A Data-Driven Analysis of Female Labor Force Participation and Reported Crimes Against Women in India (2001–2022)

Aditi Vishwakarma
Independent Researcher

Abstract

This study analyzes statistical associations between female labor force participation (FLFP) and reported crimes against women in India using publicly available national datasets from 2001 to 2022. Socio-economic datasets are affected by reporting practices, underreporting, and survey measurement limitations, making statistical relationships difficult to interpret causally.

After preprocessing and temporal alignment, the analysis applies correlation analysis, Ordinary Least Squares (OLS) regression, and regression-based machine learning models including Random Forest and Gradient Boosting. Model performance is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) with a time-aware train–test split.

The results show a consistent statistical association between FLFP and reported crime rates at the aggregated national level; however, predictive performance varies across years and is sensitive to dataset limitations. The findings demonstrate that predictive relationships in socio-economic datasets should not be interpreted as causal and highlight challenges of applying machine learning to socially sensitive observational data.

Overall, the study demonstrates how computational modeling can support exploratory analysis while requiring cautious interpretation in socially sensitive domains.

The analysis also highlights challenges of applying predictive models to observational social data, where patterns may reflect data generation processes rather than real-world mechanisms.

Keywords: machine learning; observational data; regression analysis; causal inference; gender statistics; India

1. Introduction

The increasing availability of large public datasets has expanded opportunities for computational analysis of human behavior. In machine learning, large datasets have often been found to improve predictive performance more than increases in model complexity (Halevy et al., 2009). As a result, statistical relationships identified by machine learning models may reflect characteristics of data collection rather than underlying real-world phenomena.

Crime statistics and female labor force participation rates represent particularly complex examples of such data. Reported crime counts depend not only on the occurrence of incidents but also on awareness, legal accessibility, social stigma, and reporting behavior. Such measurement and reporting effects can introduce systematic bias into machine learning analyses of social data (Mavrogiorgos et al., 2024). Similarly, labor force participation estimates are influenced by informal employment, survey methodology, and economic structure. These factors make the datasets useful for identifying broad patterns while limiting straightforward predictive interpretation.

This study examines these challenges through an applied computational analysis of two national-level datasets in India from 2001 to 2022: reported crimes against women and female labor force participation (FLFP). Rather than attempting to establish causality, the objective is to examine how statistical and machine learning methods behave when applied to socially sensitive observational data. This distinction between correlation and causation is a central issue in observational data analysis (Pearl, 2009). In particular, the work focuses on distinguishing correlation from causation and on understanding how measurement and reporting processes influence model outputs.

Using correlation analysis, Ordinary Least Squares regression, and regression-based machine learning models, the study evaluates observed relationships while accounting for dataset limitations. The findings highlight both the usefulness and the constraints of predictive modeling in real-world social data and emphasize the need for cautious interpretation when applying AI methods outside controlled environments. The study therefore serves as a methodological case study illustrating how statistical and machine learning techniques should be interpreted when applied to real-world social data.

2. Related Work

The use of computational methods for analyzing large-scale social datasets has increased substantially in recent years (Halevy et al., 2009). Prior research has shown that predictive models trained on observational data can identify statistical patterns; however, interpretation is often complicated by measurement error, reporting bias, and confounding variables (Shmueli, 2010). Unlike controlled experimental datasets, social indicators are generated through institutional and behavioral processes that influence the values being recorded.

In the context of crime statistics, several studies have emphasized that reported crime counts do not directly measure underlying incidents. Instead, they depend on reporting behavior, legal accessibility, public awareness, (Biderman and Reiss, 1967) and administrative capacity. Consequently, statistical associations observed in crime datasets may reflect changes in reporting practices rather than changes in actual crime occurrence. Similar concerns apply to labor force participation statistics, which may exclude informal or unpaid work and therefore only partially represent economic activity, particularly in developing economies.

Recent work in machine learning has also examined how predictive models behave under dataset shift and imperfect measurement (Quionero-Candela et al., 2009). Models may achieve strong predictive performance while learning proxies or artifacts present in the data rather than meaningful relationships. This has led to increased emphasis on interpretability, evaluation methodology, and the distinction between correlation and causation in applied machine learning.

The present study follows this perspective but focuses specifically on national-level gender-related indicators in India. Rather than proposing a new predictive algorithm, the study evaluates how commonly used statistical and regression-based machine learning models behave when applied to aggregated observational social data. The goal is to understand the reliability and limitations of model outputs and to highlight the challenges of drawing substantive conclusions from predictive relationships in socio-economic datasets.

3. Dataset Description

This study uses two publicly available national-level datasets for India covering the period 2001–2022. The first dataset consists of annual reported crime statistics against women obtained from the National Crime Records Bureau (National Crime Records Bureau, Government of India, various years). The records include aggregated counts of registered cases across multiple legal categories (e.g., cruelty by husband or relatives, assault on women with intent to outrage modesty, kidnapping and abduction, and rape). The data are reported annually at national level and represent officially recorded incidents.

The second dataset contains annual female labor force participation rate (FLFP) estimates obtained from the World Bank database based on International Labour Organization (ILO) modeled estimates (World Bank, 2023; International Labour Organization, 2022). The indicator measures the percentage of women aged 15 and above who are economically active, including both employed individuals and those actively seeking employment.

Both datasets were selected because they are consistently available across the full study period and are frequently used in socio-economic analysis. Since the data originate from administrative reporting systems and surveys, they represent aggregated population-level measurements rather than individual-level observations.

Before analysis, the datasets were aligned by year to ensure temporal consistency. Missing entries were checked and, where necessary, handled through removal of incomplete observations to maintain comparability across variables. The final dataset therefore consists of matched annual observations of reported crimes against women and female labor force participation rate for the period 2001–2022.

4. Methodology

This study analyzes the relationship between female labor force participation and reported crimes against women in India using statistical and machine learning approaches. The objective is not to establish causality, but to evaluate associations and predictive behavior in observational socio-economic data.

4.1 Variables

Let Yt denote the reported number of crimes against women in year $t1$ and Xt denote the female labor force participation rate (FLFP) in the same year.

The dataset consists of annual observations from 2001 to 2022, forming a time-indexed sequence.

The analysis treats crime counts as the dependent variable and FLFP as the explanatory variable:

$$Yt = f(Xt) + \epsilon t$$

Where ϵt represents unexplained variation due to measurement limitations, reporting behavior, and omitted socio-economic factors.

4.2 Data Preprocessing

Before analysis, the datasets were aligned temporally to ensure comparable yearly observations. Preprocessing included:

- handling missing observations through removal or interpolation where necessary
- standardizing year formats and variable labels
- aggregating crime categories into a single annual total
- aligning FLFP estimates with corresponding calendar years

No feature engineering or artificial transformation was introduced in order to preserve interpretability of results.

4.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand distributional properties and temporal behavior. Descriptive statistics and visualizations were used to examine trends over time and to identify potential non-stationarity or structural shifts in the series.

Pearson correlation was used to measure linear association between FLFP and crime counts :

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Spearman rank correlation was additionally computed to assess monotonic relationships less sensitive to outliers.

4.4 Statistical Modeling

An Ordinary Least Squares (OLS) regression model was used to estimate the association between the variables:

$$Yt = \beta 0 + \beta 1\ Xt + \epsilon t$$

where β0 is the intercept and β1 measures the change in reported crime associated with a one-unit change in FLFP.

Residual analysis was performed to evaluate model fit and to identify systematic deviations.

4.5 Machine Learning Models

To evaluate predictive behavior, regression-based machine learning models were applied:

- Random Forest Regressor
- Gradient Boosting Regressor

These models were selected because they can capture nonlinear relationships and interaction effects without assuming a predefined functional form.

A time-aware train-test split was used to avoid data leakage. Earlier years (training set) were used to predict later years (testing set), preserving chronological order.

4.6 Model Evaluation

Model performance was evaluated using:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum |Y_t - \hat{Y}_t|$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum (Y_t - \hat{Y}_t)^2}$$

MAE measures average prediction error, while RMSE penalizes larger deviations more heavily. Both metrics were used to compare statistical and machine learning approaches.

4.7 Interpretation Framework

Because the datasets are observational and affected by reporting behavior, model outputs were interpreted as associations rather than causal relationships. The analysis focuses on pattern detection and predictive stability rather than policy inference.

5. Experimental Results

The correlation analysis indicated a consistent association between female labor force participation and reported crimes across the observed period. OLS regression produced a non-zero coefficient for FLFP, indicating that the variables were not independent in the aggregated dataset.

Machine learning models demonstrated moderate predictive performance on the held-out time period. Both Random Forest and Gradient Boosting were able to capture general trend behavior but showed sensitivity to year-to-year fluctuations. Feature importance analysis indicated that labor participation contributed to predictions but did not fully explain variation in reported crimes (Hastie, Tibshirani, & Friedman, 2009).

These results suggest that while predictive patterns exist, substantial unexplained variability remains, suggest that additional unobserved factors may influence the recorded crime statistics.

Due to the relatively small number of yearly observations, machine learning models were used primarily for pattern exploration rather than real-world predictive deployment.

6. Bias, Ethical Considerations, and Limitations

The datasets used consist of aggregated public statistics and do not contain individual-level data. However, both variables are affected by measurement and reporting processes. Crime statistics depend on reporting behavior, legal accessibility, and institutional capacity, while labor participation estimates may exclude informal and unpaid work. Therefore, statistical relationships observed in this study should not be interpreted as causal explanations of gender inequality.

Another limitation arises from contextual bias. Quantitative data can highlight disparities in areas such as education, employment, or income, but it often fails to capture the cultural, social, and structural factors that contribute to these inequalities. Machine learning models, when applied without sufficient contextual understanding, risk oversimplifying complex social realities into numerical trends. This reinforces the idea that Machine learning models applied to observational data do not model social mechanisms directly; they capture statistical regularities present in the training data.

The analysis also emphasized the risk of misinterpretation when predictive or analytical models are applied to social data. Patterns identified through data analysis may suggest correlations, but they do not establish causation. Without careful interpretation, there is a possibility that AI-driven insights could be misused to justify existing disparities rather than challenge them (Barocas et al., 2019). This highlights an important ethical responsibility for practitioners to ensure that AI outputs are not treated as objective truths, especially in domains involving human rights and social equity.

Overall, the study highlights that the value of AI in socially sensitive contexts depends not only on technical accuracy but also on careful interpretation and ethical awareness. Machine learning models applied to socio-economic data should therefore be treated as analytical support tools rather than definitive decision systems. Performance observed in controlled or curated datasets may not generalize to real-world settings (Sculley et al., 2015).

7. Reproducibility and Data Availability

All datasets used in this study are publicly available. Crime statistics were obtained from the National Crime Records Bureau (NCRB), Government of India, and female labor force participation rates were obtained from the World Bank (modeled ILO estimates).

All data preprocessing, statistical analysis, and machine learning experiments were implemented in Python using pandas, numpy, statsmodels, and scikit-learn (Pedregosa et al., 2011). The analysis includes data cleaning, temporal alignment, correlation analysis, regression modeling, and evaluation using time-aware train–test splits.

To support transparency and reproducibility, the project structure separates raw data, processed data, and analysis scripts. The code used to generate results, figures, and model evaluations can be executed end-to-end to reproduce the findings described in this paper.

The full analysis code and processed datasets are publicly available at:
https://github.com/AditiVishwakarma01/gender-inequality-research

8. Conclusion

This study examined the relationship between female labor force participation and reported crimes against women in India from 2001 to 2022 using statistical and machine learning methods. Correlation analysis and OLS regression identified a consistent association in the aggregated national dataset, while machine learning models captured general temporal trends but exhibited sensitivity to year-to-year variation.

More broadly, the study illustrates a key limitation of applying machine learning to socially generated data. Predictive accuracy does not necessarily indicate meaningful understanding of real-world phenomena. Models may capture patterns produced by reporting practices, institutional processes, or measurement artifacts rather than underlying social mechanisms.

The results demonstrate that predictive patterns can be detected in large socio-economic datasets, but they are strongly affected by measurement processes, reporting behavior, and missing contextual variables. Consequently, statistical significance and predictive performance alone are insufficient to support causal interpretation.

The analysis highlights a broader methodological implication: when machine learning techniques are applied to observational social data, evaluation must consider data generation mechanisms and dataset limitations, not only model accuracy. Future work may incorporate additional socio-economic indicators and regional-level data to better understand variability in reported crime patterns. The study therefore emphasizes the importance of combining computational methods with domain knowledge and careful interpretation when applying artificial intelligence techniques to public policy and social datasets.

9. References

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8–12.

Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., & Kyriazis, D. (2024). Bias in machine learning: A literature review. Applied Sciences, 14(19), 8860. https://doi.org/10.3390/app14198860

National Crime Records Bureau. (2022). *Crime in India Statistics*. Ministry of Home Affairs, Government of India. https://ncrb.gov.in

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems (NeurIPS).

Vishwakarma, A. (2025). Gender Inequality Research: Reproducible research project accompanying the paper: A Data-Driven Analysis of Female Labor Force Participation and Reported Crimes Against Women in India (2001–2022).
GitHub repository - https://github.com/AditiVishwakarma01/gender-inequality-research
Accessed: February 2026.

World Bank. (2023). Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate). World Development Indicators. https://data.worldbank.org