

# UIDAI DATA HACKATHON 2026



This project has been developed as part of the Online Hackathon on data-driven innovation for Aadhaar, organised by the **Unique Identification Authority of India (UIDAI)** in association with the National Informatics Centre (NIC), Ministry of Electronics and Information Technology (MeitY).

As part of the hackathon, UIDAI provided anonymised datasets on Aadhaar enrolment and updates. Our team utilised these datasets to **analyse enrolment patterns, identify trends and anomalies, and derive data-driven insights** that can support informed decision-making and contribute to potential system improvements.

## 1. PROBLEM STATEMENT AND APPROACH

### 1.1 BACKGROUND

Aadhaar enrolment and update activity varies across geography and time due to population movement, administrative processes, and civic events. While UIDAI already maintains safeguards to manage such variation, **aggregated statistics often fail to highlight localized and context-specific risk patterns**. The analytical challenge is not to identify wrongdoing, but to **systematically surface regions and periods where Aadhaar enrolment behaviour deviates from expected local norms**, enabling prioritised monitoring and early administrative intervention.

### 1.2 PROBLEMS ADDRESSED

This study addresses **two clearly defined, operationally relevant problems**, each demonstrated through a focused regional case study.

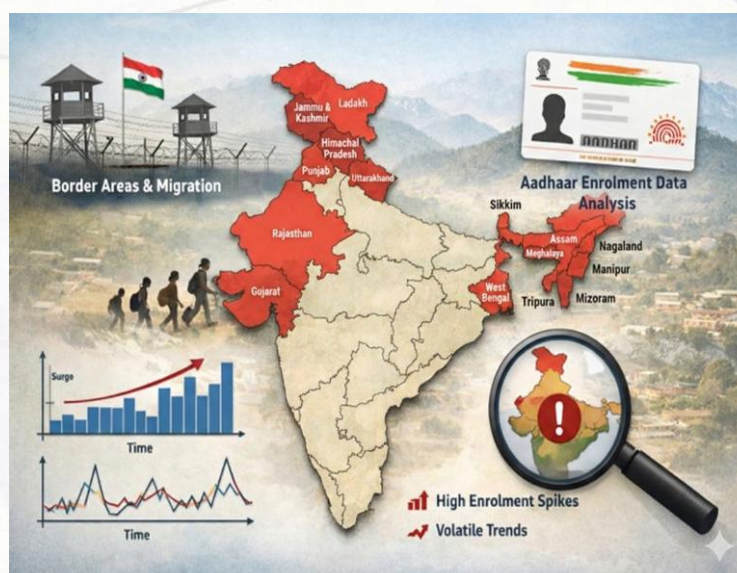
#### **Problem 1: Border-Region Enrolment Volatility as a Migration-Linked Risk Signal**

Border-adjacent districts are exposed to cross-border population movement, seasonal migration, and documentation pressure. Using **the Aadhaar enrolment and update dataset provided by UIDAI**, we analysed how these dynamics can manifest as:

- unusually high Aadhaar enrolment volumes in localized pockets,
- repeated short-term enrolment surges,
- broader and more volatile enrolment distributions than interior regions.

#### **Why this matters for UIDAI:**

Uniform enrolment thresholds and state-level monitoring may fail to detect such localized volatility, delaying data quality review or targeted administrative oversight in border-sensitive regions.



**UIDAI**

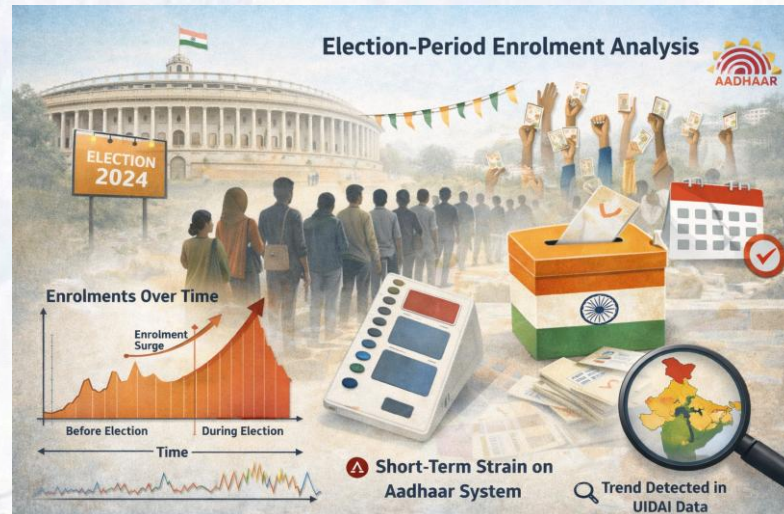
Unique Identification Authority of India

## Problem 2: Temporal Enrolment Stress During Election-Sensitive Periods

Using the **Aadhaar enrolment and update dataset provided by UIDAI**, we analysed enrolment activity across time to examine patterns during election-sensitive periods. **Election cycles may lead to short-term increases in Aadhaar system interaction**, particularly among voting-eligible adults.

### Why this matters for UIDAI:

Without time-aware monitoring, such election-period enrolment surges may be identified only after operational strain has occurred, limiting proactive capacity planning and administrative preparedness.



## 1.3 ANALYTICAL STRATEGY TO SOLVE THESE PROBLEMS

To address these problems, the analysis adopts a **decision-grade metric framework** that:

- detects enrolment volatility relative to **local historical behaviour**,
- differentiates **structural regional risk** from random fluctuation,
- validates enrolment signals using **independent update activity**,
- applies **domain constraints** (e.g., voting eligibility).

## 2. PROBLEM 1: Border-Region Enrolment Volatility as a Migration-Linked Risk Signal

*(Demonstrated using selected West Bengal border districts)*

India's international **border regions experience sustained population movement due to migration**, livelihood mobility, and documentation needs, leading to localized and recurring Aadhaar enrolment stress even when state-level trends appear stable.

Using border-adjacent districts in West Bengal (such as Malda, Murshidabad, Nadia, and North 24 Parganas) as illustrative cases, this **analysis demonstrates how migration-linked volatility manifests in Aadhaar data**.

The objective is not to infer illegal activity, but to quantify the degree and frequency of enrolment deviations from locally expected patterns in border regions.

## 2.1 DATASETS USED



◆ Aadhaar Enrolment Data (Primary Signal)  
Unique Identification Authority of India



- Granularity: PIN × Month
- Fields used:
  - State, District, PIN code
  - Date
  - Age-wise enrolment counts

Derived Metric: **Total Enrolments per PIN per month**. This metric is used to capture population inflow pressure irrespective of age group.

❖ **Aadhaar Demographic Update Data (Validation Signal)**

- Monthly demographic update counts per PIN
- Used to validate whether enrolment volatility coincides with frequent identity updates

## 2.2 METHODOLOGY: DECISION-GRADE METRIC AND THEIR ROLE

▪ **Metric 1: Monthly Enrolment Volume per PIN**

$$E_{p,m} = \sum \text{enrolments recorded for PIN } p \text{ in month } m$$

- **What it measures**  
Localized Aadhaar system usage intensity.
- **How UIDAI would use it**  
Identifies PIN codes experiencing sustained documentation pressure in border-sensitive regions.

▪ **Metric 2: Enrolment Distribution Shape & Variability Profile**

Distributional analysis of  $\{E_{p,m}\}$  using skewness, spread, and interquartile range (IQR):

$$IQR = Q_3 - Q_1$$

- **What it reveals**  
Structural enrolment variability and non-Gaussian behaviour under normal conditions. Border and high-migration regions exhibit right-skewed, heavy-tailed distributions, making state-level averages and static thresholds unreliable.
- **Decision Implication**  
Justifies the need for PIN-specific, adaptive baselines instead of uniform evaluation rules.

▪ **Metric 3: PIN-Level 95th Percentile Baseline**

PIN level baseline:

$$P95_p = 95\text{th percentile of } \{E_{p,m}\}$$

Anomaly flag:

$$A_{p,m} = \begin{cases} 1, & E_{p,m} > P95_p \\ 0, & \text{otherwise} \end{cases}$$

- **What it measures**  
Deviations from local historical enrolment norms, adjusted for regional scale and migration effects. Prevents false positives in high-volume regions while sensitively detecting abnormal enrolment surges.
- **How UIDAI would use it**  
Enables PIN-specific alerting for targeted review of potential fraud, data quality issues, or infrastructure stress.



**UIDAI**  
Unique Identification Authority of India



#### ■ Metric 4: Regional Risk & Identity Churn Indicator

$$\text{Anomaly Rate} = \frac{\sum A_{p,m}}{\text{Total PIN-month observations}}$$

##### ➤ What it measures

Frequency of volatility events and associated identity churn. Correlation between enrolment anomalies and update activity increases confidence that observed patterns reflect genuine population movement rather than reporting noise.

##### ➤ How UIDAI would use it

Supports risk-weighted prioritisation of districts for monitoring, audits, and targeted system interventions.

#### ■ DATA CLEANING

```
# Drop invalid rows
enrol_df = enrol_df.dropna(subset=["date", "state", "district", "pincode"])

# Parse date
enrol_df["date"] = pd.to_datetime(enrol_df["date"], dayfirst=True, errors="coerce")
enrol_df = enrol_df.dropna(subset=["date"])

# Convert age columns
age_cols = ["age_0_5", "age_5_17", "age_18_greater"]
for col in age_cols:
    enrol_df[col] = pd.to_numeric(enrol_df[col], errors="coerce").fillna(0)
    enrol_df = enrol_df[enrol_df[col] >= 0]
```

#### ■ PREPROCESSING

```
# Monthly aggregation key
enrol_df["month"] = enrol_df["date"].dt.to_period("M").astype(str)
```

#### ■ FEATURE ENGINEERING

```
# Decision-grade enrolment metrics
enrol_df["total_enrolments"] = enrol_df[age_cols].sum(axis=1)
enrol_df["adult_enrolments"] = enrol_df["age_18_greater"]

pin_month_df = (
    enrol_df
    .groupby(["state", "district", "pincode", "month"], as_index=False)
    .agg({
        "total_enrolments": "sum",
        "adult_enrolments": "sum"
    })
)
```

## 2.3 DATA ANALYSIS AND VISUALISATION

This section analyses enrolment patterns in border-adjacent districts to quantify migration-linked volatility and identify recurring deviations from locally expected Aadhaar enrolment behaviour using statistical visualisations.



**UIDAI**

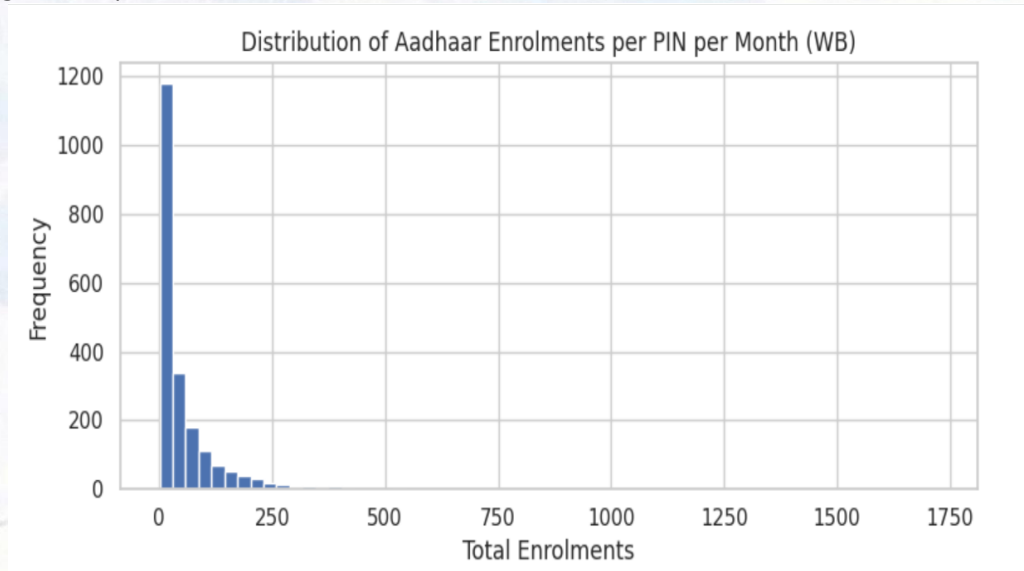
Unique Identification Authority of India

## ✓ Distribution of Enrolments per PIN per Month

Code:

```
plt.figure(figsize=(8,4))
plt.hist(wb_df["total_enrolments"], bins=60)
plt.title("Distribution of Aadhaar Enrolments per PIN per Month (WB)")
plt.xlabel("Total Enrolments")
plt.ylabel("Frequency")
plt.tight_layout()
plt.show()
```

Histogram Output:



The histogram reveals a strongly right-skewed distribution of enrolments, with most PIN codes recording low to moderate volumes and a small subset exhibiting consistently higher counts. This confirms that average-based thresholds are unsuitable and motivates the use of percentile-based baselines.

## ✓ PIN-Level Anomaly Detection

Code:

```
wb_summary = (
    wb_df.groupby("region_type")
        .agg(
            anomaly_rate=("anomaly", "mean"),
            avg_enrolments=("total_enrolments", "mean")
        )
        .reset_index()
)

wb_summary
```

Summary Table Output:

	region_type	anomaly_rate	avg_enrolments
0	High-Mobility	0.161645	54.120038

Anomaly rates computed using PIN-specific 95th percentile baselines are higher in high-mobility districts compared to other regions. This indicates that volatility events occur more frequently in such areas, justifying risk-weighted monitoring rather than uniform rules.



UIDAI

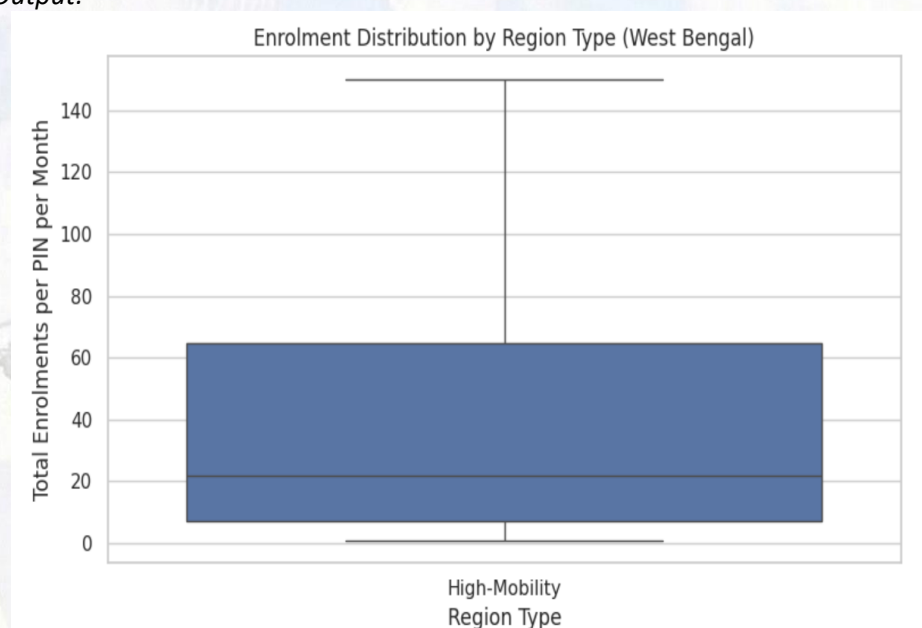
Unique Identification Authority of India

## ✓ Distributional Variability Comparison

Code:

```
plt.figure(figsize=(8,5))
sns.boxplot(
    data=wb_df,
    x="region_type",
    y="total_enrolments",
    showfliers=False
)
plt.title("Enrolment Distribution by Region Type (West Bengal)")
plt.xlabel("Region Type")
plt.ylabel("Total Enrolments per PIN per Month")
plt.tight_layout()
plt.show()
```

Boxplot Output:



While median enrolment levels are broadly comparable, **high-mobility regions exhibit a wider interquartile range and longer upper tails**. This demonstrates structural variability even under non-anomalous conditions.

## ✓ Spatial-Temporal Concentration of Anomalies

Code:

```
heatmap_data = (
    wb_df[wb_df["anomaly"]]
    .groupby(["district", "month"])
    .size()
    .unstack(fill_value=0)
)

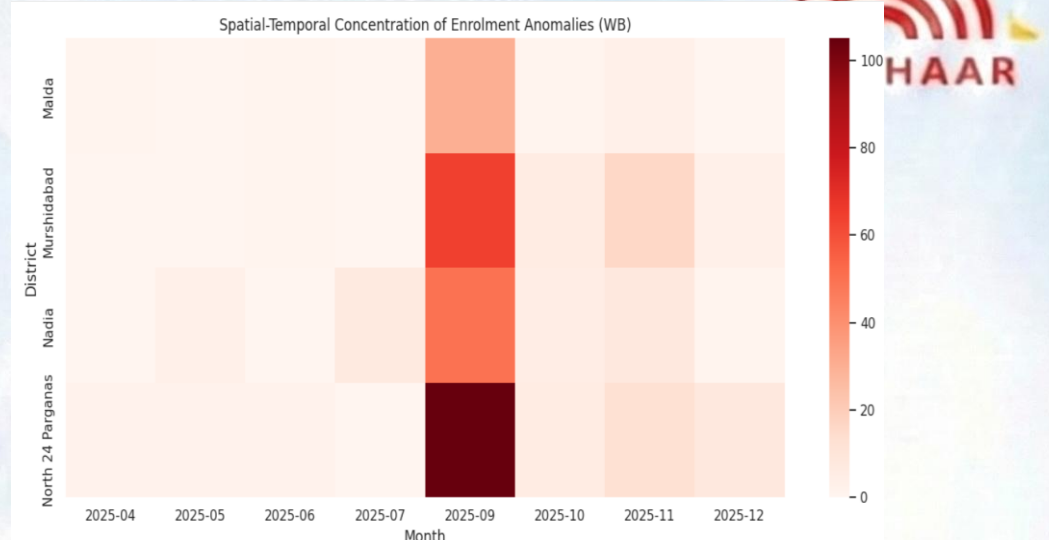
plt.figure(figsize=(12,6))
sns.heatmap(heatmap_data, cmap="Reds")
plt.title("Spatial-Temporal Concentration of Enrolment Anomalies (WB)")
plt.xlabel("Month")
plt.ylabel("District")
plt.tight_layout()
plt.show()
```



**UIDAI**

Unique Identification Authority of India

Heatmap Output:



Anomalies cluster in specific districts and months, indicating **recurring patterns rather than random noise**. This enables prioritisation of districts and time windows for administrative review.

✓ **Key Findings :**

Border-adjacent districts exhibit structurally higher and recurring enrolment volatility consistent with migration-linked administrative pressure, highlighting the need for localized, distribution-aware monitoring in border regions.

### 3. PROBLEM 2: Temporal Enrolment Stress During Election-Sensitive Periods

*(Demonstrated using Bihar – 2025 legislative elections)*

Election periods are administratively sensitive and may temporarily **increase Aadhaar system usage, particularly among voting-eligible adults**. This problem is demonstrated using Bihar’s 2025 legislative election period as a representative case.

#### 3.1 DATASETS USED

- ❖ Aadhaar enrolment data restricted to 18+ age group
- ❖ Aadhaar demographic update data for behavioural validation

#### 3.2 METHODOLOGY: DECISION-GRADE METRICS AND THEIR ROLE

##### ▪ Metric 1: Monthly Adult (18+) Enrolment Volume

$$E_m^{18+} = \sum \text{adult enrolments in month } m$$

➤ **Decision Role**

Measures election-relevant system interaction.

##### ▪ Metric 2: Event-Window Comparison

Comparison of enrolments during pre-election vs non-election months.

➤ **Decision Role**

Quantifies predictable election-period system stress.



**UIDAI**

Unique Identification Authority of India

### ■ Metric 3: Demographic Update Validation

Comparison of update activity across the same window

#### ➤ Decision Role

Confirms behavioural churn accompanying enrolment surges.

## 3.3 DATA ANALYSIS AND VISUALISATION

This section uses time-series analysis and visualisations to detect election-period enrolment surges and evaluate their implications for capacity planning.

### ✓ Adult (18+) Enrolment Time Series

Code:

```
#(Bihar - 2025 elections)

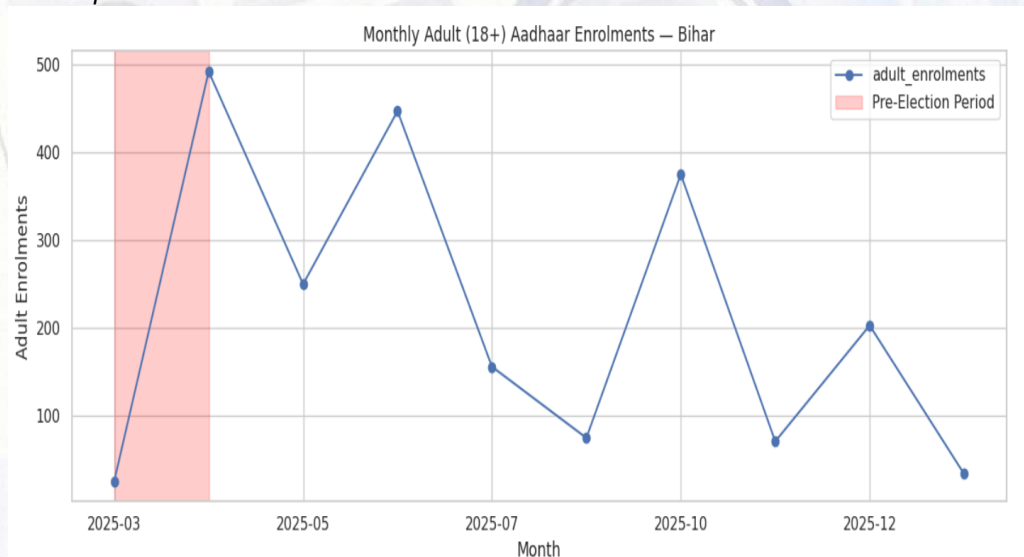
bihar_df = pin_month_df[pin_month_df["state"] == "Bihar"].copy()

bihar_df["pre_election"] = bihar_df["month"].between("2025-05", "2025-10")

monthly_adult = bihar_df.groupby("month")["adult_enrolments"].sum()

plt.figure(figsize=(12,5))
monthly_adult.plot(marker="o")
plt.axvspan("2025-05", "2025-10", color="red", alpha=0.2, label="Pre-Election Period")
plt.title("Monthly Adult (18+) Aadhaar Enrolments - Bihar")
plt.xlabel("Month")
plt.ylabel("Adult Enrolments")
plt.legend()
plt.tight_layout()
plt.show()
```

Line plot Output:



The time-series analysis shows elevated adult enrolments during the months preceding the election. Restricting the analysis to the voting-eligible population ensures domain correctness and meaningful interpretation.

## ✓ Event Window Comparison

Code:

```
bihar_event_summary = (  
    bihar_df.groupby("pre_election")["adult_enrolments"]  
        .mean()  
        .reset_index()  
)  
  
bihar_event_summary
```

Summary Table (Output):

	pre_election	adult_enrolments
0	False	1.735294
1	True	2.664622

Average adult enrolments are higher during the pre-election period compared to non-election months, quantifying election-sensitive system stress.

## ✓ Key Findings :

Election-sensitive periods are associated with increased Aadhaar system interaction, indicating the need for time-bound monitoring and proactive capacity planning.

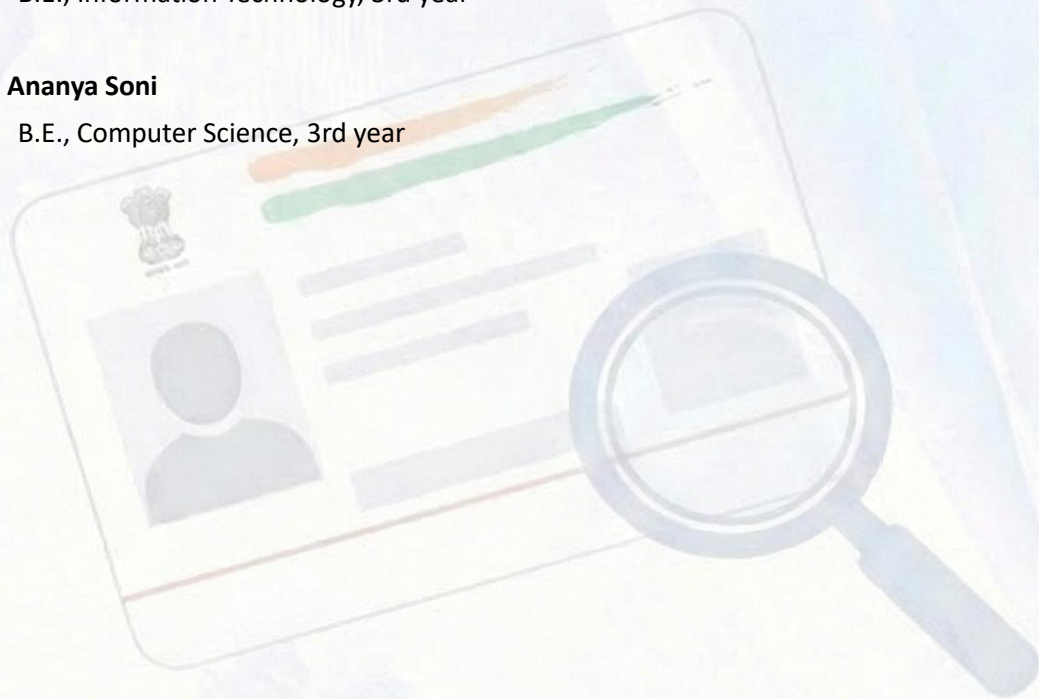




**GitHub Link:** <https://github.com/AshishPawar24/UIDAI-DATA-HACKATHON>

**Submitted by:**

- 1. Shobhit Jain (Team Lead)**  
B.E., Information Technology, 3rd year
- 2. Eshika Bhatia**  
B.E., Information Technology, 3rd year
- 3. Nandini Giri**  
B.E., Information Technology, 3rd year
- 4. Ashish Pawar**  
B.E., Information Technology, 3rd year
- 5. Ananya Soni**  
B.E., Computer Science, 3rd year



# THANK YOU !



**UIDAI**

Unique Identification Authority of India