

Crime Pattern Analysis and Prediction Using Machine Learning: A Case Study of Mumbai

Author: Sakshi Shashikant Bhosale (2705) and Darpan Gowari (2711)

Affiliation: Department of Information Technology, Pillai College, Panvel

Abstract

This study analyses real reported Indian Penal Code (IPC) crime statistics for Mumbai for 2022 and 2023 using the “Mumbai – Crime Data 2023” dataset published by OpenCity, with Mumbai Police listed as the source. The dataset provides registered cases, detected cases, and detection percentages by crime category. Total registered IPC cases decreased from 70,367 in 2022 to 45,869 in 2023 (-34.8%). The source-reported detection rate rose from 68% to 76%. This is a descriptive analysis of real data; no machine-learning prediction model is claimed because the dataset contains only two annual snapshots and no incident-level records.

1. Introduction

Background:

Mumbai is a highly populated and economically active metropolitan region. This study compares published Mumbai IPC crime statistics for 2022 and 2023. The data supports comparisons of registered cases, detected cases, and detection rates by crime category, but it does not include the date, time, or location of individual incidents.

Problems:

- Registered IPC crime totals differed between 2022 and 2023.
- Crime categories contributed differently to the registered-case totals.
- Detected-case counts and the source-reported detection rate changed differently across the two years.
- The dataset does not contain incident-level, monthly, or area-level information for detailed spatial or temporal prediction.
- The findings describe recorded statistics and do not establish the causes of change or predict future crime.

Objectives:

- To compare Mumbai’s registered IPC cases in 2022 and 2023.
- To compare detected cases and source-reported detection percentages.
- To identify the largest registered crime categories in 2023.
- To calculate absolute and percentage changes using the published values.
- To explain the limitations and data needs of future machine-learning prediction.

2. Literature Review

- Crime-pattern analysis often examines crime type and frequency to describe differences across categories and time periods.
- Aggregated category-level data supports annual comparison but cannot reveal incident-level or neighbourhood-level patterns.
- Classification requires labelled observations and a clearly defined target.
- Models such as Random Forest may be considered if a sufficiently large, structured dataset becomes available.
- Model quality depends on data consistency, validation, and careful interpretation of reporting differences.

This paper analyses a real published Mumbai crime dataset. Machine-learning methods are discussed as future work only; no model was trained because the current file has only two annual snapshots and lacks incident-level predictors.

3. Methodology

3.1 Study Area

The study area is Mumbai, Maharashtra. The dataset provides IPC crime-category figures for 2022 and 2023, without police-station, ward, or individual incident details.

3.2 Proposed Data Requirements

- Crime category or crime head.
- Registered case counts for 2022 and 2023.
- Detected case counts for 2022 and 2023.

- Source-reported detection percentages for 2022 and 2023.
- Derived absolute and percentage changes between years.
- For future ML work: a larger, consistent monthly or incident-level dataset with time/location fields.

The present analysis uses only fields available in the published CSV. Month, weekday, time-of-day, and location values are not inferred.

3.3 Data Preprocessing

- Convert registered and detected counts to numeric values.
- Preserve the published crime-category labels and source percentages.
- Calculate absolute change as 2023 value minus 2022 value.
- Calculate percentage change as $(2023 - 2022) \div 2022 \times 100$ when the 2022 baseline is non-zero.
- Keep the total IPC row separate from individual crime categories.
- Use only the same IPC CSV throughout this analysis.

3.4 Machine Learning Techniques

- Descriptive comparison: performed using the published 2022 and 2023 values.
- Clustering/classification: not performed because this dataset lacks sufficient observations and target labels.
- Time-series forecasting: not performed because two annual values are insufficient for reliable forecasting.
- Future evaluation should use more observations, an appropriate holdout strategy, and metrics suited to the prediction task.

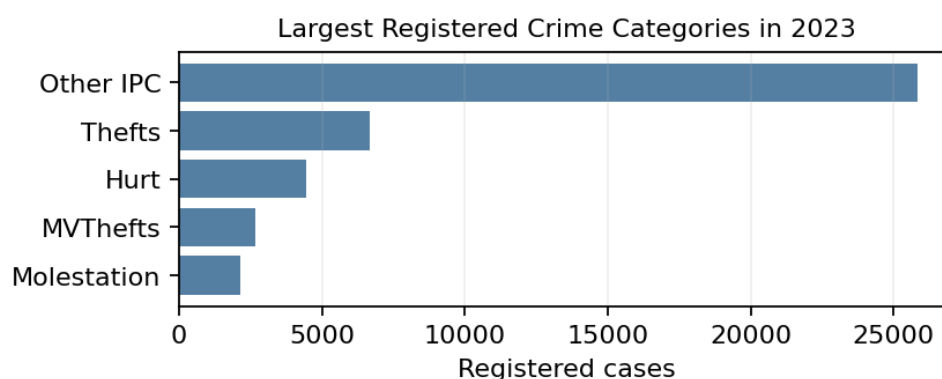
3.5 Proposed Workflow

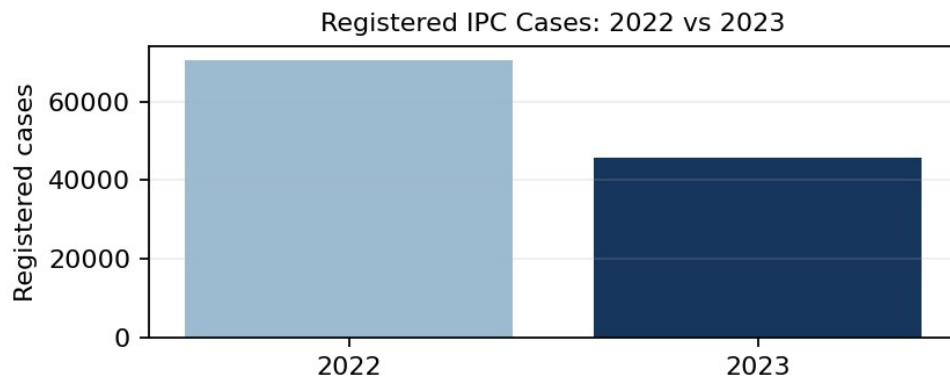
Stage	Activity
1	Download the published Mumbai IPC CSV.
2	Check fields and convert counts to numeric values.
3	Compare registered and detected cases for 2022 and 2023.
4	Calculate absolute and percentage changes.
5	Rank registered crime categories for 2023.
6	Create charts and review descriptive findings.
7	Interpret results cautiously and document limitations.
8	Identify additional data needed for spatial/temporal analysis.
9	Do not claim ML prediction until enough data is available and evaluated.

4. Case Study Analysis

4.1 Crime Pattern Analysis

The published dataset reports 70,367 registered IPC cases in 2022 and 45,869 in 2023, a change of -24,498 cases (-34.8%). Detected cases changed from 47,866 in 2022 to 34,991 in 2023. These figures describe recorded annual statistics and do not establish why the values changed.





4.2 Spatial Pattern Analysis

Spatial analysis cannot be performed from this dataset because it contains no police-station, ward, zone, or incident-coordinate fields. A future spatial study would require consistent location-level records.

4.3 Temporal Pattern Analysis

Temporal analysis is limited to comparing annual values for 2022 and 2023. The CSV does not contain monthly, weekday, or time-of-day records, so seasonal patterns cannot be measured.

4.4 Machine Learning Prediction

A machine-learning prediction model has not been trained or evaluated. The dataset has only two annual summary values per crime category and lacks detailed time/location predictors. Reliable prediction requires a larger time-indexed dataset and a defined target.

5. Proposed Model Evaluation

No model accuracy is reported because no machine-learning model was trained or tested. A future study should define its target, use sufficient historical observations, evaluate on unseen data, and report suitable metrics.

6. Expected Analytical Outputs

Actual outputs include year-to-year comparisons of registered cases, detected cases and source-reported detection rates, plus a ranking of 2023 registered crime categories. Spatial maps, monthly patterns, clustering and prediction results are not available from this dataset.

Analysis	Actual output
Crime category analysis	2023 registered cases compared by category
Annual comparison	70,367 in 2022; 45,869 in 2023 (-34.8% change)
Detected cases	47,866 in 2022; 34,991 in 2023
Detection rate	68% in 2022; 76% in 2023
Spatial analysis	Not possible: no location fields in this CSV
Temporal analysis	Annual comparison only; no monthly/time-of-day data

7. Proposed Applications

- Support deployment planning by highlighting recurring high-activity areas or periods.
- Support patrol planning and preventive measures through historical pattern analysis.
- Help identify changes in crime patterns that require further investigation.
- Provide dashboards or maps for easier interpretation of analytical findings.
- Support decision-making without replacing police investigation or human judgment.

8. Limitations

- The dataset contains annual aggregate figures for 2022 and 2023 only.
- It does not provide individual incident dates, time of day, police-station/ward location, or case-level information.
- Changes in registered cases do not establish causes or prove changes in underlying crime.
- No machine-learning prediction or accuracy score is claimed.
- Future crime analytics should consider reporting bias, privacy, fairness, and responsible use.

9. Conclusion

This study compares real published Mumbai IPC crime figures for 2022 and 2023. Registered cases fell from 70,367 to 45,869 (-34.8%), while the source-reported detection rate rose from 68% to 76%. These are descriptive findings. A valid machine-learning prediction study would require more detailed time-indexed data and model evaluation.

References

- OpenCity. “Mumbai – Crime Data 2023,” resource: “IPC Crimes for 2022 and 2023.” <https://data.opencity.in/dataset/mumbai-crime-data-2023/resource/18e13d1a-678e-4caa-96e5-2b585826f1f1>
- OpenCity. “Mumbai – Crime Data 2023” dataset page. <https://data.opencity.in/dataset/mumbai-crime-data-2023>
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.