

Mini Project Report on

**CRIME RATE PREDICTION USING MACHINE
LEARNING**

**Submitted in partial fulfillment of the requirement for the award of the
degree of**

**BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING**

Submitted by:

Student Name: Abhay Rautela

University Roll No.: 2018059

Under the Mentorship of
Ms. Sonali
Assistant Professor

**Department of Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand
January 2023**

CANDIDATE’S DECLARATION

I hereby certify that the work which is being presented in the project report entitled “**Crime Rate Prediction Using Machine Learning**” in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era Hill University, Dehradun shall be carried out by myself under the mentorship of **Ms. Sonali, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

Name: Abhay Rautela

University Roll no.: 2018059

Table of Contents

Chapter No.	Description	Page No.
Chapter 1	Introduction	1 - 2
Chapter 2	Literature Survey	3 - 5
Chapter 3	Methodology	6 - 8
Chapter 4	Result and Discussion	9
Chapter 5	Conclusion and Future Work	10
	References	11

Chapter 1

Introduction

1.1 Introduction

Crime is an act that is prohibited by law and is punishable by a fine, imprisonment, or other legal action. Every day, reports of criminal activity fill our news outlets and social media platforms, painting a picture of a world in which crime is an ever-present concern. From robberies and violent assaults to cybercrimes and white-collar fraud, there is seemingly no end to the number of ways in which criminals can cause harm. Crime has been a part of human civilization since time immemorial. It has become increasingly prominent in today's world. The rise of technology has created a variety of new crimes, while the emergence of globalization has made the world a smaller place, allowing criminals to move and operate in different countries.

Crime is uncertain and cannot be predicted. Crime prediction is significant to determine increase or decrease in crime rate from preceding years. A huge number of crimes happen every second in different places, in different patterns and in different times and the number is increasing each growing day. A good prediction technique provides a more rapid evolution of criminal data sets. It helps in predicting the correct place of crime and criminal activity, as well as aids in keeping track of resources pertaining to the analysis of crime.

Crime prediction using machine learning is an emerging field of study that uses sophisticated algorithms and data-driven methods to detect and predict criminal activities. Machine learning algorithms can be used to identify patterns in data that may indicate a future crime, such as past criminal activities, demographic information, and environmental factors. By leveraging such data, machine learning can be used to create predictive models that identify the likelihood of a certain crime occurring in a particular area or time frame. Additionally, machine learning can be used to develop insights into the behavior of criminals, helping law enforcement professionals better understand and address criminal activity.

1.2 Problem Statement

Crime is a major problem in today's world, and it is a threat to global security. The population of cities is constantly increasing, resulting in an increase in crime rates. Officials are tasked with the monumental challenge of accurately predicting future crime rates and attempting to reduce them. To help in this regard, various large datasets have been reviewed, extracting information such as location and crime type. Crime prediction utilizes various methods to identify areas that are likely to experience higher levels of crime. Given a set of historical crime data, develop a predictive model to identify areas of high risk for future criminal activity, to improve the accuracy of crime prevention and policing efforts. These methods include analyzing past crime data, identifying crime hotspots, and utilizing predictive analytics.

1.3 Objectives of the Project

The main aim of this project is to develop a system that can accurately predict crime rates and identify potential future crime trends. This information can then be used by officials to devise strategies to reduce crime rates and create a safer environment. To predict the crime rate (dependent variable) based on the year, location, and type of crime (independent variables), various types of machine learning algorithms will be applied. The system will examine how to convert the crime information into a regression problem, thus helping the officials to solve crimes faster. Crime analysis using available information to extract patterns of crime. Based on the territorial distribution of existing data and the recognition of crimes, various multi-linear regression techniques can be used to predict the frequency of crimes.

Chapter 2

Literature Survey

Various applications of crime rate prediction have been carried out, some of which are listed below:

Prediction of Crime Rate in Banjarmasin City Using RNN-GRU Model proposed by Muhammad Alkaff describes a model to predict the crime rate by using the Recurrent Neural Network (RNN) with the Gated Recurrent Unit (GRU) architecture. The model takes into consideration the inflation rate and discretionary income. GRU is a modified RNN algorithm that is simpler than the Long-Short Term Memory (LSTM) Neural Network and is more effective in adapting to different timescales and dealing with Vanishing Gradient problems. It consists of two gates, the Update gate (z_t) and the Reset gate (r_t), and is compatible with data that is not as much as LSTM, achieving optimal results even with fewer data. After collecting and normalizing the data, the model produced the best results with the lowest MAE and RMSE values of 1.7368 and 2.21, respectively, and an R-Squared value of 0.84, indicating good model performance.[1]

Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques proposed by Wajiha Safat aims to analyze crime prediction in the Chicago and Los Angeles datasets by improving the predictive accuracy with the Logistic Regression, SVM, Naïve Bayes, KNN, Decision Tree, MLP, Random Forest, and XGBoost algorithms, time-series analysis with LSTM, exploratory data analysis for visual summary, and crime forecasting for the crime rate and high-intensity crime areas for subsequent years with an ARIMA model. This paper investigated the predictive accuracy of eight different algorithms for the Chicago and Los Angeles datasets, with XGBoost performing best with an accuracy of 94% and 88%, respectively. To measure scale-dependent error, an LSTM model was implemented, and RMSE and MAE metrics were used. In addition, an ARIMA model was used to forecast future crime density areas, indicating that Chicago will continue to increase moderately, followed by a stable decline, while Los Angeles will decline sharply.[2]

Sakib Mahmud and Musfika Nuha proposed the relationship between crime and different features in the criminology literature. To reduce crimes and detect criminal activity,

the author used Z-Crime Tools and Advanced ID3 algorithms with data mining technology, K-Means Clustering and deep learning algorithms, random forest and naïve Bayes algorithms, and multi-linear regression. Additionally, the author used Apriori and Naive Bayes algorithms to identify and predict criminal trends and patterns. For classification, algorithms such as Naive Bayes were used to classify objects into predefined groups and classes. The accuracy of different algorithms is evaluated, with K-nearest neighbour providing the most precise crime rate forecast system. Linear, Naive Bayes and KNN algorithms had accuracy scores of 73.6%, 69.5% and 76.9% respectively.[3]

Gaurav Hajela proposed a clustering-based hotspot identification approach for crime prediction. The study of crime shows that it can be represented with a spatiotemporal pattern across geographical space. There are many indicators of crime such as urban or census-based indicators, streetlight and daylight, social media-based indicators, population flow indicators, and climate-based indicators. A crime hotspot is an area with a higher concentration of crime than the rest of the area. This paper proposes a crime prediction model for the dataset of San Francisco, which includes crime hotspot identification, dataset preparation, and crime prediction approach. Results show that the best accuracy is obtained when $k=4$ and when coupled with hotspot identification. The decision tree approach achieved 83.95 % and outperforms Nave Bayes.[4]

Masoomali Fatehkia used Facebook interests to improve predictions of crime rates in urban areas. This study discusses the potential for using data from the Facebook Advertising API to gain insight into the distribution of individual-level processes concerning crime rates across different neighbourhoods. It begins by describing existing theories of carcinogenesis related to factors such as poverty, social disorganization, income inequality, and impulsivity. It then outlines how the API could be used to measure the prevalence of interests among a ZIP code's Facebook population, which can be used to reflect the behavioural and attitudinal features of a population. The models used only demographic factors, only Facebook interests, or both, and controlled for each city's baseline crime rate and the age composition of the neighbourhood. Results showed that the combination of demographic factors and Facebook interests had the greatest predictive strength for all three crime types, both in-sample (using adjusted R^2) and out-of-sample prediction (using MAE).[5]

Crime Rate Prediction using KNN proposed by Ms. Vrushali Pednekar, Ms. Trupti Mahale, Ms. Pratiksha Gadhave, and Prof. Arti Gore discusses about a system that convert crime information into a data-mining problem to help detectives solve crimes faster. It focuses on crime analysis, extracting target datasets, data pre-processing, data mining, and interpretation and using discovered knowledge. The proposed model of crime analysis and prediction uses a general algorithm which takes raw data of crime from a government repository as input and produces a correlated dimensions model for crime analysis and prediction as output. It also uses various data mining techniques to predict the frequency of occurring crime based on territorial distribution of existing data. It also involves data cleaning and treating missing values to improve the quality of data for mining. With the proposed system, real-time data can be analyzed to cluster and predict crimes. The methods proposed for crime prediction do not address parameters such as outlier effects during the data mining preprocessing, the quality of training and testing data, or the value of features. [6]

Chapter 3

Methodology

3.1 Ensemble Learning

Ensemble learning is a type of machine learning technique that combines multiple individual models to produce better predictive performance than could be achieved from any of the individual models alone. It works by building multiple models from the same training data set, then combining the models to make more accurate predictions. Ensemble learning has been shown to be successful in a wide variety of applications from computer vision to natural language processing. It is popular because it can produce better results with less data and is more robust to outliers in the data. Ensemble learning is used in many areas like image recognition, natural language processing, and medical diagnosis.

3.2 Algorithm - Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees to generate a more accurate prediction. It is a versatile and powerful algorithm used in a wide variety of applications. The main idea behind the Random Forest algorithm is to combine multiple decision trees into a single model, thereby reducing the variance and improving the accuracy of predictions. This is achieved by randomly sampling data points, randomly selecting features, and then building multiple decision trees on the data. The predictions from each tree are then averaged to produce the final prediction. Random Forests have been shown to be more accurate than traditional decision trees and have become one of the most popular machine learning algorithms. They are also very robust, even when dealing with large datasets, and are resistant to overfitting. The algorithm can be illustrated as follows:

1. Begin with a dataset of observations and their associated labels.
2. Randomly select 'k' features from the dataset.
3. For each of the 'k' features, choose the best split point.
4. Create a tree using the chosen split points.
5. Repeat steps 2-4 for each of the 'k' features.
6. Combine the trees to form a forest.
7. Use the forest to make predictions on new data.
8. Calculate the accuracy of the predictions.

3.3 Proposed System

- Initially the dataset is prepared manually based on the publication available on the National Crime Rate Bureau (NCRB) official website.
- Data Preprocessing: The data is prepared in the correct format for analysis. Some columns are removed or transformed, and label encoding is used to convert the categorical data into numeric for better prediction.
- Random Sampling: After feature selection, the data has been splitted into two parts: training data (70%) and testing data (30%).
- Model Creation: The model algorithms are imported from sklearn. Model is build using `model.fit()`. The dataset has been analyzed using five different models: support vector machine, nearest neighbor, decision tree, random forest, and neural network.
- Model Selection: Based on the defined goals and model performance, random forest model has been selected. Prediction is done using `model.predict()`. The model accuracy is determined using `accuracy_score` imported from metrics.
- Model Deployment: The model has been deployed for prediction using various web technologies.

The project has undergone the following process:

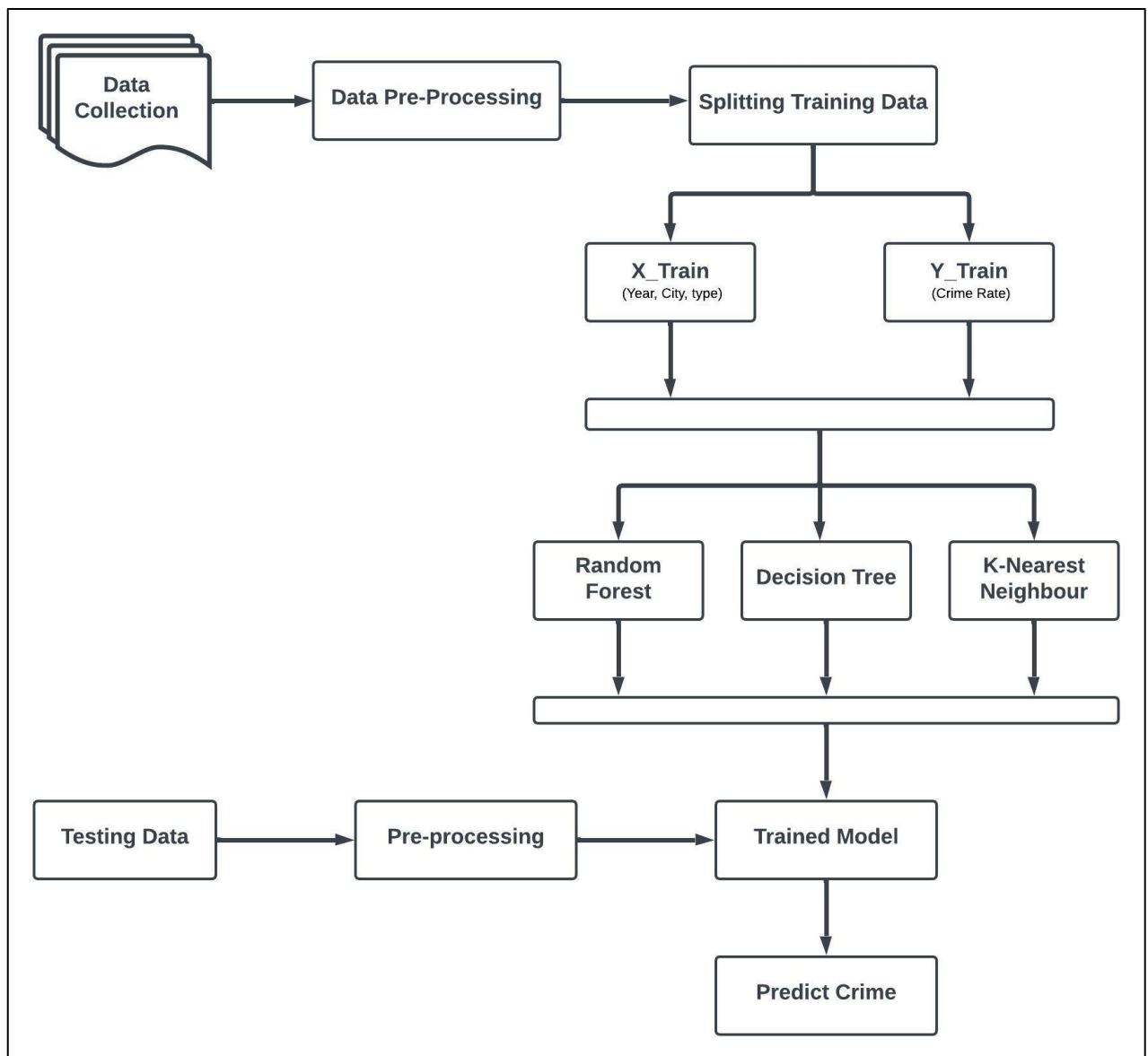


Fig 3.1 System Architecture

Chapter 4

Result and Discussion

The Random Forest Regression model demonstrates the best accuracy in predicting test data among the five selected models. The model predicts the crime rate value for 10 different categories of crimes, including Murder, Kidnapping, Crime against Women, Crime against Children, Crime Committed by Juveniles, Crime against Senior Citizens, Crime against SC, Crime against ST, Economic Offenses, Cyber Crimes that will occur in 19 Indian metropolitan cities: Ahmedabad, Bengaluru, Chennai, Coimbatore, Delhi, Ghaziabad, Hyderabad, Indore, Jaipur, Kanpur, Kochi , Kolkata, Kozhikode, Lucknow, Mumbai, Nagpur, Patna, Pune, Surat in future.

The accuracy results obtained after testing are listed below:

Algorithm	Mean Absolute Error	Mean Squared Error	R2 Score
Support Vector Machine	10.3204	371.7907	0.17886
K-Nearest Neighbor	6.58181	140.8179	0.55349
Neural Networks MLPRegressor	12.4248	307.5506	0.24823
Decision Tree Regressor	2.89024	34.95932	0.88915
Random Forest Regressor	2.49143	21.43956	0.93201

Chapter 5

Conclusion and Future Work

Crime rate prediction has become an important tool for law enforcement agencies to help them focus their resources in high-crime areas. With the help of sophisticated algorithms and data analysis, law enforcement agencies can predict when and where crimes are likely to occur. By focusing their resources in the right areas, police officers can help reduce the overall crime rate in a community. Predictive policing has already proven to be an effective tool in reducing crime rates in many areas, and it looks like it will continue to be a key tool in the future.

As a result of machine learning technology, finding relationships and patterns between various data has become easier. The project focuses primarily on predicting the crime rate given the year, city, and types of crime in the future. The training data has been cleaned and transformed to create a machine learning model using the concept of machine learning. The model predicts the crime rate with an accuracy of 93.20%. The model prediction of crime rate and data visualization helps in analysis of data set and prediction of crimes. Many graphs are created to find interesting statistics that helped in understanding different crime datasets that can be used in implementing the factors that can help in keeping society safe.

References

- [1] M. Alkaff, N. F. Mustamin, and G. A. A. Firdaus, “Prediction of Crime Rate in Banjarmasin City Using RNN-GRU Model”, *Int J Intell Syst Appl Eng*, vol. 10, no. 3, pp. 01–09, Sep. 2022.
- [2] W. Safat, S. Asghar and S. A. Gillani, “Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques”, in *IEEE Access*, vol. 9, pp.
- [3] Mahmud, S., Nuha, M., Sattar, A. (2021). “Crime Rate Prediction Using Machine Learning and Data Mining”. In: Borah, S., Pradhan, R., Dey, N., Gupta, P. (eds) *Soft Computing Techniques and Applications. Advances in Intelligent Systems and Computing*, vol 1248. Springer, Singapore. https://doi.org/10.1007/978-981-15-7394-1_5
- [4] Gaurav Hajela, Meenu Chawla, Akhtar Rasool, “A Clustering Based Hotspot Identification Approach for Crime Prediction”, *Procedia Computer Science*, Volume 167, 2020, Pages 1462-1470, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.03.357>.
- [5] Fatehkia, Masoomali & O’Brien, Dan & Weber, Ingmar. (2019). Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas. *PLOS ONE*. 14. e0211350. 10.1371/journal.pone.0211350.
- [6] Ms. Vrushali Pednekar, Ms. Trupti Mahale, Ms. Pratiksha Gadhave, Prof. Arti Gore. 2018. “Crime Rate Prediction Using KNN”. *International Journal on Recent and Innovation Trends in Computing and Communication* 6(1) : 124 - <https://doi.org/10.17762/ijritcc.v6i1.1392>.