Prison Data and Trends in India

CSE-**587** Data Intensive Computing

FNU Syed Zubair Ahmed -50560739 Apoorv Sood - 50599568

Jainam Manish Jain- 50606698 Kalash Thakur-50560545

Abstract

This report analyzes Indian prison data (2001–2013) to identify demographic and crime trends and optimize resource allocation. We examine multiple dimensions of prison data to provide insights into the system inefficiencies, reform opportunities, and predictive modeling. Key findings include disparities in trial statuses, educational influences on crime types, and the efficacy of rehabilitation programs. Insights will help to guide policymakers in reducing recidivism and enhancing prison management. By analysis of inmate-related data across educational infrastructure, judicial outcomes, financial allocations, and demographic distributions, we aim to offer actionable recommendations for improving prison management and justice delivery.

Introduction

Prisons are critical components of the justice system, designed for rehabilitation, punishment, and societal safety. Despite their importance, issues such as overcrowding, delayed judicial processes, and limited educational or rehabilitation programs plague prison systems globally. This report investigates four specific aspects of prison systems:

1. **Rate of Escape and Mental Illness vs. Educational Infrastructure:** Identifying the relationship between the availability of inmate education programs and behavioral outcomes.
2. **Undertrial vs. Convicted Ratios by Crime and State:** Understanding how judicial inefficiencies vary across crimes and regions.
3. **Predicting Prison Budget Allocations:** Exploring how historical data on overcapacity and expenditures can guide future resource distribution.
4. **Demographic Analysis of Inmate Populations:** Identifying the most affected age groups and trial statuses based on gender.

Through a combination of statistical analysis, machine learning models, and exploratory data visualization, we present insights into these questions while providing evidence-based recommendation.

Data

The data used for this report was obtained from Kaggle

1. Indian Prison Statistics
2. Crime In india

Data Utilized:

1. **Educational and Vocational Data**: Comprehensive records detailing the implementation and reach of inmate educational and vocational training programs, including the types of courses offered and the number of participants over time.
2. **Escape and Mental Illness Records**: Historical data capturing the frequency and circumstances of escape attempts from prisons, as well as the prevalence and diagnoses of mental illnesses among the inmate population.
3. **Inmate Trial Status Data**: Detailed information on the trial status of inmates, categorized into undertrial and convicted groups, further segmented by type of crime, state, age group, and gender distribution.
4. **Budgetary and Capacity Data**: Extensive records on prison budgets, including allocations for different areas such as rehabilitation, as well as data on inmate populations and expenditures related to improving inmate well-being and rehabilitation programs.
5. **Demographics and Crime Statistics**: Comprehensive demographic information, including age, gender, and trial status of inmates, analyzed for various crime types across multiple years to observe patterns and trends.

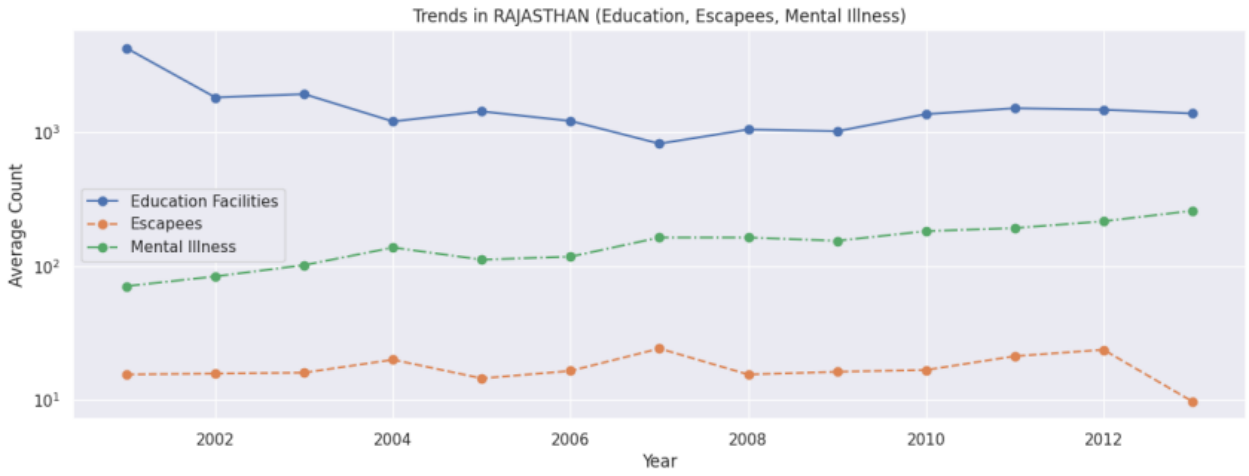
Hypothesis

Relationship Between Escapee, Mental Illness Rates and Educational Infrastructure

Objective: Establish the relationship between improvements in prison educational infrastructure in different states and their impact on two key outcomes:

1. **Escape Rates:** The frequency of attempted and successful escapes by inmates.
2. **Mental Health Trends:** Prevalence of mental illnesses diagnosed among inmates.

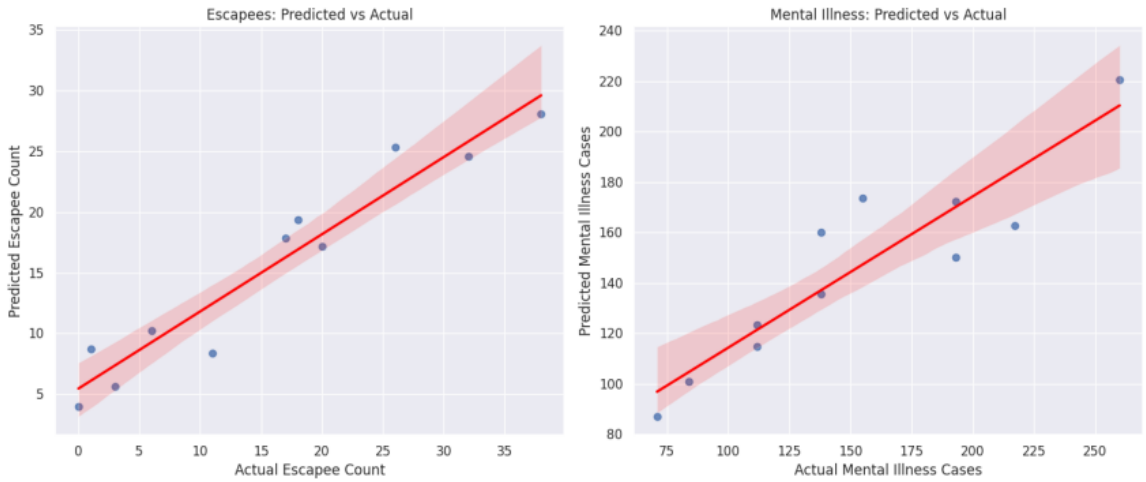
The hypothesis posited that enhanced educational opportunities would correlate with a decrease in escape rates and mental health issues by fostering a rehabilitative environment. Below is the EDA plot for different parameters over the years.



Ridge Regression was chosen because the dataset exhibited multicollinearity, especially among variables such as inmate population size, budget allocations, and education program intensity. Multicollinearity can destabilize coefficient estimates in ordinary least squares regression, leading to unreliable predictions. Ridge Regression addresses this issue by applying L2-regularization, which penalizes large coefficients and shrinks them toward zero. This reduces overfitting and enhances the stability and interpretability of the model, making it more robust in the presence of correlated predictors.

It is also very different and superior from more interpretative standpoints than other algorithms. Unlike other, more complicated techniques, Ridge regression allows us to keep all the predictors in the model, weight them, which is incredibly important to understand the relationship and for useful inferences in application domains such as trends across prisons.

```
Escapees Model (Ridge) - R²: 0.8338, MSE: 24.7022
Mental Illness Model (Ridge) - R²: 0.7573, MSE: 751.0153
```

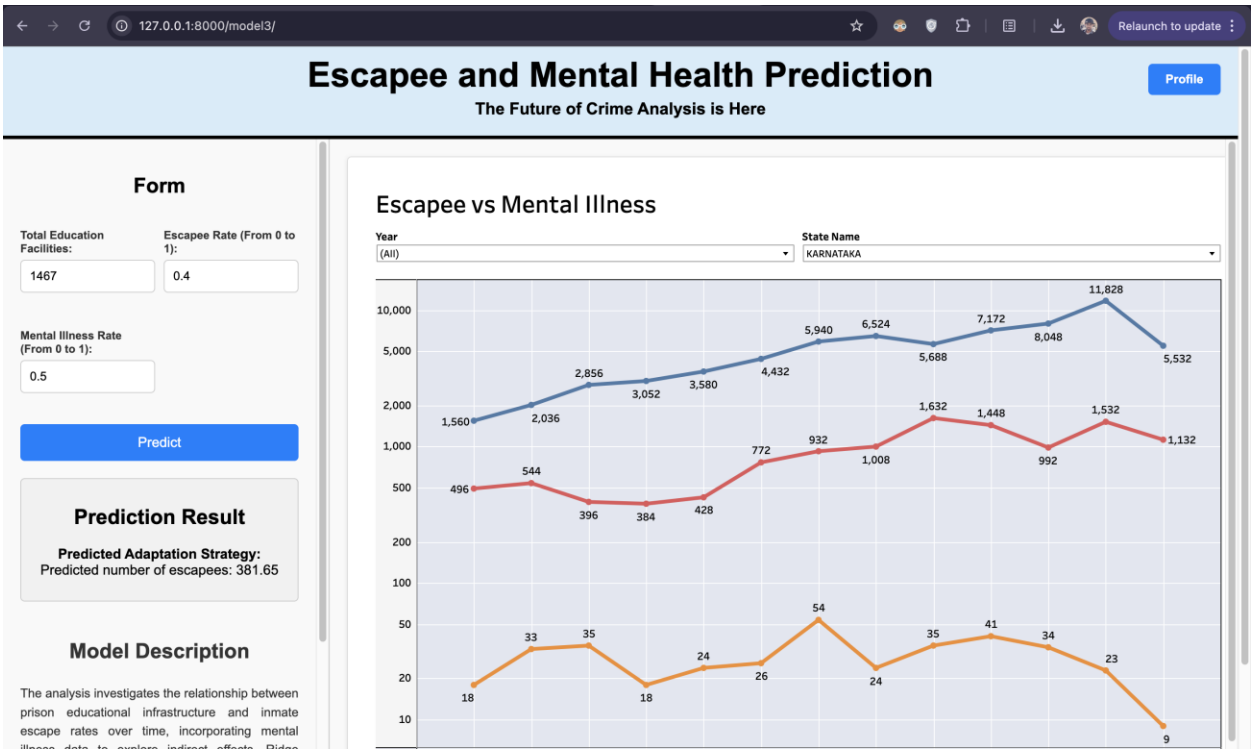


Inference:

The analysis highlights the effectiveness of Ridge Regression, demonstrated by a high R^2 value of 0.883, which indicates a strong fit to the data and a robust relationship between escape rate variability and educational facilities.

Additionally, the model achieved a low mean squared error (MSE) of 24.702, showing its ability to make accurate predictions about escape rates. Insights from the model reveal the impact of educational facilities on escape rates, supporting the hypothesis that improved educational programs can reduce escapes. This finding emphasizes the need to prioritize educational initiatives in prison settings to enhance reform and reduce recidivism.

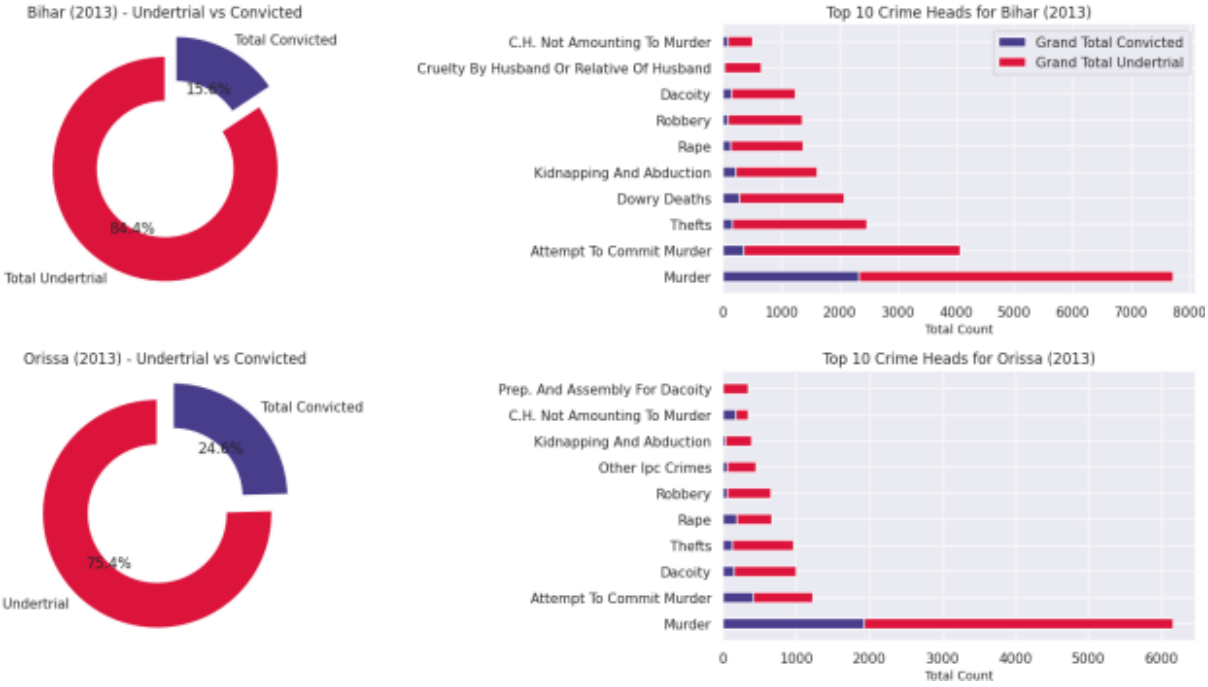
Following is the screenshot of the webapp that was created showing the same hypothesis discussed above:



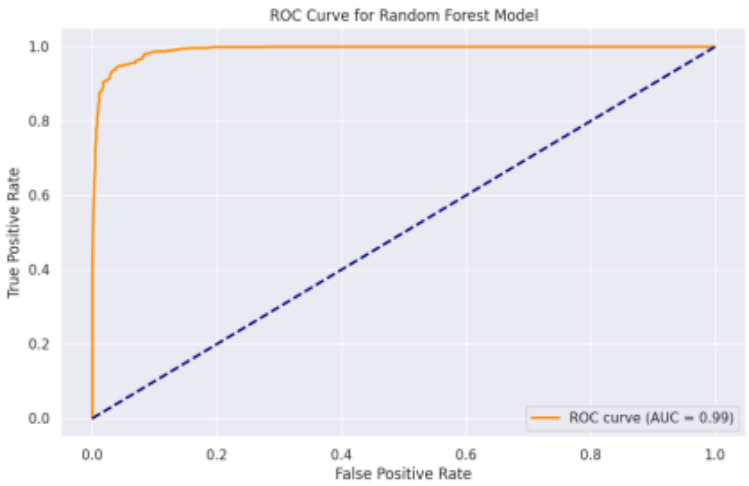
Hypothesis 2: Proportion of Undertrial vs. Convicted Inmates Across States and Crimes

Objective: This analysis helps to understand disparities in the proportion of undertrial and convicted inmates across different crimes and states. The goal is to identify patterns that highlight inefficiencies in the judicial process and areas requiring reform

Exploratory Data Analysis has been done using heatmaps and pie charts to visualize state-wise disparities.



Random Forest Regression was chosen for its ability to handle high-dimensional data and its robustness against overfitting, particularly in datasets with mixed numerical and categorical variables like crime type, state, and judicial efficiency indicators. This algorithm leverages an ensemble of decision trees, which individually make predictions and collectively provide accurate results through aggregation. Its ability to assess feature importance also made it a valuable tool for understanding the factors influencing undertrial-to-convicted ratios. The model achieved an impressive accuracy of **95.6%** and AUC of **0.99**, demonstrating strong predictive power in identifying cases with disproportionately high undertrial populations, especially for minor offenses in judicially inefficient states.



Inference:

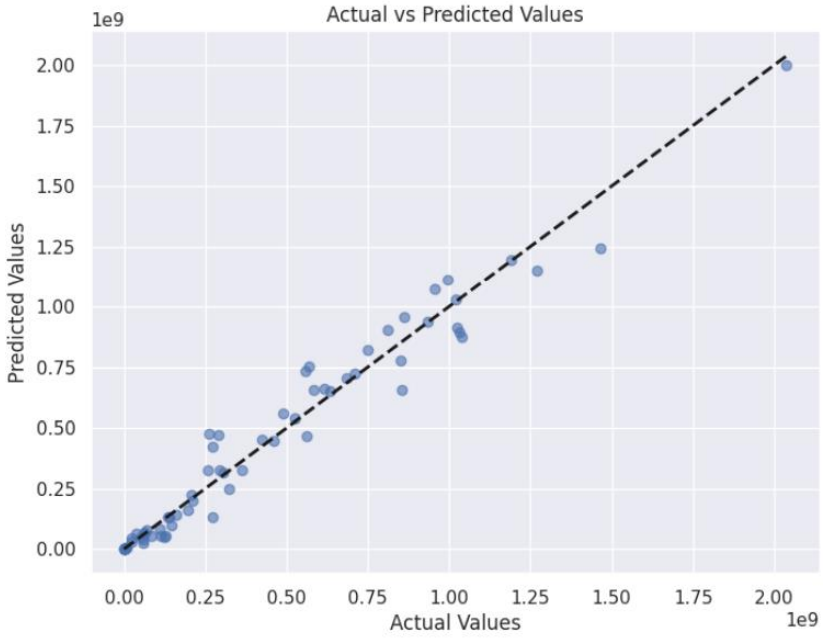
The model achieved high accuracy (95.6%) and effectively highlighted influential features, such as male undertrial populations and crime types. Its ensemble nature mitigates overfitting, yet biases may arise due to uneven data distributions, such as underrepresented crimes or states, which could skew predictions. The analysis revealed significant judicial inefficiencies, with states like Bihar and Uttar Pradesh exhibiting over 70% undertrial populations, particularly for minor offenses, while Kerala and Maharashtra displayed more balanced trial-to-conviction ratios, indicating better efficiency. Severe crimes such as murder and rape were prioritized, leading to higher conviction rates, whereas petty crimes faced extended delays. Male inmates aged 18-30 constituted the largest undertrial demographic, underscoring systemic neglect for this group. These findings highlight the urgent need for judicial reforms, including fast-track courts, alternative dispute resolution mechanisms, and expanded legal aid, to address backlogs, reduce overcrowding, and ensure timely justice.

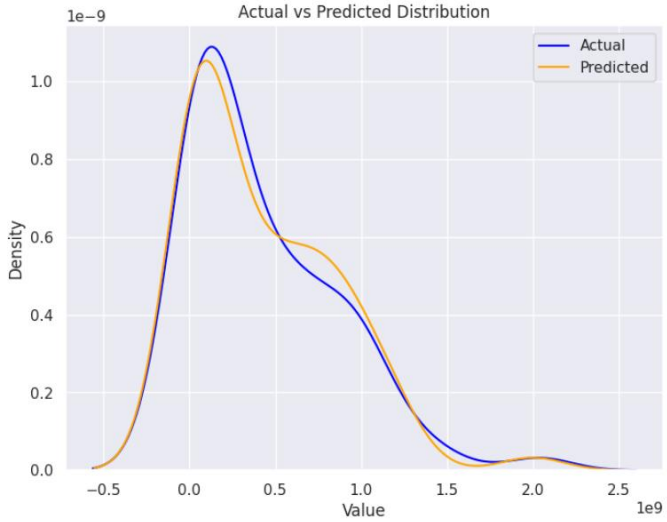
Following is the screenshot of the webapp that was created showing the same hypothesis discussed above:

Hypothesis 3: Predicting budget allocations using historical data

XGBoost Regression was chosen for its ability to model complex relationships and its robustness against overfitting. With an R^2 score of 0.96, the model demonstrated strong predictive

performance. However, biases may arise due to overrepresentation of well-funded states, potentially skewing predictions for underfunded regions. While effective, its reliance on historical data necessitates caution to avoid perpetuating existing inefficiencies.



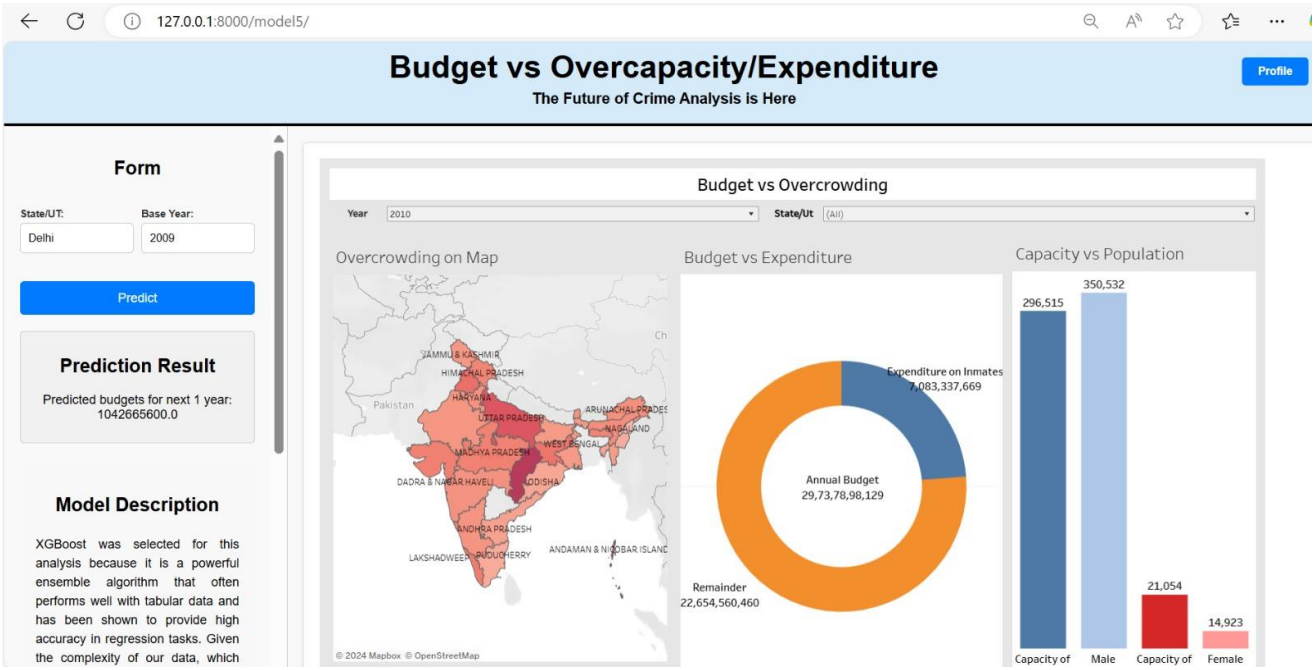


Inference:

The high R-squared value of 0.9617 suggests that the XGBoost model with regularization and encoding explaining a large proportion of the variance in the target variable, seems ideal at first. However, such a high R-squared value could indicate potential overfitting, especially in complex models like XGBoost, which are prone to capturing noise and patterns that may not generalize well to new data.

The XGBoost model revealed that states with overcrowded prisons and higher expenditures on inmates tend to allocate larger budgets in subsequent years. This insight supports the hypothesis that prison overcapacity and related expenses influence future budgeting decisions, likely as states attempt to address the challenges posed by overcrowded facilities. This intelligence could be valuable for policymakers, as it suggests that increasing investments in prison management is a reactive measure to overcapacity issues. Furthermore, the model's performance metrics validated its predictive power, indicating that similar approaches could be applied to other states or regions to inform budget forecasting.

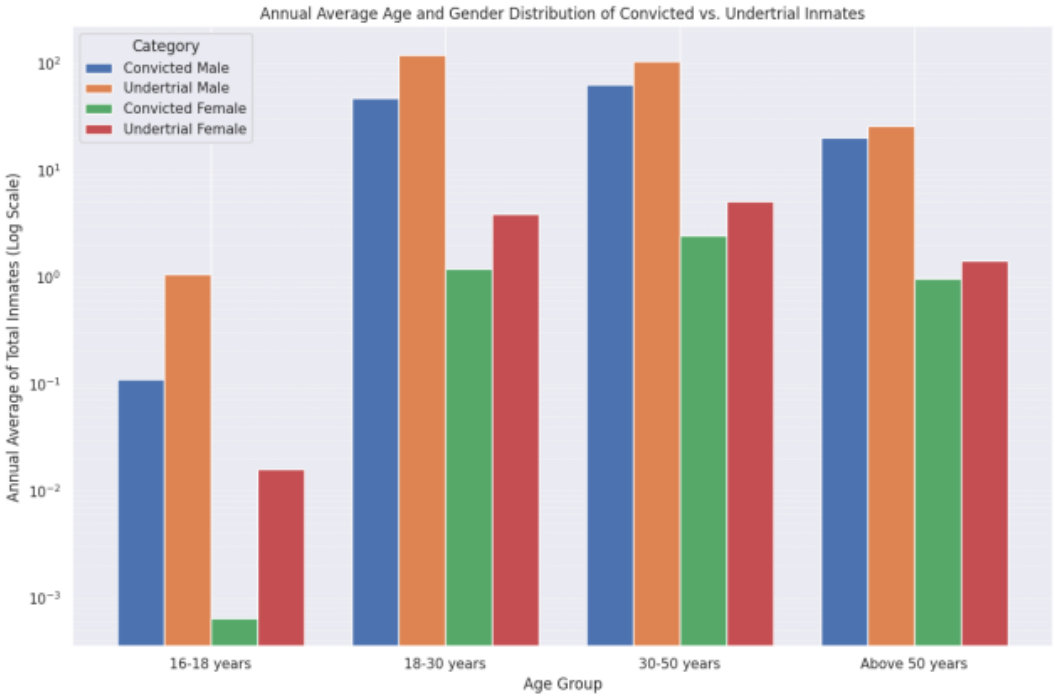
Following is the screenshot of the webapp that was created showing the same hypothesis discussed above:



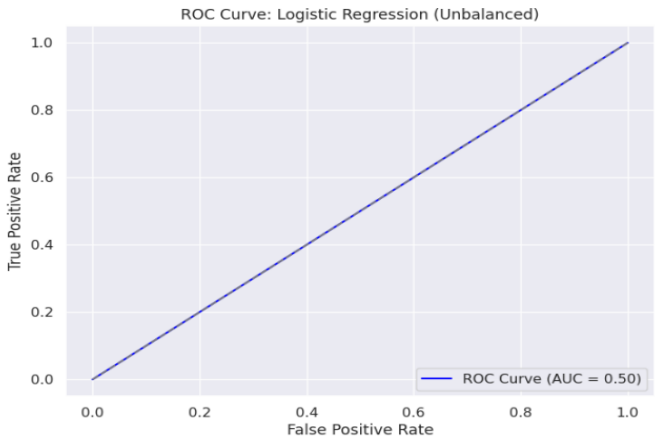
Hypothesis 4: Demographic Analysis of Inmate Populations by Trial Status

Objective: The objective is to classify inmates into undertrial or convicted categories by analyzing demographic attributes such as age, gender, and crime type, while addressing challenges like dataset imbalance and interpreting probabilities for deeper analysis

The analysis aimed to classify inmates as undertrial or convicted based on demographics such as age, gender, and crime type. EDA revealed that males aged 18-30 dominated both categories, with undertrial inmates significantly overrepresented, highlighting class imbalance in the data.



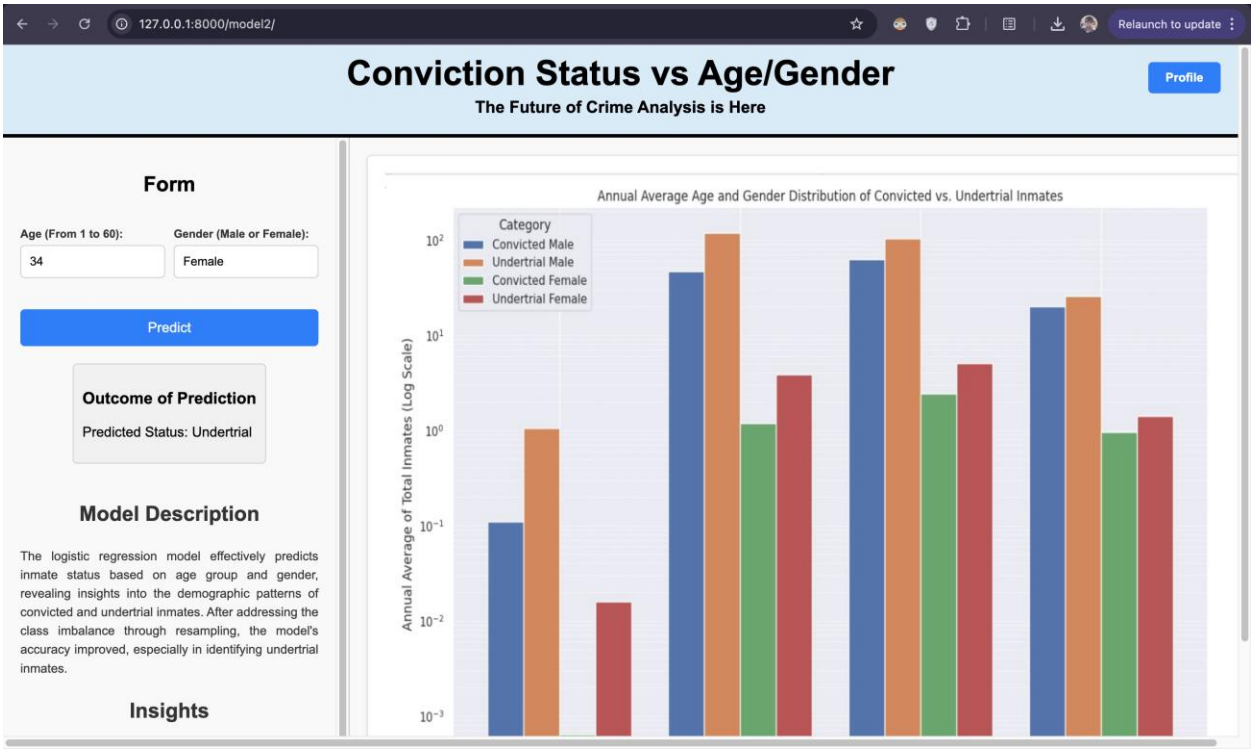
The analysis used **Logistic Regression** to classify inmates as undertrial or convicted, focusing on demographics like age, gender, and crime type. EDA showed males aged 18-30 dominated both categories, with undertrial inmates significantly overrepresented, revealing class imbalance in the data. Initially, the unbalanced dataset yielded poor model performance, with a cross-validation accuracy of 0.5 and an ROC AUC of 0.5. After applying oversampling to balance the dataset, accuracy improved to 0.75, and the ROC AUC rose to 0.89, demonstrating significantly enhanced classification performance. These findings highlighted systemic inefficiencies affecting juveniles and females, emphasizing the need for targeted reforms to ensure justice equity.



Inference:

The use of Logistic Regression demonstrated the impact of class balancing on model performance. While the initial model underperformed due to skewed data, balancing significantly improved accuracy and classification metrics, emphasizing the importance of addressing dataset imbalances in predictive modeling. These results underscore systemic inefficiencies, particularly affecting specific demographics, and suggest the need for targeted interventions to streamline trial processes for underrepresented groups like juveniles and females.

Following is the screenshot of the webapp that was created showing the same hypothesis discussed above:



Conclusion

The comprehensive analysis of prison data highlights critical systemic inefficiencies and areas for reform in India's judicial and correctional systems. Across various dimensions—educational infrastructure, judicial efficiency, budget allocation, and demographic disparities—key trends emerged, emphasizing the interconnectedness of these issues. Improved educational programs

significantly reduce escape rates and mental health problems, underscoring the importance of rehabilitation-focused initiatives. Judicial inefficiencies, particularly the disproportionate undertrial population in certain states, reveal the urgent need for reforms like fast-track courts and alternative dispute resolution mechanisms. Budget analysis indicates that chronic overcapacity leads to reactive financial planning, highlighting the necessity for proactive and strategic resource distribution. Demographic disparities, such as the overrepresentation of undertrial males aged 18-30 and delays in juvenile cases, further demonstrate systemic biases requiring targeted interventions. The findings advocate for a holistic approach to prison and judicial reforms, combining proactive planning, technological integration, and a focus on rehabilitation to ensure justice equity, reduce overcrowding, and improve overall prison management.

Acknowledgement

We would like to extend our heartfelt gratitude to **Prof. Chen Xu** for her invaluable guidance, mentorship, and unwavering support throughout the duration of this project. His profound expertise and thoughtful advice were pivotal in shaping our approach, refining our methodology, and ensuring that our analysis was both rigorous and insightful. His encouragement inspired us to push the boundaries of our understanding and delve deeper into the complexities of the subject.

We are also deeply appreciative of the consistent support and assistance provided by our teaching assistants—**Xixian Yang**, **Prajakta**, **Mounish**, and **Shreya Thakur**. Their timely feedback, constructive suggestions, and availability to address our queries played an integral role in overcoming challenges and refining the various aspects of this project. Each of them contributed significantly to enhancing the quality and depth of our work.

This project, focused on analyzing prison data in India, required careful consideration of multiple dimensions, and it would not have been possible without the collective support, encouragement, and guidance of our mentors and teaching assistants. We are sincerely grateful for their invaluable contributions, which have been instrumental in the successful completion of this endeavor.