IARJSET

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

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81135

CRIME ANALYSIS USING PREDICTIVE MODELING

DIVYA CHOPRA[1*], DEEPANSHU KAUSHIK[2]

[1,2]Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

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 in a particular crime. This research aims at providing people a thought on how crime patterns can be analyzed and can help in creating a crime free neighborhood with the help of supervised learning. K-nearest neighbor algorithm was used to find locations that are vulnerable to the classified crime. As of now, we have taken six classes of crime: Robbery, Accident, Gambling, Violence, Kidnapping, and Murder. 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.

KEYWORDS- Crime Analysis, Predictive Modelling, K-nearest neighbor, Supervised Learning

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 to 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 next predictions. Since crime rates and places change dynamically we have to train our models regularly to get an accurate result[1].

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, Random Forest** to get the best predictions possible.

TIMES SERIES ANALYSIS FOR PATTERNS

The factor which is of utmost importance in ensuring success in a business is Time. In today's world, it is very difficult to keep up with the pace of time. Time Series Modeling is a powerful method using which we can see ahead of time. A continuous list of data points listed or graphed in time order is used for time series analysis. Time series plots are usually plotted with the help of line charts. Time series is largely used in any science and engineering domain where temporal measurements are involved[2].

The time series analysis can be defined as the methodology to examine and scrutinize the time series data so that insightful statistics and valuable information can be obtained from it. This information may not be visualized as usual. Time series analysis is also the beginning to the time series forecasting where future values are predicted based upon the previously observed values. This kind of analysis finds its application in diverse types of data which includes continuous data, real-valued data, discrete numeric data and discrete symbolic data. Statistical inference is a part of time series prediction and a particular approach to such an inference is known as predictive inference. Time series analysis of the data will give a **visualization of the data** which will help us to assess the pattern of the dataset i.e. whether the particular factor we are measuring is increasing or decreasing with respect to a certain period of time.

A common notation of a time series X indexed by the natural numbers is written

$$\_X = \{X1, X2, \dots\}$$

We can get datasets for each crime for example Murder, rape, kidnapping for individual states/Union Territories. The datasets consist of the numerical value of that crime in each state according to increasing order of time. We can find the patterns in the data using time series models and then can use the pattern to predict the crime at a time and place using different regression models.

MACHINE LEARNING FOR PREDICTING FUTURE RATES

A kind of 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 is 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.

1. **KNN (K nearest neighbor):** K nearest neighbor falls under supervised learning algorithm that can implement the regression and classification problems.

2. **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

3. **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** [3][4].

IMPLEMENTATION

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 stop such activities in that state. In order to achieve the goals, we have used the **architecture/workflow** diagram as shown.

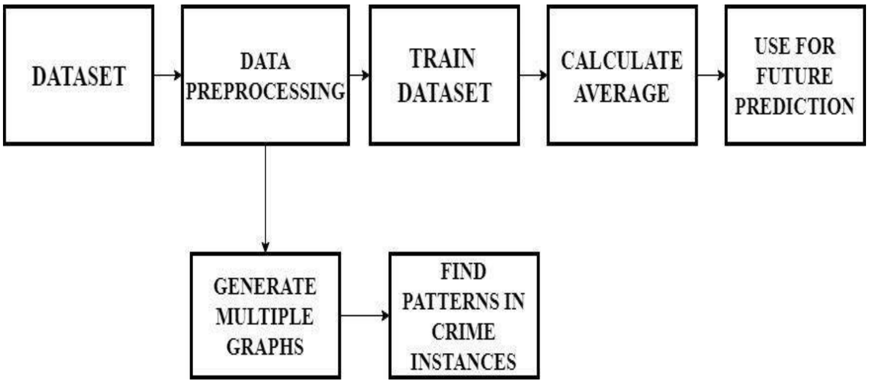
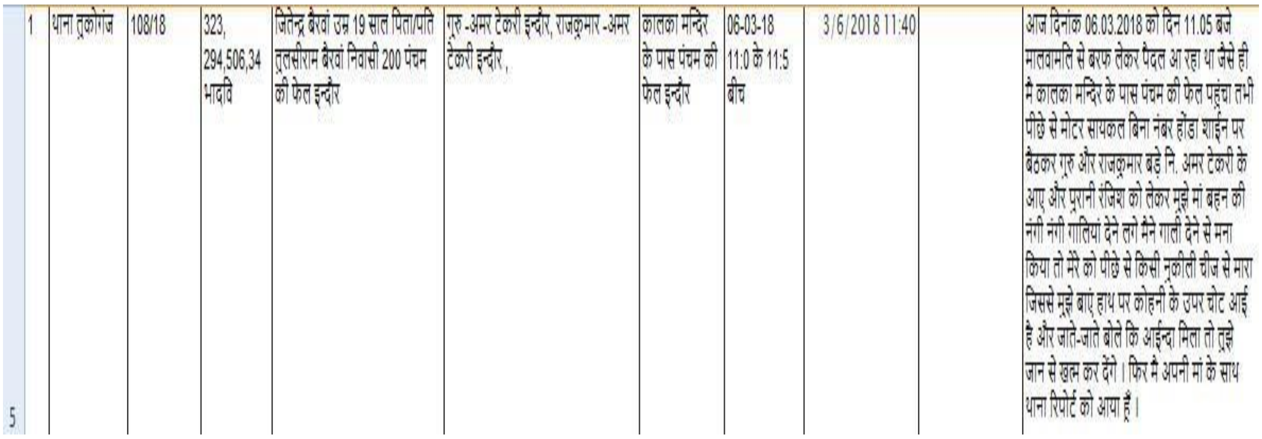


Fig 1- Workflow Diagram

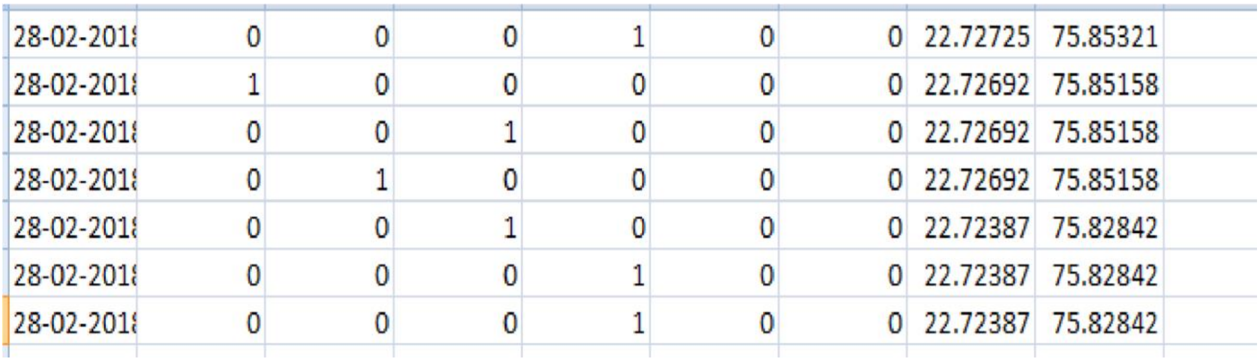
DOI: 10.17148/IARJSET.2021.81135

Multiple dataset files were obtained from government website [5] which included the section of crime , time it took place, name of the person reported the crime, location of crime. The data found was in Hindi, so the first task was to convert the data into English. After converting the data, the first step was to preprocess the data. This included two main steps: *first*, handling the missing values. The data was cleaned by removing any inconsistencies and removing the data that is not for use like who reported the crime or place since we already have the longitude and latitude. Initially we have six categories of crime that are Robbery, Accident, Violence, Gambling, Murder and Kidnapping for **act 379, 279, 323, 13, 302, 363** respectively[6].



5	1	थाना तुकोगंज	108/18	323, 294,506,34 भादवि	जितेन्द्र बैरवां उम्र 19 साल पिता/पति तुलसीराम बैरवां निवासी 200 पंचम की फेल इन्दौर	गुरु -अमर टेकरी इन्दौर, राजकुमार -अमर टेकरी इन्दौर ,	कालका मन्दिर के पास पंचम की फेल इन्दौर	06-03-18 11:0 के 11:5 बीच	3/6/2018 11:40		आज दिनांक 06.03.2018 को दिन 11.05 बजे मालवामिल से बरफ लेकर पैदल आ रहा था जैसे ही मै कालका मन्दिर के पास पंचम की फेल पहुंचा तभी पीछे से मोटर सायकल बिना नंबर होंडा शाईन पर बैठकर गुरु और राजकुमार बड़े नि. अमर टेकरी के आए और पुरानी रंजिश को लेकर मुझे मां बहन की नंगी नंगी गालियां देने लगे मैने गाली देने से मना किया तो मेरे को पीछे से किसी नुकीली चीज से मारा जिससे मुझे बाएं हाथ पर कोहनी के उपर चोट आई है और जाते-जाते बोले कि आईन्दा मिला तो तुझे जान से खत्म कर देंगे । फिर मै अपनी मां के साथ थाना रिपोर्ट को आया हूँ ।

Fig 2 - Data before Pre-processing



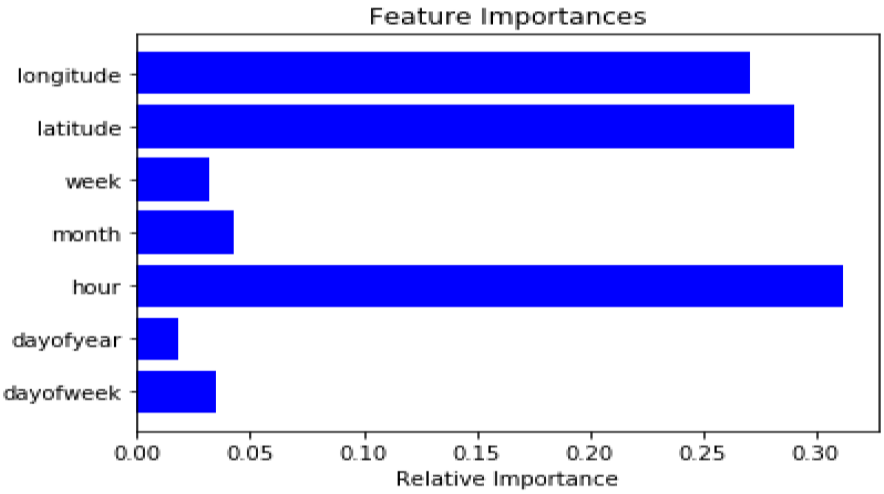
28-02-201ﾠ	0	0	0	1	0	0	22.72725	75.85321	
28-02-201ﾠ	1	0	0	0	0	0	22.72692	75.85158	
28-02-201ﾠ	0	0	1	0	0	0	22.72692	75.85158	
28-02-201ﾠ	0	1	0	0	0	0	22.72692	75.85158	
28-02-201ﾠ	0	0	1	0	0	0	22.72387	75.82842	
28-02-201ﾠ	0	0	0	1	0	0	22.72387	75.82842	
28-02-201ﾠ	0	0	0	1	0	0	22.72387	75.82842	

Fig 3-Data after Pre-processing

The data thus obtained was implemented on various algorithms for training. The supervised learning algorithms that were applied in the paper were Support Vector Regression, Decision Tree Regression and Random Forest Regression. We bifurcated the data into testing and training so as to apply algorithms on both separately, thus training the data. We used the Python programming language for each of the algorithms. **Scikit library** is used for the model building algorithms. We generated plots for every model for the training set in order to visualize the fitness of the model. Following this, we generated the predicted values for the testing data and matched the difference in the errors obtained in the actual and predicted values. The model which gave the least error was used for the prediction of the crime rates for the future rates.

RESULTS

- Out of all the tried algorithms we got the best result of **99% accuracy** with **Random forest.**
- The Factors most affecting the possibility of crime were **location and hour of the day**.
- The Factors least affecting the possibility of crime were day of year, week



FUTURE SCOPE

The Goal of a society should be to create a safe environment and crime analysis can help in achieving it

- **Predicting crime using face recognition**: We can use face recognition and crime analysis algorithms to predict if an individual will commit a crime or not. If there is any suspicious change in the behavior of anyone on a crime hotspot (predicted by the crime analysis algorithm) the system will detect it and alert the police. For example- If a camera captures the same person after most of the robberies in the neighborhood it can alert the police.
- Predicting **Crime Hotspots**: By using historical data and observing where recent crimes took place we can predict where future crimes will likely happen. For example if a snatching gang is targeting a specific area we can easily catch them using crime analysis. System can show the crime hotspots on a map.
- Using Crime Rates as a factor for house hunting- while looking for a house people usually look for a **good locality**, **property rates** etc.

REFERENCES

[1] https://www.researchgate.net/publication/322541877_SURVEY_ON_CRIME_ANALYSIS_AND_PREDICTION_USING_DATA_MINING_TECHNIQUES
[2] https://machinelearningmastery.com/time-series-forecasting-super vised-learning/
[3] Tom M. Mitchell, "Machine Learning
[4] Vojislav Kecman, "Learning and Soft computing: support vector machines, neural networks and fuzzy logic models"
[5] https://data.gov.in
[6] https://www.kaggle.com/yashraut/indore-police-crime-dataset