International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Crime Rate Prediction Using K- Nearest Neighbors (Knn) Algorithm

Y Reddemma

Computer Science and Engineering,
S V U College of Engineering, Tirupati, India.

Abstract

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. The main aim of this paper 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.

The proposed system will focus on converting the crime prediction task into a regression problem, wherein the crime rate (dependent variable) is predicted based on independent variables such as year, location, and type of crime. Various machine learning algorithms, including multi-linear regression techniques, will be applied to model the relationships between these variables and the frequency of crimes. Through rigorous analysis of past crime data, the system will identify crime hotspots and provide predictive analytics to assist in proactive crime prevention strategies. Moreover, this project will delve into the spatial and temporal distribution of crimes, using geographic information systems and clustering algorithms to identify areas that are statistically more prone to criminal activity. By recognizing patterns in crime distribution, the system will offer predictive insights that can be instrumental in strategic planning for law enforcement operations. The ultimate objective is to create a safer environment by enabling officials to not only react to crimes more effectively but also to take preemptive measures based on data-driven predictions. The findings and methodologies developed through this paper will have broad applications in urban planning, public safety, and policy-making, contributing to the broader efforts of reducing crime rates and enhancing the security of urban populations. The integration of machine learning models with real-time data will pave the way for innovative approaches to crime prevention, ultimately leading to safer and more resilient communities.

Keywords: Prediction algorithms, Machine Learning, Predictive models, K-Nearest Neighbor, Crime Prevention, Classification Task, Recurrent Neural Network, Mean Absolute Percentage Error, Machine Learning Prediction, Neighborhood Crime, Crime Data, Crime Patterns, Crime Reports, Anomaly Data, Crime prediction, crime datasets, smart policing, Mean Squared Error**.**

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.

2. LITERATURE REVIEW

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 (zt) and the Reset gate (rt), 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 R2) 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]

3. METHODOLOGY

K-Nearest Neighbor-Based Model for Crime Rate Prediction

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 dataset, 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 languageprocessing. 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.

Algorithm - K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, yet effective, machine learning algorithm used for classification and regression tasks. The core idea behind KNN is to identify the ‘k’ closest data points to a given input and make predictions based on the majority class (in the case of classification) or average value (in the case of regression) of these neighbors. The algorithm is non-parametric, meaning it does not make any assumptions about the underlying data distribution, which makes it versatile across various types of datasets.

1. Begin with a labeled dataset: Each data point has associated features and a label.

2. Select the number of neighbors (k): The value of ‘k’ is a hyperparameter that determines how many neighbors influence the prediction.

3. Calculate the distance: For a new input, compute the distance (commonly using Euclidean distance)

between this input and every point in the training dataset.

4. Identify the nearest neighbors: Select the ‘k’ closest data points based on the computed distances.

5. Classify the new input: In classification, the majority class among the neighbors determines the predicted class. In regression, the average of the neighbors' values is taken as the prediction.

6. Make a prediction for the new data point: Based on the identified neighbors, the final prediction is made.

7. Calculate the accuracy of the predictions: Accuracy or other relevant metrics are computed to evaluate the model’s performance.

The proposed methodology begins with data collection, where a dataset related to crime rates in various cities is sourced from a reliable repository, such as an official website. The dataset undergoes preprocessing to ensure it is in the correct format for analysis. This step includes removing or transforming certain columns and applying label encoding to convert categorical data into numerical values, which facilitates better predictions. The data is then divided into two subsets: training data (70%) and testing data (30%) using random sampling.

For the model creation, a K-Nearest Neighbors (KNN) regression algorithm is utilized. The dataset is analyzed to predict crime rates across different cities and crime types. After fitting the model with the training data using model.fit(), the model's performance is evaluated on the test data. Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score are computed to assess the model's accuracy.

Following the model selection, where the KNN model is chosen based on its performance, the trained model is saved using Python's pickle module for future use. This saved model is later deployed in a Tkinter-based graphical user interface (GUI) application. The application allows users to predict crime rates by selecting a city, crime type, and year. The model predicts the crime rate, which is then categorized into "Very Low Crime Area," "Low Crime Area," "High Crime Area," or "Very High Crime Area," based on predefined thresholds. The GUI provides dynamic feedback and displays the estimated crime rate, predicted number of cases, and the crime status, making it an intuitive tool for crime rate prediction and analysis.

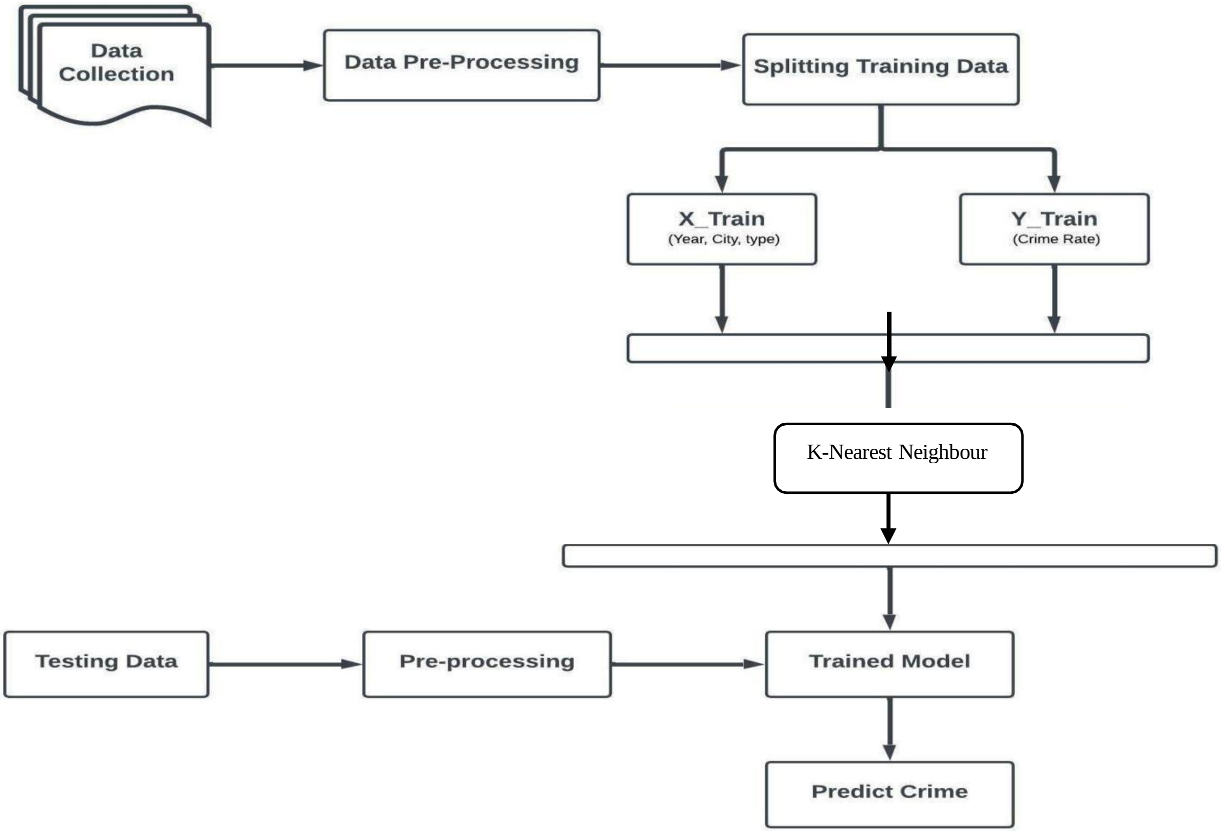


Fig 1: System Design

4. IMPLEMENTATION AND RESULT

The K-Nearest Neighbor-Based Model for Crime Rate Prediction was implemented using Python, leveraging various libraries and methodologies. The implementation process was structured into several key stages: data acquisition, preprocessing, feature scaling, model training, evaluation, and saving the model.

- **Data Preparation**: The dataset is initially prepared manually based on publications available on the National Crime Rate Bureau (NCRB) official website.

dataset = pd.read_excel("Dataset/crp.xlsx", sheet_name="Sheet1") for i in range(0, 11): plt.figure(figsize=(12, 6))
plt.barh(dataset['City'], dataset[dataset.columns[i+2]], 0.6, color='Salmon') plt.title('City vs ' + dataset.columns[i+2], fontsize=16) plt.xlabel(dataset.columns[i+2], fontsize=14)
plt.ylabel('City', fontsize=14) plt.tick_params(axis='x', labelsize=12) plt.tick_params(axis='y', labelsize=12) plt.show()

- **Data Preprocessing**: The data is cleaned and transformed to be in the correct format for analysis. This includes removing or transforming some columns and using label encoding to convert categorical data into numeric values for better prediction.

```
new_df = pd.DataFrame(columns=['Year', 'City', 'Population (in Lakhs) (2011)+', 'Number Of Cases',
'Type']) for i in range(3, 13):
temp = dataset[['Year', 'City', 'Population (in Lakhs) (2011)+']].copy() temp['Number Of Cases'] =
dataset[[dataset.columns[i]]] temp['Type'] = dataset.columns[i]
```

```
new_df = pd.concat([new_df, temp])
new_df['Crime Rate'] = new_df['Number Of Cases'] / new_df['Population (in Lakhs) (2011)+'] new_df =
new_df.drop(['Number Of Cases'], axis=1) new_df.to_excel("Dataset/new_dataset.xlsx", index=False,
sheet_name ='Sheet1')
```

- **Random Sampling**: After feature selection, the dataset is split into two parts: 80% for training and 20% for testing.

```
x = new_dataset[new_dataset.columns[0:4]].values y = new_dataset['Crime Rate'].values
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=50)
```

- **Model Creation**: The KNN algorithm is imported from the sklearn library, and the model is built using KNeighborsClassifier().fit(). The algorithm has been tested with different values of ‘k’ to identify the optimal

number of neighbors for the model.

```
model1 = KNeighborsRegressor(n_neighbors=2) model1.fit(x_train, y_train)
y_pred = model1.predict(x_test)
```

- **Model Selection**: Based on the defined goals and model performance (e.g., accuracy, precision, recall), the KNN model with the best ‘k’ value is selected. Predictions are made using model.predict(), and the model's accuracy is determined using accuracy score from the metrics module.

- **Model Deployment**: The selected KNN model is deployed for prediction using python tkinter, allowing for practical use in predicting crime rates.

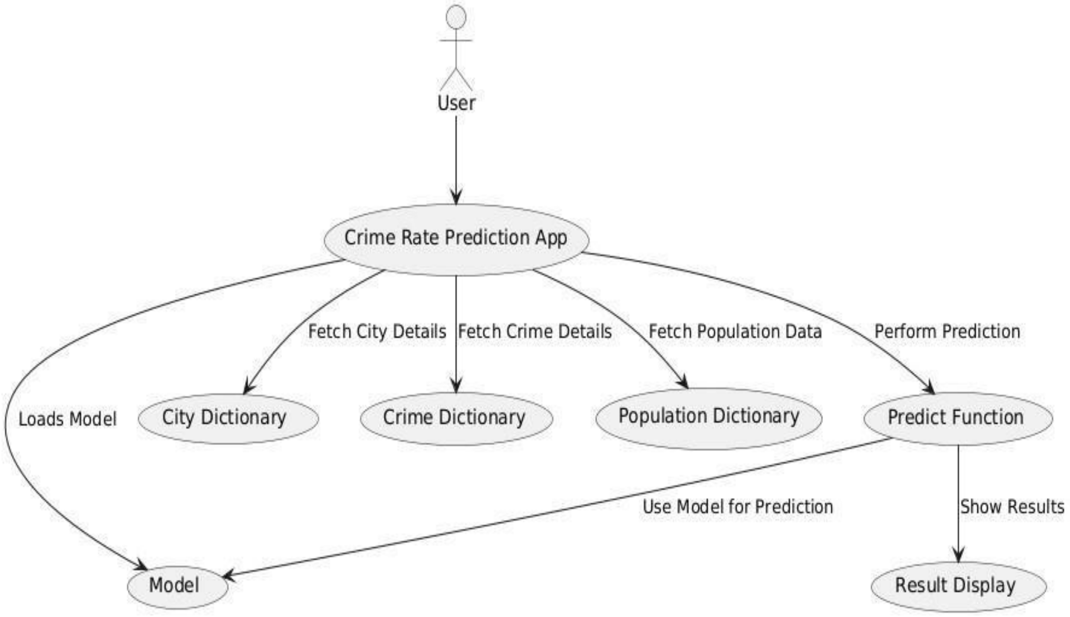


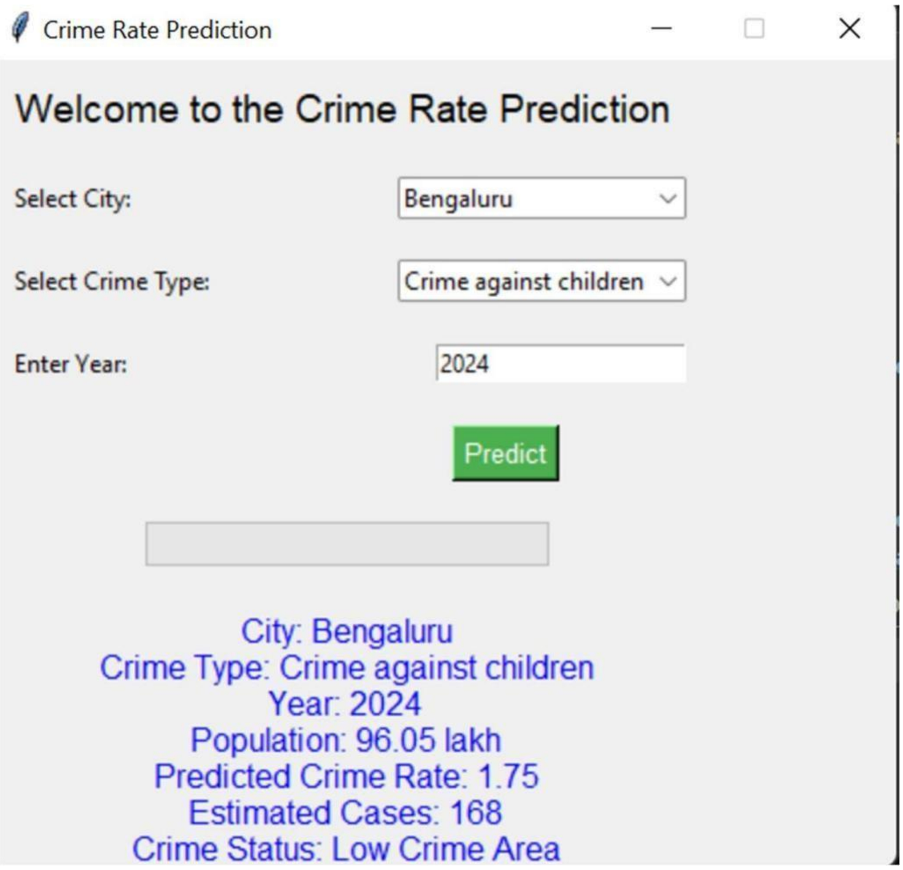
Fig 2: Data flow

5. Result

The The K-Nearest Neighbor-Based Model demonstrates the best accuracy in predicting test data. 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.

Algorithm	***K-Nearest Neighbor***
Mean Absolute Error Mean Squared Error R2 Score	6.58181
	140.8179
	0.55349

Table: accuracy results obtained after testing



Welcome to the Crime Rate Prediction

Select City: Bengaluru

Select Crime Type: Crime against children

Enter Year: 2024

City: Bengaluru
Crime Type: Crime against children
Year: 2024
Population: 96.05 lakh
Predicted Crime Rate: 1.75
Estimated Cases: 168
Crime Status: Low Crime Area

Fig 3 : output

6. CONCLUSION

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 found 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.