International Journal of Multidisciplinary Engineering in Current Research - IJMEC

Volume 10, Issue 5, May-2025, http://ijmec.com/, ISSN: 2456-4265

Analysis Crime Data Using Python

Pathan Mahaboobi
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
A.Naga Raju
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract

This project focuses on the analysis and prediction of crime trends across various states and union territories in India using machine learning techniques. The dataset comprises crime-related statistics categorized by state, district, and year. Initial data preprocessing steps include handling missing values and removing duplicates to ensure data quality. Exploratory Data Analysis (EDA) is conducted through various visualizations to highlight crime patterns, identify states with high and low crime rates, and observe temporal trends in Indian Penal Code (IPC) crimes.A machine learning model using Random Forest Regressor is trained to predict the total number of IPC crimes based on state, district, and year as input features. Label encoding is used to convert categorical variables into numeric format suitable for model training. The model's performance is evaluated using the R-squared metric, and predictions are visualized to compare actual versus forecasted crime numbers.Furthermore, a user interface component is incorporated, allowing users to input a specific state, district, and year to receive a crime forecast along with a safety classification (e.g., "Safest City", "Medium Safe City", or "Not Safe City"). This application can serve as a decision-support tool for policymakers and law enforcement agencies to proactively address crime trends.

I. Introduction

In the rapidly evolving digital age, data-driven decision-making is becoming an integral part of governance, law enforcement, and policy formulation. Crime, as a persistent societal challenge, affects public safety, social well-being, and economic development. Therefore, understanding crime patterns and forecasting potential criminal activities using historical data is critical for formulating effective strategies to maintain law and order. The increasing availability of public datasets and advancements in machine learning and data analytics have created opportunities to harness computational models for crime trend analysis and prediction. This project presents a comprehensive approach to analyzing crime statistics in India and predicting future crime occurrences using a supervised machine learning model—Random Forest Regressor.

India, with its diverse demography, varying socioeconomic conditions, and complex law enforcement mechanisms, presents a unique landscape for crime analysis. Each state and district has distinct patterns of criminal activities influenced by cultural, political, and economic factors. Traditional methods of crime analysis rely heavily on manual data inspection, which is both time-consuming and prone to errors. Moreover, it lacks predictive capability. With the emergence of machine learning techniques, it is now possible to automate crime data analysis and develop models that can accurately predict future crime trends, thus enabling proactive policing and resource allocation.The dataset used in this project comprises crime records from across the country, categorized by state/union territory (UT), district, and year. The dataset includes various crime types such as murder, rape, kidnapping, robbery, theft, dowry deaths, and other Indian Penal Code (IPC) crimes. The rich granularity of the dataset allows for comprehensive exploratory data analysis (EDA), revealing valuable insights into crime distribution, frequency, and trends over time. The initial stage of the project involves rigorous data cleaning and preprocessing steps, such as identifying and removing missing values, detecting duplicates, and encoding categorical variables for model training.

Visualization plays a crucial role in uncovering hidden patterns in crime data. Through various plots—bar charts, pie charts, and line graphs—this project identifies states with the highest and lowest crime rates, determines which crimes are most

prevalent, and examines how crime rates have changed over the years. For instance, bar graphs comparing total IPC crimes across states help in visualizing the relative safety of different regions, while line plots of average crimes per year offer a temporal perspective of criminal activity in the country.

Following EDA, the project employs a Random Forest Regressor model, a powerful ensemble learning technique that constructs multiple decision trees and merges their results to improve accuracy and control overfitting. The model is trained using selected features: state, district, and year, to predict the target variable, which is the total number of IPC crimes. The dataset is split into training and testing sets, ensuring a reliable evaluation of the model's generalization performance. The accuracy of the model is assessed using metrics such as R-squared score, mean absolute error, and mean squared error, which quantify how well the model captures the relationship between the input features and the crime rates.

An interactive component of this project allows users to input specific values for state, district, and year to receive a crime forecast. This feature uses the trained model to predict the total number of IPC crimes and classifies the forecast into one of three safety levels: "Safest City", "Medium Safe City", or "Not Safe City", based on predefined thresholds. This classification provides an intuitive understanding of the predicted crime levels and can be useful for both general public awareness and official decision-making.

II. Literature Survey

Crime analysis and prediction have long been subjects of interest for researchers in criminology, sociology, computer science, and data science. The increasing availability of digitized crime records, combined with advances in machine learning (ML) and data analytics, has led to a surge of studies aimed at modeling and predicting criminal behavior. The primary goal of these studies is to enable law enforcement agencies and policymakers to understand crime patterns and proactively design effective intervention strategies. This literature survey highlights the key contributions, methodologies, and technologies that have influenced the development of crime prediction systems.

1. Crime Pattern Recognition and Statistical Approaches

Early efforts in crime analysis primarily relied on statistical methods such as regression analysis, time-series analysis, and clustering. These techniques were used to identify trends, seasonality, and hotspots in crime occurrences.

For instance, **Chainey and Ratcliffe (2005)** emphasized the significance of Geographic Information Systems (GIS) in mapping and analyzing crime. Their work demonstrated how spatial data visualization could help identify high-risk areas. Similarly, **Piquero and Weisburd (2009)** discussed the application of longitudinal analysis and crime mapping to understand the temporal and spatial dynamics of crime.

However, while statistical methods provided foundational insights, they were limited in handling complex, non-linear relationships and large-scale multidimensional datasets, which paved the way for machine learning approaches.

2. Machine Learning in Crime Prediction

Machine learning has been increasingly adopted to overcome the limitations of traditional statistical techniques. Algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Neural Networks have shown promise in classifying and predicting criminal activities.

Kianmehr and Alhajj (2009) developed a framework using decision trees to detect and predict crime hot spots. Their system effectively identified high-crime areas by learning from historical data.

Similarly, **Mohler et al. (2015)** introduced the use of self-exciting point process models to predict the locations of future crimes based on past incidents, an approach that has since been adopted in several predictive policing tools.

S. Wang et al. (2017) applied Random Forest and SVM algorithms to predict different categories of crimes using features such as time, location, and weather. Their findings revealed that Random Forest offered high accuracy and was robust against overfitting, which supports the use of this algorithm in the current project.

3. Deep Learning and Neural Networks

Recent research has explored deep learning models for crime prediction due to their ability to capture complex patterns in large datasets. **Xu et al. (2018)** proposed a deep learning approach using Recurrent Neural Networks (RNNs) to model crime trends over time. Their model was able to capture sequential dependencies in crime data and improved forecasting accuracy compared to traditional models.

Similarly, **Zhao et al. (2019)** used Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for spatiotemporal crime forecasting. These models incorporated both spatial and temporal features and demonstrated improved performance in urban crime prediction tasks.

While deep learning models are highly accurate, they often require large amounts of labeled data and are computationally expensive. For projects with limited computational resources or where interpretability is a concern, ensemble models like Random Forests remain a practical and effective choice.

III. Proposed System

The proposed system is a **machine learning-based interactive crime forecasting application** that not only analyzes historical crime data but also predicts the **future number of IPC crimes** in any given district/state and year in India. It combines **data preprocessing**, **exploratory data analysis (EDA)**, **visualization**, and a **Random Forest regression model** to offer both insights and predictive capabilities.

Key Features of Proposed System:

Automated data processing: Handles missing values, duplicates, and encodes categorical data.

Dynamic visualizations: Displays state-wise and crime-wise trends through interactive plots and charts.

Predictive model: Uses supervised learning (Random Forest Regressor) to forecast IPC crime totals based on inputs.

User input capability: Accepts user queries for specific state, district, and year and returns a forecast along with safety classification.

Safety categorization: Classifies regions into “Safest”, “Medium Safe”, or “Not Safe” based on crime predictions.

Advantages:

Helps law enforcement anticipate future crime trends.

Supports data-driven policy and resource allocation decisions.

Improves public awareness of crime patterns.

Scalable and adaptable to other regions or updated datasets.

IV. Result

V. Conclusion

The **Crime Analysis and Prediction System** presented in this project provides a robust framework

for understanding crime patterns and forecasting future criminal activities using machine learning. By leveraging a dataset containing historical crime records across various Indian states and districts, the system not only analyzes but also predicts the likely crime rate for a given region and year.

Using the **Random Forest Regressor** as the core predictive model, the system achieves a high level of accuracy, thanks to its ability to handle non-linear data and multiple input features. With the integration of **Label Encoding**, the model efficiently interprets categorical data like State and District names, transforming them into a format suitable for numerical modeling.

References

1. **National Crime Records Bureau (NCRB), India**
   Crime in India Reports.
   Available at: https://ncrb.gov.in/en/crime-india
   (Used for crime data collection and analysis framework)
   a. **Scikit-learn Documentation**
      Scikit-learn: Machine Learning in Python.
      Available at: https://scikit-learn.org/stable/
      (Used for Random Forest Regressor, LabelEncoder, model evaluation, and data preprocessing)
   b. **Pandas Documentation**
      Pandas: Python Data Analysis Library.
      Available at: https://pandas.pydata.org/
      (Used for data manipulation and analysis)
   c. **NumPy Documentation**
      NumPy: The fundamental package for scientific computing with Python.
      Available at: https://numpy.org/doc/
      (Used for numerical operations and data handling)
   d. **Matplotlib & Seaborn**
      i. Hunter, J.D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering.
      ii. Waskom, M.L. (2021). *Seaborn: statistical data visualization*. Journal of Open Source Software.
      (Used for data visualization and exploratory data analysis)
   e. **Joblib Library**
      Joblib: Tools for lightweight pipelining in Python.
      Available at: https://joblib.readthedocs.io/
      (Used for saving and loading the trained machine learning model)
   f. **Tkinter GUI Documentation**
      Tkinter: Python's standard GUI package.
      Available at: https://docs.python.org/3/library/tkinter.html
      (Used for basic GUI elements in the CLI input system)
   g. **Kaggle Crime Datasets (if applicable)**
      Example: *Crime in India (NCRB)* – Public dataset on Kaggle.
      Available at: https://www.kaggle.com/
      (Alternative or supplemental dataset used for training or validation)
   h. **Bishop, C. M. (2006)**
      Pattern Recognition and Machine Learning, Springer.
      (Reference for machine learning principles and model evaluation)
   i. **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013)**
      An Introduction to Statistical Learning, Springer.
      (Used to understand regression models and evaluation techniques)