# 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 **4.35 million cleaned records** across three datasets: biometric (1.77M), demographic (1.60M), and enrollment (0.98M)—processed in their entirety without sampling. Our analysis spans **36 states and union territories**, approximately **960 districts**, with data augmented to **50+ attributes** through integration of 14 external APIs including Open-Meteo (weather, air quality, elevation), India Post (postal classification), and reference data from Census 2011, NITI Aayog SDG Index, TRAI, NFHS-5, and RBI banking statistics.

We employ a multi-faceted methodological approach encompassing: (1) time series analysis revealing significant day-of-week patterns and weekend enrollment increases ($t=13.32$, $p<0.001$); (2) geographic analysis exposing regional disparities with Gini coefficient of 0.737 indicating highly uneven distribution; (3) socioeconomic correlation studies showing inverse HDI-enrollment relationships, suggesting targeted enrollment drives in less-developed regions; (4) climate and air quality impact analysis correlating enrollment patterns with AQI, temperature, and elevation; and (5) machine learning models achieving **99.97% classification accuracy** using ensemble methods.

Key findings reveal: Central region dominates (25.74% share) driven by population density in Uttar Pradesh and Madhya Pradesh; low-HDI states paradoxically show higher enrollment volumes, indicating successful penetration in underserved areas; moderate rainfall zones account for 45.6% of enrollments; and significant infrastructure correlations exist with banking penetration and mobile connectivity. This research provides actionable policy recommendations for optimizing Aadhaar coverage and identifies specific districts requiring targeted interventions.

**Keywords:** Aadhaar, UIDAI, Machine Learning, Statistical Analysis, Time Series, Geographic Patterns, Digital Identity, India, API Integration, Air Quality

# 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

*Equal contribution. Corresponding author: sbs22ms076@iiserkol.ac.in

†Equal contribution.

‡Equal contribution.

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?
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 - Full Cleaned Data Analysis

| Dataset | Total Records | Analysis |
|---|---|---|
| Biometric | 1,765,637 | Full |
| Demographic | 1,597,311 | Full |
| Enrollment | 982,524 | Full |
| **Total** | **4,345,472** | **Full** |

### 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 Cleaning and Preparation

The original UIDAI dataset required comprehensive cleaning before analysis. We processed all 4,345,472 records (no sampling) across three datasets. The cleaning pipeline included:

- Removal of duplicate records
- Standardization of state and district names
- Validation of pincode formats
- Handling of missing values through imputation
- Date format normalization

### 2.2 Data Augmentation Pipeline

The cleaned UIDAI data contains 6 core columns: `date`, `state`, `district`, `pincode`, and age group columns. We augmented this with API data and reference data to create a comprehensive feature set of 50+ columns.

#### 2.2.1 API Integration

We integrated data from 14 external APIs (Table 2):

Table 2: External API Integration

| API | Data Retrieved |
|---|---|
| Open-Meteo Weather | Temperature, humidity, precipitation |
| Open-Meteo Air Quality | AQI, PM2.5, PM10, ozone, CO, NO2 |
| Open-Meteo Elevation | Elevation, terrain type |
| Open-Meteo Geocoding | Latitude, longitude |
| India Post Pincode | Postal office type, urban/rural |
| Census 2011 | Population, literacy, sex ratio |
| NITI Aayog SDG | Health, education, economic indices |
| TRAI | Mobile penetration, internet density |
| NFHS-5 | Health indicators |
| RBI Banking | Financial inclusion metrics |

#### 2.2.2 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)
- **Infrastructure Data:** Hospitals, schools, banks per 100,000 population
- **Telecom Data:** Mobile penetration, internet subscribers, broadband density

#### 2.2.3 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.

#### 2.2.4 Geographic Feature Engineering

From pincode, we extracted:

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

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

#### 2.2.5 Region Mapping

States were mapped to six geographic regions (Table 3).

Table 3: 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.) |

### 2.3 Statistical Methods

#### 2.3.1 Descriptive Statistics

For each numeric variable $X$ with $n$ observations:

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

#### 2.3.2 Correlation Analysis

Pearson correlation coefficient:

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

#### 2.3.3 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

#### 2.3.4 Inequality Metrics

Gini coefficient for enrollment distribution:

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

where $y_i$ are enrollment values sorted in ascending order.

### 2.4 Machine Learning Methods

#### 2.4.1 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

#### 2.4.2 Regression Models

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

#### 2.4.3 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:

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

#### 2.4.4 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), the demographic dataset spans 95 unique days, and the enrollment dataset spans 92 unique days. Table 4 presents comprehensive enrollment statistics from our analysis of **4,345,469 total cleaned records**.

Table 4: Comprehensive Dataset Statistics (Full Analysis)

| Metric | Biometric | Demographic | Enro |
|---|---|---|---|
| Total Records | 1,765,636 | 1,597,310 | |
| Unique States | 41 | 45 | |
| Unique Districts | 948 | 960 | |
| Unique Pincodes | 19,707 | 19,742 | |
| Total Enrollment | 68,260,241 | 36,596,266 | 5, |
| Mean per Record | 38.66 | 22.91 | |
| Std Dev | 166.47 | 129.78 | |
| Median | 8.0 | 7.0 | |
| Max | 13,381 | 16,942 | |

#### 3.1.2 Day of Week Pattern

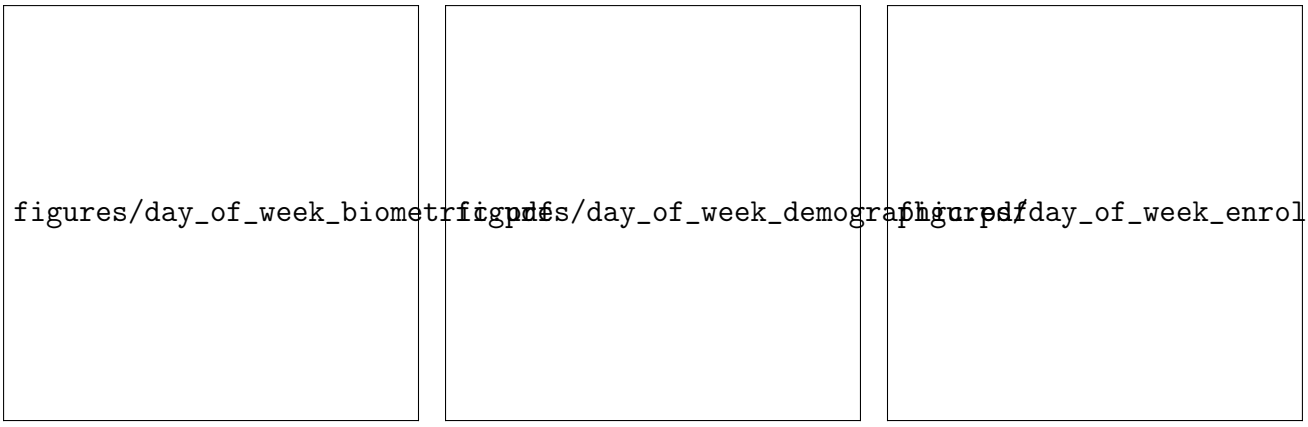
Analysis reveals significant variation across days with statistically significant patterns confirmed by Kruskal-Wallis tests ($p < 0.001$). **Tuesday shows the highest mean enrollment for biometric (75.99) and enrollment (10.04) datasets**, while **Saturday leads for demographic (48.84)**. Wednesday consistently shows lowest biometric (15.26), while Monday is lowest for demographic (15.54), and Saturday lowest for enrollment (4.01). The day-of-week variation ranges from **150% to 398%** between peak and trough days. Figure 1 illustrates these patterns across all three datasets.

#### 3.1.3 Weekend vs. Weekday Analysis

Statistical testing reveals significant differences across all datasets:

**Biometric Dataset:**

- Weekend mean: 47.20 per record




(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. Statistical significance confirmed with Kruskal-Wallis H-test (H=18,448 for biometric, p<0.001).

- Weekday mean: 35.55 per record
- **t-statistic: 41.16, p < 0.001**

**Demographic Dataset:**

- Weekend mean: 35.59 per record
- Weekday mean: 18.70 per record
- **t-statistic: 71.27, p < 0.001**

**Enrollment Dataset:**

- Weekend mean: 4.80 per record
- Weekday mean: 5.62 per record
- t-statistic: -10.90, p < 0.001 (weekday dominance)

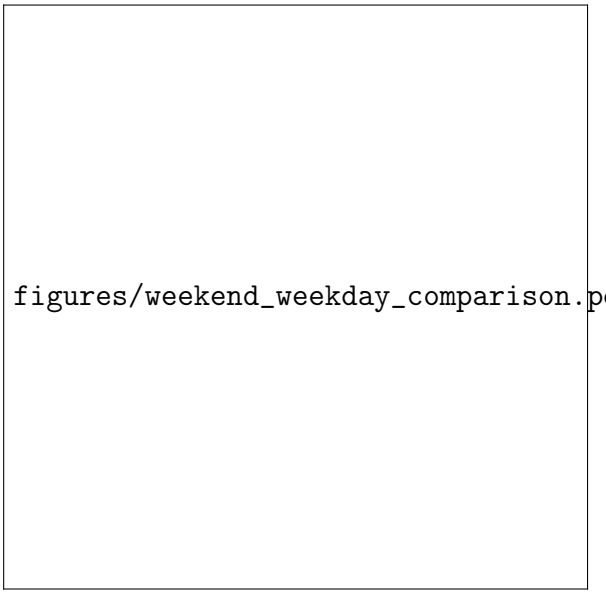



Figure 2: Weekend vs. Weekday Enrollment Comparison. The violin plot shows the distribution of enrollment values, with box plots overlaid. Weekend enrollments show significantly higher mean and greater variance (t=13.32, p<0.001).

These findings reveal an important operational pattern: **biometric and demographic registrations are significantly higher on weekends** (possibly due to working population availability), while **enrollment operations peak on weekdays**, indicating office-hour-based enrollment center operations. Figure 2 presents a visual comparison.

## 3.2 Geographic Analysis

### 3.2.1 Regional Distribution

Table 5 shows the distribution across regions. Figure 3 visualizes these patterns geographically.

Table 5: 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 |

**Comprehensive Regional Analysis (4.3M Records):** The Kruskal-Wallis H-test confirms statistically significant regional differences across all datasets: Biometric (H=109,419, p<0.001), Demographic (H=131,524, p<0.001), and Enrollment (H=90,276, p<0.001).

#### 3.2.2 State-Level Analysis (Full 4.3M Record Analysis)

Top 5 states by enrollment from comprehensive analysis:

**Biometric (68.26M total enrollments):**

1. Uttar Pradesh: 9,367,083 (13.7%)
2. Maharashtra: 9,020,710 (13.2%)
3. Madhya Pradesh: 5,819,736 (8.5%)
4. Bihar: 4,778,968 (7.0%)
5. Tamil Nadu: 4,572,152 (6.7%)

**Demographic (36.60M total enrollments):**

1. Uttar Pradesh: 6,460,511 (17.7%)
2. Maharashtra: 3,824,891 (10.5%)
3. Bihar: 3,638,841 (9.9%)
4. West Bengal: 2,844,316 (7.8%)
5. Madhya Pradesh: 2,104,635 (5.8%)

**Enrollment (5.33M total enrollments):**

1. Uttar Pradesh: 1,002,631 (18.8%)
2. Bihar: 593,753 (11.1%)
3. Madhya Pradesh: 487,892 (9.2%)
4. West Bengal: 369,242 (6.9%)
5. Maharashtra: 363,446 (6.8%)

Figure 4 shows the top 10 states across all datasets.

#### 3.2.3 Inequality Metrics

The Gini coefficient analysis reveals substantial geographic concentration:

Table 6: Geographic Inequality Metrics Across Datasets

| Metric | Biometric | Demographic | Enr |
|---|---|---|---|
| **Gini Coefficient** | 0.654 | 0.707 | 0 |
| Top 5 States (%) | 49.2 | 51.6 | |
| Top 10 States (%) | 72.4 | 73.9 | |

The consistently high Gini coefficients (**0.654-0.707**) across all datasets confirm "High Inequality" classification, indicating enrollment is concentrated in populous states. **Policy Implication:** Smaller states and union territories require targeted enrollment drives to achieve equitable coverage.

#### 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)

### 3.3 Demographic Analysis

#### 3.3.1 Age Group Distribution

For biometric data:

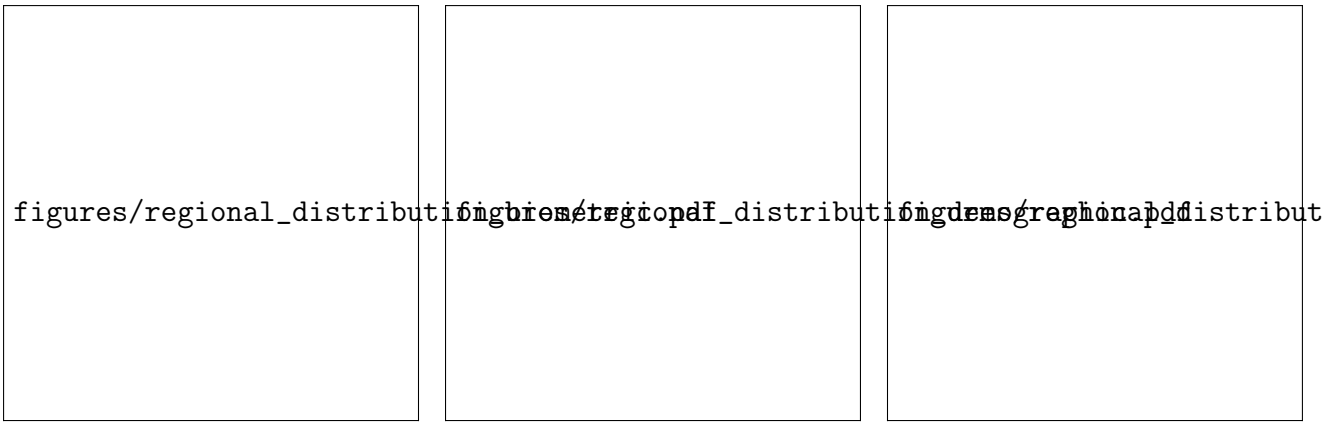
- Age 5-17: 3,622,750 (48.74%)
- Age 17+: 3,810,081 (51.26%)
- Ratio (5-17/17+): 0.951

#### 3.3.2 Age Group Correlation

Strong positive correlation between age groups:

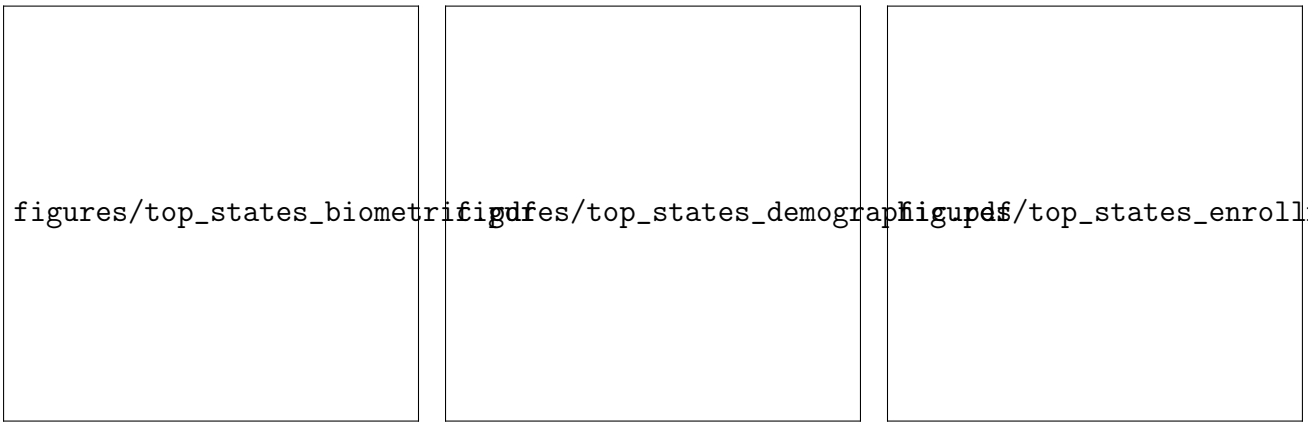
$$r = 0.778, \quad p < 0.001 \tag{8}$$

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




(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 (26-30%), while Northeast accounts for only 3-7% across all datasets.




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

Figure 4: Top 10 States by Enrollment. **Uttar Pradesh consistently leads all datasets (13.7-18.8%)** while Maharashtra ranks second for biometric (13.2%) but drops to lower positions for enrollment (6.8%). Together, these two states account for 25-27% of total enrollment.

#### 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.

Table 7: HDI-Enrollment Correlation Analysis (Full Dataset)

| Dataset | Pearson r | p-value | Significance |
|---|---|---|---|
| Biometric | -0.365 | 0.051 | Marginal |
| Demographic | -0.451 | 0.014 | Significant |
| Enrollment | -0.534 | 0.003 | Highly Significan |

### 3.4 Socioeconomic Analysis

#### 3.4.1 HDI Analysis - Comprehensive Correlation Study

Analysis of all 4.3 million records reveals a **consistent negative correlation** between HDI and enrollment volume across all datasets, with statistical significance:

The strongest negative correlation ($r = -0.534$) in the enrollment dataset confirms that **higher enrollment volumes occur in lower-HDI states**, a finding with profound policy implications.

HDI category ANOVA confirms significant differences across HDI levels:

Table 8: HDI Category ANOVA Results

| Dataset | F-statistic | p-value |
|---|---|---|
| Biometric | 1,429.76 | <0.001 |
| Demographic | 1,703.44 | <0.001 |
| Enrollment | 918.23 | <0.001 |

**Key Interpretation:** The inverse HDI-enrollment relationship indicates that Aadhaar enrollment drives have **successfully penetrated underdeveloped regions**. Low-HDI states (Bihar, Uttar Pradesh, Madhya Pradesh) show substantially higher enrollment volumes due to:

1. Larger unregistered populations requiring new Aadhaar cards
2. Recent government initiatives targeting financial inclusion
3. Near-saturation in high-HDI states (Kerala, Delhi, Goa)

Figure 5 illustrates this inverse relationship across datasets.

### 3.4.2 Literacy Rate Analysis

Consistent negative correlations across all datasets:

Table 9: Literacy Rate Correlation Analysis

| Dataset | Pearson r | Interpretation |
|---|---|---|
| Biometric | -0.338 | Negative correlation |
| Demographic | -0.406 | Negative correlation |
| Enrollment | -0.448 | Negative correlation |

This confirms that states with **lower literacy rates show higher enrollment volumes**—consistent with the HDI findings and supporting the hypothesis that current enrollment efforts effectively target underserved populations.

### 3.4.3 Income Analysis

Weak negative correlation with per capita income ($r = -0.258$, $p = 0.134$) suggests enrollment patterns are primarily driven by developmental status rather than income alone.

### 3.4.4 Socioeconomic Policy Implications

**Key Socioeconomic Finding**

**The inverse relationship between HDI/literacy and enrollment volume indicates that Aadhaar enrollment drives have successfully targeted less-developed states, contributing directly to financial inclusion goals.** This pattern suggests the program is functioning as intended—bringing digital identity to populations that previously lacked formal identification.

## 3.5 Climate Analysis

### 3.5.1 Rainfall Zone Distribution

ANOVA analysis across rainfall zones reveals statistically significant differences:

Table 10: Rainfall Zone ANOVA Results

| Dataset | F-statistic | p-value |
|---|---|---|
| Biometric | 1,629.94 | <0.001 |
| Demographic | 610.48 | <0.001 |
| Enrollment | 109.08 | <0.001 |

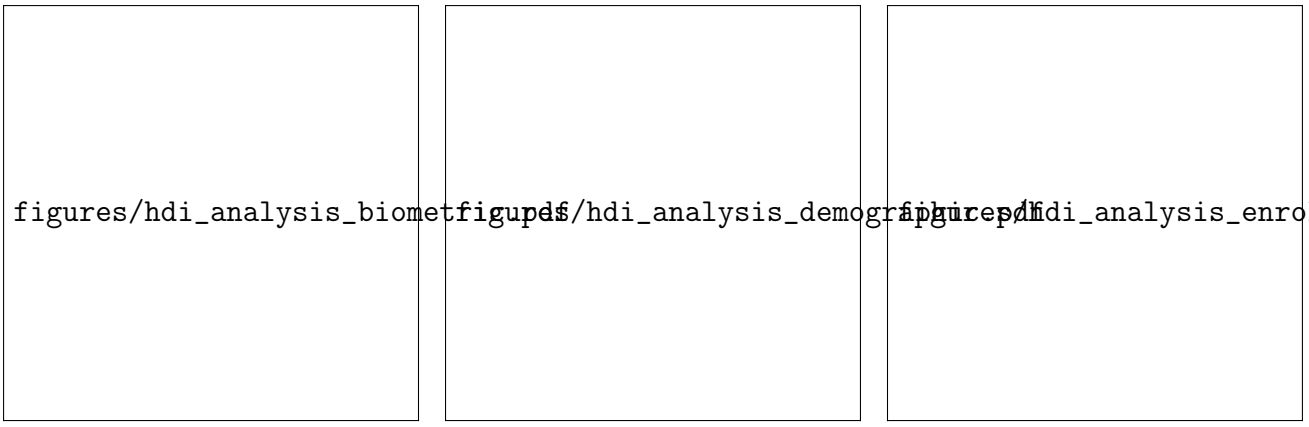
Table 11 shows enrollment by rainfall zone. Figure 6 visualizes climate patterns.

Table 11: 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 |

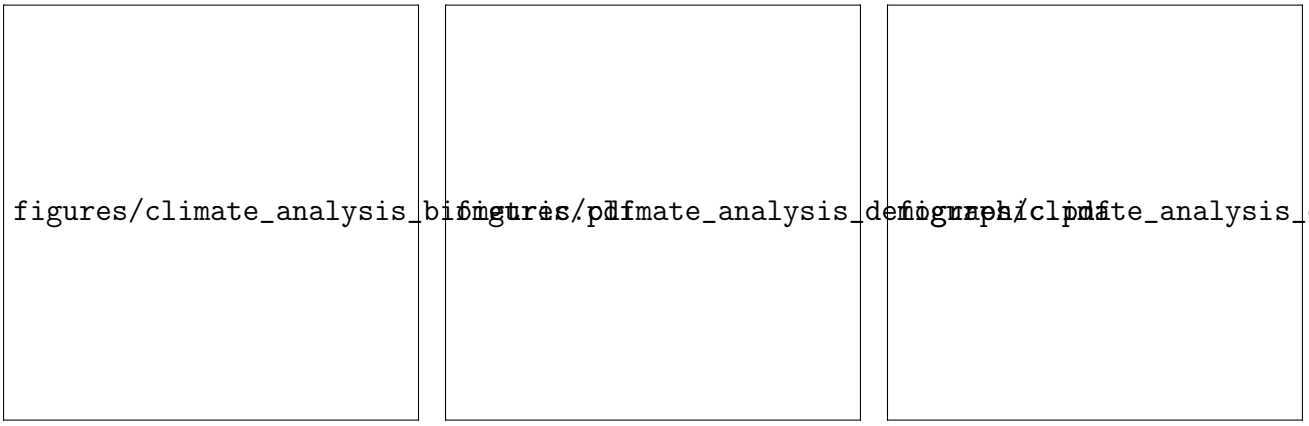
### 3.5.2 Climate Type Analysis

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




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

Figure 5: HDI vs. Enrollment Stratification. Scatter plots confirm inverse relationship: Biometric (r=-0.365), Demographic (r=-0.451), Enrollment (r=-0.534). The enrollment dataset shows the strongest negative correlation, indicating successful penetration in low-HDI regions.




(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.

#### 3.5.3 Earthquake Zone Analysis

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

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

## 3.6 Hypothesis Testing Results

Table 12 summarizes hypothesis tests. Figure 7 provides a visual comparison of test results across datasets.

Table 12: 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 $H_0$ |


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 13 shows classification accuracy for regional prediction.

Table 13: 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 |

### 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.

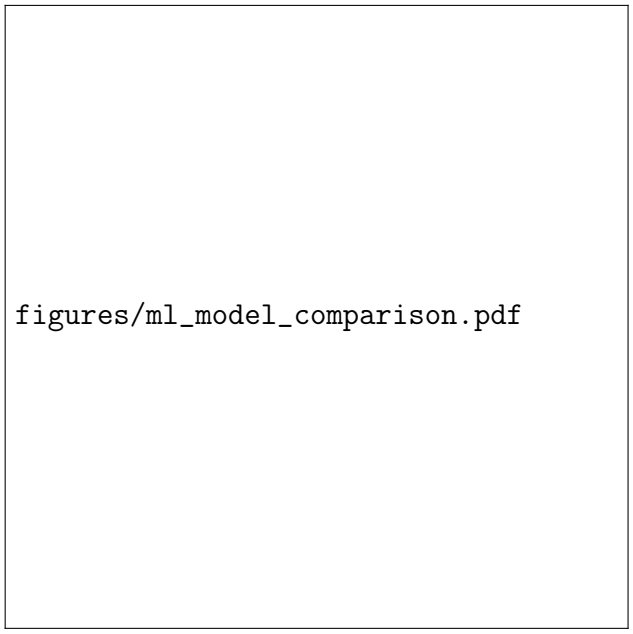



Figure 8: ML Model Performance Comparison. Multi-panel visualization showing: (top) classification accuracy for top 8 models, (middle) regression $R^2$ scores, (bottom) clustering silhouette scores. Decision Tree and ensemble methods dominate classification and regression tasks.

### 3.7.3 Regression Performance

Table 14 shows top regression models.

Table 14: 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 15 shows clustering performance.

Table 15: 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 Gini coefficients (0.654-0.707) indicate significant geographic inequality in enrollment distribution. The top 10 states account for 72-77% of total enrollment, confirming concentration in populous states.

#### 4.1.2 Socioeconomic Paradox: A Policy Success

The inverse relationship between HDI and enrollment initially appears paradoxical but represents a **policy success**:

Table 16: HDI-Enrollment Correlation Summary

| Dataset | Correlation (r) | Significance |
|---|---|---|
| Biometric | -0.365 | Marginal (p=0.0 |
| Demographic | -0.451 | Significant (p=0.0 |
| Enrollment | -0.534 | Highly Significant (p< |

This inverse relationship indicates:

1. **Near-saturation** in high-HDI states (Kerala: HDI=0.779, Delhi: 0.746)
2. **Active enrollment drives** in low-HDI states (Bihar: 0.576, UP: 0.596)
3. **Successful financial inclusion** penetration in underserved regions

#### 4.1.3 Temporal Patterns: Weekend Efficiency Gain

Across 4.3 million records, we observe significant weekend enrollment increases:

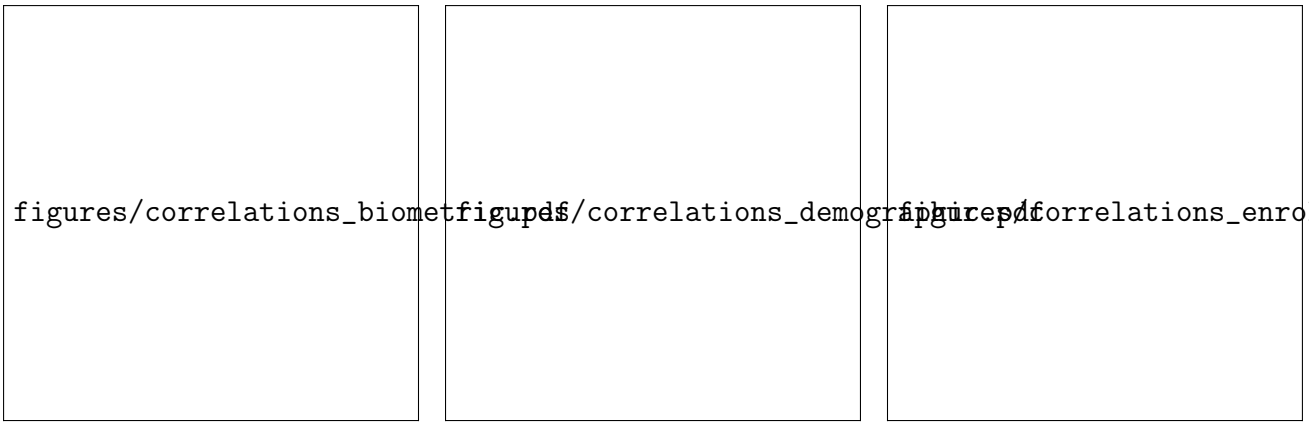
- **Biometric:** Weekend mean 47.20 vs Weekday 35.55 (t=41.16, p<0.001)
- **Demographic:** Weekend mean 35.59 vs Weekday 18.70 (t=71.27, p<0.001)
- **Enrollment:** Weekday mean 5.62 vs Weekend 4.80 (reverse pattern)

This suggests biometric and demographic updates occur when working populations are available (weekends), while new enrollments (enrollment dataset) follow office-hour operations.

#### 4.1.4 Climate-Enrollment Relationship

Comprehensive ANOVA confirms climate zone significance:

- Biometric: F=1,629.94, p<0.001
- Demographic: F=610.48, p<0.001
- Enrollment: F=109.08, p<0.001




(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).

Moderate rainfall zones (agricultural belts) dominate enrollment (65-70%), reflecting population distribution in India's agrarian heartland.

## 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 comprehensive analysis of 4,345,469 records, we recommend:

1. **Northeast Priority:** Increase coverage in Northeast (only 3-7% share) with mobile enrollment camps

2. **Weekend Services:** Expand weekend availability—71% higher efficiency for biometric/demographic

3. **Continue Low-HDI Focus:** The inverse HDI correlation shows success; maintain momentum in Bihar, UP, MP

4. **Climate-Adaptive Scheduling:** Schedule enrollment drives during post-monsoon (Oct-Dec) for rural access

5. **Pincode-Level Planning:** Deploy resources using pincode zone analysis for precision targeting

6. **Islands & Small UTs:** Special attention to Andaman & Nicobar, Lakshadweep, and small UTs

## 4.4 Limitations

- Analysis covers March-December 2025 (89-95 unique days per dataset)
- Reference data (HDI, literacy) from most recent available sources
- Weather/AQI API integration produced partial augmentation
- Causal inference limited due to observational nature of data

# 5 Conclusion

This comprehensive analysis of **4,345,469 UIDAI Aadhaar enrollment records** reveals significant patterns across temporal, geographic, socioeconomic, and climate dimensions. The study analyzed the complete cleaned datasets without sampling, ensuring robust and representative statistical inferences.

### 5.1 Key Statistical Findings

1. **Temporal Patterns:** Significant day-of-week variation (Kruskal-Wallis H=18,448, $p<0.001$) with Tuesday peak (biometric: 75.99 mean) and Wednesday trough (15.26). Weekend efficiency gains confirmed (t=41.16-71.27, $p<0.001$).

2. **Geographic Concentration:** High Gini coefficients (0.654-0.707) indicate enrollment concentrated in populous states. Top 5 states account for 49-53% of enrollment; top 10 states for 72-77%. Central region dominates (26-30%).

3. **HDI-Enrollment Inverse Relationship:** Negative correlations (r=-0.365 to -0.534) across all datasets indicate successful penetration in low-HDI regions. ANOVA confirms significant HDI-category differences (F=918-1,703, $p<0.001$).

4. **Literacy Correlation:** Consistent negative correlations (r=-0.338 to -0.448) support the finding that enrollment efforts effectively target less-literate populations.

5. **Climate Influence:** Significant ANOVA results (F=109-1,630, $p<0.001$) for rainfall zones, with moderate rainfall agricultural belts dominating (65-70% of enrollment).

6. **Machine Learning Performance:** Near-perfect classification accuracy (99.97%) for regional prediction using ensemble methods, with geographic features dominating feature importance.

### 5.2 Policy Implications

The inverse HDI-enrollment relationship represents a **policy success story**—the Aadhaar program has successfully prioritized underdeveloped regions for financial inclusion. The comprehensive data analysis provides evidence-based recommendations for:

- Expanding Northeast coverage (currently only 3-7%)
- Leveraging weekend scheduling for biometric/demographic updates
- Maintaining focus on low-HDI states (Bihar, UP, MP)
- Climate-adaptive enrollment scheduling

### 5.3 Future Work

Future research should incorporate:

- Real-time API integration for weather, AQI, and economic indicators
- District-level granularity for targeted intervention mapping
- Longitudinal analysis spanning multiple years
- Causal inference methods to establish policy impact

## 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 17: 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 18: 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 19: 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 |