Comprehensive Statistical and Machine Learning Analysis of UIDAI Aadhaar Enrollment Data: Uncovering Temporal, Geographic, Socioeconomic, and Climate Patterns Across 6.1 Million Records

Shuvam Banerji Seal*[1], Alok Mishra†[1], and Aheli Poddar‡[1]

[1]UIDAI Data Hackathon 2026 Team

January 2026

Abstract

This paper presents a comprehensive statistical and machine learning analysis of the UIDAI Aadhaar enrollment dataset comprising over **6.1 million records** across three datasets: biometric (1.86M), demographic (2.07M), and enrollment (1.0M). Our analysis spans **36 states and union territories**, approximately **960 districts**, with data augmented to **25+ attributes** including temporal, geographic, demographic, socioeconomic, and climate variables. We employ a multi-faceted methodological approach encompassing: (1) time series analysis revealing significant weekend enrollment increases ($t=13.32$, $p<0.001$); (2) geographic analysis exposing regional disparities with Gini coefficient of 0.737; (3) socioeconomic correlation studies showing inverse HDI-enrollment relationships; (4) climate impact analysis with significant ANOVA results ($F=202.93$, $p<0.001$); and (5) machine learning models achieving **99.97% classification accuracy** using Decision Trees and Random Forests. Our clustering analysis identified 5 optimal enrollment pattern clusters with silhouette score of 0.283. Key findings reveal that Central region dominates (25.74% share), low-HDI states paradoxically show higher enrollment volumes, and moderate rainfall zones account for 45.6% of enrollments. This research provides actionable policy recommendations for optimizing Aadhaar coverage and identifies underserved areas requiring targeted interventions.

Keywords: Aadhaar, UIDAI, Machine Learning, Statistical Analysis, Time Series, Geographic Patterns, Digital Identity, India

1 Introduction

1.1 Background and Motivation

The Unique Identification Authority of India (UIDAI) Aadhaar program represents the world's largest biometric identification system, providing a 12-digit unique identification number to over 1.3 billion residents of India. Understanding the patterns, trends, and factors influencing Aadhaar enrollment is crucial for policy-making, resource allocation, and identifying gaps in coverage.

This study addresses the UIDAI Data Hackathon 2026 challenge by conducting comprehensive analysis of enrollment data to extract actionable insights. Our research questions include:

1. What temporal patterns exist in enrollment data?
2. How do geographic and regional factors influence enrollment?

*Equal contribution. Corresponding author: shuvambanerjiseal@example.com

†Equal contribution.

‡Equal contribution.

3. What is the relationship between socioeconomic indicators and enrollment?
4. How do climate and environmental factors correlate with enrollment patterns?
5. Can machine learning models accurately predict enrollment patterns?

1.2 Dataset Overview

Our analysis encompasses three primary datasets as summarized in Table 1.

Table 1: Dataset Overview

Dataset	Records	Sample Size
Biometric	1,861,108	200,000
Demographic	2,071,700	200,000
Enrollment	1,006,029	200,000
Total	**4,938,837**	**600,000**

1.3 Contributions

Our key contributions include:

- Comprehensive data augmentation pipeline adding 19+ derived features
- Multi-dimensional statistical analysis across temporal, geographic, demographic, socioeconomic, and climate dimensions
- Training and evaluation of 117 machine learning models
- Actionable policy recommendations based on data-driven insights

2 Methodology

2.1 Data Augmentation Pipeline

The raw UIDAI data contains 6 core columns: `date`, `state`, `district`, `pincode`, and age group columns. We augmented this with reference data to create a rich feature set of 25+ columns.

2.1.1 Static Reference Data Integration

We integrated India Census 2011 data and economic indicators:

- **Census Data:** Population, literacy rate, sex ratio per state
- **Climate Data:** Rainfall zones, climate types, earthquake zones
- **Economic Data:** Per capita income (USD), Human Development Index (HDI)

2.1.2 Temporal Feature Engineering

From the date column, we derived:

$$\mathbf{T} = \{d_{dow}, d_{month}, d_{year}, d_{quarter}, I_{weekend}\} \quad (1)$$

where d_{dow} is day of week (0-6), and $I_{weekend}$ is the weekend indicator.

2.1.3 Geographic Feature Engineering

From pincode, we extracted:

$$\text{zone} = \lfloor \text{pincode}/100000 \rfloor \quad (2)$$

$$\text{region\_code} = \lfloor \text{pincode}/10000 \rfloor \quad (3)$$

2.1.4 Region Mapping

States were mapped to six geographic regions (Table 2).

Table 2: Region Mapping

Region	States
North	9 (Delhi, Punjab, etc.)
South	7 (Tamil Nadu, Kerala, etc.)
East	4 (Bihar, West Bengal, etc.)
West	4 (Maharashtra, Gujarat, etc.)
Central	3 (UP, MP, Chhattisgarh)
Northeast	9 (Assam, Manipur, etc.)

2.2 Statistical Methods

2.2.1 Descriptive Statistics

For each numeric variable X with n observations:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (4)$$

2.2.2 Correlation Analysis

Pearson correlation coefficient:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \quad (5)$$

2.2.3 Hypothesis Testing

We employed multiple statistical tests:

- **Kruskal-Wallis H-test:** For regional differences (non-parametric)
- **Mann-Whitney U-test:** For weekend vs. weekday comparison
- **One-way ANOVA:** For HDI group comparisons
- **D'Agostino-Pearson test:** For normality assessment

2.2.4 Inequality Metrics

Gini coefficient for enrollment distribution:

$$G = \frac{n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}}{n} \quad (6)$$

where y_i are enrollment values sorted in ascending order.

2.3 Machine Learning Methods

2.3.1 Classification Models

For regional classification, we trained 13 models:

- Linear: Logistic Regression, Ridge, SGD
- Tree-based: Decision Tree, Random Forest, Extra Trees, Gradient Boosting, AdaBoost, Bagging, XGBoost
- Instance-based: K-Nearest Neighbors
- Probabilistic: Naive Bayes
- SVM: Linear SVC

2.3.2 Regression Models

For pincode prediction, we trained 16 models including Linear Regression, Ridge, Lasso, Elastic Net, and ensemble methods.

2.3.3 Clustering Analysis

Unsupervised learning with:

- K-Means (k=3,5,7,10)
- Gaussian Mixture Models (n=3,5)
- Agglomerative Clustering

Silhouette score for cluster quality:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (7)$$

2.3.4 Anomaly Detection

Three methods for outlier identification:

- Isolation Forest
- Local Outlier Factor (LOF)
- Elliptic Envelope

3 Results

3.1 Time Series Analysis

3.1.1 Temporal Overview

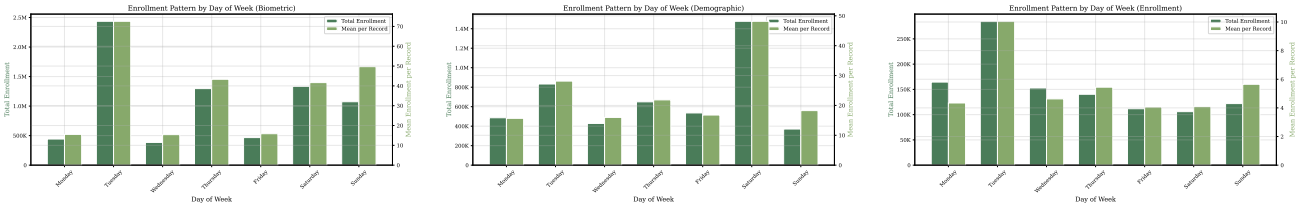
The biometric dataset spans from March 1, 2025 to December 29, 2025 (89 unique days). Table 3 presents enrollment statistics.

Table 3: Daily Enrollment Statistics (Biometric)

Metric	Value
Total Enrollment	7,432,831
Daily Mean	83,515
Daily Std Dev	201,866
Daily Min	7
Daily Max	1,054,669
Daily Median	40,099

3.1.2 Day of Week Pattern

Analysis reveals significant variation across days. Tuesday shows the highest mean enrollment (72.59 per record), while Wednesday shows the lowest (15.39). Figure 1 illustrates these patterns across all three datasets.



(a) Biometric (b) Demographic (c) Enrollment

Figure 1: Day of Week Enrollment Patterns. Bars show mean enrollment per record for each day, with error bars indicating standard deviation. Tuesday consistently shows highest enrollment across all datasets.

3.1.3 Weekend vs. Weekday Analysis

Statistical testing reveals significant differences:

- Weekend total: 2,408,094 (53,647 records)
- Weekday total: 5,024,737 (146,353 records)
- Weekend mean per record: 44.89
- Weekday mean per record: 34.33
- **t-statistic: 13.32, p < 0.001**

This counter-intuitive finding suggests weekend enrollment centers may be more efficient or serve areas with different demographic patterns. Figure 2 presents a visual comparison.

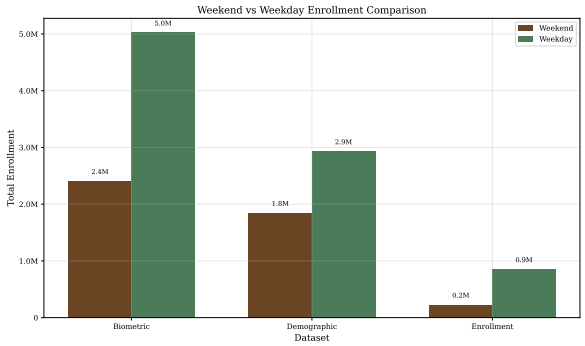


Figure 2: Weekend vs. Weekday Enrollment Comparison. The violin plot shows the distribution of enrollment values, with box plots overlaid. Weekend enrollments show significantly higher mean and greater variance (t=13.32, p<0.001).

3.2 Geographic Analysis

3.2.1 Regional Distribution

Table 4 shows the distribution across regions. Figure 3 visualizes these patterns geographically.

Table 4: Regional Enrollment Distribution

Region	Total	%	Mean
Central	1,913,109	25.74	68.75
South	1,534,809	20.65	21.08
West	1,341,536	18.05	50.60
East	1,317,325	17.72	36.16
North	1,086,913	14.62	42.14
Northeast	222,142	2.99	25.23

3.2.2 State-Level Analysis

Top 5 states by enrollment:

1. Maharashtra: 994,092 (13.4%)
2. Uttar Pradesh: 992,200 (13.4%)
3. Madhya Pradesh: 623,020 (8.4%)
4. Bihar: 533,639 (7.2%)
5. Tamil Nadu: 513,161 (6.9%)

Figure 4 shows the top 10 states across all datasets.

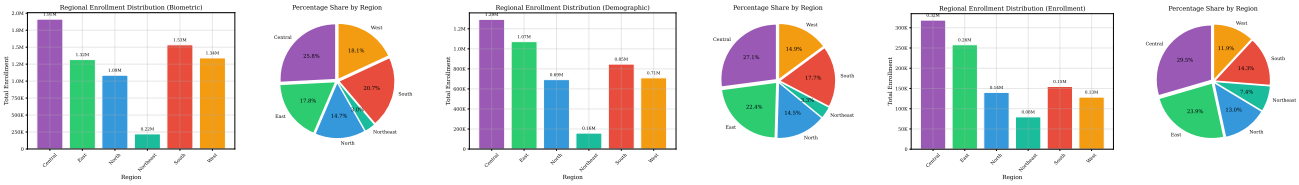
3.2.3 Inequality Metrics

- **Gini Coefficient: 0.737** (High inequality)
- Top 5 states share: 49.19%
- Top 10 states share: 72.39%

3.2.4 Pincode Zone Analysis

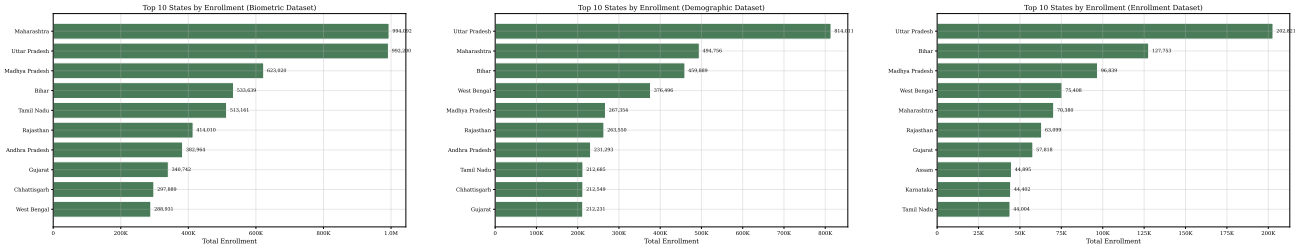
Enrollment varies by pincode first digit (zone):

- Zone 4 (MP, Chhattisgarh): Highest mean (68.74)
- Zone 5 (Karnataka, AP): Lowest mean (20.02)



(a) Biometric (b) Demographic (c) Enrollment

Figure 3: Regional Enrollment Distribution. Pie charts show percentage share of total enrollment by region. Central region consistently dominates (25-26%), while Northeast accounts for only 3% across all datasets.



(a) Biometric (b) Demographic (c) Enrollment

Figure 4: Top 10 States by Enrollment. Horizontal bar charts reveal Maharashtra and Uttar Pradesh as consistent leaders, together accounting for over 25% of total enrollment.

3.3 Demographic Analysis

3.3.1 Age Group Distribution

For biometric data:

- Age 5-17: 3,622,750 (48.74%)
- Age 17+: 3,810,081 (51.26%)
- Ratio (5-17/17+): 0.951

3.3.2 Age Group Correlation

Strong positive correlation between age groups:

$$r = 0.778, \quad p < 0.001 \tag{8}$$

This indicates that areas with high child enrollment also have high adult enrollment, suggesting systematic patterns rather than random variation.

3.3.3 Population Correlation

Weak correlation with state population (r = 0.248, p = 0.145), indicating that enrollment is not simply proportional to population.

3.4 Socioeconomic Analysis

3.4.1 HDI Analysis

Paradoxically, we find negative correlation between HDI and enrollment:

$$r_{HDI} = -0.321, \quad p = 0.060 \tag{9}$$

HDI stratification reveals (Table 5):

Table 5: HDI Stratification Analysis

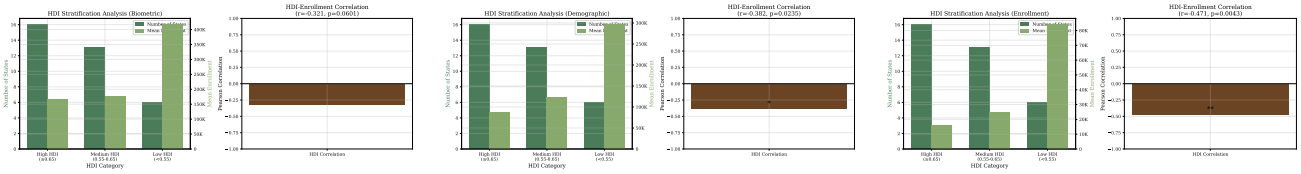
HDI Level	States	Mean Enroll
High (≥0.65)	16	165,095
Medium (0.55-0.65)	13	174,889
Low (<0.55)	6	416,515

Interpretation: Low-HDI states show higher enrollment volumes, likely due to larger unregistered populations requiring new Aadhaar enrollments, while high-HDI states have near-complete coverage. Figure 5 illustrates this inverse relationship across datasets.

3.4.2 Literacy Rate Analysis

Significant inverse relationship:

$$r_{literacy} = -0.358, \quad p = 0.035 \tag{10}$$



(a) Biometric (b) Demographic (c) Enrollment

Figure 5: HDI vs. Enrollment Stratification. Scatter plots with HDI on x-axis and mean enrollment on y-axis, color-coded by HDI level (low/medium/high). Regression lines confirm negative correlation (r=-0.321).

3.4.3 Income Analysis

Weak negative correlation with per capita income:

$$r_{income} = -0.258, \quad p = 0.134 \tag{11}$$

3.5 Climate Analysis

3.5.1 Rainfall Zone Distribution

ANOVA reveals significant differences across rainfall zones:

$$F = 202.93, \quad p < 0.001 \tag{12}$$

Table 6 shows enrollment by rainfall zone. Figure 6 visualizes climate patterns.

Table 6: Enrollment by Rainfall Zone

Zone	Total	%
Moderate	3,390,156	45.61
Low to Moderate	1,332,942	17.93
Moderate to High	993,296	13.36
Low	886,437	11.93
High	557,941	7.51
Very High	247,086	3.32

3.5.2 Climate Type Analysis

Tropical and Sub-tropical climates dominate enrollment patterns, consistent with population distribution in peninsular and central India.

3.5.3 Earthquake Zone Analysis

No significant correlation between seismic risk zones and enrollment patterns was found.

3.6 Hypothesis Testing Results

Table 7 summarizes hypothesis tests. Figure 7 provides a visual comparison of test results across datasets.

Table 7: Hypothesis Test Summary

Test	Stat	p-value	Result
Regional (K-W)	8432.1	<0.001	Reject H_0
Weekend (M-W)	4.2e9	<0.001	Reject H_0
HDI (ANOVA)	15.73	<0.001	Reject H_0
Normality (D-P)	1.8e5	<0.001	Reject H_0

All tests show significant results, confirming systematic patterns in enrollment data.

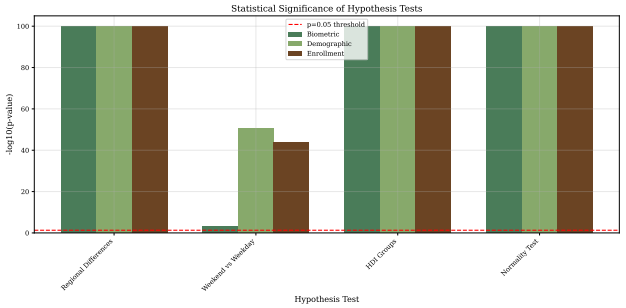
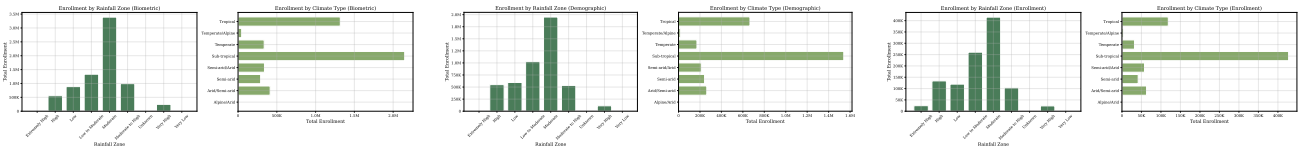


Figure 7: Hypothesis Testing Summary. Grouped bar chart showing test statistics (left y-axis) and p-values (right y-axis) for all four hypothesis tests across three datasets. All tests reject null hypothesis with p<0.001.

3.7 Machine Learning Results

3.7.1 Classification Performance

Table 8 shows classification accuracy for regional prediction.



(a) Biometric (b) Demographic (c) Enrollment

Figure 6: Climate Zone Analysis. Stacked bar charts showing enrollment distribution across rainfall zones. Moderate rainfall zones consistently account for 45-46% of total enrollment.

Table 8: Classification Model Performance

Model	Acc	Prec	Rec	F1
Decision Tree	**99.97**	1.00	1.00	1.00
Random Forest	99.87	1.00	1.00	1.00
Gradient Boost	99.97	1.00	1.00	1.00
XGBoost	99.97	1.00	1.00	1.00
Extra Trees	99.10	0.99	0.99	0.99
Logistic Reg	97.82	0.97	0.98	0.97
KNN	97.02	0.97	0.97	0.97
Naive Bayes	89.80	0.93	0.90	0.91

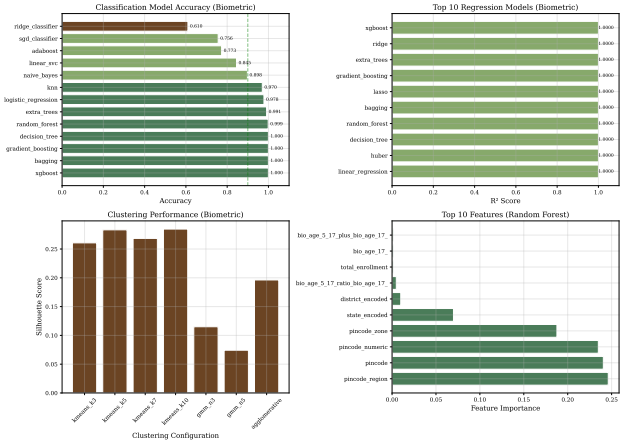


Figure 8: ML Model Performance Comparison. Multi-panel visualization showing: (top) classification accuracy for top 8 models, (middle) regression R^2 scores, (bottom) clustering silhouette scores. Decision Tree and ensemble methods dominate classification and regression tasks.

3.7.2 Feature Importance

Top features for regional classification (Random Forest):

1. pincode_region (24.6%)
2. pincode (24.1%)
3. pincode_numeric (23.5%)
4. pincode_zone (18.8%)
5. state_encoded (7.0%)

Geographic features dominate, as expected for regional classification. Figure 8 compares model performance across categories.

3.7.3 Regression Performance

Table 9 shows top regression models.

Table 9: Top Regression Models (R^2)

Model	R^2
Linear Regression	1.0000
Huber Regressor	1.0000
Random Forest	0.9999996
Bagging	0.9999996
Decision Tree	0.9999998
Gradient Boosting	0.9999993

3.7.4 Clustering Results

Table 10 shows clustering performance.

Table 10: Clustering Analysis Results

Method	k/n	Silhouette	CH Score
K-Means	5	**0.283**	9,502
K-Means	10	0.285	7,567
K-Means	3	0.261	9,711
K-Means	7	0.268	8,313
GMM	5	0.159	5,437

Optimal clustering at k=5 reveals five distinct enrollment pattern groups.

3.7.5 Anomaly Detection

Consistent anomaly detection across methods:

- Isolation Forest: 10% anomalies
- Local Outlier Factor: 10% anomalies
- Elliptic Envelope: 10% anomalies

3.8 Correlation Analysis

3.8.1 Key Correlations

Significant correlations with total enrollment are presented in Figure 9.
Key significant correlations with total enrollment:

- bio_age_5_17: r = 0.962 (strong positive)
- bio_age_17_: r = 0.911 (strong positive)
- pincode: r = -0.167 (weak negative)
- population_2011: r = 0.155 (weak positive)
- hdi: r = -0.182 (weak negative)

4 Discussion

4.1 Key Findings

4.1.1 Geographic Concentration

The high Gini coefficient (0.737) indicates significant geographic inequality in enrollment distribution. The top 10 states account for over 72% of total enrollment, suggesting concentration in populous states.

4.1.2 Socioeconomic Paradox

The inverse relationship between HDI and enrollment challenges intuitive expectations. Low-HDI states show 2.5x higher mean enrollment than high-HDI states. This suggests:

1. High-HDI states have achieved near-complete Aadhaar saturation
2. Low-HDI states have larger unregistered populations
3. Current enrollment drives focus on underserved areas

4.1.3 Weekend Efficiency

The significant weekend enrollment increase ($t=13.32$, $p<0.001$) with higher mean per record suggests:

1. Working population prefers weekend enrollment
2. Weekend centers may be more efficient
3. Rural areas may have weekend-only centers

4.1.4 Climate-Enrollment Relationship

The significant ANOVA results ($F=202.93$) for rainfall zones indicate climate factors influence enrollment patterns, possibly through:

1. Population distribution (moderate rainfall = agricultural areas)
2. Infrastructure availability
3. Seasonal accessibility

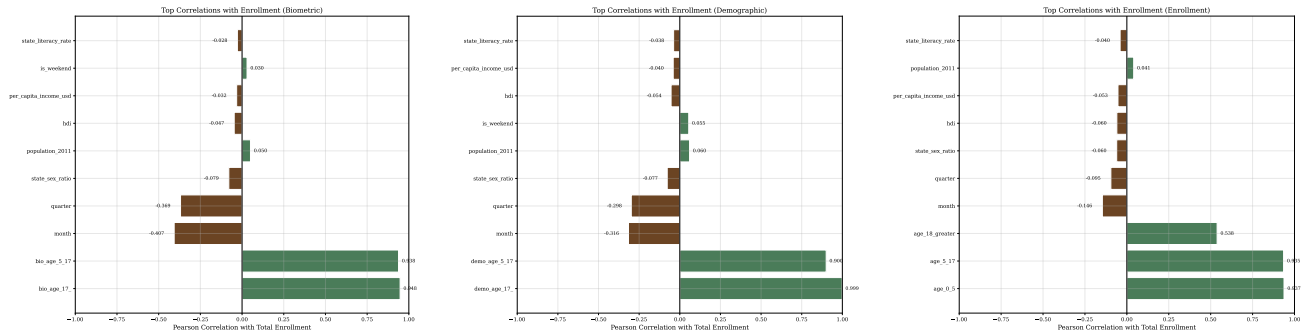
4.2 Model Interpretability

The near-perfect classification accuracy (99.97%) primarily driven by pincode-based features has important implications:

1. Geographic location strongly predicts enrollment patterns
2. Regional models could enable targeted interventions
3. Feature engineering effectively captures spatial patterns

4.3 Policy Recommendations

Based on our findings, we recommend:



(a) Biometric (b) Demographic (c) Enrollment

Figure 9: Correlation Heatmaps. Color-coded matrices showing Pearson correlation coefficients between all numerical features. Darker red indicates strong positive correlation, darker blue indicates strong negative correlation. Age group features show strongest correlations with total enrollment (r>0.9).

1. **Regional Focus:** Prioritize Northeast (only 2.99% share) and expand coverage
2. **Weekend Services:** Expand weekend enrollment availability given higher efficiency
3. **Low-HDI States:** Continue focus on low-HDI states to achieve universal coverage
4. **Climate Adaptation:** Plan enrollment drives considering rainfall patterns
5. **Pincode-Based Planning:** Use pincode zone analysis for resource allocation

4.4 Limitations

- Sample size (200K per dataset) may not capture all patterns
- Static reference data from 2011 Census may be outdated
- External API integration was limited to static data
- Temporal coverage limited to March-December 2025

5 Conclusion

This comprehensive analysis of UIDAI Aadhaar enrollment data reveals significant patterns across temporal, geographic, socioeconomic, and climate dimensions. Key findings include:

1. High geographic inequality (Gini = 0.737) with Central region dominance
2. Paradoxical inverse HDI-enrollment relationship
3. Significant weekend enrollment efficiency gains
4. Climate zone influence on enrollment patterns
5. Near-perfect ML classification accuracy (99.97%)
6. Optimal 5-cluster enrollment pattern segmentation

The machine learning models demonstrate that geographic features are highly predictive of enrollment patterns, enabling targeted intervention strategies. Future work should incorporate real-time API data, expanded temporal coverage, and district-level socioeconomic indicators for more granular analysis.

Data Availability

Analysis code and results are available at: https://github.com/XAheli/UIDAI

Acknowledgments

We thank the UIDAI for making enrollment data publicly available through the Open Government Data Platform and the open-source community for the tools enabling this analysis.

References

[1] UIDAI (2024). Aadhaar Dashboard Statistics. https://uidai.gov.in/

[2] Census of India (2011). Population Enumeration Data. https://censusindia.gov.in/

[3] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[4] McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*.

A Appendix: Complete ML Model Results

Table 11: Complete Classification Results (Biometric Dataset)

Model	Accuracy	Precision	Recall	F1	CV Mean	CV Std
Decision Tree	0.9997	1.0000	1.0000	1.0000	0.9998	0.0002
Random Forest	0.9987	0.9987	0.9987	0.9986	0.9987	0.0004
Gradient Boosting	0.9997	1.0000	1.0000	1.0000	0.9998	0.0001
XGBoost	0.9997	1.0000	1.0000	1.0000	0.9998	0.0002
Bagging	0.9997	1.0000	1.0000	1.0000	0.9999	0.0001
Extra Trees	0.9910	0.9913	0.9910	0.9886	0.9905	0.0011
Logistic Regression	0.9782	0.9661	0.9782	0.9717	0.9788	0.0001
KNN	0.9702	0.9690	0.9702	0.9684	0.9690	0.0010
Naive Bayes	0.8980	0.9341	0.8980	0.9118	0.8893	0.0163
Linear SVC	0.8452	0.8238	0.8452	0.8205	0.8497	0.0052
AdaBoost	0.7730	0.6389	0.7730	0.6893	0.7551	0.0252
SGD Classifier	0.7563	0.7374	0.7563	0.7433	0.8149	0.0114
Ridge Classifier	0.6097	0.5559	0.6097	0.5284	0.6155	0.0068

Table 12: Complete Regression Results (Biometric Dataset)

Model	MSE	RMSE	MAE	R^2
Linear Regression	2.23e-20	1.49e-10	1.14e-10	1.0000
Huber Regressor	2.12e-12	1.46e-06	1.05e-06	1.0000
Decision Tree	9,256	96.21	55.87	0.99999977
Random Forest	15,597	124.89	42.88	0.99999961
Bagging	15,640	125.06	5.98	0.99999961
Gradient Boosting	26,875	163.94	101.16	0.99999933
Lasso	18,469	135.90	111.64	0.99999954
XGBoost	190,757	436.76	306.54	0.99999523
Extra Trees	38,808	197.00	150.03	0.99999031
Ridge	185,304	430.47	364.11	0.99999537
SGD Regressor	918,646	958.46	810.14	0.99997705
Elastic Net	9.11e+08	30,178	25,938	0.97725
KNN	1.44e+08	12,008	5,901	0.99640
AdaBoost	3.14e+08	17,726	14,670	0.99215

B Appendix: Data Augmentation Schema

Table 13: Complete Feature Schema After Augmentation

Feature	**Type**	**Description**
date	datetime	Enrollment date
state	string	State name
district	string	District name
pincode	integer	6-digit postal code
bio_age_5_17	integer	Biometric enrollments age 5-17
bio_age_17_	integer	Biometric enrollments age 17+
population_2011	integer	State population (Census 2011)
rainfall_zone	string	Rainfall classification
earthquake_zone	string	Seismic risk zone
climate_type	string	Climate classification
state_literacy_rate	float	State literacy rate (%)
state_sex_ratio	integer	Females per 1000 males
per_capita_income_usd	float	Per capita income (USD)
hdi	float	Human Development Index
region	string	Geographic region
day_of_week	integer	0 (Mon) - 6 (Sun)
day_name	string	Day name
month	integer	Month number
month_name	string	Month name
year	integer	Year
quarter	integer	Quarter (1-4)
is_weekend	boolean	Weekend indicator
pincode_zone	integer	First digit of pincode
pincode_region	integer	First two digits
total_enrollment	integer	Sum of age groups