Crime Type Prediction Using Machine Learning

Rohith S [1], Subha Indu S [2]

[1]Student, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India.

[2]Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India.

Corresponding Author: sksrohith07@gmail.com

Abstract: - Crime is now a critical problem in today's culture, resulting in major disruptions and instabilities. It is crucial to comprehend crime patterns as a way to take preventive measures considering the rise in unlawful behavior could possibly result in gaps in society. Using free to download data on crimes from Kaggle, this research forecasts the frequency of crimes. The primary objective of this research is to determine the most common violations of law, and to figure out the times and places they typically occur. Various patterns of crime are classified using algorithms based on machine learning, such as the Random Forest classifier, with significantly greater precision rates compared to previous research.

Key Words: - *Crime, Prediction, Kaggle, Random Forest.*

I. Introduction

The issue of crimes, which affects society, will continue to get worse. Anything that is both highly offensive and against the law is regarded as a crime. To understand criminal trends, one must be knowledgeable in criminal psychology and possess pattern identification capabilities. The government has to devote a lot of time and energy in technology deployment to combat illegal behavior. Machine learning methods and data analysis are essential for predicting the types and patterns of crime. Studies in the past have looked at crime patterns and how they connect to certain locations. By identifying hotspots, authorities can now categorize criminal tendencies more quickly and take action more quickly. In this research, we made use of a dataset from the open-source Kaggle platform that considers the time and location of incidents over a specific time range.

Manuscript revised May 20, 2023; accepted May 21, 2023. Date of publication May 23, 2023.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

To determine the type of crime and hotspots where criminal activity happens at a certain time and day, we used a categorization system. Based on statistical and spatial information, this approach utilizes machine learning techniques for identifying identical crime behavioral patterns.

II. Proposed system

The data undergoes pre-processing using machine learning techniques such as filtering and wrapper to remove irrelevant and duplicate data values, thereby reducing dimensionality and cleaning the data. The data is then split into a test and trained dataset, followed by mapping of crime type, year, month, time, date, and location to an integer for ease of classification. Crime features are labeled, enabling the analysis of crime occurrence at specific times and locations, ultimately revealing the most frequently occurring crimes with spatial and temporal information. The prediction model's performance is evaluated by calculating the accuracy rate.
Python programming language is used to design the prediction model.

2.1 Advantage

- The proposed algorithm is highly effective in detecting crime patterns as it takes into account various featured attributes that depend on the time and location of the crime.

- This algorithm is able to overcome the issue of analyzing the independent effects of attributes.
- The proposed algorithm does not require an optimal value initialization and can handle both real-valued and nominal attributes, making it more flexible and adaptable to various datasets.
- Compared to other machine learning prediction models, the proposed algorithm has shown significantly higher accuracy rates, indicating its effectiveness in crime prediction.

III. Methodology

Random Forest works by creating a set of decision trees, each trained on a randomly sampled subset of the training data, and randomly selecting a subset of features at each node of the tree. The trees are then combined to produce a final prediction. The random sampling and feature selection help to prevent overfitting, making Random Forests more robust and accurate than single decision trees. In random forest classification, voting refers to the process of combining the predictions of multiple decision trees to arrive at a final prediction. Each decision tree in the random forest independently makes a prediction based on the features of the input data, and these predictions are then aggregated to make a final prediction.

There are two types of voting in random forest classification: hard voting and soft voting. Hard voting is used for classification problems where the class label with the highest number of votes is chosen as the final prediction. Soft voting is used when the outputs of the decision trees are probability estimates, and the final prediction is based on the average probability of each class across all the trees.

Random forest classification uses voting to reduce overfitting and increase the accuracy of the model. By aggregating the predictions of multiple decision trees, the model can reduce the variance of the predictions and provide a more stable and reliable prediction. Additionally, the use of multiple decision trees in a random forest reduces the risk of bias, as the final prediction is based on the consensus of many individual predictions.

3.1 Data pre-processing

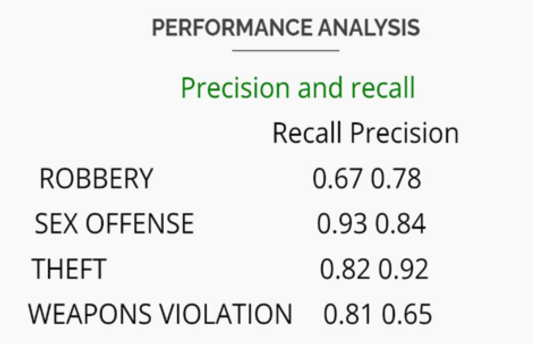
Pre-processing of data is crucial for accurate prediction of crime activities. The pre-processing involves the use of filter and wrapper techniques to identify missing values in specific attribute values and remove irrelevant context from the dataset. This helps in measuring the significance of the features and their correlation with dependent values. The data is then split into test and trained attributes to train a prediction model.

3.2 Mapping

The suggested technique includes separating crime attributes, including crime type, date, and time of occurrence, and transforming them into integers for easy classification. After labelling the data, it is analyzed and visualized using the matplotlib package in Python, which is suitable for machine learning applications. By plotting graphs, criminal activities' frequency can be visualized, and the most common crime can be identified and used for future prediction processes.

3.3 Evaluation

To improve accuracy compared to pre-existing models, the performance of the implied prediction is evaluated. Based on the analysis of crime type and occurrence, this model achieved an accuracy rate of **82.07%.** This indicates that the model is effective in identifying and predicting the occurrence of different types of crimes. This paper utilized **Random forests classifier** algorithm, to predict the occurrence of crimes. The dataset used for this study contained information on different types of crimes and their occurrence in Chicago.



PERFORMANCE ANALYSIS

Precision and recall

	Recall	Precision
ROBBERY	0.67	0.78
SEX OFFENSE	0.93	0.84
THEFT	0.82	0.92
WEAPONS VIOLATION	0.81	0.65

Fig.1.Performance analysis for random forest classifier

IV. Results And Discussion

Finding from the analysis of this paper is that crimes tend to occur more frequently on weekends compared to weekdays. This pattern is consistent across different types of crimes, including theft, burglary, and assault. This finding suggests that weekends may be a particularly vulnerable time for criminal activity, and law enforcement agencies may need to allocate more resources to monitor and prevent crimes during these times. Additionally, it may be useful for individuals and businesses to take extra precautions to protect themselves and their property during weekends.

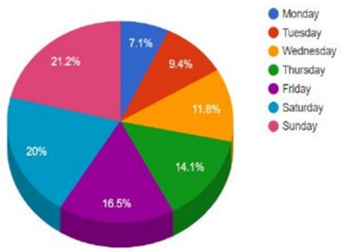


Fig.2. Percentage of crime with particular day of the week

Another finding from the analysis of this paper is that robbery is the most frequent type of crime, while weapon violations are the least frequent. The study analyzed data on different types of crimes reported in various regions, and the results showed that robbery occurred more frequently than any other type of crime. This finding is consistent with previous research that has identified robbery as a major crime concern in many areas. On the other hand, weapon violations were found to be the least frequent type of crime. This suggests that the prevalence of weapon-related crimes may not be as high as other types of crimes, and that efforts to prevent and reduce these types of crimes may be more focused on specific subpopulations or regions.

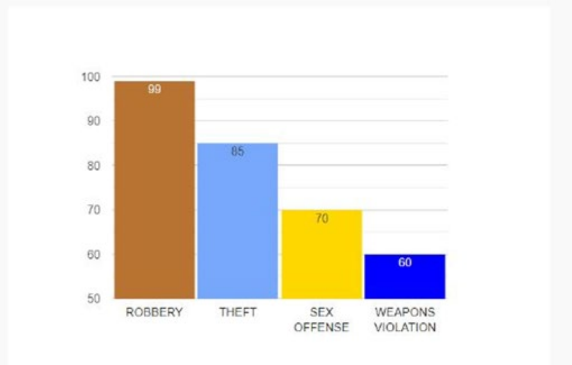


Fig.3. Highest crime type

Overall, these findings suggest that understanding the patterns of crime occurrence can be useful in developing targeted Strategies for crime prevention and law enforcement. Machine learning can be a powerful tool in identifying these patterns and predicting the likelihood of future crimes, ultimately helping to improve public safety and reduce crime rates.

In conclusion, this paper highlights the potential of machine learning in predicting the occurrence of crimes and identifying the factors that contribute to their occurrence. The high accuracy rate achieved by the model indicates that it can be a useful tool for law enforcement agencies in preventing and solving crimes. However, further research is needed to improve the accuracy of the model and make it more robust.

V. Conclusion

It is apparent that important details regarding criminal activities in a neighborhood can serve as indicators for machine learning agents to classify criminal activities based on location and date. While training, imbalanced categories within the dataset can pose challenges, but this paper proposes oversampling and under sampling techniques as a solution. Using Python as the primary language, with inbuilt libraries such as Pandas and NumPy, this study predicts crime data by inputting crime types and outputting the areas where these crimes are likely to occur. Additionally, the Scikit library provides processes on how to use different Python libraries. Prediction results vary depending on the algorithm used, and the Random Forest Classifier was found to have good accuracy, with a rate of 82.07.

References

[1]. Crime Type and Occurrence Prediction Using Machine Learning Algorithm" by K. N. Kanimozhi, N. Keerthana, G. Pavithra, M. Ranjitha, and S. Yuvarani, published in the International Journal of Innovative Technology and Exploring Engineering in 2021.

[2]. "A Comparative Study on Machine Learning Algorithms for Crime Prediction" by S. Prabhakar and R. Venkatesan was also published in 2021 in the International Journal of Recent Technology and Engineering.

[3]. "Crime Prediction in Smart Cities using Deep Learning Techniques" by A. Biswas and R. Sarkar was also published in 2021, in the International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

[4]. "Prediction of Crime Hotspots in a Metropolitan City in India Using Spatial Analysis" by S. Gupta and S. Mukherjee, published in 2015 in the Journal of Indian Society of Remote Sensing.

[5]. "Crime Prediction in Urban India using Regression Analysis" by M. Thakur, A. Sharma, and D. Dahiya, published in 2020 in the International Journal of Research in Engineering, Science and Management.

[6]. "Crime prediction and analysis using machine learning algorithms" by N. A. Shaikh and P. Y. Chavan, published in 2021 in the International Journal of Emerging Technologies and Innovative Research.

[7]. A comprehensive study on crime prediction with open data using machine learning algorithms" by F. Peralta, P. Latorre,

D. Ortiz-Arroyo, and J. A. Gamez, published in 2019 in the Journal of Ambient Intelligence and Humanized Computing.

[8]. "Predicting Crime Hotspots in Los Angeles: A Comparison of Machine Learning Algorithms" by Daniel O'Keefe and Yanqing Ji, published in 2021 in the Journal of Criminal Justice.

[9]. Predictive policing in practice: a review of the available tools" by E. L. Piza, published in 2018 in the Journal of Experimental Criminology.

[10]. "Exploring the potential of social media for crime prediction" by M. Shcherbakov, E. Othman, and J. Ellis, published in 2020 in the Journal of Big Data.