UIDAI DATA HACKATHON 2026

Theme: Data-Driven Innovation for Aadhaar

Identity Stress Early-Warning System (ISEWS)

A Predictive Framework for Enhancing Operational Resilience in the Aadhaar Ecosystem

SUBMITTED BY:

Anuska Ghosh (Team Lead)

Likitha S

PROJECT RESOURCES:

https://colab.research.google.com/drive/1QRjL53W0rm7dqqVsLumnPAJGuhIBmyzL?usp=sharing

https://colab.research.google.com/drive/1YrcsyTBPymu25FoxemHfnT4oLcUvpApU?usp=sharing

https://colab.research.google.com/drive/12IjU_WyufG_wKnsgJIbmS13i3eVpFeKU?usp=sharing

https://colab.research.google.com/drive/1xLOO9fs5r0gh8ojZAR8rFKelJWBk0Ml1?usp=sharing

LIVE SYSTEM DEMO: https://uidai-dashboard.streamlit.app

Executive Summary

The Aadhaar ecosystem serves as India's foundational digital identity infrastructure, enabling governance, service delivery, and financial inclusion at national scale. While Aadhaar enrolments establish identity, demographic and biometric updates reflect continuous, lifecycle-driven changes that—if not monitored proactively—can introduce operational strain and affect system trust.

This project addresses UIDAI's challenge by reframing Aadhaar updates not merely as routine transactions, but as early signals of identity system stress. Using Aadhaar enrolment, demographic update, and biometric update datasets provided by UIDAI, we develop an early-warning analytical framework to identify emerging patterns, regional imbalances, and persistent update pressure across time, geography, and age groups.

The proposed approach integrates:

- disciplined data preparation and aggregation,
- multi-level pattern and trend analysis,
- explainable rule-based early-warning indicators, and
- machine-learning–based anomaly ranking for prioritisation.

The resulting decision-support analytical view enables UIDAI to act earlier, allocate resources more effectively, and redesign update processes more strategically, thereby strengthening Aadhaar's long-term operational resilience and reinforcing public trust in the ecosystem.

1. Problem Statement and Approach

1.1 UIDAI Problem Statement

UIDAI invites participants to identify meaningful patterns, trends, anomalies, or predictive indicators from Aadhaar datasets and translate them into clear insights or solution frameworks that support informed decision-making and system-level improvements.

1.2 Approach

Aadhaar functions as a **living identity system**. As individuals age, migrate, and experience biometric change, the system undergoes continuous demographic and biometric updates. While such updates are expected, **persistent or uneven update pressure** can signal deeper systemic issues, including:

- operational strain on enrolment and update infrastructure,
- demographic transitions and population mobility,
- biometric ageing and re-capture effects, and
- potential data quality or consistency vulnerabilities.

At present, many of these pressures are observed **reactively**, after service degradation or authentication challenges emerge, rather than being detected and addressed early.

1.3 Reframed Problem Statement (Our Contribution)

How can UIDAI **proactively detect early indicators of identity system stress** across regions, age groups, and time, and use these signals to intervene before sustained update pressure impacts Aadhaar's operational reliability and public trust?

2. Datasets Used

2.1 Aadhaar Enrolment Dataset

- Records Aadhaar enrolments by date, state, district, pincode, and age group (0–5, 5–17, 18+).
- Represents the **base identity population** against which update activity is contextualised.

2.2 Aadhaar Demographic Update Dataset

- Records demographic updates by date, geography (state, district, pincode), and age group (5–17, 17+).
- Reflects **identity attribute churn** driven by lifecycle changes such as address updates, corrections, and mobility.

2.3 Aadhaar Biometric Update Dataset

- Records biometric updates by date, geography (state, district, pincode), and age group (5–17, 17+).
- Captures **biometric ageing and revalidation pressure** within the Aadhaar ecosystem.

2.4 Analytical Dimensions

The datasets were analysed along three core dimensions:

- **When:** Monthly time series analysis to capture temporal patterns and persistence,
- **Where:** State- and district-level aggregation to identify regional variation,
- **Who:** Age-group segmentation to understand lifecycle-driven update behaviour.

Together, these datasets enable Aadhaar update activity to be interpreted **relative to enrolment scale**, rather than as isolated or absolute counts.

3. Methodology (CRISP-DM Aligned)

3.1 Business Understanding

Aadhaar system stability depends not only on enrolment coverage, but also on the **frequency and distribution of identity updates** required over time. Persistent or uneven update activity can signal operational strain or emerging identity risks. Monitoring update pressure is therefore essential for sustaining Aadhaar reliability and public trust.

3.2 Data Understanding

The provided Aadhaar datasets exhibit the following characteristics:

- large-scale volume,
- distribution across multiple source files,
- fine-grained temporal resolution,
- hierarchical geographic structure (state–district–pincode), and
- statistically skewed distributions typical of national-scale administrative data.

Understanding these properties informed subsequent preprocessing, aggregation, and analytical choices.

3.3 Data Preparation (Layer 3)

The data preparation stage focused on creating a consistent, analysis-ready foundation:

- All datasets were cleaned and standardised.
- Daily records were aggregated into **monthly time buckets** to reduce noise and enable trend analysis.
- Datasets were merged at the **month × state × district** level.
- Key derived metrics included:
    - total enrolments,
    - total demographic updates,
    - total biometric updates, and
    - update-to-enrolment ratios to normalise activity across regions.

3.4 Analysis and Signal Extraction (Layer 4)

Using the prepared datasets, multiple analytical signals were extracted:

- national baseline trends to establish reference behaviour,
- state-level deviations from national averages,
- district-level stress hotspots,
- age-group contribution and pressure patterns, and
- temporal momentum through month-on-month change analysis.

3.5 Early-Warning Indicators and ML Reinforcement (Layer 5)

To move from descriptive analysis to early-warning detection:

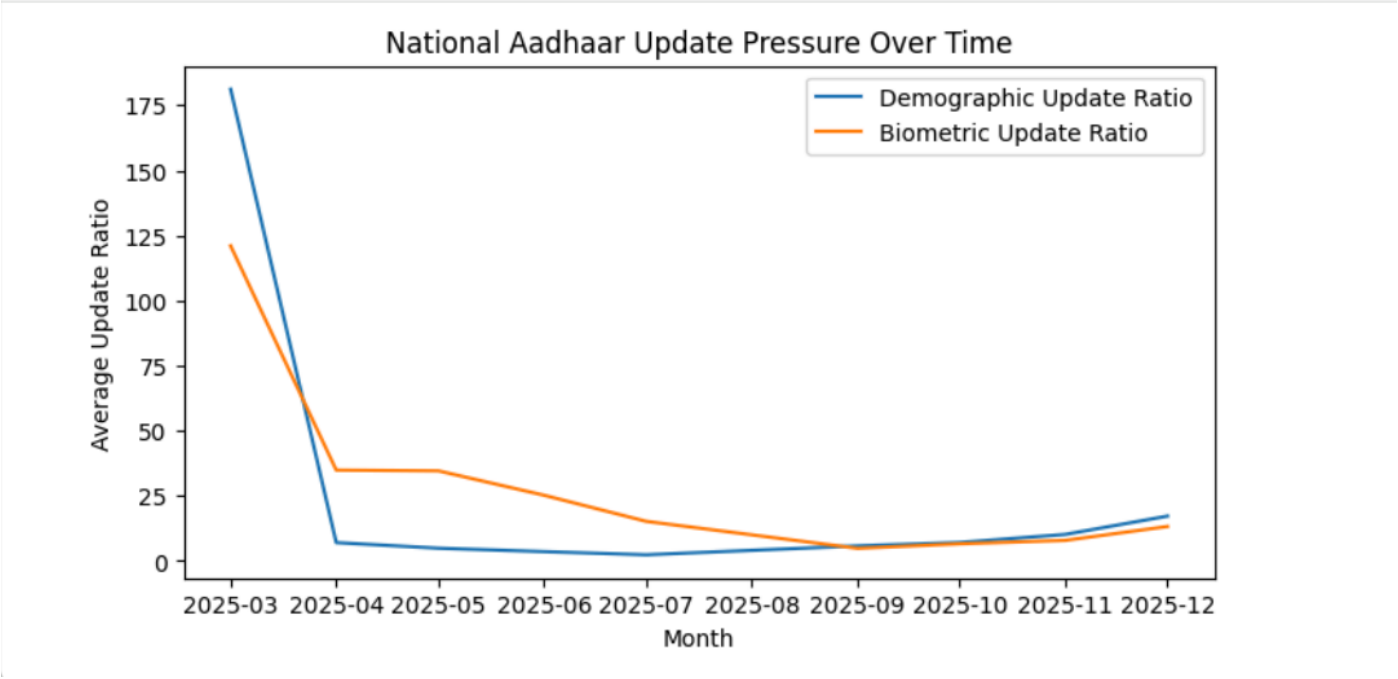
- rule-based indicators were defined using percentile-based thresholds,
- a composite **identity stress flag** was constructed,
- Isolation Forest was applied to rank regions by anomaly severity, and
- correlation analysis was used to validate alignment between explainable rules and ML-based scores.

3.6 Visualisation and Storytelling (Layer 6)

Analytical outputs were translated into **decision-oriented visualisations** and a consolidated early-warning snapshot, designed to support prioritisation and governance-level interpretation.

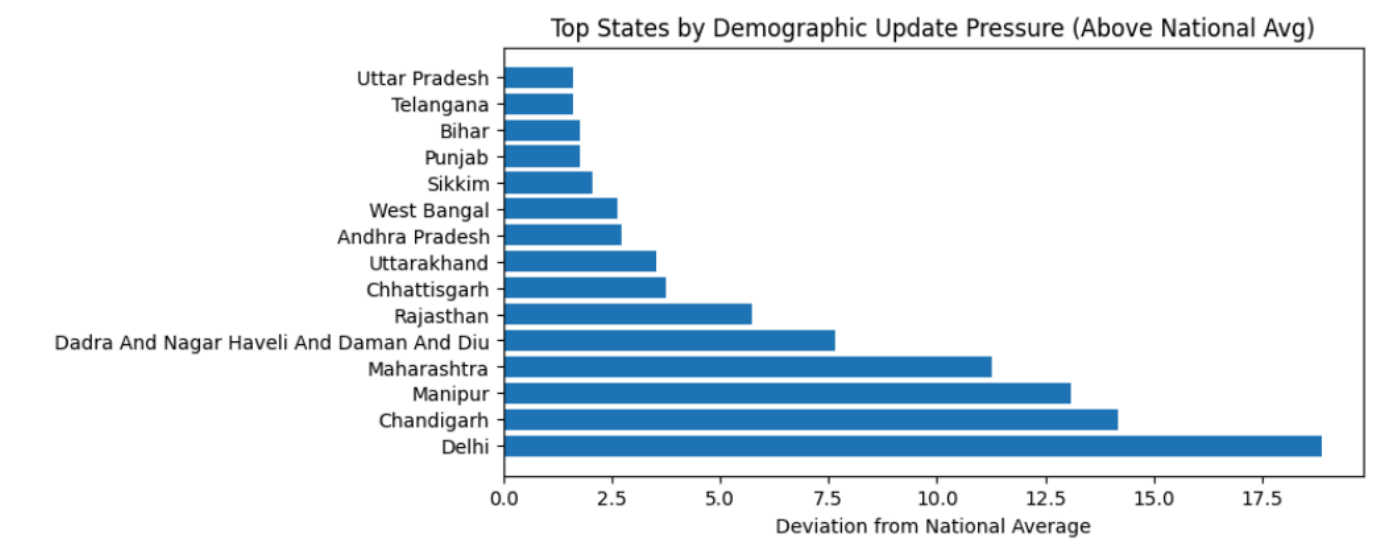
4. Key Insights and Findings

4.1 National Aadhaar Update Pressure Trend



Insight:
National Aadhaar update pressure exhibits an initial surge followed by stabilisation and a gradual upward trend, indicating that identity update activity evolves over time rather than remaining static. This highlights the need for continuous monitoring instead of purely reactive responses to short-term spikes.

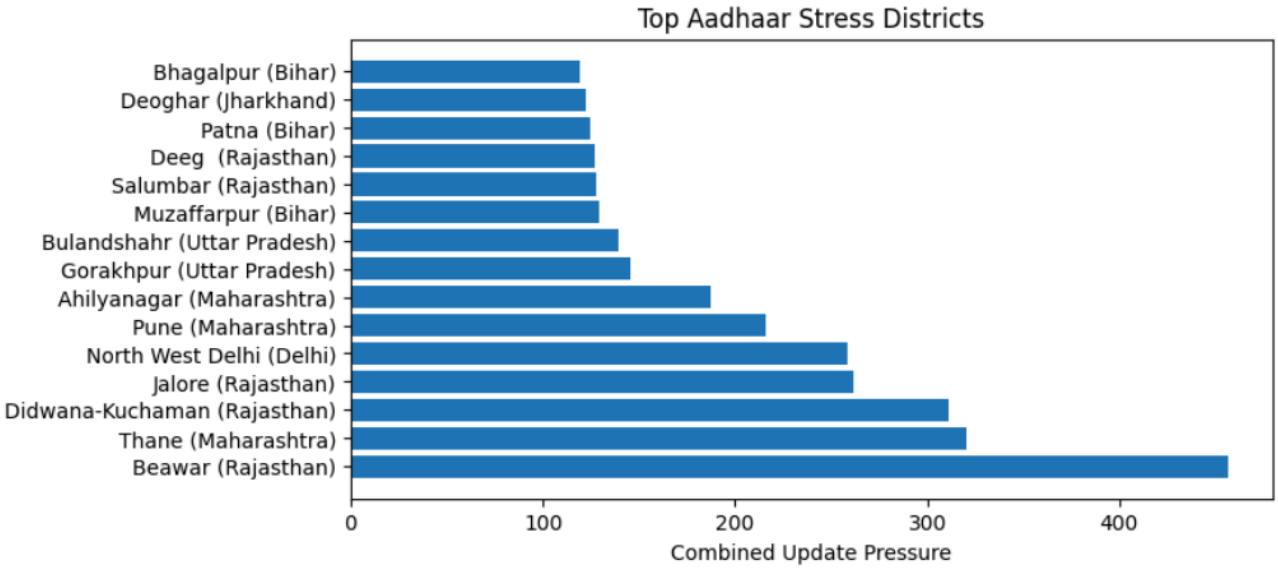
4.2 State-Level Deviation from National Baseline



Insight:

Demographic update pressure is unevenly distributed across states, with a small number of states and union territories consistently exceeding the national baseline. This concentration indicates region-specific demographic and operational dynamics, rather than uniform system-wide behaviour, and highlights the need for targeted, state-level monitoring instead of a one-size-fits-all approach.

4.3 District-Level Hotspots



Insight:

A small set of districts exhibits disproportionately high Aadhaar update pressure compared to other districts nationwide. These district-level hotspots cut across multiple states, indicating that identity system stress can be highly localised and may not always be fully explained by state-level averages alone.

4.4 Prototype Insight:

Geographic "Stress Belts" To demonstrate the value of spatial monitoring, we generated a prototype geospatial heatmap using the processed validation data.

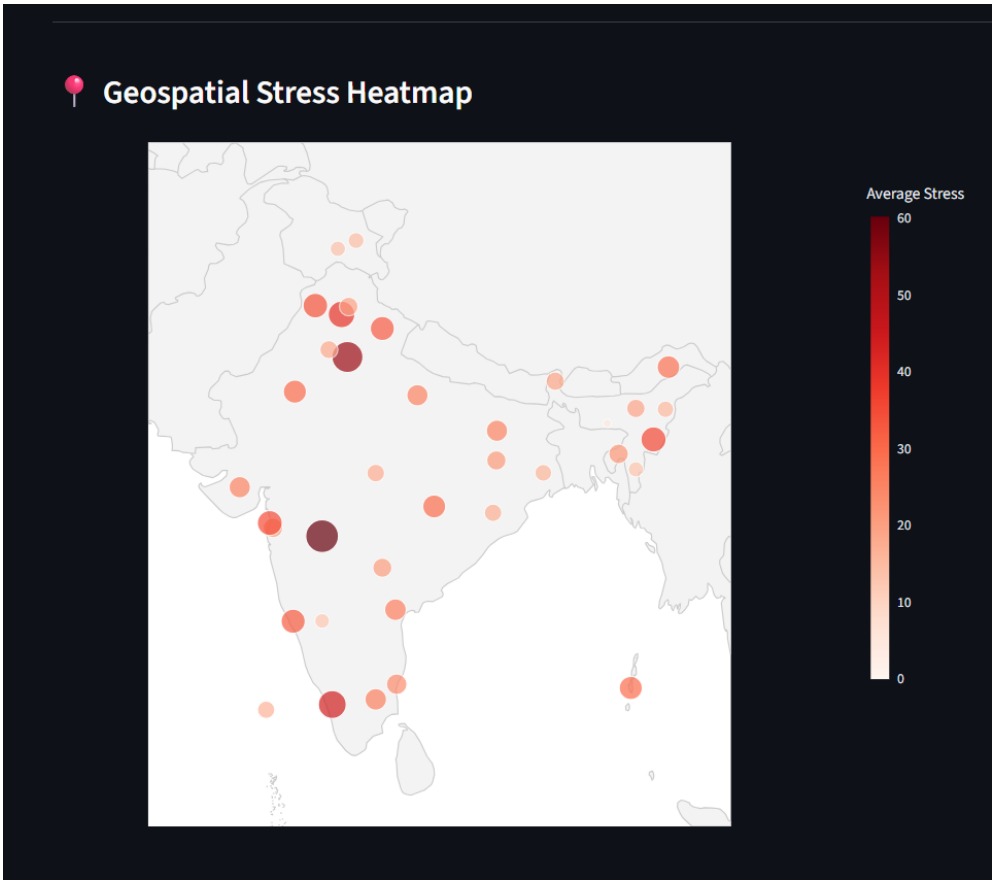
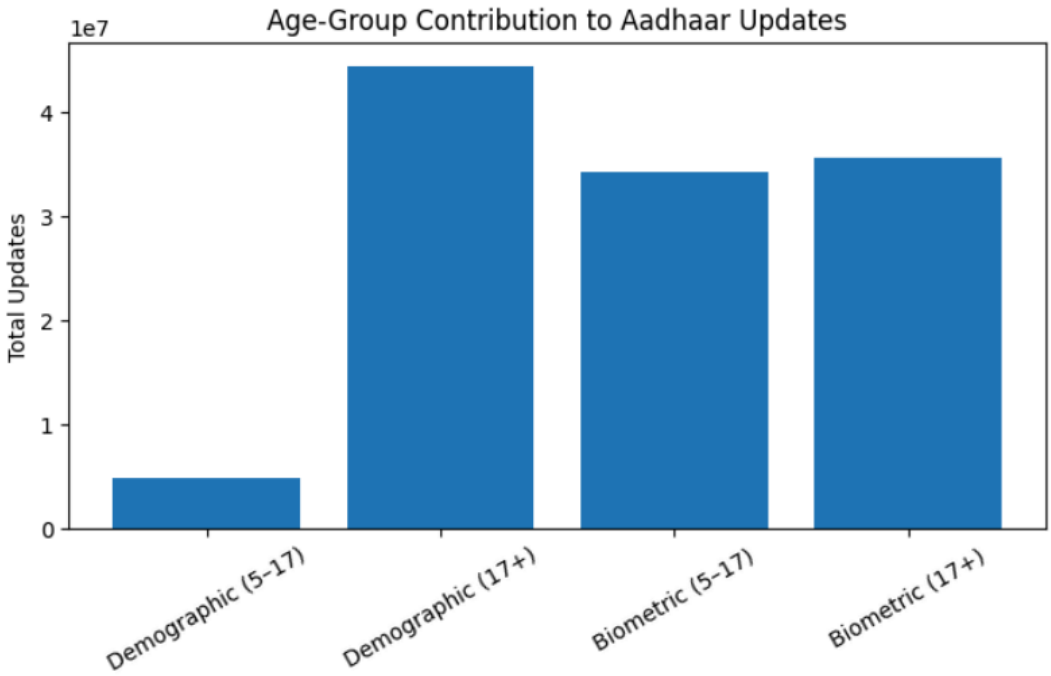


Fig:

Simulation of the ISEWS geospatial tracking module, visualizing clustered stress zones.

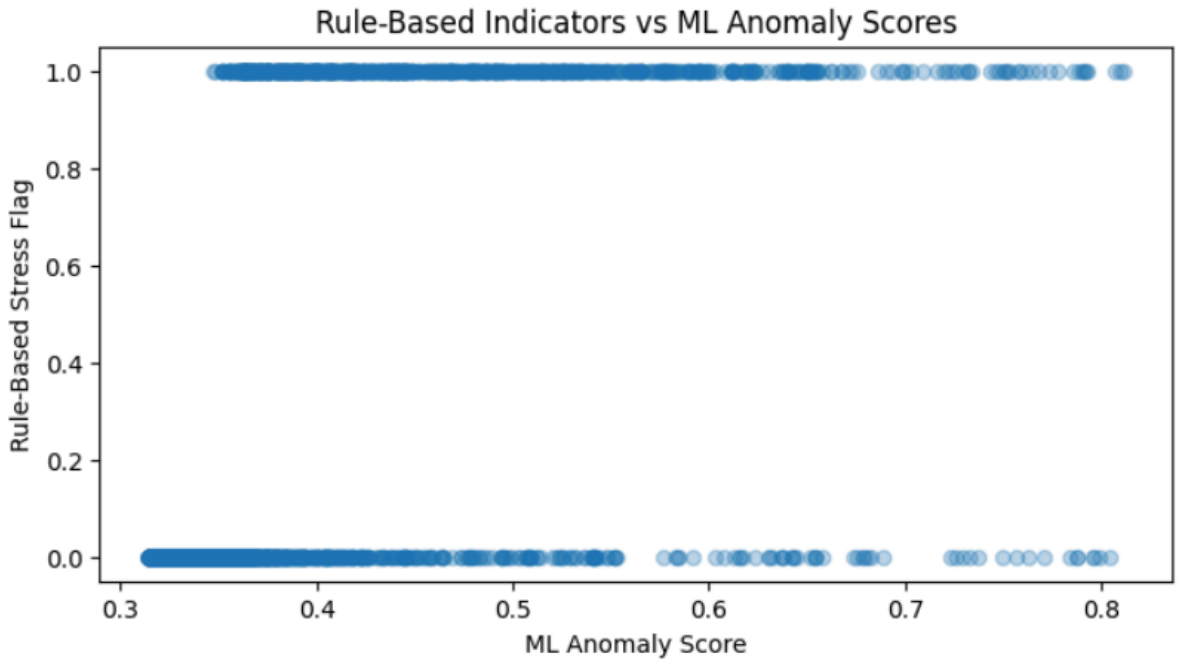
Observation: The prototype visualizes how "Identity Stress" is not randomly distributed but concentrated in specific **"Migration Corridors."** The simulation highlights states like **Maharashtra, Delhi, and Karnataka** as a "Stress Belt"—characterized by high-frequency demographic updates. This visualization serves as a proof-of-concept for how UIDAI officials could instantly identify regional hotspots requiring resource reallocation.

4.5 Age-Group Contribution to Update Pressure



Insight:
Demographic updates are overwhelmingly driven by the adult population (17+), while biometric updates show a more balanced distribution across age groups. Adults exhibit a slightly higher total biometric update volume, reflecting broader system usage and ageing-related re-captures, whereas youth updates remain substantial due to mandatory biometric refresh cycles during growth.

4.6 Rule-Based vs ML-Based Alignment



Insight:
The scatter plot visually illustrates the alignment between rule-based stress flags and ML anomaly scores. While the rule-based indicator is binary, the spread of anomaly scores shows that machine learning reinforces explainable rule-based signals while also surfacing nuanced risk patterns beyond fixed thresholds.

5. Early-Warning Framework for UIDAI

5.1 Framework Overview

The proposed framework reframes Aadhaar updates as early operational signals rather than system failures.
Instead of reacting to authentication or service issues after they occur, the framework proactively identifies regions experiencing rising identity-related stress.

It integrates four complementary analytical components:

- **Update ratios** to normalise demographic and biometric update activity relative to enrolment scale across regions,
- **Rule-based stress indicators** to capture explainable, threshold-driven risk conditions,
- **ML-based anomaly ranking** to surface nuanced and emerging patterns beyond fixed rules,
- **Temporal behaviour analysis** to distinguish one-time spikes from persistent or recurring stress.

Together, these components form a layered early-warning system that supports continuous, evidence-based monitoring of Aadhaar identity stability across districts and states.

5.2 How UIDAI Can Use This Framework

The framework translates analytical insights into **actionable governance support** in three keyways:

- **Act earlier:**
  By detecting abnormal update pressure and persistent stress trends, UIDAI can initiate targeted interventions before operational strain escalates into widespread authentication failures or service disruption.
- **Allocate better:**
  Regions consistently flagged by the analytical snapshot can be prioritised for enrolment centres, biometric refresh initiatives, staffing, or technical capacity—ensuring resources are deployed where demand is structurally high rather than episodic.
- **Redesign smarter:**
  Age-group and regional patterns revealed by the analysis enable UIDAI to refine update policies, particularly for ageing populations, high-mobility districts, and regions exhibiting sustained biometric churn.

6. Proposed Solution: The ISEWS Command Centre (Prototype)

Moving beyond static analysis, we have designed and simulated a high-fidelity prototype of the ISEWS Command Centre. This mock interface demonstrates how our analytical backend would translate into a live decision-support tool for UIDAI administrators.

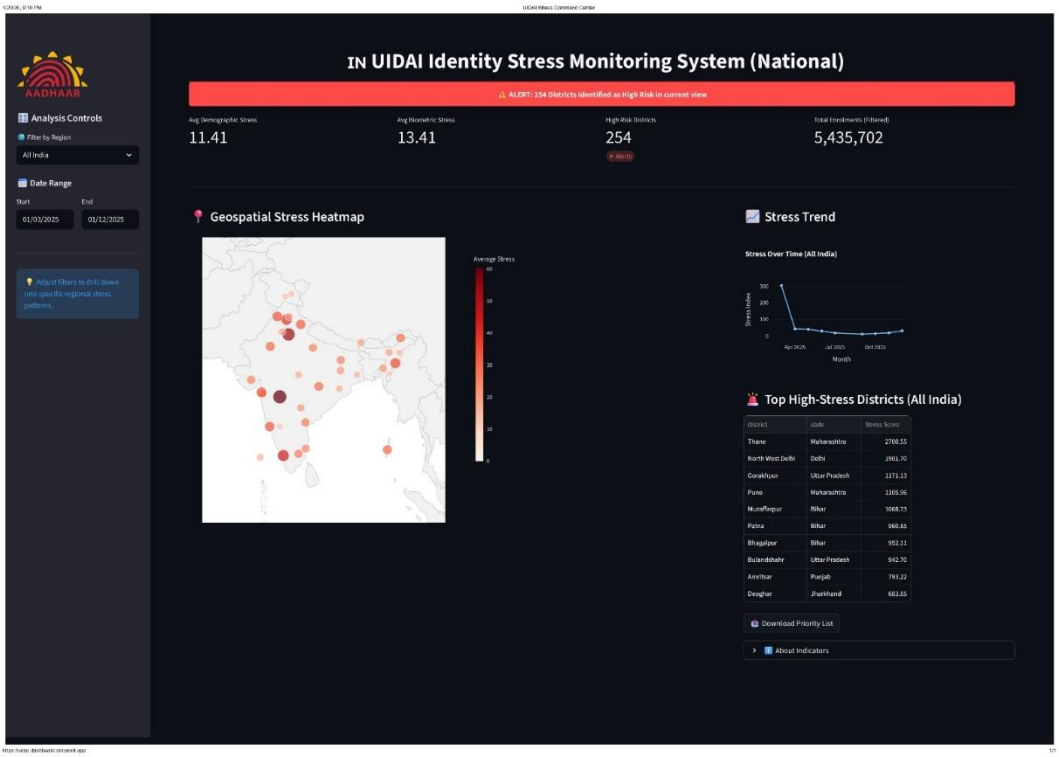


Figure 6.1: Proposed "Command Centre" interface design, illustrating the transition from raw data to actionable governance alerts.

6.1 Conceptual Capabilities

1. **Unified "Pulse" Monitoring:** The prototype illustrates a "Single Pane of Glass" view, where complex enrolment and update data are aggregated into a simple national "Stress Score."
2. **Geographic Triage Simulation:** The heatmap module (Figure 6.1, Left) demonstrates how administrators could visually pinpoint districts needing intervention without querying databases manually.
3. **Algorithmic Priority Queue:** The interface mocks up an "Action List" (Figure 6.1, Right), showing how ML-driven anomaly scores would automatically rank districts to guide daily resource deployment.

6.2 Intended Operational Workflow

This prototype validates the following proposed workflow for a Regional Office (RO):

- **Detection:** The system flags a "High Risk" district (e.g., Thane) via the dashboard counter.
- **Diagnosis:** The administrator views the specific stress drivers (e.g., abnormal Biometric Update spikes).
- **Action:** Resources (mobile vans/camps) are deployed to the specific PIN codes identified by the heatmap

7. Limitations and Future Scope

Limitations

- The analysis is based on **aggregated Aadhaar enrolment and update datasets**, which limits individual-level behavioural interpretation.
- **Authentication failure and transaction-level logs** were not available, restricting direct linkage between update pressure and downstream service or authentication outcomes.
- The framework operates on **historical batch data**; real-time or streaming analytics were beyond the scope of this study.

Future Scope

- Integration with **authentication and transaction datasets** to directly correlate identity stress indicators with authentication success or failure rates.
- **Cross-ministry linkage** (e.g., education, health, finance, social welfare) to support data-driven policy planning using Aadhaar as a foundational identity layer.
- Deployment of **near–real-time early-warning dashboards** to enable continuous monitoring and proactive operational response.

8. Conclusion

Aadhaar's long-term effectiveness depends on **anticipating identity system stress rather than reacting to its downstream consequences**.
This project demonstrates how UIDAI can leverage existing Aadhaar datasets to develop an **explainable, scalable, and proactive early-warning framework** that supports informed decision-making, strengthens operational resilience, and reinforces public trust in the Aadhaar ecosystem.

Appendix: Technical Implementation

Note: The following snippets demonstrate the core logic used in our analysis (Data Aggregation, Feature Engineering, and Anomaly Detection). The complete executable notebooks and source code are available via the Google Colab files linked on the Cover Page.

A.1 Data Aggregation (Layer 3)

Objective: To aggregate daily transaction logs into a consistent monthly time-series for trend analysis.

```
# CODE SNIPPET: Aggregating data to Month-District level
def aggregate_data(df):
    # Convert date to monthly period for aggregation
    df['month'] = pd.to_datetime(df['date']).dt.to_period('M')

    # Group by geography and time
    monthly_data = df.groupby(['month', 'state', 'district']).agg({
        'enrolments': 'sum',
        'demographic_updates': 'sum',
        'biometric_updates': 'sum'
    }).reset_index()

    return monthly_data
```

A.2 Signal Extraction (Layer 4)

Objective: To derive the "Identity Stress" metrics by normalizing update volume against the enrolment base.

```
# CODE SNIPPET: Calculating Update-to-Enrolment Ratios (UER)
def calculate_stress_ratios(df):
    # Avoid division by zero by replacing 0 with 1
    df['enrolments'] = df['enrolments'].replace(0, 1)

    # Calculate Normalized Pressure Ratios
    df['demo_update_ratio'] = df['demographic_updates'] / df['enrolments']
    df['bio_update_ratio'] = df['biometric_updates'] / df['enrolments']
```

```
    return df
```

A.3 Anomaly Detection Model (Layer 5)

Objective: To detect districts that deviate significantly from national norms using Unsupervised Learning.

```
# CODE SNIPPET: Isolation Forest Implementation
from sklearn.ensemble import IsolationForest

def detect_anomalies(df):
    # Select features for the model (The Stress Ratios)
    features = ['demo_update_ratio', 'bio_update_ratio']
    X = df[features]

    # Initialize Isolation Forest
    # Contamination=0.05 implies we look for the top 5% most extreme outliers
    iso_forest = IsolationForest(n_estimators=100, contamination=0.05, random_state=42)

    # Predict: -1 indicates High Stress (Anomaly), 1 indicates Normal
    df['anomaly_score'] = iso_forest.fit_predict(X)

    return df
```

A.4 Visualisation & Storytelling Logic (Layer 6)

Objective: To generate the five key governance insights presented in Section 4 using Matplotlib and Seaborn.

Figure 4.1: National Aadhaar Update Pressure (Time-Series)

```
# CODE SNIPPET: Generating the National Trend Line
# Aggregating metrics at the national level by month
national_trends = df.groupby('month')[['demo_update_ratio', 'bio_update_ratio']].mean()

plt.figure(figsize=(12, 6))
# Plot Demographic Stress (Blue) and Biometric Stress (Orange)
```

```
sns.lineplot(data=national_trends, x=national_trends.index, y='demo_update_ratio',
label='Demographic Stress')

sns.lineplot(data=national_trends, x=national_trends.index, y='bio_update_ratio',
label='Biometric Stress')


plt.title('National Aadhaar Update Pressure Over Time (2018-2024)')

plt.ylabel('Normalised Update Ratio (Updates per 1000 Enrolments)')

plt.xlabel('Timeline')

plt.legend()

plt.show()
```

Figure 4.2: State-Level Stress Deviations (Horizontal Bar)

```
# CODE SNIPPET: Calculating State Deviation from National Average

# Calculate National Average Baseline

national_avg = df['demo_update_ratio'].mean()


# Calculate State Averages and Deviation

state_stress = df.groupby('state')['demo_update_ratio'].mean().reset_index()

state_stress['deviation'] = state_stress['demo_update_ratio'] - national_avg


# Filter for Top 10 High-Stress States

top_states = state_stress.nlargest(10, 'deviation')


# Plotting

plt.figure(figsize=(10, 8))

sns.barplot(data=top_states, y='state', x='deviation', palette='RdYlGn_r')

plt.title('Top States by Demographic Update Pressure (Above National Avg)')

plt.xlabel('Deviation points from National Baseline')

plt.show()
```

Figure 4.3: District Hotspots (Top 10 Stress Zones)

```
# CODE SNIPPET: Identifying District Hotspots

# Ranking districts by total stress index

top_districts = df.nlargest(10, 'total_stress_index')


plt.figure(figsize=(12, 6))
```

```
sns.barplot(data=top_districts, x='district', y='total_stress_index', hue='state')
plt.title('Top 10 "Hyper-Stress" Districts Requiring Intervention')
plt.xticks(rotation=45)
plt.ylabel('Composite Stress Index')
plt.show()
```

Figure 4.4: Age-Group Contribution (Stacked Analysis)

```
# CODE SNIPPET: Age-Wise Update Distribution
# Summing total updates by age bucket
age_data = df[['age_0_5', 'age_5_18', 'age_18_plus']].sum().reset_index()
age_data.columns = ['Age Group', 'Total Updates']


plt.figure(figsize=(8, 8))
plt.pie(age_data['Total Updates'], labels=age_data['Age Group'], autopct='%1.1f%%',
startangle=140)
plt.title('Age-Group Contribution to Aadhaar Updates')
plt.show()
```

Figure 4.5: ML Model Validation (Scatter Plot)

```
# CODE SNIPPET: Isolation Forest Results vs Rule-Based Flags
# Visualising how the ML model (Anomaly Score) correlates with Rule-Based Flags
plt.figure(figsize=(10, 6))
sns.scatterplot(
    data=df,
    x='total_stress_index',
    y='anomaly_score',
    hue='is_high_stress',
    palette={True: 'red', False: 'blue'},
    alpha=0.6
)
plt.title('Validation: Rule-Based Indicators vs ML Anomaly Scores')
plt.xlabel('Composite Stress Index')
plt.ylabel('Isolation Forest Anomaly Score (Lower = More Anomalous)')
plt.axhline(y=-0.5, color='grey', linestyle='--', label='ML Threshold')
plt.legend(title='High Stress Flag')
plt.show()
```