MANTECH Publications

Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 7, Issue 1, January-June 2022

Crime Prediction Using Machine Learning Approach

Suyash Koule[1], Shreyash Kudache[2], Saksham Dhere[3], Aniket Keru[4], N.V Shaha[5]

Students[1, 2, 3, 4], Professor[5]

Department of Information Technology

Sharad Institute of Technology Polytechnic Yadrav, India

Email: *suyashkoule13@gmail.com[1]*

Abstract

Crime is one of the serious issues in our society. It is the most predominant aspect of our society. It is also predominant in society. So, the prevention of crime is one of the important tasks. The crime analysis should be done in a systematic way as the analysis makes it important in the detecting and prevention of crime. The analysis detects the investigating patterns and helps in the detection of trends in crime. The main of this paper is the analysis of the efficiency of the crime investigation. The model is designed for the detection of crime patterns from inferences. The inferences are collected from the crime scene, and these inferences, the paper demonstrates the prediction of the perpetrator. The paper gives the research way for the prediction of perpetrator age and gender. This paper gives two major aspects of crime prediction. One is perpetrator gender, and the other is perpetrator age. The parameters used are analysis of the various factors like the year, month, and weapon used in the unsolved crimes. The analysis part identifies the number of unsolved crimes. The prediction task involves the description of the perpetrator's age, sex, and relationship with the victim. The dataset used in this paper is taken from the Kaggle. The system predicts the output using Multilinear regression, K-Neighbor's classifier, and neural networks. It was trained and tested using a machine learning approach.

Keywords: *Crime Prediction, KNN, Decision Tree. Multilinear Regression; K-Neighbors Classifier, Artificial Neural Networks.*

INTRODUCTION

A crime is nothing, but it's an action. It constitutes an offence. It's punishable by law. The identification and analysis of hidden crime is a very difficult task for the police department. Also, there is voluminous data of the crime available. So, there should be some methodologies that should help in the investigation. So, the methodology should help to solve the crime.

The machine learning approach can better help in the prediction and analysis of the crime. The machine learning approach provides regression algorithms. The classification techniques provide help to fulfil the purpose of the investigation. Regression techniques such as multilinear regression are statistical methods. This method helps to find the relationship between two quantitative values or variables. This approach predicts the values of the dependent variables based on the independent variables. The classifier techniques such as K-Neighbor's classifier. These classifiers are used to classify the multiclass target variables. The neural networks are used to improve accuracy. The neural network has an input layer dense and has an output layer. Based on the above algorithms, the perpetrator description such as sex, age, and relationship is predicted. The model is thus expected to help to remove the burden of the police investigation. Thus, it helps to solve homicide cases. This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

LITERATURE REVIEW

Ling Chen, Xu Lai (2011) [1] has compared the experimental result that is obtained by the ANN (Artificial Neural Network). Jyoti Agarwal, Renuka Nagpal, et al., (2013) [2] has studied the crime analysis using K-means clustering on the crime dataset. They have developed this model using the rapid miner tool. The clustered results are obtained and analyzed by plotting the values over the years. This model gives the result of the analysis that the number of homicides decreased from 1990 to 2011.

Shiju Sathyadevan, Devan M. S, et al., (2014) [3] have predicted the regions where there is a high probability of the crime occurring. They have visualized crime-prone areas also. They have

classified the data using Naive Bayes classifiers. This algorithm is a supervised learning algorithm that also gives the statistical method for classification. This classification gives an accuracy of the 90%.

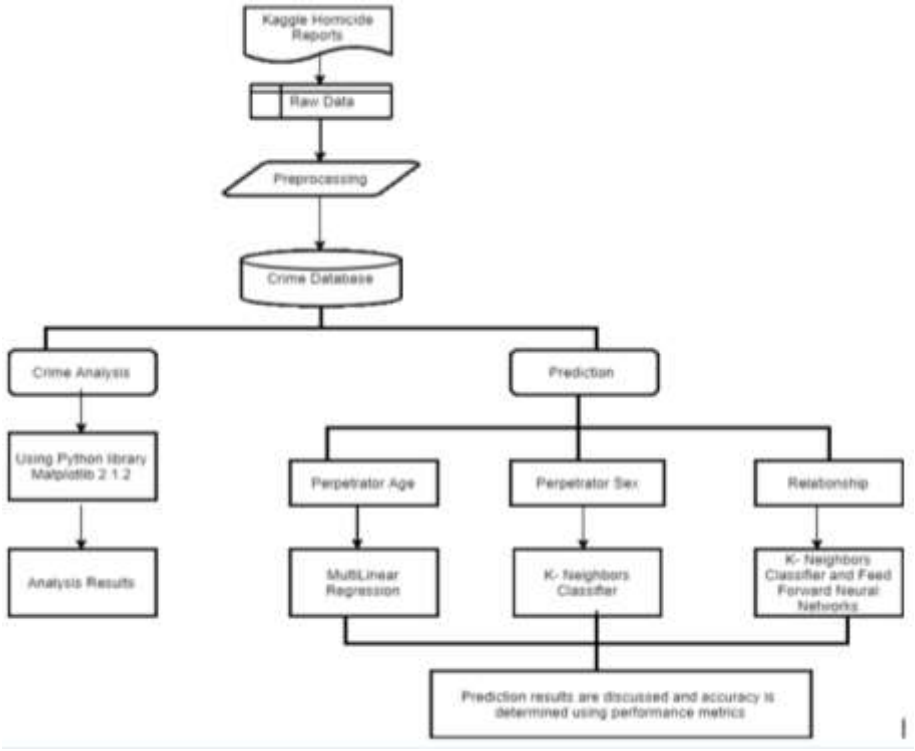
Lawrence McClendon and Natarajan Meghanathan (2015) [4] have used Linear Regression, Additive Regression, and Decision Stump algorithms using the same set of input (features) on the Communities and Crime Dataset. Overall, the linear regression algorithm gave the best results compared to the three selected algorithms.

Chirag Kansara, Rakhi Gupta, et al., (2016) [8] proposed a model which analyzes the sentiments of the people on Twitter and predicts whether they can become a threat to a particular person or society. This model is implemented using the Naive Bayes Classifier, which classifies the people by sentiment analysis.

LIMITATIONS OF THE EXISTING SYSTEM

The existing system gives an accuracy of only 65 %. The model is used only using linear regression. The multiple approaches of machine learning are not implemented. Also, the model has used the dataset of limited crimes.

PROPOSED SYSTEM MODEL



IMPLEMENTATION AND ANALYSIS

The dataset we have used contains almost 63000 values. The dataset is taken from the Kaggle website, where the dataset is freely available. It has entries from 1980 to 2014.

The analysis includes the number of unsolved crimes the weapons used in the crimes. The month when the maximum crime took place. The places and occurrence of the crime. The state where the crime rate is high.

METHODOLOGY

The dataset is obtained from the Kaggle repository. This is the domain for the various research-oriented dataset. The dataset contains homicide entries collected from the FBI's Supplementary Homicide Report.

The dataset consists of 638454 rows and 17 columns and column metadata. From the dataset, the significant features like State, Year, Month, Crime Type, Crime Solved, Victim Gender, Victim Age, Victim Race, Victim Count and Weapon are chosen as the input features for the system. The features Perpetrator Age, Perpetrator Sex and Relationship of the perpetrator with the victim are chosen as the target variable to be predicted by the system.

We have used two algorithms for the prediction one is multilinear regression, and the other is K-neighbors classifier.

MultiLinear Regression

This algorithm gives the mathematical approach to finding the relationship between the dependent variable with the given set of independent variables. In our research, the perpetrator's age is a dependent variable, and the independent variables are pieces of evidence collected from the crime scene. This algorithm predicts the perpetrator's age based on input features such as state, year, month, place, crime solved, etc.

The equation for the Multilinear Regression line is given as:

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp$$

Where,

Y is the dependent variable,

x is the independent variable,

βi are coefficients of the regression equations.

K-Neighbors Classification

This classification algorithm is used when the target variable has more than two classes to classify [11]. In our dataset, the target variable is nothing but its

perpetrator sex, and it has been classified namely as male, female and unknown. Also, the target variable relationship has 27 unique values such as friend, wife, nephew, etc., so the K-Neighbors classifier is used to classify these target variables. The target variables are perpetrator sex and relationship.

Pseudo Code

K_Nearest_Classifier (input variables);
Assign K -> the number of clusters
A set of K instances are chosen to be centred for the clusters.

For each data point in the input:

Calculate the Euclidian distance
Assign the cluster which is near the data point. Recalculate the centroids and reassign the variables in the clusters.

IMPLEMENTATION DETAILS

The implementation details include the machine learning approach.

Data-collection: The data collection for the implementation is from the Kaggle. The dataset is freely available. The record collected is almost 63000.

Preprocessing: Once the dataset is collected, it must be preprocessed to get the clean dataset. The pandas and NumPy libraries are available in python for the preprocessing. It is removing empty values from the dataset, or repeated records should be removed.

Analysis: The analysis includes the graphical representation of different values to analyze the dataset property. The different graphs are plotted by Mathplotlib libraries. The graphical analysis gives a direction towards the prediction.

Training and Testing

The dataset is divided into training and testing. Generally, 70 % dataset is kept for training and 30% for testing. The dataset ratio can be 70: 30 or 80:20.

Validation

Once the model is created, it should be validated with the real-time data values. This is called validation. The validation is nothing but its predicted value, and it's also called the output value.

RESULTS AND COMPARATIVE STUDY

Our model gives an accuracy of 85 %. The previous model gives an accuracy of 65%. The below graph gives the comparison of the model with the previous results.

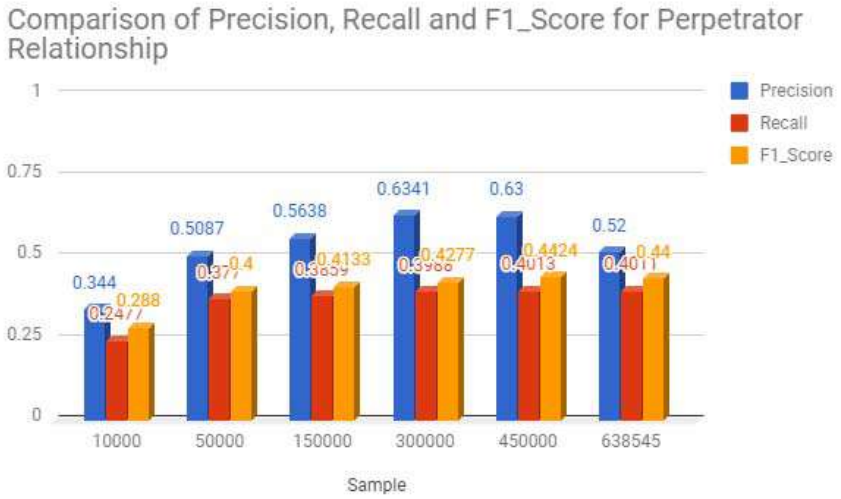


Fig. 1 Comparison of Precision, Recall and F1_Score for Perpetrator Relationship

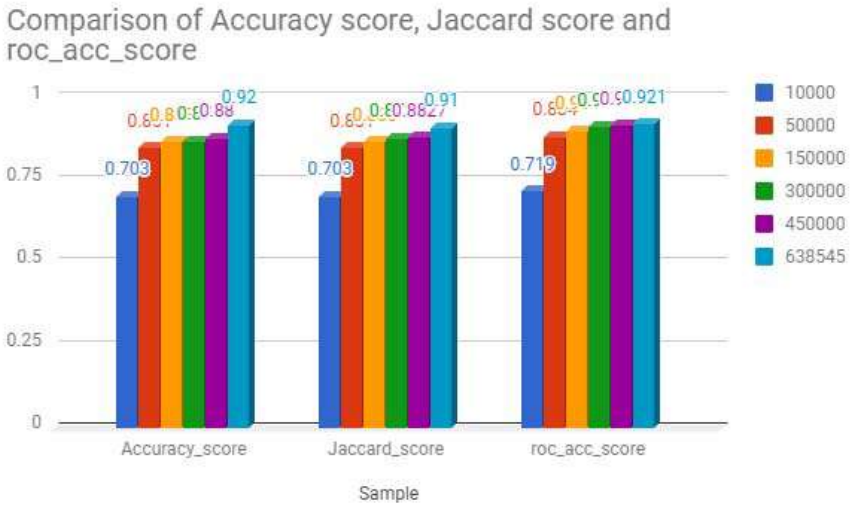


Fig. 2 Comparison of Accuracy score, Jaccard score and roc_acc_score

The ratio estimated through calculation of recall is found to outscore those of precision and F1 score. However, in the case of a set of 10,000 samples, the values of precision and F1 score are observed to be greater than the recall score. This can be inferred as an indication of a larger number of false negatives present in the sample set as opposed to the number of false positives predicted by the model.

CONCLUSION

This model helps to predict crime. The perpetrator's age, perpetrator sex, and relationship can be predicted using a machine learning approach. The regression and classifier are used here give almost 80 % accuracy.

The dataset can be enhanced and can be used in other countries if the scenario is

almost the same. The model gives the overall prediction of any crime. This model can be enhanced by using deep learning techniques.

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered. Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template.

REFERENCES

1. Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." Power and Energy Engineering Conference (APPEEC), 2011 Asia-Pacific. IEEE, 2011.

2. Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using K-means clustering." International Journal of Computer Applications 83.4 (2013).

3. Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014

4. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." Machine Learning and Applications: An International Journal (MLAIJ)2.1 (2015).

5. Kiani, Rasoul, SiamakMahdavi, and Amin Keshavarzi. "Analysis and prediction of crimes by clustering and classification." Analysis 4.8 (2015).

6. Heartfield, Ryan, George Loukas, and Diane Gan. "You are probably not the weakest link: Towards

practical prediction of susceptibility to semantic social engineering attacks." IEEE Access 4 (2016): 6910-6928.

7. Sivaranjani, S., S. Sivakumari, and M. Aasha. "Crime prediction and forecasting in Tamil Nadu using clustering approaches." Emerging Technological Trends (ICETT), International Conference on. IEEE, 2016.

8. Kansara, Chirag, et al. "Crime mitigation at Twitter using Big Data analytics and risk modelling." Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on. IEEE, 2016.

9. Tsunoda, Masateru, Sousuke Amasaki, and Akito Monden. "Handling categorical variables in effort estimation." Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement. ACM, 2012.

10. Su, Ya, et al. "Multivariate multilinear regression." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42.6 (2012): 1560-1573.

11. Viswanath, P., and T. Hitendra Sarma. "An improvement to K nearest neighbour classifier." Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE. IEEE, 2011.

12. Palocsay, Susan W., Ping Wang, and Robert G. Brookshire. "Predicting criminal recidivism using neural networks." Socio-Economic Planning Sciences 34.4 (2000): 271-284