Grenze International Journal of Engineering and Technology, Jan Issue

Comparative Study of Machine Learning Models for Crime Analysis

Yashvi Mehta[1], Sakshi Mahadik[2], Manan Shah[3], Jaiwin Shah[4] and Dr. Sudhir Dhage[5]
[1-5]Computer Engineering Sardar Patel Institute of Technology, Mumbai, India
Email: yashvi.mehta@spit.ac.in, sakshi.mahadik@spit.ac.in, manan.shah@spit.ac.in, jaiwin.shah@spit.ac.in, sudhirdhage@spit.ac.in

***Abstract*—In today's world where criminal activity is increas- ing daily, it's necessary to curb these crimes. Analysis of crime is a procedural approach to identifying crimes. Crimes can reduce if the criminal hotspots get identified. If these hotspots are pinned down based on the crime type, crime detection and analysis can. With the machine learning techniques applicable to big datasets, crime investigators could use these approaches to narrow down their search and solve the cases. We intend to preprocess thedata and find these criminal hotspots to improve our search. In this paper, we have implemented several machine learning algorithms such as decision tree regression, linear regression, random forest, and other algorithms to analyze crime patternsin the United States. A comparative study is done to see which algorithm would yield better results in terms of accuracy.**

***Index Terms*— Crime Analysis,Decision Tree Regression, K- Nearest Neighbors , Linear Regression, Support Vector Machine,Random Forest.**

I. Introduction

Criminal activities are gradually rising around us. To reducethem, high crime rate areas need to be determined. The first step is to segregate different categories of crimes committedin that neighborhood. The main purpose of crime analysis is to reduce these organized criminal activities. Crime prediction can be employed by identifying these hotspots and reporting them to the concerned authority, thereby helping to decrease the crime rate. It can be done by taking the help of various machine learning models. It would utilize the existing crime data and predict the crime type and its occurrence. The major goal behind this research is to identify crimes that can get predicted once the required data gets filtered out. This filtered data would lead to the finding of patterns that would come in handy to identify the criminal activities in a state.

The dataset taken consists of a variety of attributes. After gathering the data, the data is cleaned and divided into train and test sets. Supervised techniques like K-nearest neigh- bour, Linear Regression, Decision Tree, Random Forest, and Support Vector Machine are used for crime prediction. The results from this prediction can be very useful for the police department in investigating cases and reducing the crime rate by taking appropriate measures which will secure the people and the region.

II. Literature Survey

Akash Kumar, Aniket Verma, Gandhali Shinde, Yash Sukhdeve, and Nidhi Lal [1] proposed a crime prediction

Grenze ID: 01.GIJET.9.1.31

workbased on the supervised learning algorithm, K-nearest neigh- bors. The model consisted of observing crime rates by the KNN prediction method that predicts the type, place, and timeof a particular crime. The crime data taken for analysis consistsof the type of crime, the place, and the time of the crime that happened in the past. Based on the assumption that similar cases exist in close proximity, the KNN algorithm classifies new cases based on similarity standards. The minimum erroris obtained when the range of value of k is between 1 to 15. The error was reduced thus giving a good accuracy of 99%for crime prediction.

B. Sivanagaleela, and S. Rajesh [2], proposed a new systemof crime analysis over an existing system that used the Naïve Bayes classification algorithm. The new system uses the fuzzyC-Means algorithm to form clusters of the crime data for total perceptible crimes such as robbery, kidnapping, and other such crimes. The old data is used to identify crime locations. Then, data mining techniques are used to greatly enhance the speed, performance, and accuracy, of predicting the crime. Based on current and old data, data mining techniques analyze crime data. According to data items' natural relationships or logical groupings, crime data is clustered and a whole structure is generated. Real-life crime data utilizes Fuzzy clustering. The crime data clusters are formed and classified or grouped based on the dataset, using the Fuzzy clustering technique. It is an efficient technique to analyze crime data and classify various crime patterns.

Sunil Yadav, Rohit Vishwakarma, Nikhilesh Yadav, Ajit Yadav, and Meet Timbadia [3], proposed a crime pattern detection system that detects, analyzes, and predicts the crime.The proposed model is trained by adding past years' criminal records and it uses semi-supervised, unsupervised, and super- vised learning techniques on the criminal pieces of evidence for learning information and to assist in enhancing the ac- curacy of the prediction of the offense. K-means algorithm is applied to form clusters based on the sum of the total of people concerned with a crime. The association mining technique, theApriori algorithm, then, finds out the association between the number of attributes, for which the input is obtained from the k-means result. Naive Bayes classifier is used to understand and analyze the existing dataset and also predict the behavior of a particular dataset based on the classification principles. Finally, correlation and linear regression are applied to find and predict a significant relationship between the sum of people exonerated from the crime against the trials held that year.

Kanimozhi N, Keerthana N V, Pavithra G S, Ranjitha G, and Yuvarani S [4], proposed a system that analyzes crime patterns and estimates the most frequent crime type and its contribution along a certain period and its location. Naive Bayes, a machine learning algorithm is implemented for the classification of different crime patterns and enhances the accuracy of the model. They have implemented some extendedbayesian techniques such as Bernouille Na¨ıve Bayes, Gaussian Na¨ıve Bayes, and Multi-nominal Na¨ıve Bayes to predict the crime type. The proposed system was successful in achieving an accuracy of 93.07% and a precision of 92.53%, which is high compared to the previous system.

III. Motivation

The crime rate is gradually increasing not only in the UnitedStates but all over the world which has a severe negative social impact. Machine learning algorithms can assist to detect various patterns between datasets and reveal new pieces of information as well. Crime prediction and solving cases canbe complicated and it's one of the top problems for the police. Our motivation was to make this easier and more efficient using machine learning algorithms for crime prediction.

IV. Problem Domain

The domain of our paper is Machine Learning. We have made use of 5 models - Decision Tree Regression, K-nearest neighbors, Linear Regression, Random Forest Classification and Support Vector Machine Algorithm to predict Property Crimes and Violent Crimes.

V. Problem Definition

The number of crimes per day continues to accelerate continuously and the crime patterns change indefinitely. This constant change makes it difficult for the police to explain and understand the behavior of these patterns. If the patterns are carefully analyzed, it can be of great help to the police to prevent crime and take measures to increase the security of a region.

VI. Statement

Our problem statement is analyzing different types of crimestaking place in localities and predicting the crimes that could take place in the future.

VII. Innovative Content

This project's premise is that crimes are largely predictable;all that is needed is the ability to sort through a vast amount of data to identify trends that are helpful to law enforcement. A few decades ago, this type of data analysis was technologicallyimpractical, but current advancements in machine learningshould make it possible.

VIII. Methodology

A. Dataset

We selected a dataset with a variety of features in order to employ our models successfully and obtain thorough findings. The dataset includes data on crimes committed in the United States, sorted by state. It includes the year of the crime andthe city's population. It has both violent crimes and property crimes, such as car thefts, larcenies, and burglaries (assaults, murders, rapes, and robberies). The CORGIS Dataset is wherethe information came from. It comprises 21 characteristics andabout 3110 rows, of the 50 states of the United States.

	State	Year	Data.Population	Data.Rates.Property.All	Data.Rates.Property.Burglary
0	Alabama	1960	3266740	1035.4	355.9
1	Alabama	1961	3302000	985.5	339.3
2	Alabama	1962	3358000	1067.0	349.1
3	Alabama	1963	3347000	1150.9	376.9
4	Alabama	1964	3407000	1358.7	466.6

Fig. 1. A glimpse of the dataset used

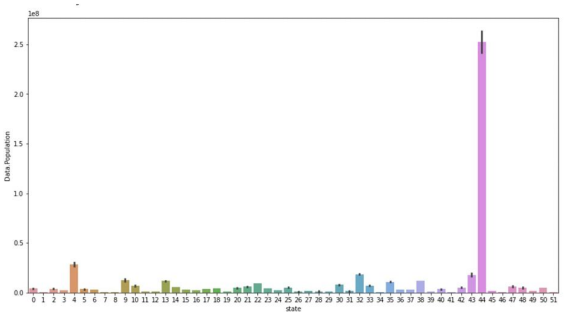

Fig. 2. State-wise population in the United States

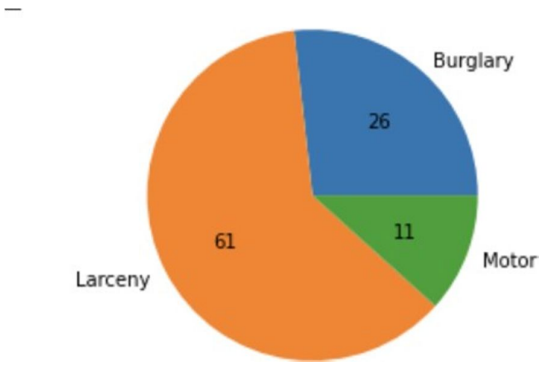

Fig. 3. Property crimes

It is evident from the chart that in the case of Property Crimes, Larceny has the highest percentage followed by Bur- glary and Motor crimes.

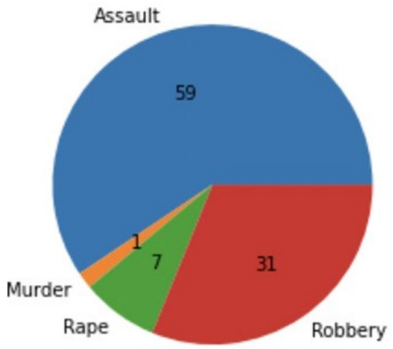

Fig. 4. Violent crimes

From the pie chart, it can be said that Assault is the most occurring violent crime in the United States. Robbery is almosthalf of Assault, followed by Rape and Murder.

B. Data Cleaning

Once the data is collected, it is pre-processed which isessential to remove the unknown values, null figures, and inconsistent data. This is needed to improve the efficiency of the model. The data after pre-processing is split equally into training and test sets.

C. Models Used

We have made use of 5 models - Linear Regression, Deci- sion Tree Regression, Random Forest, K-Nearest Neighbors, and Support Vector Machine.

1) Linear Regression: The goal of linear regression is to model a variable Y as a function of a variable(s) X. When X variable is all that is known, this is helpful in predicting Y.The arithmetic equation is :

$$Y = C1 + C2X \quad (1)$$

Where C1 is the intercept and C2 is the slope. Collectively, they are called the regression coefficients.

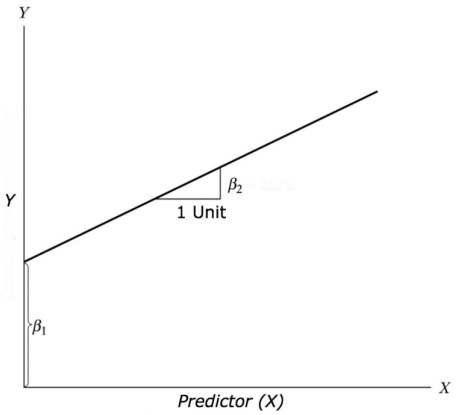

Fig. 5. Sample regression diagram

The term ”dependent variable” refers to the variable you wish to forecast. The independent variable is the one you’re using to make a prediction about the value of the other variable. In our case, independent variables are the rates of Burglary, Larceny, Motor, Assault, Robbery, Rape, andMurder. The dependent variable includes totals of Property Crimes and Violent Crimes.

By applying the Logistic Regression model, the training andthe test scores both are 99.99. Higher the test score, the more generalized the model is.

2) Decision Tree Regression: The Decision Tree model is used for classification and prediction. It can be used for large datasets and be categorized by the type of crime. In decision tree regression, the model first notes the features, and then it trains it in the form of a tree so that it can predict meaningful data in the future. We have applied the DecisionTreeRegressor to our training set and predicted the values of Property Crimes and Violent Crimes.

After running the model, it is found that the train score is 100, while the test score is 99.79.

3) Random Forest: This technique for supervised learning employs the ensemble learning approach for regression. With this method, predictions are combined from many machine learning algorithms and are therefore more accurate than predictions from a single model. As a result of all the trees in a Random Forest running concurrently without interacting with one another, the prediction of the Random Forest is given by the mean of all the classes. Because there are more trees in the forest, accuracy is higher and overfitting is drastically reduced.

Here, k random data points are selected and decision trees associated with them are built. N number of such trees are built. Every single tree of the N trees helps to forecast the value of y for each new data point, and this point is categorized under the mean of every single projected y values.

When a model using Random Forest was constructed with our training set to predict Property and Violent crimes, the training score was found to be 97.77 while the test score was 97.33.

4) K Nearest Neighbors: A straightforward approach, KNN uses a similarity metric and all of the pertinent samples to store and predict the numerical target. (e.g., distance functions). In other words, the new point's value is determined by how nearly it resembles the points in the training set. By averaging the data in the same area, this method naturally approximates the relationship between the X variables and the continuous outcome. The first step in this to determine the distance between each point that is trained and the new point. There are two ways to do this:

$$Euclidean: \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (2)$$

$$Manhattan: \sum_{i=1}^{k}|x_i - y_i| \quad (3)$$

Once the distance between the new point and the training point is measured, the next step is to select the points that are near to the new point. When we assign a value to any new observation, the parameter k determines how many neighbors we consider.

On applying KNN on the training set, train score was found to be 99.69, and the test score was 99.61.

Support Vector Machine: To perform classification, a straightforward linear SVM classifier constructs a straight line that separates two classes. Hence, the set of data points on each of the opposing sides of the line would fall under the same division. This denotes that the number of possible lines is unlimited. The SVM model seeks to identify the optimal decision boundary for categorizing the n-dimensional space so that future classification of additional data points will be simple. Because it may be found with more than two characteristics, not only two, this optimum choice boundary is known as a hyperplane. SVMs are popular because they may detect complicated correlations in your data without requiring you to do several modifications. It's excellent for working with smaller datasets that include a huge range of features.

When this model was applied, the train score was 100, and the test score was 98.41.

IX. RESULTS

TABLE I. R-SQUARED SCORE OF THE MODELS

Sr. No.	Models	SquaredScore
1	Linear Regression	99.968
2	ecision Tree Regression	99.55
3	Random Forest	97.03
4	KNN	98.18
5	SVM	96.32

A number of models were considered as specified in the literature review the decision tree classifier model gave the best results because of its performance in adapting to the data collected. Both the regression as well as the classification problems can be solved with the help of the decision tree model, and it is easy to interpret, understand, and visualize. The outliers or the misplaced values do not largely affect the Decision trees, and they can handle both quantitative and categorical variables.

X. Conclusion

This research paper proposed a way of analyzing criminal activities and frauds that can happen within a state. The main objective was to create a crime pattern analyzing tool that can be highly useful for law enforcement. This paper compares different machine learning algorithms in terms of analyzing crime patterns and performance. The decision tree regression algorithm was successful in analyzing the crime with better performance.

Future Scope

Although successful models have been made to analyze and predict crime, our aim is to extend our work to more crimes that are generally overlooked such as domestic violence. Our current model does not take into account the state of global economics and crises, for example, with the outbreak of Covid-19, a whole new category of crime has emerged such as the supply of fake vaccination doses and these categories need to be taken into account in real-time. Therefore, in the next iteration, the above-mentioned factors must be included to build an accurate predictor.

Acknowledgement

This work was fully supported by Sardar Patel Institute of Technology. We would like to thank Dr. Sudhir Dhage for his constant support and guidance and for pushing us to write this research paper.

References

[1] Kumar, A. Verma, G. Shinde, Y. Sukhdeve and N. Lal, "Crime Predic- tion Using K-Nearest Neighboring Algorithm," 2020 International Con- ference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.155.

[2] Sivanagaleela and S. Rajesh, "Crime Analysis and Prediction Using Fuzzy C-Means Algorithm," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 595-599, doi: 10.1109/ICOEI.2019.8862691.

[3] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma and N. Yadav, "Crime pattern detection, analysis & prediction," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 225-230, doi: 10.1109/ICECA.2017.8203676.

[4] N. Kanimozhi, N. V. Keerthana, G. S. Pavithra, G. Ranjitha and S. Yuvarani, "CRIME Type and Occurrence Prediction Using Ma- chine Learning Algorithm," 2021 International Conference on Artifi- cial Intelligence and Smart Systems (ICAIS), 2021, pp. 266-273, doi: 10.1109/ICAIS50930.2021.9395953.

[5] S. G. Krishnendu, P. P. Lakshmi and L. Nitha, "Crime Analysis and Pre- diction using Optimized K-Means Algorithm," 2020 Fourth International Conference on Computing Methodologies and Communication (IC- CMC), 2020, pp. 915-918, doi: 10.1109/ICCMC48092.2020.ICCMC- 000169.

[6] Mahmud, Sakib & Nuha, Musfika & Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining. 10.1007/978-981- 15-7394-1 5.

[7] VarshithaD, N., VidyashreeK., P., Aishwarya, P., JanyaT, S., Gupta, K.R., & Sahana, R. (2018). Paper on Different Approaches for Crime Prediction system. International journal of engineering research and technology, 5.

[8] O. Llaha, "Crime Analysis and Prediction using Machine Learn- ing," 2020 43rd International Convention on Information, Communi- cation and Electronic Technology (MIPRO), 2020, pp. 496-501, doi: 10.23919/MIPRO48935.2020.9245120.

[9] David. H, Benjamin & Suruliandi, A.. (2017). SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. ICTACT Journal on Soft Computing. 7. 1459-1466. 10.21917/ijsc.2017.0202.

[10] Shah, N., Bhagat, N. & Shah, M. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Vis. Comput. Ind. Biomed. Art 4, 9 (2021). https://doi.org/10.1186/s42492- 021-00075-z