# Crime Prediction Using Machine Learning

**Statistical Machine Learning CSET211**

A Project Report Submitted by: -

ANSHUMAN MATHUR

ARYAN SINGH RAJPOOT

AKSHAT PANDEY

Submitted to

**DR. NITIN ARVIND SHELKE (Dr. Sanchali Das)**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr.No | Content | Page No |
|  |  |  |
|  |  |  |

* **Table of Contents**

1. Introduction
2. Objectives of the Project
3. Dataset Overview
4. Data Preprocessing
5. Exploratory Data Analysis
6. Methodology
   * Visualization
   * Model Creation
     + Forecasting Crime Trends
     + Clustering Hotspots
     + Classification Model for High-Risk Areas
7. Results
8. Conclusion
9. References

**1. Introduction**

* Crime is a significant social issue that affects millions globally. Understanding historical crime patterns can aid in better allocation of law enforcement resources and improve public safety. Machine learning provides tools to predict potential future crime trends and identify high-risk regions. This project leverages historical crime data to create predictive models using visualization, clustering, and classification techniques.
* **2. Objectives of the Project**
* To analyze historical crime data for identifying patterns and trends.
* To develop machine learning models for predicting high-risk areas.
* To assist law enforcement agencies in resource allocation.
* To create community awareness and improve public safety.
* **3. Dataset Overview**

The dataset includes records of various crimes against women in India from 2001 to 2014. Key attributes in the dataset:

* **STATE/UT**: Name of the state or union territory.
* **Year**: Year of the recorded incident.
* **Crime Types**: Includes Rape, Kidnapping and Abduction, Dowry Deaths, etc.
* **Data Source**: Publicly available government records and databases.
* **Summary of the dataset**:
* Total records: 676
* Features: State, Year, and 9 types of crimes.
* **4. Data Preprocessing**

Preprocessing ensures the dataset is clean, consistent, and ready for analysis.

* **Handling Missing Values**: Missing values were replaced with 0.
* **Column Adjustments**: Renamed columns for better readability (e.g., "Kidnapping and Abduction" renamed to "Kidnapping Abduction").
* **Unnecessary Columns Removed**: Dropped unnamed index columns.
* **5. Exploratory Data Analysis**

EDA was conducted to understand crime patterns and trends across states and years.

1. **Trend Analysis**:
   * Visualized average rape cases per state over the years using line plots.
   * Key Finding: A gradual increase in average rape cases was observed.
2. **Crime Distribution**:
   * Bar plot showcasing total cases for each crime type.
   * Finding: Rape, Kidnapping, and Dowry Deaths were the most reported crimes.
3. **Heatmap Analysis**:
   * Heatmap highlighting rape cases per state and year.
   * Finding: Certain states consistently reported higher cases, making them high-risk regions.

* **6. Methodology**
* **6.1 Visualization**

Data was visualized using Matplotlib and Seaborn to uncover insights.

* **Line Plot**: Showed trends in average rape cases over years.
* **Bar Chart**: Illustrated total cases by crime type.
* **Heatmap**: Highlighted state-wise crime distribution across years.
* **6.2 Model Creation**
* **6.2.1 Forecasting Crime Trends**
* **Objective**: Predict future trends for rape cases in Andhra Pradesh.
* **Algorithm**: ARIMA (Auto-Regressive Integrated Moving Average).
* **Results**:
  + Predicted crime counts for the next 5 years.
  + Insight: Crime rates show a consistent trend over the years.
* **6.2.2 Identifying Crime Hotspots**
* **Objective**: Cluster regions based on crime rates.
* **Algorithm**: K-Means Clustering.
* **Features Used**: Rape, Kidnapping, Dowry Deaths.
* **Results**:
  + Data divided into 5 clusters representing different crime levels.
  + High-risk regions were identified visually.
* **6.2.3 Classification Model for High-Risk Areas**
* **Objective**: Classify regions into high-risk and low-risk categories.
* **Algorithm**: Random Forest Classifier.
* **Features Used**: Rape, Kidnapping, Dowry Deaths.
* **Threshold**: Regions with more than 100 rape cases labeled as high-risk.
* **Model Evaluation**:
  + Accuracy: 87%
  + Metrics: Precision, Recall, F1-Score.
* **7. Results**

1. **Trend Analysis**:
   * Crime trends are increasing over time for most states.
2. **Clustering Results**:
   * High-risk regions successfully identified.
3. **Prediction Results**:
   * Forecasted crime numbers and identified areas requiring urgent attention.
4. **Classification Model**:
   * Model achieved high accuracy and successfully categorized high-risk areas.

* **8. Conclusion**

This project demonstrates how machine learning can be effectively used for crime analysis and prediction. By leveraging historical data:

* Forecasted future trends to prepare law enforcement for potential spikes.
* Identified high-risk areas to optimize resource allocation.
* Created actionable insights to enhance public safety.
* **9. References**
* Government crime records.
* Documentation for ARIMA and Random Forest in Python.
* Research articles on crime prediction using machine learning.

**GitHub Link of Your Complete Project**

https://github.com/ANSHUUMAAN5/Crime-prediction-project