International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

CRIME PREDICTION AND VISUALISATION USING DATA ANALYTICS

Ram Sridhar[1], Chakradhar.D[2], Thiagaraj.S[3], Dr.P.Mohamed Fathimal[4]

[1,2,3]Student, Dept. of Computer Science & Engineering, SRM IST., Tamil Nadu, India

[4]Assistant Professor, Dept. of Computer Science & Engineering, SRM IST., Tamil Nadu, India

---

Abstract - *Crime analysis and prevention is a structured and methodical application which is utilised to identify and analyse models, patterns and trends in crime. The system analyses the crime data, evaluates the likelihood for crime occurrence in each area and builds visualisations for the same. The ubiquitous use of computerized systems for a wide array of applications paves way for crime data analysts to assist the law in pacing the process of solving and thereby preventing crimes. There has been far-reaching implications of crime across India in recent times developing to being a huge menace for the government and the citizens of the country[1]. Thus we aim to study these patterns of crime based on the data acquired from Indian Government website. The algorithms such as Linear Regression, Random Forest and Support Vector Machine were employed to achieve the objective. The output is generated using simple visualisation charts.*

Key Words: **data analytics, crime analysis, linear regression, decision tree, random forest, support vector machine**

1. INTRODUCTION

The purpose of the project is to better facilitate the monitoring and regulation of crime occurrences by furnishing an intricate and detailed outlook of the patterns that could be noticed in the forthcoming years. There is also the additional option to visualise the crime reporting of the previous years to discern critical areas that require to be looked upon. The process of implementing predictive analytics over crime data comprises of the factors- population, geographical areas, social issues, etc. The entire system is established through a website - "Indian Crime Analysis". The website provides three main services - predictions, visualizations and observations. Under the predictions service, four major sectors of crime is considered - crime against women, crime against children, crimes under the Indian Penal Code (IPC) and crimes under Special and Local Laws (SLL).

The model is designed to predict crime rates for the next few years. The visualisation makes use of factors like population, literacy rate corresponding to each demographic area. The model is built to produce a clear and lucid representation of the various ills of society. The observations will be of critical use to the concerned authorities to execute appropriate actions to alleviate wrongdoing. The application of visualisation charts will be useful analytical tools in observing the changing crime trends and ratios over time and other factors like population and demographic.

Linear Regression is the most common predictive model to identify the relationship among the variables. Apart from univariate or multivariate data types the concept is linear. Linear regression can be either simple linear or multiple linear regression. The linear regression is described as:-

$$y=x\beta + \varepsilon$$

Regression is an unpredicted change that occurs whenever the code changes. It can be used to automate testing activities in order to deal with efficiency and speed constraints.

The Random Forest (RF) is an "ensemble learning" technique consisting of the aggregation of a large number of decision trees, resulting in a reduction of variance compared to the single decision trees[2]. Random forests consist of 4 –12 hundred decision trees, each of them built over a random extraction of

the observations from the dataset and a random extraction of the features. Not every tree sees all the features or all the observations, and this guarantees that the trees are de-correlated and therefore less prone to over-fitting. Each tree is also a sequence of yes-no questions based on a single or combination of features.

When decision tree is evaluated, we must partition the data into training and testing dataset. Train- test process is repeated using cross- validation which determines the average success rate. It establishes whether testing activity will create high or low accuracy outputs.

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. The objective of the support vector machine algorithm is to find a n-dimensional space hyperplane(N — number of features) that separately classifies the data points. Hyperplanes are boundaries for decision making which help to distinguish data points. Data points which fall on either side of the hyperplane can be attributed to various groups. An SVM model is a representation of the examples as space points, mapped in such a way as to distinguish the examples of different categories by a simple distance that is as wide as possible. Then, new examples are mapped into the same space and projected to belong to a group depending on the side of the gap they fall through.

1.1 RELATED WORK

Different methodologies for crime analysis have been developed to effectively represent and understand crime. Researches so far have analysed the relation among crimes and socio-economic factors like unemployment, earnings level, level of schooling and so forth.

Bharati et al worked-on Crime Prediction and Analysis Using Machine Learning where the author used machine learning and data mining for prediction of crimes in Chicago [4]. The datasets factors vicinity description, type of crime, date, time, range, etc. Linear regression and K-Nearest Neighbours (KNN) are both predominently used in this paper to test for crime prediction. This model tried to figure out crime trends based on crime zone thereby implementing more datasets with features such as season, date, time proving to be beneficial for police task force.

In[5] Alves(2018),"Crime prediction through urban metrics and statistical learning" linear regression on two pre-processed data sets with Bonferroni correction after cross validation yielded an accuracy of 88.21%

Kang et al. analyzed crime occurrences by the usage of multimodal records in which deep learning techniques was applied [7].

A conclusion that DNN version provides greater precision values in predicting crime prevalence than other methods was drawn. Chandrasekar et al implemented Gradient Boosted trees and Support Vector Machines. The paper has discovered that the maximum suitable classifiers that possible practice on those sorts of datasets can be tree-based methods.

In[3], Nejdetdogru (2018),"Traffic Accident Detection Using Random Forest Classifier",The system proposed used data collected from vehicular adhoc networks (VANETs) based on the speeds and coordinates of the vehicles and sends alerts to the driver. The model developed used SVM and RF to distinguish between traffic accidents and normal cases.RF algorithm hasshowed better performance with 92.54% accuracy than SVM with 89.72%.

Crime Prediction using Ensemble Approach by Almaw et al investigates on hidden crime data mining [6]. Bayes classification and artificial network to test for crime analysis. Crime trends and patterns are identified in a specific location with co-ordinates in a specific time frame factoring means of entry (front door, window, etc.), day of the week, characteristics of the property (apartment, house), and geographic proximity to other break-ins.

1.2 SYSTEM DESIGN

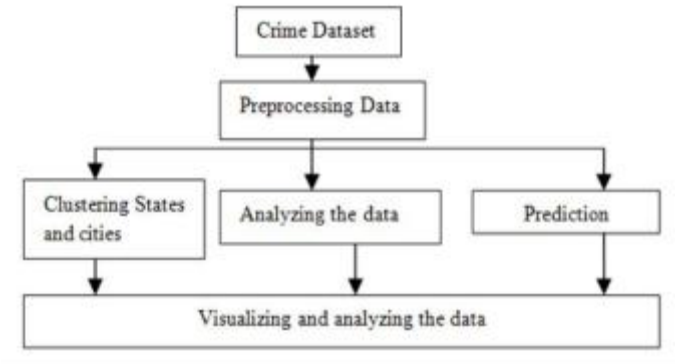

Fig -1: Architecture Diagram

Fig.1 demonstrates the process involved in developing the model. The block diagram represents the key steps involved in the development of the proposed model. The key phases are data acquisition, data analysis and prediction.

- Information gathering : Data extraction of various Crime Report against women, children, senior citizens, Indian Penal Code and Special and local laws crimes.
- Data Cleaning : Data is converted to suitable format using data mining techniques such as eliminating missing values, eliminating redundant data, data transformation, etc.
- Clustering states : Grouping the states based on factors like population, literacy rate, demographics, etc.
- Algorithm Implementation : The extracted data is fed to the suitable algorithm. The algorithms implemented are Linear regression, Support vector machine and Random forest regression. We implemented the algorithms and found Linear Regression as the best fit. The system also predicts the rate of crime for the next five years.

Decision trees were constructed using bagging and feature randomness to create and uncorrelated forest of trees that were used to predict future trends using random forest based on the training data.

Linear regression was used to represent the dependent variable as a sum of the Y-intercept and the product of the slope co-efficient with the independent variable. Error component was taken into consideration.

● Visualisation : The observations and patterns that were deduced by the model was represented in the form of bar graphs, line graphs and scatter plots. Generated output was to compare data across different demographic and population for State-level comparative analysis.

2. METHODOLOGY

Future Crime Rate Prediction and Visualisation : The crime patterns of the next four years are predicted under this module. The system in built using python. The analysis is performed using the python packages numpy, pandas, seaborn, sklearn,etc. Plotly and seaborn packages are employed for the visualisations.

To quantitatively forecast crime different data mining techniques can be used. The associated task for the dataset we have used in this paper is linear regression, svm and random forest. The linear regression model provides description of how the scalar response affects the output per se. A variable Y (target variable) is predicted as a mathematical linear function of another variable X (input variable/features), given n training examples of the form (x1,y1), (x2,y2), ..., (xn,yn), where xi ε X and yi ε Y. The formula used is J(theta_0, theta_1) = 1/(2m) * sum_(i=1)^m [h_theta(x^i) - y^i]^2 = 1/(2m) * sum_(i=1)^m [h_theta(x^i) - y^i]^2.

The objective of using linear regression is to minimize the cost function so that the function J(theta_0, theta_1) is proximate to y for all the training examples (x,y). way to minimize the cost function is to use the batch gradient descent algorithm. To improve the efficiency of the model, every feature is normalized in the range [-1,1].

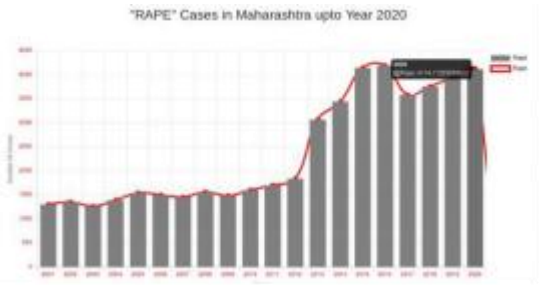

Chart -1: Prediction

Random forests are a learning method used for classification, regression etc and operate by constructing a plethora of individual decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

The technique of bagging was used in our project to create thousands of decision trees with minimal correlation In bagging, the random subset of training data is selected to train each tree. Furthermore, our model randomly restricts the variables which may be used at the splits of each tree. Hence, the trees that dissimilar, retain predictive power.

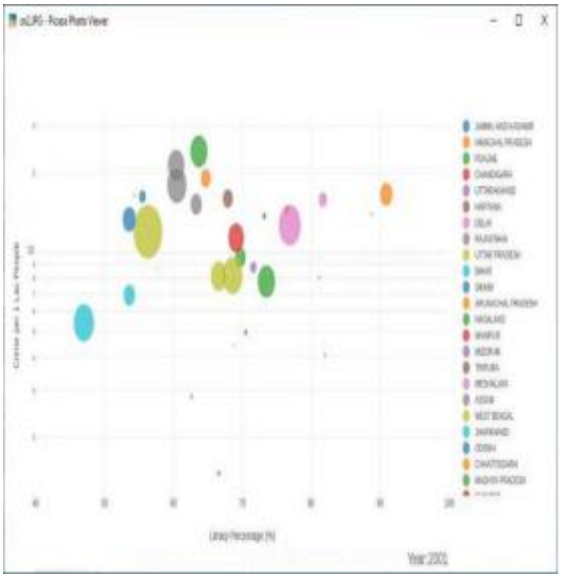

Chart -2: Visualisation

Random forests are a learning method used for classification, regression etc and operate by constructing a plethora of individual decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

The technique of bagging was used in our project to create thousands of decision trees with minimal correlation In bagging, the random subset of training data is selected to train each tree. Furthermore, our model randomly restricts the variables which may be used at the splits of each tree. Hence, the trees that dissimilar, retain predictive power.

Mean squared error(MSE) was mainly incorporated to analyse data branch from each node as follows

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

To make a prediction on a futuristic basis, random forest aggregates the predictions from its set of individual decision trees each pertaining to various factor. The data gets split into many subsets and it compares the train and test data to arrive at a conclusion. In accordance to this process, each subset has its own predicted class. The final result of the model is a thorough comparison of the individual predicted class that incorporates the factors taken into consideration as such.

3. CONCLUSIONS

The result is to build a system for analyzing, predicting and correlating crimes from the humongous data available. The output will be in the form of correlation between the types of crimes and the state/city, location,etc. Various techniques and algorithms are employed for prediction.

It was observed that Random Forest overfitted the data at various point thereby yielding results that were not in consideration to actual trends observed in the year 2016-2020. Algorithms like Random forest regressor tend to have a higher accuracy rate than normal for inconsistent and unpredictable data rendering it ineffective. It can be concluded that Random forest using subset of various parameters and characteristics took into account several anomalies thereby resulting in inaccurate predictions. Excessive bagging and feature randomness could be the cause for the same.

In comparison, linear regression using Mean Squared Error(MSE) yields effective results even with error correction co-efficient implemented. Predictors along with its subset was used to find out the best (most appropriate)combination between the dependant and independent variable thereby more accurately analysing future trends. MSE was made further negligible on performing lasso regression, making it much more feasible than RF. The other two algorithms tends to overfit the data and provide results even when the data wasn't suitable for prediction.

The accuracy of each algorithms are listed below:-

Algorithm	Accuracy
Linear Regression	0.8094431949335529
SVM Regression	0.8089961972390511
Random Forest	0.8096154847789507

Though RF has a greater accuracy than Linear Regression, the former overfits the data, thus becoming unviable. Keeping in mind the consistency of the data and the type of results expected Linear regression is employed in for prediction results.

The system is designed keeping in mind the effort that is taken by legal force departments to maintain law and order throughout society. Future improvements may include secure access to the website by linking to Aadhaar to uniquely verify the user, periodic and dynamic extraction of the dataset from the government authorities.

REFERENCES

[1] Ahamed Shafeeq B M1, Dr. Binu V S2, Spatial Patterns of Crimes in India using Data Mining Techniques, Certified International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 11, ISSN: 2277-3754 ISO 9001:2008, May 2014

[2] Feature Selection Using Random forest-Towards Data Science

[3] NejdetDogru,"TrafficAccident Detection Using Random Forest Classifier",978-1-5386-2659-7/18/$31.00 ©2018 IEEE.

[4] Alkesh Bharati and Dr Sarvanaguru RA.K. 2018. Crime Prediction and Analysis Using Machine Learning.

[5] L. G. A. Alves, H. V Ribeiro, and F. A. Rodrigues, "Crime prediction through urban metrics and statistical learning," Physica A, vol. 505,pp. 435–443, 2018

[6] Ayisheshim Almaw and Kalyani Kadam. 2018. Survey Paper on Crime Prediction using Ensemble Approach, International Journal of Pure and Applied Mathematics, 118 (8), 133-13

[7] Hyeon-Woo Kang and Hang-Bong Kang. 2018. Prediction of crime occurrence from multimodal data using deep learning. DOI= https://doi.org/10.1371/journal.pone.0176244