International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Machine Learning Based Prediction of Crime Rates in Districts in India

[1]Prof. Mayur Nanekar , [2]Vidya Kathole, [3]Divya Rahangdale , [4]Niranjan shinde, [5]Prathamesh Savant, [6]Prem Gaikwad
[1,2,3,4,5,6]Department of Computer Engineering, P.K. Technical Campus
Pune- 410501, India
[2]Katholevidya07@gmail.com, [3]divyarahangdale555@gmail.com ,
[4]niranjanshinde313@gmail.com
, [5]Prathameshsavant503@gmail.com, [6]premgaikwad1708@gmail.com

Abstract
Crime is a significant problem in India and around the world. Understanding trends in crime rates can help with understanding crime patterns in different districts. This project presents a system called CrimeSense which uses Machine Learning to predict crime rates in India based on historical crime data from different districts from 2014 to 2023.
The dataset was cleaned and analysed using 4 different regression models; Linear Regression, Decision Tree, Random Forest and Gradient Boosting. The models were compared using several metrics including MAE, RMSE and R^2 score. Initial results show that Gradient Boosting and Linear Regression both provided similar results. Both having an R^2 score of 0.4898 and 0.4896 respectively.
CrimeSense also provides ranked crime data for each district. The system was built using Python and Streamlit and makes crime data from previous years easier to analyse and view the results of the prediction. The aim of this project is to show that Machine Learning can be used to understand crime trends in different districts in India.
Keywords: Crime rate prediction ;Machine learning ;Hotspot analysis; Indian metropolitan cities.

1.Introduction

1.1Background and Motivation

Crime is an important social problem that can affect the safety of the public, growth of the economy and the quality of life. Understanding crime trends is important to understand which areas have a higher crime rate. In a country with a high level of geographic and social diversity like India crime can vary greatly between states, districts and years.
Traditionally crime data has been analysed using historical data, statistical analysis and manually analysing the crime trends. This can provide useful information but does not always capture

complex patterns and associations from large datasets. However due to the availability of historical crime data and advancements in machine learning techniques, data driven methods can be used to analyse crime trends and predict future crime rates.

Machine learning techniques can be used to identify associations between historical data and crime rate trends. Regression models are useful for predicting a continuous numeric value such as the crime rate per 100,000.

The CrimeSence project aims to use Machine Learning techniques to analyse historical crime data from districts in India. The system will analyse crime data from 2014 to 2023, compare different regression models, predict future crime rates and provide visualised crime data for districts.

The aim of this research is to create a data driven system to make historical crime data easier to analyse.

1.2Problem Statement

There is a large amount of historical crime data in India however it can be difficult to analyse. This is because there are many differences between crime types, population, reporting and geography.

The main problem with this research is how to predict future district level crime rates using historical crime data. While traditional descriptive analysis can be used to understand previous trends, it does not provide Machine learning based predictions of future crime rates.

This research looks at using regression based machine learning models to analyse historical crime data from districts. The models can then be compared to see how well they can predict crime rates.

1.3Research Objectives and Scope

The main objective of this research is to create a Machine Learning based approach to predict crime rates in districts in India.

Specifically the project will;

1. Collect and analyse historical crime data from districts in India.
2. Clean data to remove missing values, duplicates, incorrect data types.
3. Undertake exploratory data analysis to understand crime trends.
4. Use regression models to predict future crime rates.
5. Compare the performance of Linear Regression, Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor.
6. Compare models using MAE, RMSE and R^2 score.
7. Identify districts with higher crime rates using hotspot analysis.
8. Provide interactive visualisations of crime related data using CrimeSense.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Scope of Research

This research looks at using historical crime data from districts in India from 2014-2023. It focuses on using numerical techniques to predict crime rates and comparing the performance of regression models. The project also includes ranking and visualising crime data for districts.

The system is for analysis and research purposes only. It does not attempt to predict individual criminal behaviour or guarantee that criminal activity will be prevented.

1.4Research Contributions

The main contributions to this research are;

A Machine Learning based system to predict crime rates in districts in India.

Comparison of 4 regression models using standard metrics.

Analysis of historical crime trends in districts.

Crime rate prediction using district level hotspot ranking.

Crime rate prediction using district level hotspot ranking.

CrimeSense, a web application to provide crime analysis and prediction.

Geographic visualisation to make district level crime data easier to understand.

2. Literature Review

2.1 Crime Prediction Research

Crime prediction has become increasingly popular in research due to availability of historical crime data. Many researchers have used different statistical and machine learning techniques to analyse crime data and identify areas with a higher risk of crime.

Previously crime data would be analysed using historical reports and statistical analysis.

However with the advancement of Machine Learning researchers have began using algorithms such as Decision Trees, Random Forest, Support vector machines and Neural Networks to analyse crime related data.

Some research has been carried out to predict individual crimes, however other research has looked at crime trends in cities or geographical areas. This shows that Machine Learning can be useful to identify patterns in large sets of data. However the results depend on the quality of the data, which features are used and what geographical area is analysed.

The CrimeSene project looks at analysing historical crime data in districts in India. Regression techniques are used to predict numerical crime rates and provide visualisations for districts.

2.2Machine Learning for Regression

Regression is a Machine Learning technique used to predict outputs as a numerical value. In this case the target variable is the crime rate.

Four different regression models were used in this project.

1. Linear Regression

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Linear Regression is a basic regression model used to predict numerical values. It identifies the relationship between the input features and output. In this project it was used to understand how different factors relate to crime rates and predict future crime rates. It was also used as it is an easy model to understand how it works.

2. RandomForest Regressor

Random Forest Regressor is a Machine Learning model that uses multiple decision trees to make predictions. Each tree produces a result and they are combined to produce the final result.

3. Decision Tree Regressor

Decision Tree Regressor is an algorithm that splits the data into smaller groups based on different conditions. It works like a flowchart where different decisions are made leading to the next step until the final prediction is reached. This helped identify different patterns in the crime data and predict crime rates based on the information available.

4. Gradient Boosting Regressor

Gradient Boosting Regressor is an algorithm that creates multiple decision trees. Each tree focuses on fixing the problems made by the previous tree. This helps improve the results to make more accurate predictions. In this project it was used to predict future crime rates by learning from patterns in the historical rime data.

2.3Research Gap

After researching previous research it was found that many projects focus on analysing crime in certain cities rather than districts throughout India. There is still a need to understand historical crime trends and use Machine Learning to predict future crime rates.

The main research gaps identified were;

1. Geographic coverage - Understand crime trends in different districts throughout India rather than certain cities.
2. Historical analysis - Understand crime trends by analysing data from multiple years.
3. Compare models - Compare different regression models to see which performs better.
4. Time based - Use previous years as training data and future years as testing data.
5. Combined - Combine crime prediction, hotspot identification and visualisations into one application.

CrimeSense aims to cover all of these areas by analysing historical crime data, predicting future crime rates and providing visualisations for crime trends in districts.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

3.Methodology

3.1Research Design And System Overview

This project follows a step by step process to analyse historical crime data and predict future crime rates. The first step was to collect and clean the crime data. Then the data was analysed to understand crime trends in different districts and years.

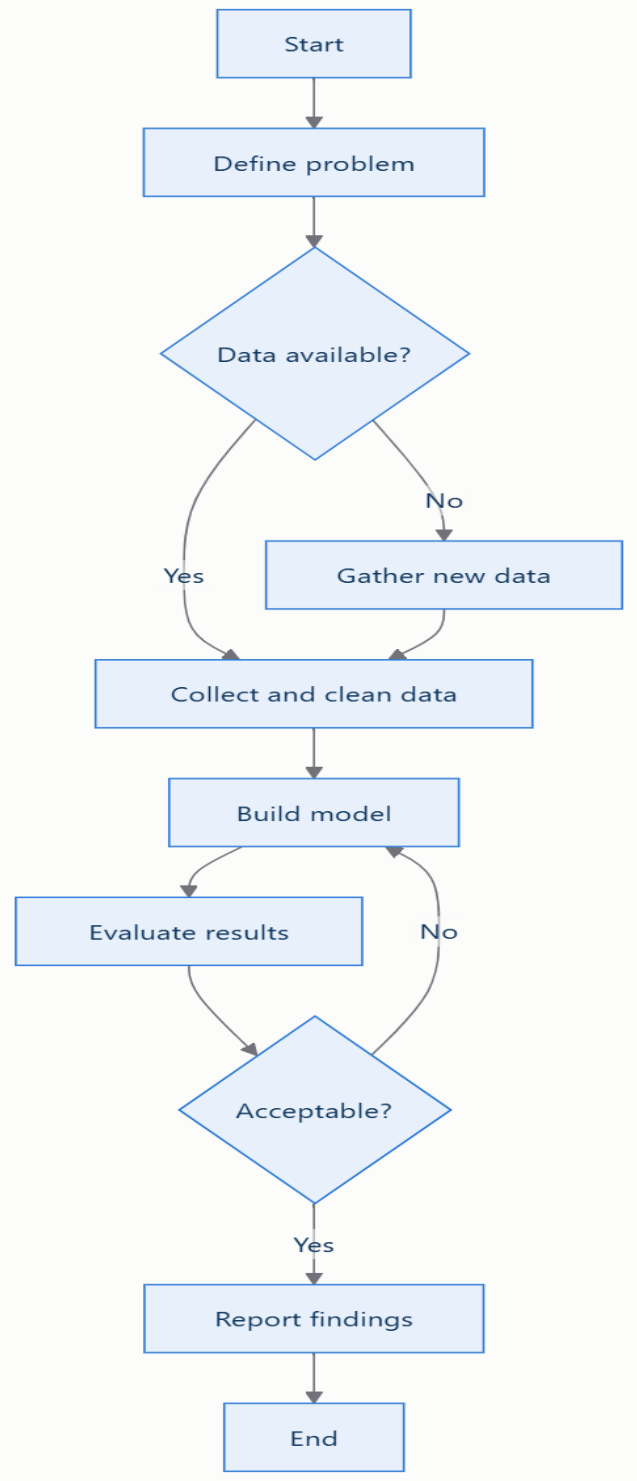
Regression models were then applied and the results compared using evaluation metrics. Finally the crime prediction and hotspot analysis were provided using the CrimeSense web application.

The main steps in this project were;

1. Collect historical crime data.
2. Clean and prepare data.
3. Analyse crime trends.
4. Select and prepare key features.
5. Apply regression models.
6. Evaluate and compare models.
7. Identify districts with higher crime rates.
8. Provide results using CrimeSense.

Figure 1: Overall CrimeSense system.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |



3.2Data Preprocessing and Feature Engineering

Data Preprocessing

Once the data had been collected, it was checked to ensure it was in the correct format. Some data had missing information, duplicates and varying district names. This was resolved before continuing with the analysis.

The following was undertaken;

1. Check data - Opened the dataset and checked the columns, data and values.
2. Missing data - Identified empty cells and filled in where needed.
3. Remove duplicates - Removed duplicate data so the data wasn't double counted.
4. Correct data types - Checked year, population, crime cases and crime rate columns and made corrections where needed.
5. District names - Corrected spelling of district names.
6. Check for outliers - Checked for very high/low values.
7. Select features - Selected the relevant columns, state, district, year, population and crime records. The target was crime rate.

Regression Problem

In this project crime rate prediction is a regression problem. Historical crime data and other related information is used to predict the crime rate.

The input data is represented using X and the actual crime rate is represented using Y. The model analyses the data to try and predict the crime rate as accurately as possible.

The prediction can be shown using;

\[\hat{Y} = f(X) \]

Where,

X = Input features (District, State, Year and Population).

Y = Actual crime rate.

$\hat{Y}$ = Predicted crime rate.

f = Trained Machine Learning model.

The data was split into training and testing data. The training data was used to train the model and the testing data was used to test the model.

As this project aims to predict future crime rates the data was split by year. The older data was used as training data and the newer data was used as testing data. This helps to understand how well the model can predict future crime rates.

3.3Machine Learning Algorithms

Four different regression models were used to predict crime rates. As each model works differently the results were compared to see which provided the best predictions.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Linear Regression - Used to understand the relationship between selected features and crime rates. It is also one of the simplest models to compare.
Decision Tree Regressor - Splits the data into smaller groups based on different conditions to make predictions. Helps identify patterns in the crime data.
Random Forest Regressor - Uses multiple decision trees and combines the results to make a prediction. This helps reduce errors that may occur from one tree.
Gradient Boosting Regressor - Creates decision trees one after the other. Each tree focuses on improving the prediction by using the errors from the previous tree.
All models were tested using the same evaluation metrics.

Experimental Setup

3.4 Model Training and Evaluation

At this stage all 4 Machine Learning models were trained and tested to see how well they could predict crime rates.
The data was split into two sections. The training data was used to train the models and the testing data was used to test the models. In Crime Sense earlier the data from 2016-2021 was used as training data and 2022-2023 as testing data. This should be checked in the final code for the project as it is split by year.
To compare all 4 models the following metrics were used;
1. Mean Absolute Error (MAE)
This measures the average difference between the actual crime rate and predicted crime rate. The lower the MAE the closer the predictions are to the actual values.
$$ MAE=\frac{1}{n}\sum |y_i-\hat{y}_i| $$
2. Root Mean Squared Error (RMSE)
This measures the differences between actual and predicted values. Larger errors are weighted more so a lower value is preferable.
$$ RMSE=\sqrt{\frac{1}{n}\sum(y_i-\hat{y}_i)^2} $$
3. R^2 Score
This indicates how well the model explains variations in the crime rate in the dataset. Generally the higher the R^2 score the better the model is performing on the testing data.
$$ R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar y)^2} $$
These models help to compare the models.
Implementation of CrimeSense
CrimeSense is a web application that provides crime data, predicts future crime rates and analyses crime trends in the different districts in India.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Python was used to clean the data and create the Machine Learning models. The trained model was then added to the application to predict crime rates.

Streamlit was used to create the application which makes it easier to view the crime data, predictions and graphs.

Key features of CrimeSense include;

View crime data from previous years.

Analyse crime trends in different districts.

Predict crime rates for future years.

Compare results from different Machine Learning models.

View districts with higher crime rates using rankings.

View crime hotspots on maps.

The project uses data on district boundaries to provide crime information on maps. District names must match those in the map data or some locations may not be in the correct place.

4.Results

4.1Dataset Analysis

An exploratory data analysis was carried out to better understand the crime data.

The data was analysed to see how crime rates varied between different districts and years. Missing data, duplicates and outliers were checked.

Bar charts, line graphs and district comparisons were used to make the data easier to understand.

The analysis included;

Crime data for different years.

Crime rate comparison between districts.

Crime rate distribution.

Crime trends in different states and districts.

This helps to understand the data available and select the key features needed to predict crime rates. Actual findings from the graphs are included based on the results from the dataset.

4.2Regression Model Results

All four Machine Learning models were trained using the ready data. The results were compared to see how accurate the models were at predicting crime rates.

The results from the previous experiment are shown below;

Algorithm MAE RMSE R^2 Score

Linear Regression 9.9176 13.7356 0.4896

Decision Tree Regressor 11.6082 Approx. 17.4 0.1797

Random Forest Regressor 10.0865 14.1221 0.4605

Gradient Boosting Regressor 9.9202 13.7322 0.4898

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

The results show that Linear Regression and Gradient Boosting both provided very similar results. Gradient Boosting had a slightly higher R^2 score while Linear Regression had very similar error values.
Random Forest also provided reasonable results but was slightly worse than the first two models. Decision Tree had the highest prediction error and lowest R^2 score.
Overall Gradient Boosting performed slightly better in this experiment.
Model Comparison
Visualisation of Predictions
Graphs and charts were used to make the results of the project easier to understand. This shows how much difference there is between the actual and predicted crime rates and how crime rates change over the years.
Line graphs were used to show crime trends over the years. Bar charts were used to compare crime rates between districts. Maps were used to show areas with higher crime rates.
All results were shown in CrimeSense to make it easy to view and understand.

5.Discussion

5.1Interpretation of Results

All four Machine Learning models provided different results when tested.
Gradient Boosting provided slightly better results with an R^2 score of 0.4898. Linear Regression provided very similar results with an R^2 score of 0.4896. Random Forest also provided similar results. Decision Tree provided lower results.
The results from Gradient Boosting and Linear Regression were very similar, showing that using a more advanced model does not always provide much better results.
From this experiment it can be understood that all models should be tested using the available data.

5.2Implications for Crime Analysis

CrimeSense helps to analyse historical crime data and predict future crime rates using Machine Learning. It also allows users to compare crime rates between different districts and identify areas with more crimes reported.
The graphs, maps and prediction results make it easier to understand crime trends. This information can be useful to students, researchers and anyone interested in analysing crime. The predictions are based on historical data and so may not always match real life.

5.3Limitations:

Despite providing an efficient analysis of the crime data, the following limitations could happen:
1. Dataset: There could be missing or incorrect data within the dataset.
2. Reporting: Crime data may not be reported in the same way in all districts.
3. Limited Data: Not all the factors affecting the crime rate are shown in the dataset.

4. Maps: There could be differences in the spelling and boundaries of the districts causing issues when plotting locations on maps.
5. Model: The models do not fully explain the variations in the crime rate meaning there is room for improvement.
6. Changes: Crime rates may change due to social, economic and other factors that aren't within the dataset.
This should be considered when analysing the results and predictions.

6.Conclusion

6.1Summary:

This project looks at analysing crime data from different districts in India and predicting future crime rates. Historical crime data has been used from 2014 to 2023.
Four different models have been analysed; Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. The dataset has been cleaned and analysed before training the models. The models were compared to see which performed best using MAE, RMSE and R2 Score.
Gradient Boosting and Linear Regression both performed very similarly whereas Random Forest and Decision Tree produced different results. This helped to understand how each model performed with the data.
CrimeSense also provides a way to analyse crime data on a district by district basis, predict crime rates and visualise hotspots. This has all been combined into one web application using Python and Streamlit.
This project has demonstrated how machine learning can be used to analyse historical crime data to gain an idea of future crime trends. Future improvements could be made by using more detailed data, adding more useful features and testing other machine learning models.

6.2Future Scope:

There are many ways in which CrimeSense can be improved.
1. Dataset: More up to date and detailed crime data could be added to provide better results.
2. Features: More data could be added such as population, education, employment and other social factors to gain more insight into crime patterns.
3. Algorithms: Other machine learning models and time series models could be tested to improve the predictions.
4. Model: Different model parameters and feature selection methods could be tested to improve the predictions.
5. Maps: Up to date district maps and correct location data could be used to improve the accuracy of identifying hotspots.

6. Predictions: New features could be added to show which factors have the most impact on the predictions.
7. Application: More filters, up to date crime data and interactive graphs could be added to make the CrimeSense application more user-friendly.

References:

[1] S. V. Tikore, K. Chaudhari, O. Rohamare, A. Nikam, K. Kalne, "A Comprehensive Computational Framework for Crime Rate Prediction Using Machine Learning in Indian Metropolitan Cities", Journal of Smart Sensors and Computing, 2026.
[2] K. Jenga, C. Catal, G. Kar, "Machine learning in crime prediction", Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 2887–2913, 2023.
[3] N. Shah, N. Bhagat, M. Shah, "Crime Forecasting: A Machine Learning and the Computer Vision Approach the to Crime Prediction and Prevention", Visual Computing for Industry, Biomedicine, and Art, vol. 4, 2021.
[4] M. A. Shafee, N. A. Shelke, R. Agrawal, C. K. Choubey, "Crime Prediction Before, During and After the COVID-19 by Using Machine Learning and RNN-LSTM Models", Discover Artificial Intelligence, 2026.
[5] National Crime Records Bureau, Crime in India, Ministry of Home Affairs, Government of India. [Online]. Available: https://ncrb.gov.in/