CRIME RATE PREDICTION USING MACHINE LEARNING

Mrs. Bindu KP*1, Mohammed Yaseen*2, Nagendra BN*3, Nitesh A*4, Prajwal R*5

*1Assistant Professor, Department Of CSE, K S School Of Engineering And Management, Bengaluru, Karnataka, India.

*2, 3, 4, 5Department Of CSE, K S School Of Engineering And Management, Bengaluru, Karnataka, India.

ABSTRACT

The context of crime analysis, the investigative process is often time-consuming, particularly in developing countries like India. This research aims to explore techniques that can facilitate the prediction of prevalent crime types in specific regions. Crime is a harmful and social issue that affects individual people all around the world. The rate of crime has risen extremely in recent years and going to be more in upcoming years. The purpose of this study is to provide the police department and people with proper crime forecasting so they can better delegate their resources in response to future crime hot spots. Therefore, this research paper proposes machine learning algorithm to detect the crimes based on the data collected and to show the extent of crime in a particular area. Previous studies are lacking to achieve crime forecasting and prediction accuracy based on learning models. Therefore, this study applied different machine learning algorithms, namely, the k-nearest neighbors (KNN), decision tree, support vector machine (SVM), eXtreme Gradient Boosting (XGBoost) and random forest, the predictable and totally the algorithm which furnish excessive accuracy is be chosen for evaluation the result.

**Keywords:** Crime Rate, Random Forest, Crime Prediction, Machine Learning, Analysis And Forecast.

I. INTRODUCTION

Crime analysis and prediction is a systematic approach for detecting the crime. This system can predict region which have high probability for crime occurrences, crime prone areas. Crimes mainly affect the quality of life, economical growth of nation. With the aim of securing the society from crimes, there is a need for advanced systems and new approaches for improving the crime analytic for protecting their communities. We propose this system which can detect, and predict various crime probability in given regional areas. Crime, as a societal threat and violation, poses a considerable chance to mankind. The spectrum of crimes encompasses a huge variety, consisting of but no longer confined to battery, rape, murder, theft, assault, kidnapping, and drug offenses. Beyond those number one classes, crimes make bigger to embody visitors violations, damage to private and public residences, and different offenses. Crime analysis, an important facet of criminology, entails the systematic examination of styles in criminal activities to determine relationships between the crimes and incidents of the particular regions.

These difficulties stimulated a lot of research work in recent times related to predicting future crimes to help the police department allocate their resources. The machine learning techniques were used to extract information from these enormous databases and discover newly undiscovered connections. This information was then used to report and uncover crime tendencies. This might help crime analysts study these crime networks using various interactive visualizations for crime prediction, as a result, it’s helpful in preventing crimes. S. Kim et al. proposed different machine learning methods to predict crimes in Vancouver from crime data in the last 15 years and obtained an accuracy of 39% and 44% for the K-nearest neighbour and boosted decision tree algorithms, respectively. P. Kumari et al. applied Extra Tree Classifier, K-Neighbour, Support Vector Machines, Decision Tree Classifier, and k-nearest neighbors algorithms to predict the probability of different types of crimes in different locations and time around the city.

According to data released by the Indian National Crime Records Bureau, the information presents an image of the country security situation indicating a rise in crime rates every year. Crimes today come in forms, including offenses like theft and assault as well as emerging challenges such as digital crimes, which necessitate a comprehensive approach. The data serves as a reminder of the importance of enhancing crime prevention strategies strengthening law enforcement capabilities and investing in technologies to effectively combat the growing number of criminal incidents, in India. Addressing this issue not ensures the safety and well being of citizens. Also underscores the urgency of adopting innovative solutions to stay ahead of evolving criminal
methods. The continuous monitoring and analysis of such data are critical for informed decision-making and the development of targeted initiatives to address specific challenges within the Indian criminal landscape.

II. LITERATURE SURVEY

Literature survey is the main advance in programming improvement measure. Before embarking on the construction of the instrument, it is crucial to carefully determine the time constraints. When these things are fulfilled, at that point the subsequent stage is to figure out which working framework and language can be utilized for building up the device. When the developers begin assembling the apparatus the software engineers need parcel of outer help. This help can be gotten from senior developers, from book or from sites. The major part of the project development sector considers and fully survey all the required needs for developing the project. Literature survey is the main area in programming improvement measure. Prior to building up the instruments and the related planning it is important to decide and survey the time factor, asset prerequisite, labor, economy, and friends strength. When these things are fulfilled and completely surveyed, at that point the following stage is to decide about the product details in the separate framework, for example, what kind of working framework the venture would require and what are largely the important programming are expected to continue with the subsequent stage like building up the apparatuses, and the related activities. After going through different types of surveys of different authors and students mentioned some of the the fundamental problems and issues with respect to their research.


Frequent occurrences of crimes inside a society could have a massive impact on companies and institutions. Therefore, there is a compelling want to investigate the reasons, elements, and relationships associated with numerous crimes, aiming to become aware of effective strategies for manipulate and prevention. This research focuses on the classification of clustered crimes based totally on their prevalence frequencies throughout unique years. Utilizing statistics mining strategies for analysis, research, and sample discovery, we implemented a theoretical model to actual crime datasets recorded by means of the police in England and Wales from 1990 to 2011. To enhance the version's great, we assigned weights to the functions, emphasizing their significance and getting rid of much less valuable ones. The Genetic Algorithm (GA) become employed to optimize the parameters of the Outlier Detection operator using the Rapid Miner tool. The model has the data of limited crime and can not be suitable for the large amount of the crime data.


Crime and violations pose a significant challenge to the administration of justice and demand effective control measures. The computational enhancement of metropolitan safety relies on accurate crime prediction and forecasting trends. Despite substantial research efforts, there remains a gap in achieving optimal predictive algorithms that can guide police patrols toward potential criminal activities. This study addresses this gap by applying various machine learning algorithms, including logistic regression, support vector machine (SVM), Naive Bayes, k-nearest neighbors (KNN), decision tree, multi-layer perceptron (MLP), random forest, and extreme Gradient Boosting. Additionally, time series analysis is conducted using long short-term memory (LSTM) and auto-regressive integrated moving average (ARIMA) models to better fit crime data. The study found that LSTM performed reasonably well in time series analysis, as evidenced by the magnitude of root mean square error (RMSE) and mean absolute error (MAE) on both datasets. This comprehensive approach to leveraging machine learning and time series analysis contributes to advancing the accuracy and effectiveness of crime forecasting and prediction models. The overall system will now not expect the time in which the crime is occurring and time consuming.

2.3 “Crime Prediction using KNN” Vrushali Pednekar, Trupti Mahale and Arti Gore (IJRITCC-2018)

The increase in reporting of criminal cases, advances in data mining, led to the development of research methods aimed at extracting insights from criminal records and gaining a deeper understanding of criminal behavior and they have actively prevented future criminal activity although many methods in this field clustering and association deal There are methods that do not use mining methods and yet focus on predictive crime models which is completely absent. This study examines predictive models to estimate the number of
crime types based on LSOA codes (Lower Layer Super Output Areas—the operational framework used by the UK police) and the frequency of antisocial behaviour. Three algorithms representing different approaches—pattern-based learning, regression, and decision trees—are used. The data set from UK police includes more than 600,000 records before it is processed. The findings, considering both predictive accuracy and processing time, suggest that decision trees (specifically the M5P algorithm) demonstrate reliability in forecasting crime frequency, including the occurrence of anti-social behavior crimes.

III. MACHINE LEARNING ALGORITHMS

Several different machine learning models were applied to the datasets. The detailed architecture of the models and their parameters are discussed in this section.

3.1 Support Vector Machine (SVM)

The Support Vector Machine (SVM) stands out as a popular supervised learning algorithm known for its versatility in solving the challenges of regression and classification. Although SVM is often used for machine learning and classification jobs, however, its variability extends to regression conditions. Originally developed for binary classification problems, SVM has become one of the most revolutionary techniques in the machine learning landscape. Its widespread adoption is evident in fields such as computational biology and handwriting recognition. In addition to the classification function, SVM has found applications in image processing, banking analysis, weather forecasting, and especially crime forecasting. The ability of algorithms to process complex data and identify patterns makes them valuable tools for predicting and understanding criminal activity, helping to improve law enforcement and crime prevention strategies.

3.2 Gaussian Naive Bayes

The classifier being referred to is the Naive Bayes classifier, a probabilistic model that relies on strong independence assumptions. Known for its simplicity and effectiveness, this classifier operates on the principles of the Bayes theorem from Bayesian statistics, earning it the moniker of an "independent feature model." Despite its naive assumption of feature independence, which seldom holds true in real-world scenarios, the Naive Bayes classifier remains remarkably efficient. Its ability to perform well and rapidly adapt to new information makes it a popular choice in various supervised classification tasks. This classifier’s strength lies in its computational efficiency and ease of implementation, making it particularly advantageous when dealing with large datasets or in situations where real-time learning and prediction capabilities are crucial.

3.3 Decision Tree Based Algorithm

Decision trees are a type of supervised learning algorithms employed for both classification and regression problems. A Decision tree is a non-parametric supervised gaining knowledge of algorithm, which is utilized for both class and regression tasks. It has a hierarchical, tree shape, which consists of a root node, branches, internal nodes and leaf nodes.

3.4 K Nearest Neighbor (KNN)

The algorithm described here is the k-nearest neighbor (KNN) algorithm, a distance-based approach that excels when all attribute values are continuous. One of the key strengths of KNN lies in its adaptability to categorical attributes, making it suitable for a wide range of datasets. In this method, the classification of an unknown instance is approximated by considering the class of the instances that are most similar to it. The "k" in KNN refers to the number of nearest neighbors considered when making a classification or regression prediction. Essentially, the algorithm determines the class of a given instance by assessing the classes of its k-nearest neighbors in the problem space. This technique is widely used in pattern classification and categorization, offering a flexible and intuitive approach to classifying elements based on their proximity to training examples in the feature space.

3.5 Random Forest

Random Forest serves as a flexible tool in machine learning, perfect for classification and regression jobs. It's part of the supervise-d learning method and utilizes ensemble learning to boost the model's effectiveness. The algorithm works by making a 'forest' of many decision trees. Training for these trees is through bagging or bootstrap aggregating. This method is meant to increase a machine learning model's accuracy. Essentially, Random Forest stands out as a classifier that harnesses the collective intelligence of diverse decision trees to
improve predictive accuracy. By embracing the principle of ensemble learning, it enhances the model's resilience and generalization capabilities, making it a powerful tool in the realm of machine learning.

The working diagram of the Random Forest algorithm:

![Structure of Random Forest](image)

**Figure 1: Structure of Random Forest**

Here are some important points that explain to choose Random Forest algorithm:

1. It takes less training time as compared to other algorithms.
2. It predicts output with high accuracy, even for the large datasets it runs efficiently.
3. It can also maintain accuracy when a large amount of data is missing.

IV. CONCLUSION

The primary objective of this study is to estimate crime rates in various locations, with a specific focus on factors such as population density, country, crime rate, and centrality. The chosen destination for this investigation is India, and a Random Forest model has been employed to forecast crime rates. The study presents a graph following the implementation of the Random Forest model. Recognizing that crime is a significant global issue impacting communities and societies worldwide, the research emphasizes the need to address and manage this challenge effectively. The study acknowledges the complexity of predicting crime and extracting meaningful insights from extensive crime datasets. The potential reduction of crime through advanced forecasting is highlighted, emphasizing the ongoing efforts to enhance prediction systems through data gathering and mining technologies. The survey aims to improve crime predictions by identifying trends and patterns, anticipating the types of violations likely to occur in specific districts during specific time periods and seasons. The practical application of crime prediction is underscored, emphasizing its role in helping individuals make informed decisions about their living and visiting choices based on the anticipated crime scenarios in different neighborhoods at various times.

ACKNOWLEDGEMENTS

We would like to express my sincere gratitude to our project guide, Mrs. Bindu K P, Assistant Professor for their invaluable guidance, unwavering support, and insightful feedback throughout the course of this research which have significantly contributed to the successful completion of this research project.

V. REFERENCES


