Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model

Al Amin Biswas
Dept. of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka
alaminbiswas.cse@gmail.com

Sarnali Basak
Dept. of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka
sarnali.cse@juniv.edu

***Abstract*—Last few years in Bangladesh, the crime rate has increased rapidly. Hence it is an essential task to analyze and predict the crime so that the authority can minimize or prevent the crimes easily. In this situation, machine learning can perform a notable role to reveal the crime trends and patterns of Bangladesh. Here, various machine learning regression models i.e. linear regression, polynomial regression, and random forest regression are used to forecast the trends and patterns of crime in Bangladesh. Dataset used in this research is available for the public which is gathered from the Bangladesh police's website. The dataset comprises record about various crime types i.e. dacoity, robbery, kidnapping, murder, women & child repression, theft, burglary, arms act, explosive, narcotics, and smuggling of Bangladesh. Firstly, training of regression models is done on the training dataset. After completion of the training, forecasting of crime is performed on the test data by the different regression models. Then we compare the forecasting results with the actual results and calculate the model evaluation metrics for the different applied regression models. After comparing the result, it is possible to find out the best-suited regression model for the crime-related data among all the applied regression models. Finally, it is observed that polynomial and random forest regression are better to predict the crime trends and patterns than the linear regression.**

Keywords— Crime Forecasting, Machine Learning, Linear Regression, Polynomial Regression, Random Forest Regression.

I. INTRODUCTION

Crime is social oppression which is considered harmful to the welfare of the citizens. It is an impediment to development for any country. So, it is essential to minimize the crime rate to speed up the development of a country. Rate of crime has increased rapidly within the last few years in Bangladesh which is very much alarming for the advancement of a country like Bangladesh. Hence, it is necessary to analyze the crime data to forecast the crime patterns and trends so that the law enforcement authorities can take some outstanding action to minimize the crime. Because of the rapid development of technology, a vast amount of crime data are available in the various authorized website. Here, various crime types in Bangladesh i.e. dacoity, robbery, kidnapping, murder, women & child repression, theft, burglary, arms act, explosive, narcotics, and smuggling are counted. For the computational purpose, we have used real dataset which is accessible on the Bangladesh Police's official website [1,2,3,4]. The dataset comprises the various type of crimes classified by the department of Bangladesh police and are used to be predicted. Here, three regression algorithms are used to forecast the crime trends and patterns for the Bangladesh context. All the regression models are trained on the crime dataset. After completion of the training, crime prediction is performed for the various crime types in the context of Bangladesh.

This research paper is arranged as follows: section II illustrates the related existing works entitled as a literature review; Research methodology is described in section III; Results and discussions of this research are described in section IV; lastly, section V concludes this research including future works.

II. LITERATURE REVIEW

Domingos et al. [5], stated the elementary goal of machine learning is to generalize beyond the samples in the training set. So it is possible to predict something with a very high level of accuracy from the training set.

Awal et al. [6], applied only the linear regression model to forecast the trends of crime in Bangladesh. Dataset used in this research is gathered from Bangladesh police's website. They have made a prediction about crime for the various regions of Bangladesh.

McClendon et al. [7], showed a comparative analysis between the actual statistical crime data and a given dataset of the Mississippi state. Three different machine learning algorithms were performed and claimed that the linear regression outperforms among them.

One Nearest Neighbor, Decision Tree, Support Vector Machine, Naïve Bayes, and Neural Network were utilized to predict the hotspot of crime, stated by Yu et al. [8]. They ensemble these classification models to obtain the most realistic results.

Stec et al. [9], stated that crime prediction is taking advantage of deep neural networks to get crime count of next day in a fine-grain city partition. The counts of crime are divided into ten bins and neural network forecast the most suitable bin. Training of Chicago and Portland crime data is done using increasingly complex neural network structures. With best, it was possible to forecast the right bin for the overall count of crime with 65.3% correctness for Portland and 75.6% correctness for the Chicago.

978-1-7281-1711-9/19/$31.00 ©2019 IEEE

Iqbal et al. [10], applied Naïve Bayesian and Decision Tree to a dataset to forecast the crime category for several states in the USA. Between two, Decision Tree exceeded the Naïve Bayesian and attained accuracy of 83.95%.

Nath et al. [11], used a clustering technique for a data mining strategy which assists to identify the patterns of crimes and speed up the crime-solving procedure. Some enhancements of k-means clustering were implemented to help in the process patterns identification of crime and validate obtained results.

Saeed et al. [12], applied Decision Tree and Naïve Bayes classifier to the criminal activity dataset to forecast the attributes and also event outcomes. After comparing, they stated that the Naïve Bayes is much stable and more precise on analysis of crime and can extract rules for classification.

Tayal et al. [13], suggested a technique for the design and implementation of detection of crime and identification of criminal in Indian cities. Detection of crime is examined using the k-means clustering technique, which iteratively produces two clusters of crime that are based on related attributes of crime. Identification of criminal and prophecy are examined utilizing KNN classification. Finally, 93.62% and 93.99% accuracy is measured, respectively, in the production of two clusters of crime using chosen attributes of crime.

Yu et al. [14], designed and developed the Cluster-Confidence-Rate-Boosting algorithm to effectively decide appropriate local spatio-temporal patterns to form a global pattern of crime from the training dataset which later used to predict future crime. The outcomes revealed that the recommended CCRBoost algorithm has gained nearly 80 percent on accuracy in prediction.

Das et al. [15], used various types of classification algorithm i.e. K-Nearest Neighbor, Decision Tree, Naive Bayes, Random Forest, and AdaBoost to predict various types of crime i.e. kidnapping, rape, murder, and dowry death. After analyzing the crime, classifiers are applied to forecast the forthcoming trends of crime in Indian states and union territories.

Yadav et al [16], performed an experimental analysis on a dataset to forecast the resolution. A standard dataset was used here which was collected from San Francisco Police Department Crime Incident Reporting System. CART, Gaussian Naive Bayes, K-NN, and Multilayer Perceptron (MLP) were used to predict. This analysis showed that the CART algorithm was presenting the best accuracy to predict (i.e. 80.86%).

In [17], crime prediction was done based on the machine learning model. Immediate fifteen years of crime data of Vancouver was examined by two distinctive data processing strategies. Boosted Decision Tree and K-nearest-neighbor were performed and the accuracy of crime prediction was obtained within 39% to 44%.

A crime data mining study was presented in [18] reviewing ANN, Rule Induction, Decision Tree, Genetic Algorithm, and Nearest-Neighbor method.

Ajao et al. [19], presented cubic polynomial least square regression as a variant approach of performing forecast in the business sector rather than the typical linear regression. This research presents that polynomial regression is a beneficial option with a really great determination coefficient.

III. METHODOLOGY

To predict the number of crime, several machine learning models can be applied. Here, three regression models named as linear regression, polynomial regression, and random forest regression is considered to predict. Several regression models are mainly varied based on the various types of relationships within independent and dependent variables. Regression models are done forecasting the results based on the independent variables. This is mainly applied to find out the relationship between the variables and predicting. This section is partitioned into four parts, namely, A) Dataset Description B) Different Types of Regression Model Selection C) Model Evaluation Metrics and D) Forecasting Procedure

A. Dataset Description

The dataset contains records about various types of crime (i.e. dacoity, robbery, kidnapping, murder, women & child repression, theft, burglary, arms act, explosive, narcotics, and smuggling), collected from Bangladesh Police's website [1, 2,3,4]. For example, according to the dataset, the total number of murder in Bangladesh in 2014 is 4514. In this paper, 94% of data are utilized for the training purpose and 6% of data are utilized for testing purpose. The dataset contains records about various types of crime from 2002 to 2018.

B. Different Types of Regression Model Selection

a) Linear Regression Model: Linear regression is a supervised learning algorithm that performs a regression task. Linear regression performs prediction of the value of the dependent variable (y) from the independent variable (x). Hence, this model determines a linear relation between x and y.

Hypothesis function for linear regression can be represented as

$$y = \theta_0 + \theta_1 x \tag{1}$$

x (input variable) and y (target variable) are given at the time of performing the training of the model and it suits the best line to forecast y (target variable) for a given x (input variable). Linear regression model generates the best regression fit line by determining the best values of θ_0 and θ_1. Then by using our model, we can forecast the value of y (target variable) for the value of input variable x.

Where θ_0 is intercept and θ_1 is the coefficient of x. After determining the best values of θ_0 and θ_1, we get the best-fitted line. So, while we are eventually using our model to forecast, the model will forecast the value of y (target variable) for the input value of x.

b) Polynomial Regression Model: Hypothesis function for polynomial regression can be expressed as

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_n x^n \tag{2}$$

Where, x = independent variable, y = dependent variable, and n is the polynomial degree.

$\theta_{1,}\ \theta_{2,}\ \ldots,\ \theta_n$ are the coefficient which determines how a unit change in x will cause a change in y and θ_0 is intercept. The value of the degree of the polynomial n can affect the results significantly.

c) Random Forest Regression Model: Random Forest regression is a version of ensemble learning. Ensemble learning method combines various base models in order to produce one predictive model which is optimal and more powerful. Steps of building a Random Forest are listed below.

1. Select k data points randomly from the training set.
2. Build a decision tree associated with the selected k data points.
3. Repeat step 1 and 2 for n times to produce n trees.
4. Predict the value of a data point using n trees and find out the average of all the predicted values.

C. Model Evaluation Metrics

a) Explained Variance Score (EVS):

Explained Variance Score is a type of metric which helps us to calculate the ratio between the variance of error and true values. Mathematically, Explained Variance Score (EVS) can be calculated using the following formula:

$$\text{explained\_variance}(y,\hat{y}) = 1 - \frac{\mathrm{Var}\{y-\hat{y}\}}{\mathrm{Var}\{y\}} \qquad (3)$$

Where, y =true value, ŷ = predicted value and Var = variance. From the formula, the possibility of the best score as well as the highest score of EVS is 1.0 and the lower value of EVS is worse.

b) Mean Absolute Error (MAE) :

If $\hat{y}_k$ is the predicted value of the kth sample and y_k is the corresponding true value, then the Mean Absolute Error (MAE) is expressed as

$$\mathrm{MAE}(y,\hat{y}) = \frac{1}{N}\sum_{k=1}^{N} |y_k - \hat{y}_k| \qquad (4)$$

Where N= Number of samples.

c) R Squared Score (R^2Score):

If $\hat{y}_k$ is the predicted value of the kth sample and y_k is the corresponding true value, then the score of R² is expressed as

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{k=1}^{N}(y_k-\hat{y}_k)^2}{\sum_{k=1}^{N}(y_k-\bar{y})^2} \qquad (5)$$

Where, y = true value, ŷ = predicted value, ȳ = average of all the true values, and

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k \qquad (6)$$

N = Number of samples.

From the formula, the possibility of the best score as well as the highest score is 1.0 and in the rare case, it can be less than 0.0 which represents that the model is arbitrarily worse.

D. Forecasting Procedure

This section mainly describes the forecasting procedure of crime in Bangladesh. Crime forecasting procedure consists of several steps.

Forecasting procedure to find the trends and patterns of crime in Bangladesh is described in Fig. 1.

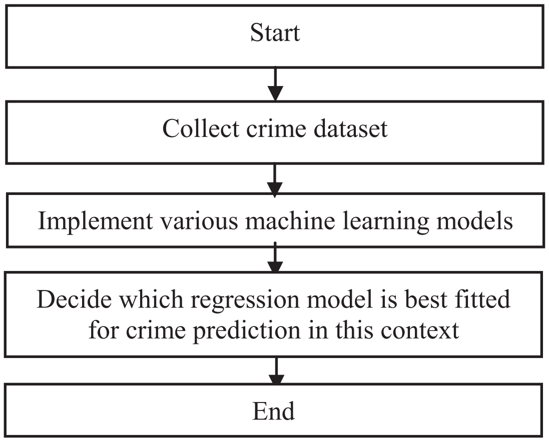
Start
Collect crime dataset
Implement various machine learning models
Decide which regression model is best fitted for crime prediction in this context
End

Fig. 1. Forecasting procedure to find the trends and patterns of crime in Bangladesh.

Implementation of the various machine learning models is further divided into the following step shown in Fig. 2.

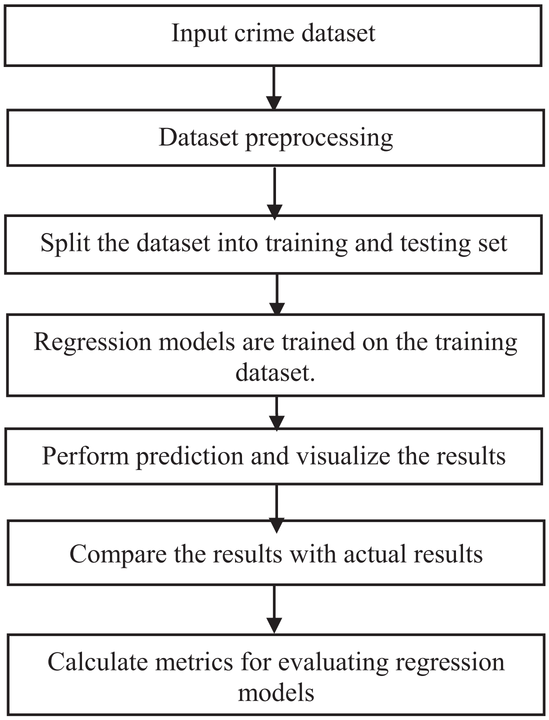
Input crime dataset
Dataset preprocessing
Split the dataset into training and testing set
Regression models are trained on the training dataset.
Perform prediction and visualize the results
Compare the results with actual results
Calculate metrics for evaluating regression models

Fig. 2. Implementation of the various machine learning models.

IV. RESULTS AND DISCUSSIONS

This section mainly shows the experimental results (tabular form and graph) obtained from the implementations of all the working regression algorithms. All the working regression models are applied to forecast each of the following crime: dacoity, robbery, kidnapping, murder, women & child repression, theft, burglary, arms act, explosive, narcotics, and smuggling. Table I shows how accurately the regression models are able to predict the trends of crime in the context of Bangladesh.

TABLE I. COMPARISON BETWEEN ACTUAL RESULTS AND PREDICTED RESULTS BY VARIOUS REGRESSION MODELS FOR DIFFERENT TYPES OF CRIME IN 2018 IN BANGLADESH

Crime Category	Actual Number of Crime	Predicted Number of Crime		
		Linear Regression	Polynomial Regression	Random Forest Regression
Dacoity	262	387	286	384
Robbery	562	832	637	690
Murder	3830	4114	3544	3705
Woman & Child Repression	16253	21063	21653	17988
Kidnapping	444	696	646	596
Burglary	2137	2269	1759	2196
Theft	5561	6972	4205	5972
Arms Act	2515	1703	2756	2263
Explosive	1310	486	708	445
Narcotics	112549	68926	90304	78625
Smuggling	4501	6611	4844	5140

We also calculate model evaluation metrics of the regression models for the overall crime prediction which is shown in TABLE II. It is a measure which helps us to understand the best-fitted regression model.

TABLE II. MODEL EVALUATION METRICS OF DIFFERENT REGRESSION MODELS FOR CRIME PREDICTION

Name of Regression Model	EVS	MAE	R^2 Score
Linear Regression	0.83	4968	0.82
Polynomial Regression	0.95	2832	0.95
Random Forest Regression	0.90	3492	0.89

The following line graph shows the actual trends versus forecasting trends. In order to compare, we have shown actual and predicted trends of three significant crimes (Robbery, Arms Act, and Smuggling) from 2002 to 2018. Fig. 3, Fig. 4, and Fig. 5 represents comparative analysis among 3 different regression models for the year 2018 and actual trends from 2002 to 2018 respectively.

We have used the crime data from the year 2002 to 2017 for the training purpose. To test the performance of regression algorithms, we have considered the most recent year (2018) data for which actual crime data is also available. For the simplicity of understanding the trends of crime, we have selected three crime categories here.

Comparison between actual trends and forecasting trends of robbery is depicted in Fig. 3.

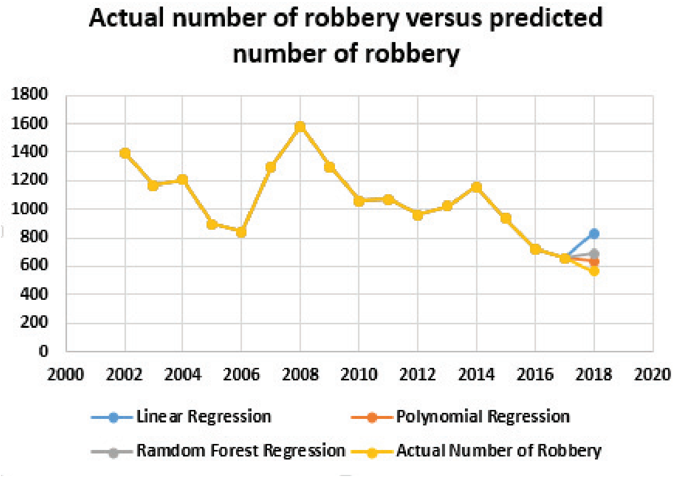

Fig. 3. Actual trends versus forecasting trends of robbery.

From the result, the actual number of robbery for the forecasting year (2018) is 562. In contrary, the predicted result is 832, 637, and 690 for the linear, polynomial and the random forest regression respectively. So, it indicates that the polynomial regression outperforms than other two regression algorithm and the random forest predicts better than linear regression.

Fig. 4 shows the comparison between the actual trends and the forecasting trends of arms act.

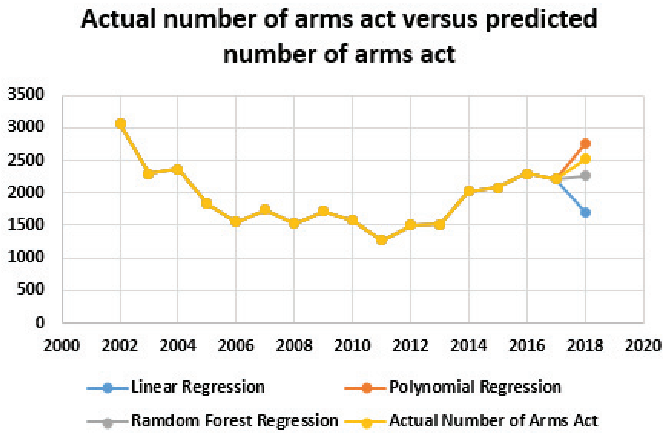

Fig. 4. Actual trends versus forecasting trends of arms act.

From the result, the actual number of arms act for the year 2018 is 2515. Here, it indicates that both the polynomial regression and random forest regression performs very close to the actual number of arms act. So, these two regression algorithm outperforms the linear regression algorithm.

Fig. 5. illustrates the statistical difference between the actual result and the forecasting result of smuggling.

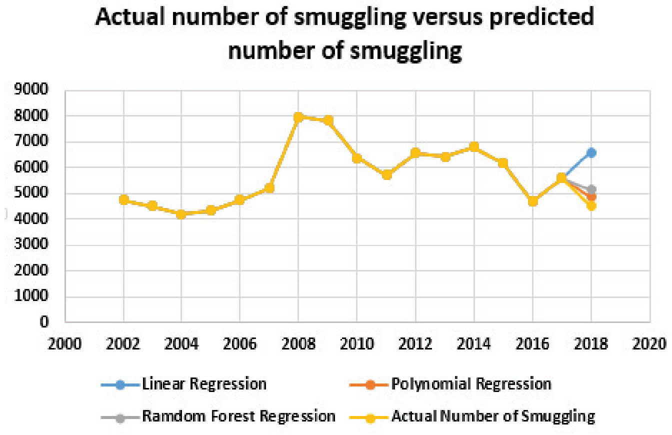

Fig. 5. Actual trends versus forecasting trends of smuggling.

For 2018, the actual number of smuggling is 4501. Here, it is observable that the polynomial regression predicts very close to the actual result and also outperforms than other two regression algorithms.

V. CONCLUSION AND FUTURE WORK

In this work, three machine learning models are applied to predict the crime trends and patterns of Bangladesh. This analysis may help the department of Bangladesh police as well as many agencies of law enforcement to minimize the crime by taking some emerging steps. Different regression models are trained on crime dataset of earlier years. After completion of the training, regression models are used to forecast for the year of 2018. All the predicted outcomes alongside with actual crime occurred has shown in this paper. The above-mentioned tables show the accuracy and evaluation metrics result for the various regression models to forecast the trends of crime in Bangladesh. From the obtained result, it is observed that polynomial and random forest regressions perform better for this particular dataset when to predict. In contrary, random forest regression outperformed than the linear regression, but it does not scale particularly well regarding training set for time-series data. In this paper, we perform the entire prediction for a year only because taking more years as test data do not vary the outcome for the random forest regression. On the other hand, linear and polynomial regression do vary in outcomes in terms of multiple years. Therefore, we should have to choose linear regression in the time of solving the linear problem. In contrary, polynomial regression and random forest regression could be chosen for the non-linear relationship problem. This analysis work can be extended by considering multiple years as testing data and also it is possible to apply many others regression model into the given dataset and predict the crime trends for a series of years.

ACKNOWLEDGMENT

We would like to thanks the department of Bangladesh Police for providing access to the dataset for all that helps us to accomplish this research.

REFERENCES

[1] 'Comparative Crime Statistics: 2002 – 2015', Available online: https://www.police.gov.bd/en/comparative_crime_statistics:__2002_-_2015 [Last accessed on March 01, 2019]

[2] 'Crime Statistics 2016', Available online: https://www.police.gov.bd/en/crime_statistic/year/2016 [Last accessed on March 01, 2019]

[3] 'Crime Statistics 2017', Available online: https://www.police.gov.bd/en/crime_statistic/year/2017 [Last accessed on March 01, 2019]

[4] 'Crime Statistics 2018', Available online: https://www.police.gov.bd/en/crime_statistic/year/2018 [Last accessed on March 01, 2019]

[5] P. Domingos, "A few useful things to know about machine learning," Commun. ACM, vol. 55, no. 10, pp. 78–87, 2012.

[6] M. A. Awal, J. Rabbi, S. I. Hossain, and M. M. A. Hashem. "Using linear regression to forecast future trends in crime of Bangladesh," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 333-338, 2016.

[7] L. McClendon and N. Meghanathan, "Using machine learning algorithms to analyze crime data, " Machine Learning and Applications: An International Journal (MLAIJ), vol. 2, no. 1, pp. 1-12, 2015.

[8] C. H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 779-786, 2011.

[9] A. Stec and D. Klabjan, "Forecasting crime with deep learning," arXiv preprint arXiv:1806.01486, 2018.

[10] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. S. Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian Journal of Science and Technology, vol. 6, no. 3, pp. 4219-4225, 2013.

[11] S. V. Nath, "Crime pattern detection using data mining," 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pp. 41-44, 2006.

[12] U. Saeed et al. , "Application of machine learning algorithms in crime classification and classification rule mining," Research Journal of Recent Sciences, vol. 4, no. 3, pp. 106-114, 2015.

[13] D. K. Tayal et al. , "Crime detection and criminal identification in india using data mining techniques," AI & SOCIETY, vol. 30, no. 1, pp. 117–127, 2015.

[14] C. H. Yu, W. Ding, P. Chen, and M. Morabito. "Crime forecasting using spatio-temporal pattern with ensemble learning," Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer International Publishing, pp. 174-185, May 2014.

[15] P. Das and A. K. Das, "Application of classification techniques for prediction and analysis of crime in India," Computational Intelligence in Data Mining, pp. 191-201, 2019.

[16] N. Yadav, A. Kumar, R. Bhatnagar, and V. K. Verma, "City crime mapping using machine learning techniques," The International Conference on Advanced Machine Learning Technologies and Applications, Springer, pp. 656-668, 2019.

[17] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime analysis through machine learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 415-420, 2018.

[18] S. Prabakaran and S. Mitra, "Survey of analysis of crime detection techniques using data mining and machine learning," In Journal of Physics: Conference Series, vol. 1000, no. 1, p. 012046, IOP Publishing, 2018.

[19] I. O. Ajao, A. A. Abdullahi, and I. I. Raji, "Polynomial regression model of making cost prediction in mixed cost analysis," Mathematical Theory and Modeling, vol. 2, No. 2, pp. 14-23, 2012.