USING A MACHINE LEARNING APPROACH, PREDICT CRIMES

Nitin Kumar, Prof. Sapna Jain Choudhary

Abstract: One of the major problems in our society is crime. It dominates every area of our culture. Additionally, it rules society. So, one of the most crucial tasks is crime prevention. An organised approach should be taken to the criminal analysis. It is significant in identifying and preventing crime, according to the report. The analysis identifies patterns in the investigation process and aids in the identification of criminal trend. The analysis of the effectiveness of the criminal investigation is the primary focus of this research. The model is developed to identify criminal trends based on conclusions. The conclusions drawn from the crime scene are used in the paper to show how the offender was predicted. The report outlines a research strategy for predicting the age and gender of the perpetrator. Two key components of crime prediction are presented in this essay. The first is the perpetrator's gender, and the second is their age. Analysis of numerous factors, including the year, month, and weapon used in the unsolved crimes, is one of the criteria considered. The quantity of unresolved crimes is determined by the analysis. The description of the perpetrator's age, sex, and relationship with the victim is part of the prediction task. The Kaggle website provided the dataset for this paper. Multi-linear regression, K-Neighbor's classifier, and neural networks are all used by the system to predict the output. It underwent testing and training using a machine learning strategy.

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

I. INTRODUCTION

Nothing but an action constitutes a crime. An offence has been committed. It is a criminal offence. For the police department, locating and analysing hidden crime is a highly challenging task. Additionally, there is a wealth of information about the crime. Therefore, some approaches ought to be able to aid in the inquiry. Therefore, the approach should aid in the crime's resolution.

The prediction and analysis of the crime can be improved with the use of machine learning. Regression methods are provided by the machine learning method. The investigation's goal is served by the classification procedures. A statistical method is regression techniques, such as multilinear regression. This technique aids in determining the connection between two numerical numbers or variables. Based on the independent variables, this method forecasts the values of the dependent variables. approaches for classifiers like K-Neighbor's classifier. The multiclass target variables are classified using these classifiers. The accuracy is increased by using neural networks. The neural network has a dense input layer and a layer for its output. The perpetrator description, including sex, age, and relationship, is predicted using the aforementioned algorithms. Thus, it is anticipated that the model will assist in easing the strain of the police investigation. So, it aids in the resolution of homicide cases.

This document serves as a model. The conference website offers a download for an electronic version. Please get in touch with the conference publications committee as listed on the conference website if you have any issues about the paper guidelines. On the conference website, you may find information about submitting your final work.

II. LITERATURE REVIEW

Ling Chen and Xu Lai (2011) [1] have contrasted the experimental outcomes produced by artificial neural networks (ANNs). The crime dataset was used in a study by Jyoti Agarwal, Renuka Nagpal, et al. (2013) [2] to examine the crime analysis using K-means clustering. They used the fast miner tool to create this model. By graphing the values across time, the clustered results are discovered and examined. According to this model's data, homicide rates dropped between 1990 and 2011.

Researchers Shiju Sathyadevan, Devan M. S., et al. (2014) [3] identified the areas where there is a high likelihood that a crime will occur. They also depicted places that are prone to crime. They used Naive Bayes to classify the data.

classifiers. This algorithm provides the statistical method for classification and is a supervised learning algorithm. The accuracy of this classification is 90%.

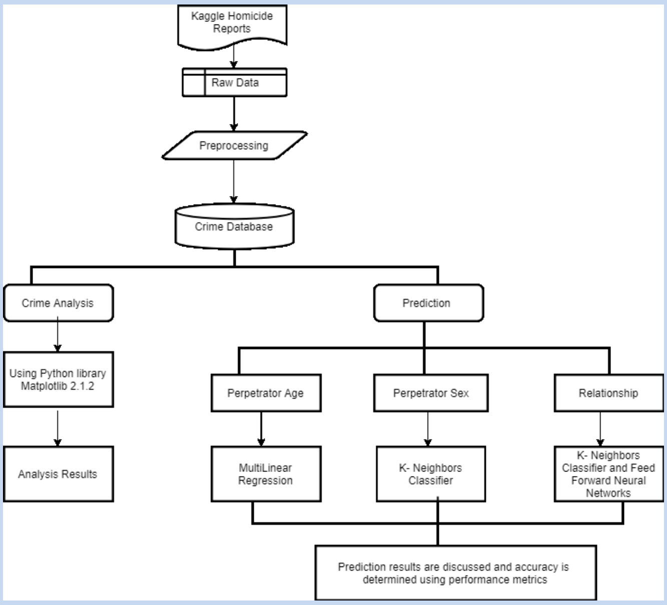
On the Communities and Crime Dataset, Lawrence McClendon and Natarajan Meghanathan (2015) [4] employed the Linear Regression, Additive Regression, and Decision Stump algorithms with the identical set of input (features). Comparing the three chosen algorithms, the linear regression algorithm produced the best results overall.

A methodology was put up by Chirag Kansara, Rakhi Gupta, et al. in 2016 [8] and examines Twitter users' sentiments to determine whether they pose a threat to a specific person or society. The Naive Bayes Classifier, which uses sentiment analysis to categorise the individuals, is used to implement this model.

III. 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 the limited crimes.

IV. Proposed system model



V. 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.

VI. 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 the 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.

a) MultiLinear Regression

This algorithm gives the mathematical approach to find 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, and 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.

b) 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 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 centres for the
clusters
For each data point in the input:
Calculate the Euclidian distance
Assign the cluster which is near to the data point
Recalculate the centroids and reassign the
variables in the clusters.
```

VII. 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.

Pre-processing:
Once the dataset is collected, it must be pre-processed to get the clean dataset. The pandas and NumPy libraries are available in python for the pre-processing. it is removing of empty values from the dataset or repeated records should be removed.

Analysis:
The analysis includes the graphical representation of different values to analyse the dataset property. The different graphs are plotted by Matplotlib 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.

VIII. 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.

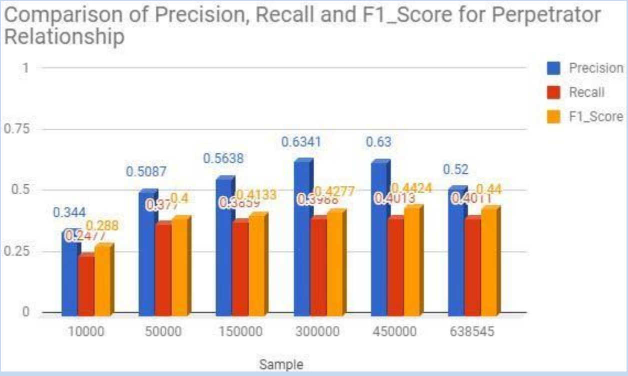


Fig. 1Comparison of Precision, Recall and F1_Score for Perpetrator Relationship.

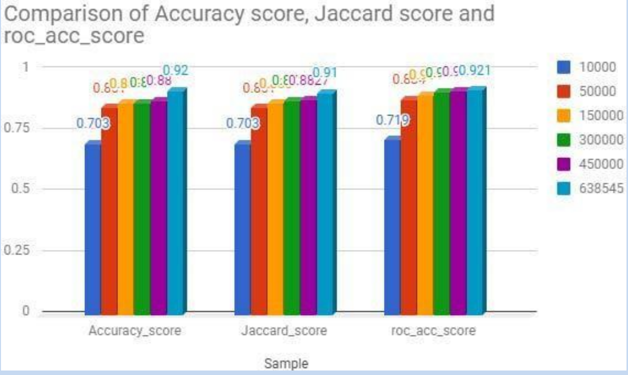


Fig. 2Comparison of Accuracy score, Jaccard score and roc_acc_score.

The ratio estimated through calculation of recall in 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.

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

X. FUTURE WORK

This model gives an accuracy of almost 80 % for the perpetrator age, 82 % for the perpetrator sex, and 85 % for the relationship. The accuracy can be improved by using a complex neural network such as the recurrent neural network. Also, the deep learning approach can be used to enhance the accuracy of the model.

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 analyse crime data." Machine Learning and Applications: An International Journal (MLAIJ)
2.1 (2015).
[5]. Kiani, Rasoul, Siamak Mahdavi, 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 TamilNadu 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