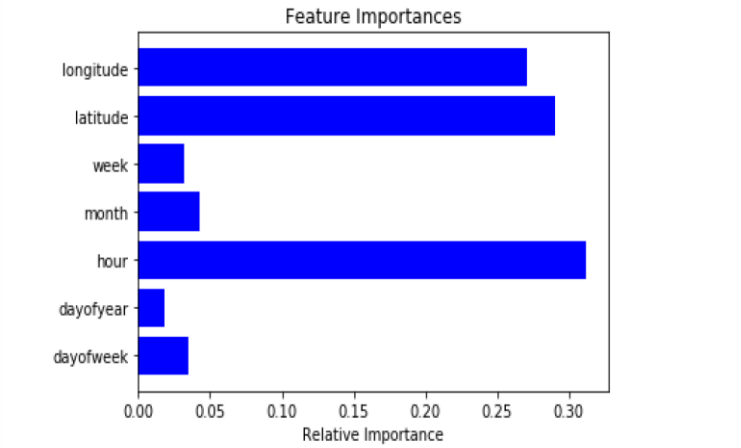
DOI: 10.17148/IARJSET.2022.94103

IV. RESULTS

- Out of all the tried algorithms we got the best result of 99% accuracy with Random Forest.

The testing score from the given data came out to be 98% and the training score is 99%.

- The Factors most affecting the possibility of crime were location and hour of the day as concluded from the feature importance graph shown below.



- The Factors least affecting the possibility of crime were the day of the year, and week.

V. CONCLUSION

The proposed model aims at predicting the total number of incidents of a particular crime in a state. We approached the problem with three different approaches, KNN, Decision Tree Regression and Random Forest Regression. The original dataset had to be trained and processed differently for each algorithm. This was accomplished by implementing python scripts. The data was then fed into the method and a test model was generated. This model was then tested with the help of test values and the model was then made to predict the number of a particular crime that will occur in that state in the given year. The error in the test predictions were very high which led to further training of the models. From the test run of the algorithms it was evident that the Random ForestTree Regression algorithm gave the predictions.

Once an acceptable model was ready for each algorithm, it was made to predict the number of crime instances in 2022. Government bodies can use this data and predictions to come up with different policies and laws for each state depending on the type of crime which is expected to be prevalent in the next year. More police force can be deployed in the areas which have a history of crimes being committed around them. The people living in these states can also use this data to learn how to defend themselves from these threats. A little more effort from everyone and we can bring down this number to 0 in the years to come.

In our analysis high R-squared values could not achieved. This is not due to wrong training of the model but rather because this is a human psychological analysis. There are numerous factors which affect a person's psychological state, which could lead to committing a crime. The values predicted are based on the past trends in the numbers of crimes committed and does not put a hard value on the numbers predicted.

REFERENCES

[1] Pratibha, A. Gahalot, Uprant, S. Dhiman and L. Chouhan, "Crime Prediction and Analysis," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170731.
[2] Vojislav Kecman, "References," in Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models , MIT Press, 2001, pp.531-538.
[3] https://machinelearningmastery.com/time-series-forecasting-super vised-learning/
[4] https://data.gov.in
[5] https://www.kaggle.com/yashraut/indore-police-crime-dataset
[6] https://www.researchgate.net/publication/322541877_SURVEY_ON_CRIME_ANALYSIS_AND_PREDICTION_USING_DATA_MINING_TECHNIQUES
[7] Mitchell, T. M. (1997), Machine Learning , McGraw-Hill , New York .

