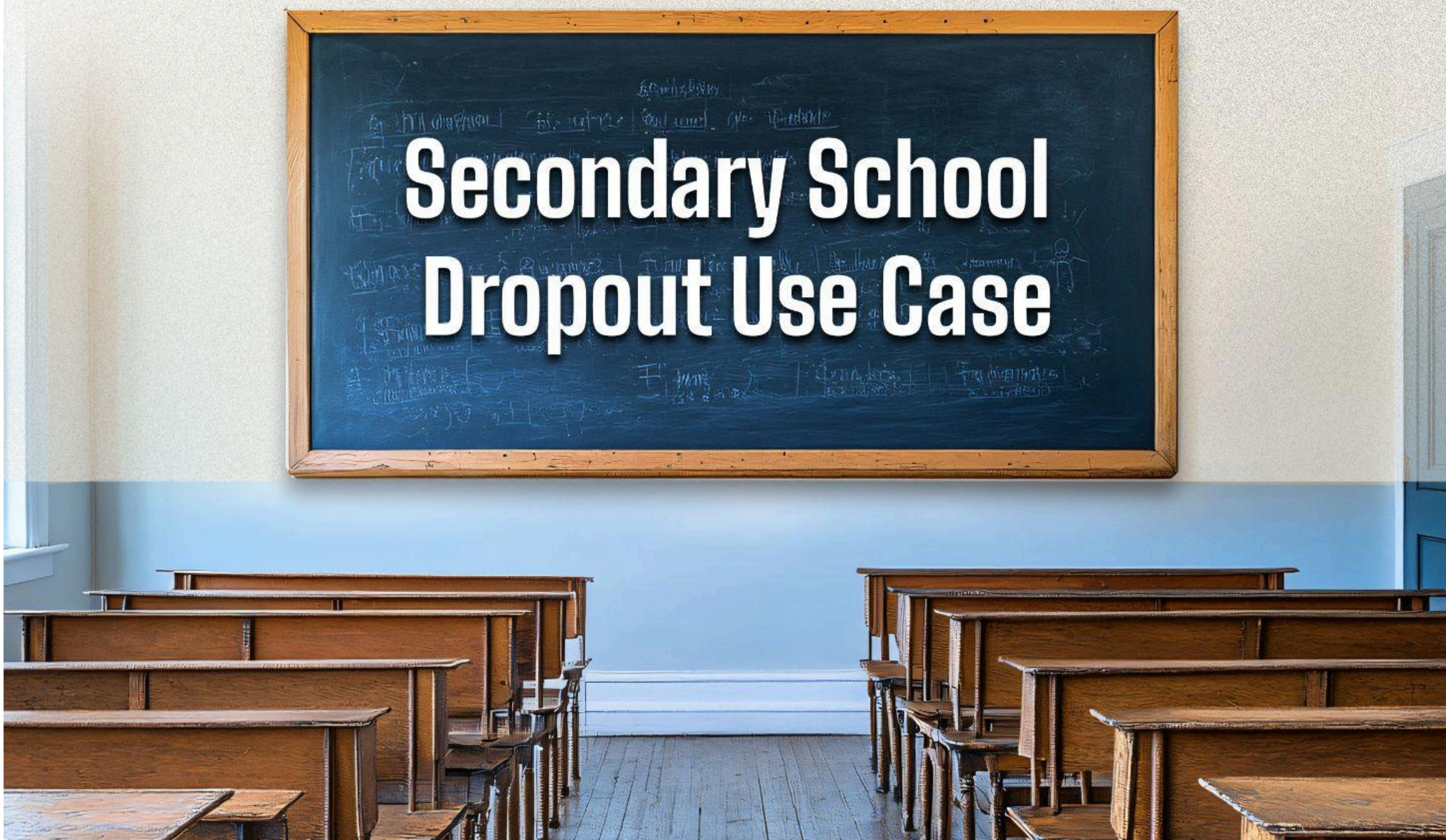
International
Institute of Information
Technology Bangalore

NITI Aayog

Lead Knowledge Institution (LKI) under NITI Aayog's State Support Mission (SSM)

Secondary School Dropout Use Case



Acknowledgement

IIITB expresses its sincere gratitude to the NITI Aayog State Support Mission (SSM) team for their strategic oversight and unwavering support of LKI activities. We are particularly grateful for their role in aligning this initiative with the national vision of Viksit Bharat@2047. We also extend our deep appreciation to Sattva Consulting for their valuable partnership and technical expertise.

Finally, we would like to thank the Department of School Education & Literacy (DoSEL). The insights and data shared by the department were vital to the development of this dropout use case. We also acknowledge the continued support of various stakeholders and advisors whose contributions enriched the depth and reach of both the use case and the report.

Executive Summary

The International Institute of Information Technology Bangalore (IIITB) serves as a Lead Knowledge Institution (LKI) for NITI Aayog's State Support Mission. In this capacity, IIITB develops data driven approaches to decision-making at state and district levels. This initiative employs Machine Learning (ML) models to address high impact use cases in sectors such as education and health. The program develops a structured framework for use case development and scaling while assisting NITI Aayog in refining use cases received from the ground. Furthermore, it enables states to utilize district level data for designing customized interventions.

While current data analysis often relies on a backward looking perspective that hinders effective change, policymakers require foresight and prescriptive strategies. These tools allow them to accurately identify high impact intervention areas and determine the necessary steps to achieve specific targets. The "Do-Board" approach moves beyond standard dashboards by utilizing statistical analysis and ML models to provide enhanced, prescriptive insights.

The first use case in education focuses on reducing secondary school dropout rates for grades 9-10. The Do-Board utilizes sensitivity analysis to rank primary factors contributing to these rates and prescriptive intervention to outline the specific changes required to achieve targeted reductions. The process identifies key contributing factors, including socio-economic, infrastructural, and institutional indicators, based on the sensitivity of these factors for each district. Consequently, the model prescribes improvements in these factors to achieve the desired reduction in dropouts.

The primary outcome is targeted action, which enables districts and states to channel resources toward the top contributing factors. These prescriptions guide policymakers as they realign priorities, select effective interventions, and allocate budgets.

Table of Contents

1. Problem Statement and Approach

1.1 Problem Statement

The problem statement for this use case focuses on reducing secondary school dropouts (Grades 9-10). Secondary education represents a critical developmental bottleneck within the national educational framework. Although enrollment at the primary level is nearly universal, the transition to and completion of secondary school remain significant challenges.

From a decision making prescriptive, we attempt to address the following questions:

1. For any district, how significant is the contribution of each factor in improving the dropout rate (for grades 9-10)?
2. How does a selected factor contribute to improving the dropout rate in other districts of the same state?
3. For any district or state, given the current values of relevant factors, what are the expected factor values required to ensure the targeted dropout rate is achieved, while considering user priorities?
4. For any state and a selected factor, what is the impact of a small (10%) change in that factor on the dropout rates of each district in the state?
5. What is the relative level of support or action required by each state?
6. What is the effort required in improving each factor to attain the targeted dropout rate?

1.2 Use Case Approach

The methodology for each use case begins with acquiring relevant, publicly available datasets from various government portals and data repositories.

In the secondary school dropout use case, the primary data source is the Unified District Information System for Education (UDISE) from the Ministry of Education's Department of School Education and Literacy. This is further supplemented with socio-economic indicators from the National Family Health Survey 5 (NFHS-5).

The data processing and analysis involve the following stages:

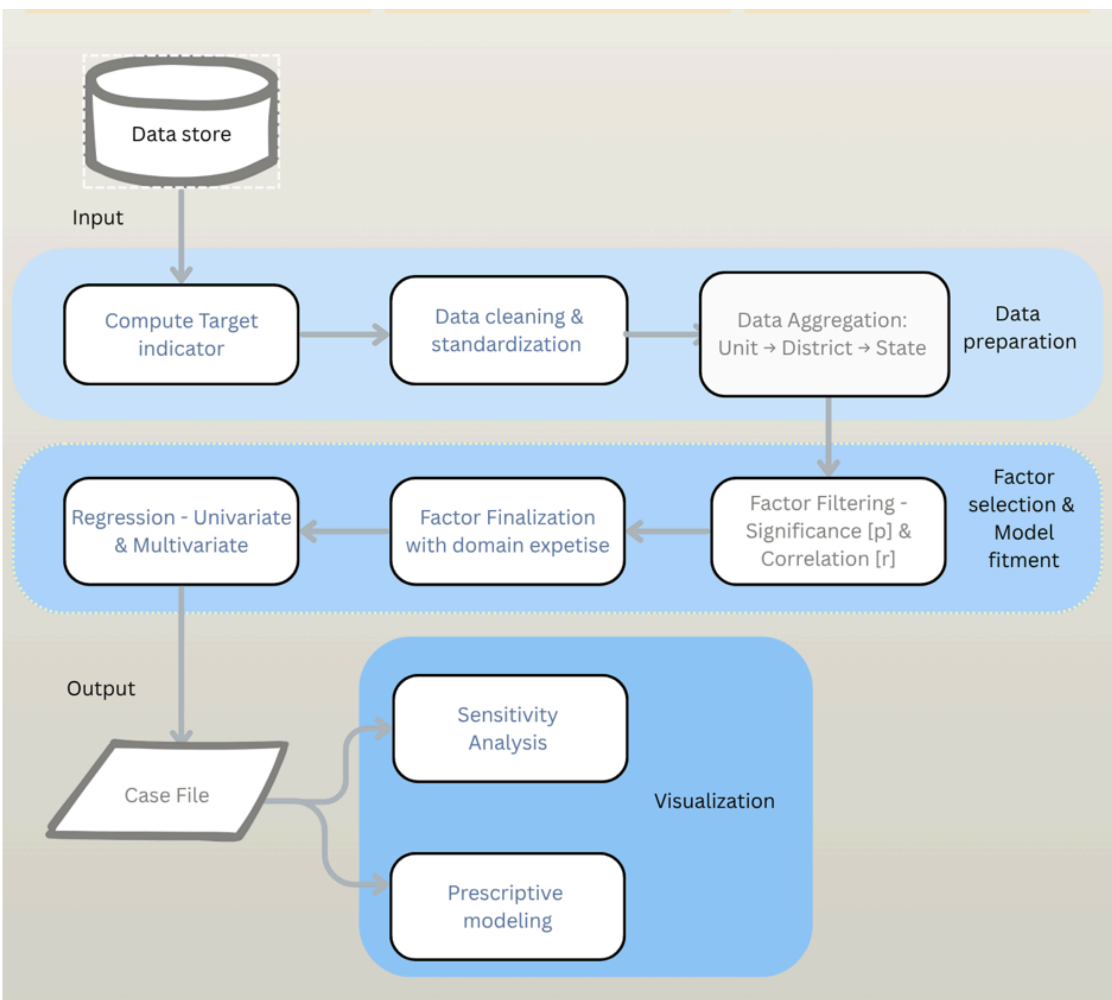
1. **Data Integration and Cleaning:** The process begins by using UDISE data to create a master file of school-level indicators. This dataset undergoes cleaning, the addition of calculated fields, and aggregation at both the district and state

levels. Subsequently, the system integrates socio-economic indicators from the district-level NFHS dataset and computes the target indicator.

2. **Statistical Filtering:** A threshold file is developed based on statistical significance (p-value < 0.05) and pearson correlation coefficients (|r| >= 0.4).
3. **Integrity Checks and Optimization:** Initial consistency checks ensure data integrity. This stage includes removing spurious correlations, optimizing redundant factors, and classifying remaining factors into specific bucket categories.
4. **Modeling and Visualization:** Once final factors are confirmed through domain expertise, both univariate and multivariate regression models are executed. The resulting case files serve as inputs for the visualization dashboards, which feature sensitivity analysis and prescriptive intervention models.

The use case workflow is depicted below:

Figure 1.1: Use case workflow



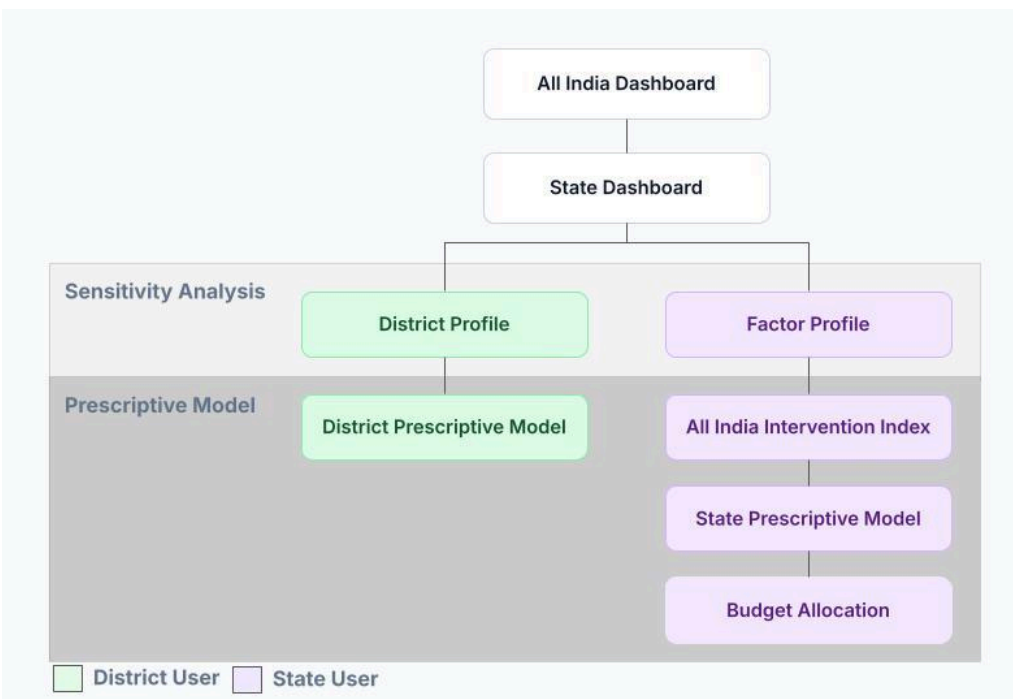
2. Dashboard Walkthrough

2.1 User Personas

The dashboard walkthrough section illustrates the two distinct navigation flows designed for the district user and the State user personas. Both user workflows initiate at the All India Dashboard and navigate to a selected state level view dashboard.

For the district user, the flow leverages sensitivity analysis to generate a district profile. The 'what-to-do analysis' (prescriptive intervention) steps then generate actionable insights, culminating in a district prescription. For the state user, the flow leverages sensitivity analysis to generate a factor profile. The prescriptive intervention steps then generate an all india intervention index followed by a state prescription together with a budget allocation.

Figure 2.1: User Persona - Tableau Dashboard

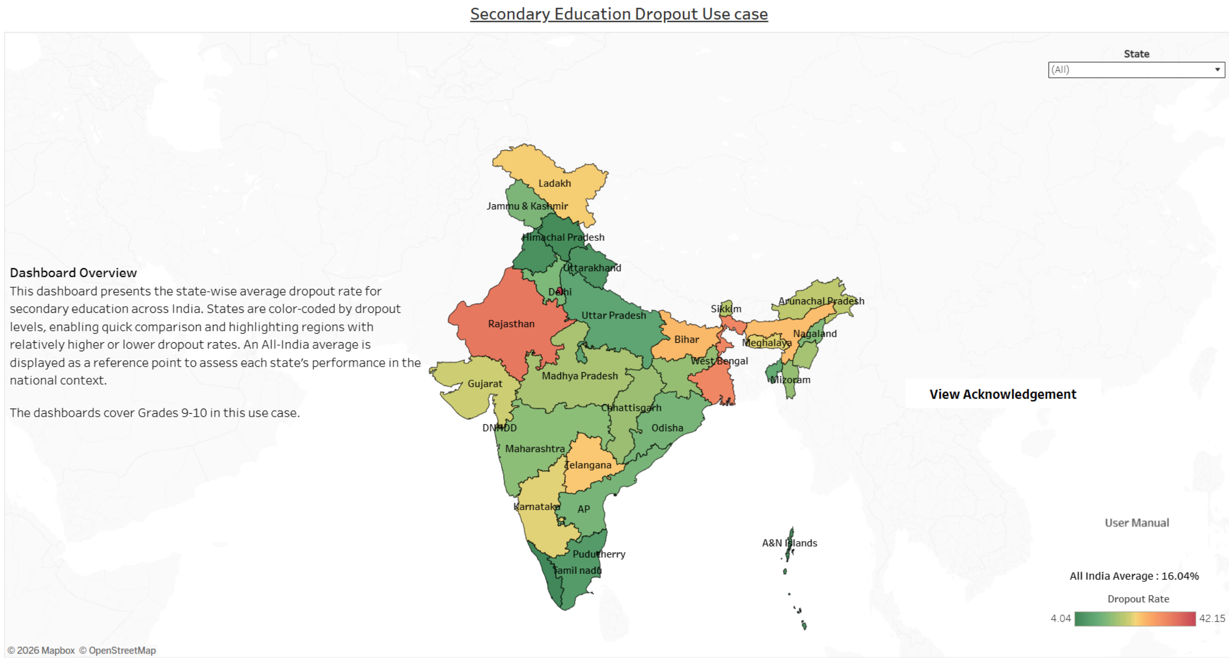


2.2 All India Dropout Dashboard

Purpose

This dashboard presents an All India perspective of dropout rates categorized by state.

Figure 2.2: Tableau - All India Dropout rate Dashboard



Key Metrics

The interface illustrates state-wise average dropout rates utilizing a color-coded performance scale (green-to-red):

- Green: Better performing states (low risk)
- Yellow/Orange: Moderate-risk states
- Red: High-risk states (need the most attention)

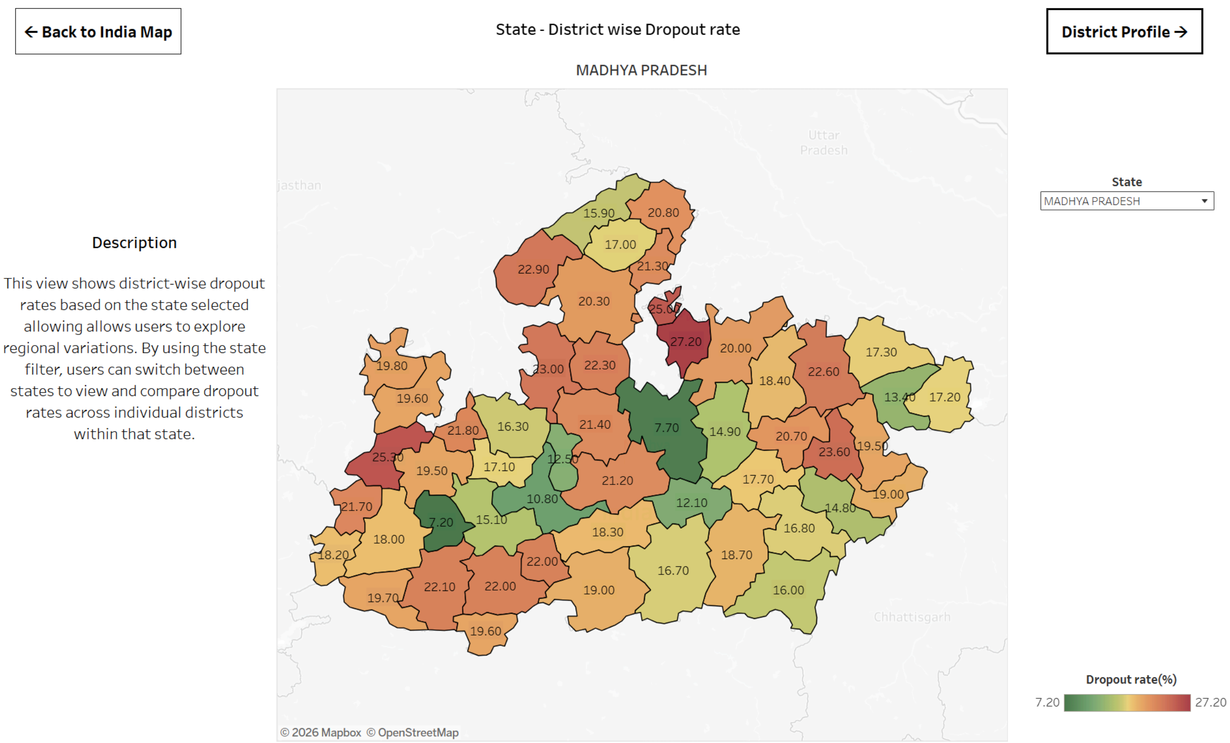
Interactive Navigation: Users may select a specific state to access its Factor Profile dashboard for more granular analysis.

Categorization

- Better-performing States/UTs (Green): Kerala, Lakshadweep, Himachal Pradesh
- Moderate-risk States/UTs (Yellow/Orange): Telangana, Bihar, Karnataka
- High-risk States/UTs (Deep Orange/Red): Rajasthan, Delhi, West Bengal

2.3 State Dropout Dashboard

Figure 2.3: Tableau - State District wise Dropout rate Dashboard



Key Metrics

The dashboard visualizes district-level average dropout rates within a selected state using a standardized color-coded risk assessment (green-to-red):

- Green: Better performing districts (low risk)
- Yellow/Orange: Moderate-risk districts
- Red: High-risk districts (need the most attention)

2.3 District Profile Dashboard

Purpose
The District Profile dashboard enables users to analyze how specific educational and socio-economic factors contribute to dropout sensitivity within a selected district.

Key Metrics

District Sensitivity Ratio:

This metric represents each factor's proportional contribution to the dropout rate relative to all other analyzed factors, totaling 100%.

Step-by-Step Workflow

Step 1: Select District

Users select a district to generate a stacked bar chart that displays the normalized Sensitivity Ratio (%).

Each segment of the bar corresponds to a specific factor, illustrating its relative influence on dropout rates in that district.

Step 2: Interpret the Stacked Bar Chart

The stacked bar chart provides a proportional view of factor aggregation:

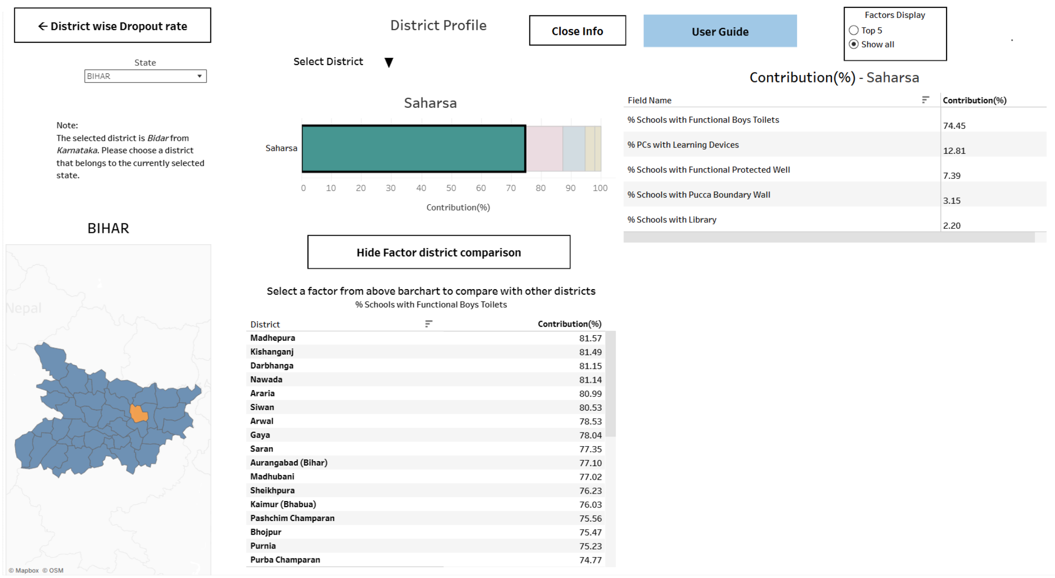
- The full bar represents 100% of the district's sensitivity.
- Individual colored segments indicate each factor's share, allowing users to identify primary drivers.

Step 3: Compare with Other Districts

To examine factor behavior across different regions, users can select a specific segment in the chart. The dashboard then compares that factor's contribution in the current district against its significance in others.

This comparison helps determine whether a factor is uniquely critical to one district or consistently significant across the state.

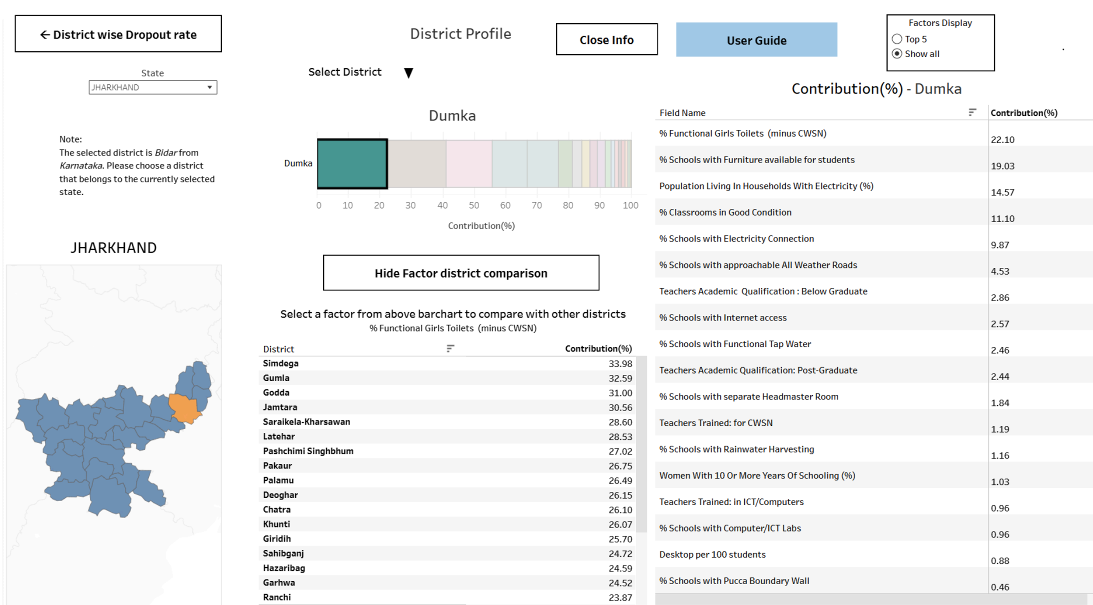
Figure 2.4: Tableau - District Profile Dashboard - Saharsa district, Bihar



The analysis for Saharsa district indicates the following:

- The availability of functional boys' toilets contributes 74.45%, representing the largest share of sensitivity
- This suggests that improving sanitation facilities is the highest priority intervention for Saharsa due to its significant relative impact.

Figure 2.5: Tableau - District Profile Dashboard - Dumka district, Jharkhand



In Dumka (Jharkhand), the sensitivity is distributed across several factors rather than being dominated by a single variable.

Three key factors account for nearly 56% of total sensitivity, indicating a multifaceted driver for dropouts:

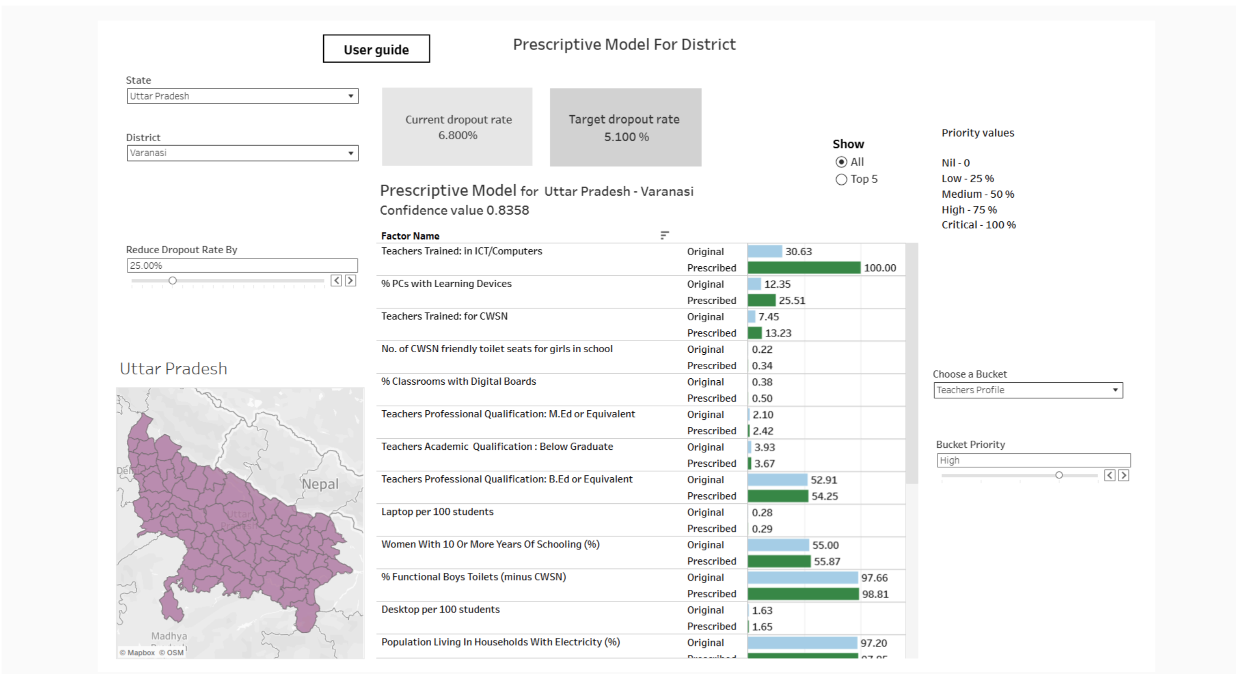
- % Functional Girls Toilets (minus CWSN)
- % Schools with Furniture available for students
- Population living in households with Electricity (%)

Unlike Saharsa, districts like Dumka show that dropout rates are influenced by a combination of school infrastructure and household socio-economic conditions. Consequently, interventions must address systemic issues rather than isolated factors.

2.4 Prescriptive Model for District

The Prescriptive Model identifies the adjustments required in key educational and socio-economic factors to reach target dropout reductions. It integrates data modeling with policy priorities to provide strategic recommendations for state and district-level actions.

Figure 2.6: Tableau - Prescriptive Model for District Dashboard - Varanasi district, Uttar Pradesh



Step-by-Step Workflow

Step 1: Select the District

Users select the relevant district to load current metrics and establish reduction targets.

Step 2: Set the Dropout Reduction Target

Adjusting the reduction slider sets the model's target dropout rate. The system subsequently calculates prescribed values for each factor that collectively achieve the desired target.

Step 3: Review Prescribed vs. Original Values

The dashboard displays a comparative view for each factor:

- Original Value: Represents the current factor value.
- Prescribed Value: Represents the computed factor value after target reduction

Factors are categorized into thematic buckets, such as Infrastructure and Facilities, Digital/ICT, and Teacher Profile. This view assists in identifying which areas need the most improvement and where efforts should be focused.

The system assigns a default priority to each factor bucket, generating recommended levels for each factor. These values work together to reach the dropout target by guiding where change is most needed.

Step 4: Adjust Bucket Priorities

Users can adjust the priority level for specific buckets on the right side of the dashboard. The remaining effort is then distributed proportionally among the other buckets as follows:

- Nil (0%)
- Low (25%)
- Medium (50%)
- High (75%)
- Critical (100%)

Users may adjust these settings based on policy focus. For example, if teachers recruitment/training is a priority, setting the Teachers Profile bucket to *High* results in the remaining 25% being distributed proportionally among the other buckets. Each bucket's share is then distributed proportionally among its individual factors.

The model utilizes these weighted priorities to refine its prescriptions, giving greater weight to higher-priority buckets and proportionally less to lower-priority ones.

Step 5: View Confidence value

The dashboard also displays the following:

- **Confidence Value**: 0.8358 — this indicates a strong fit between the model's predictions and the current dropout rate.

This metric serves as an indicator of model reliability for the selected state. Higher confidence values represent a closer alignment between model projections and empirical data.

2.5 Factor Sensitivity Profile

Purpose
The Factor Sensitivity Profile evaluates the relative influence of a specific educational or socio-economic factor on dropout rates across all districts within a selected state.

Key Metrics

Factor Sensitivity Index:

Quantifies the projected contribution of each district toward reducing dropout rates based on a 10% incremental change in the selected factor.

Step-by-Step Workflow

Step 1: Select State

Users begin by selecting a state, which populates the dashboard with granular district-level metrics relevant to that administrative region.

Step 2: Select Bucket (Optional)

Optionally, users may select a specific Factor Bucket to filter variables by thematic categories. If a comprehensive view is desired, the selection may be maintained as "All."

Step 3: Select Factor

Upon selecting a specific variable for analysis, the dashboard visualizes the Sensitivity Index (%) across districts, highlighting the comparative intensity of influence for the chosen factor.

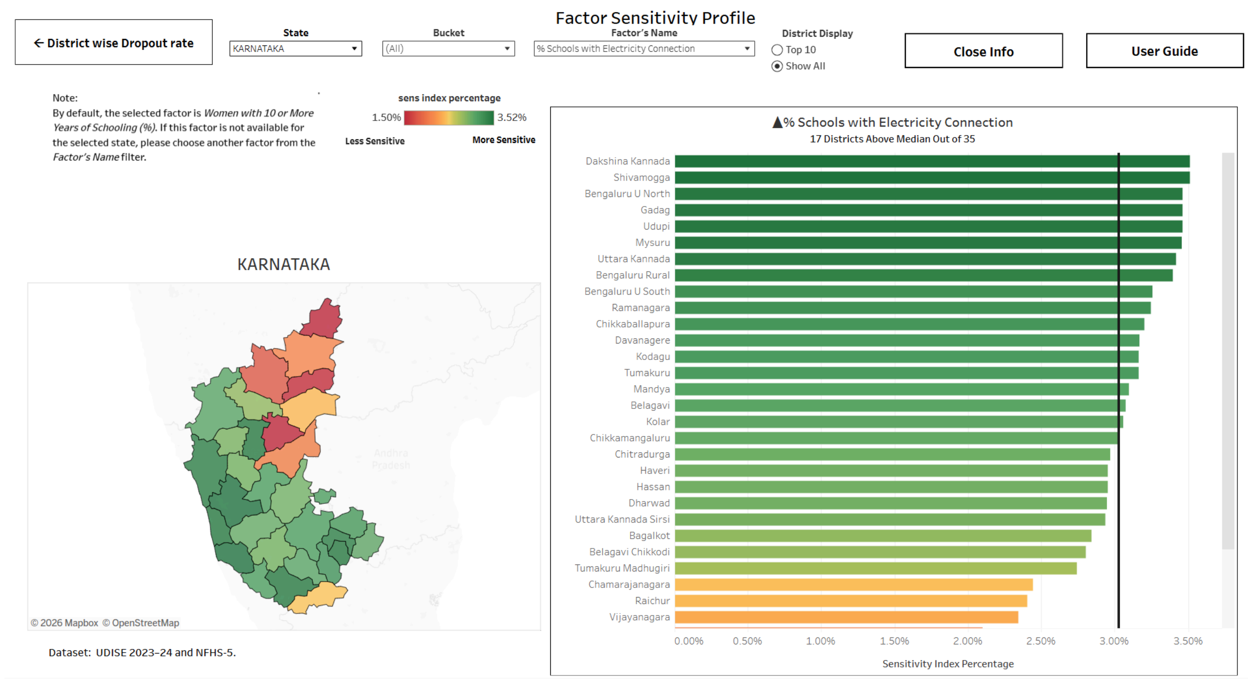
Step 4: Select District Display

The **District Display** filter allows users to customize the visual output:

- **Top 10**: Isolates the districts demonstrating the highest sensitivity levels.
- **Show All**: Provides a complete visualization of all districts within the selected state.

Example: Electricity connection available in school

Figure 2.6: Tableau - Factor Sensitivity Profile - %Schools with Electricity Connection

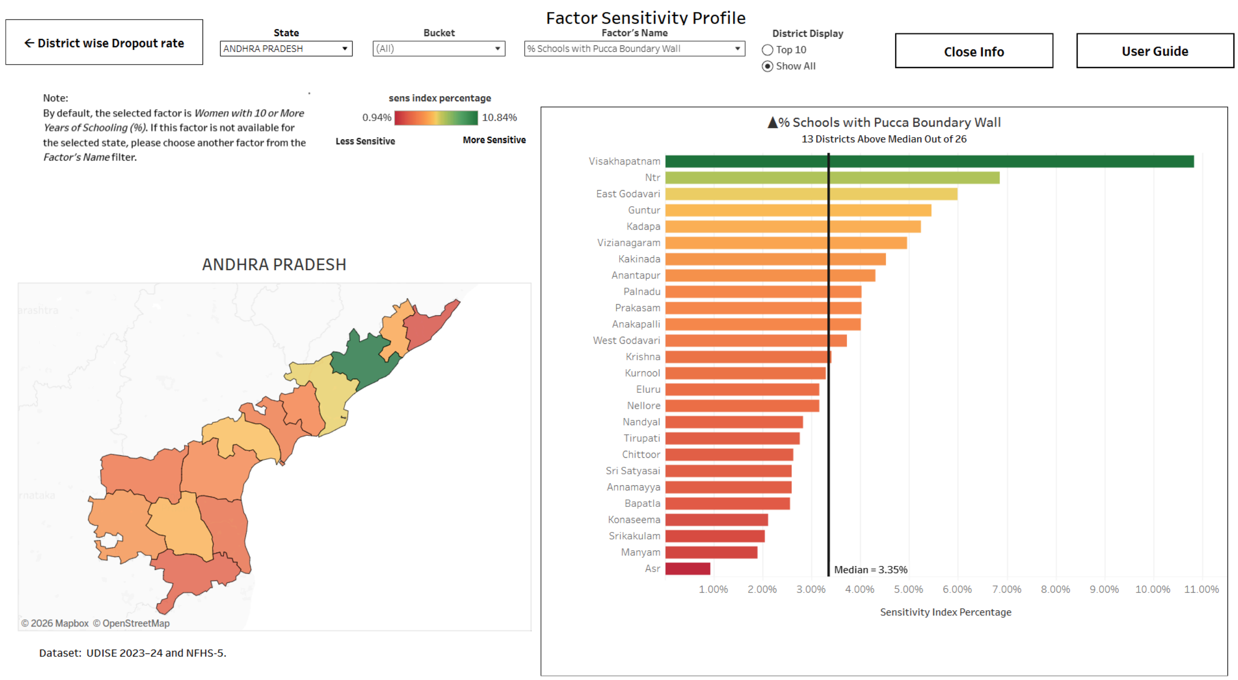


The following analysis examines "% Schools with electricity connection," a factor demonstrating a negative correlation with dropout rates.

The Sensitivity Index reveals that the impact is distributed broadly across most districts, indicating that systemic improvements in school electrification can yield widespread state-level benefits.

Example: % Schools with Pucca boundary wall

Figure 2.7: Tableau - Factor Sensitivity Profile - %Schools with Pucca Boundary Wall

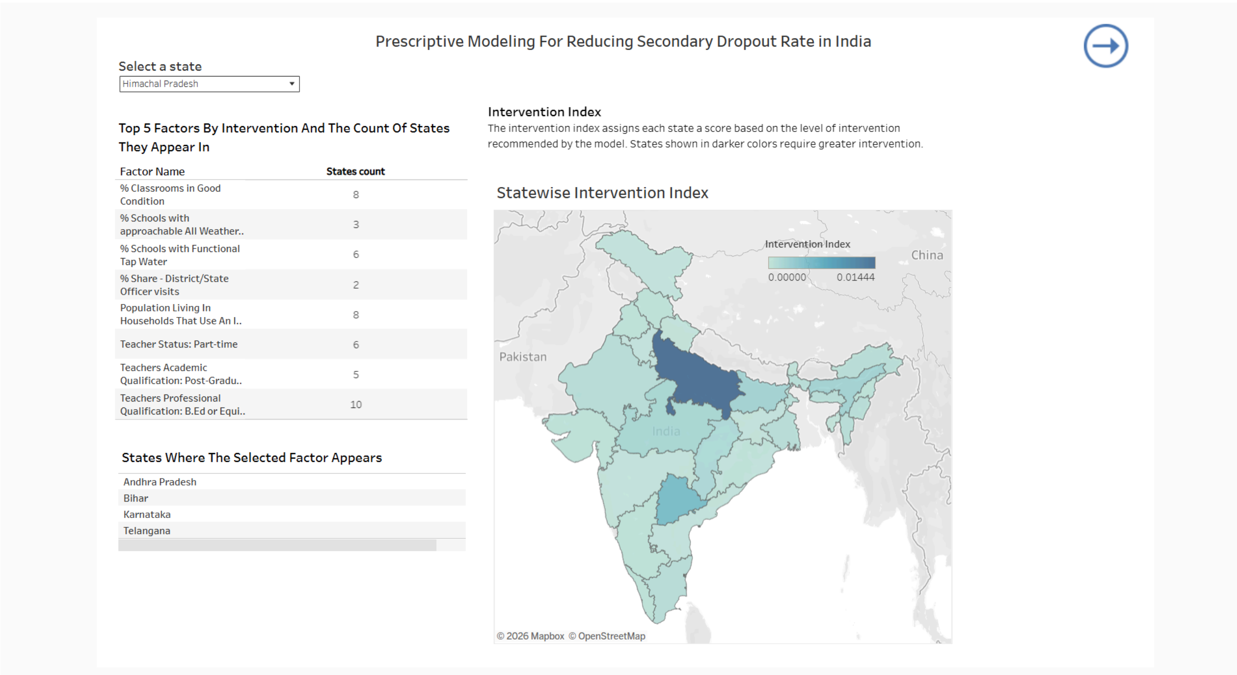


In this scenario, "% Schools with Pucca boundary wall" is selected, showing a negative correlation with dropout rates.

The analysis suggests that this factor influences dropout rates in a skewed manner across the districts of Andhra Pradesh. The potential for impact is concentrated in specific regions, such as Visakhapatnam and Ntr, rather than being uniform across the state.

2.6 All India Intervention Index

Figure 2.8: Tableau -All India Intervention Index - Himachal Pradesh



The heat map of India illustrates the Intervention Index for each state.

Intervention Index
The Intervention Index is a score that indicates the level of intervention required in a state based on model recommendations.

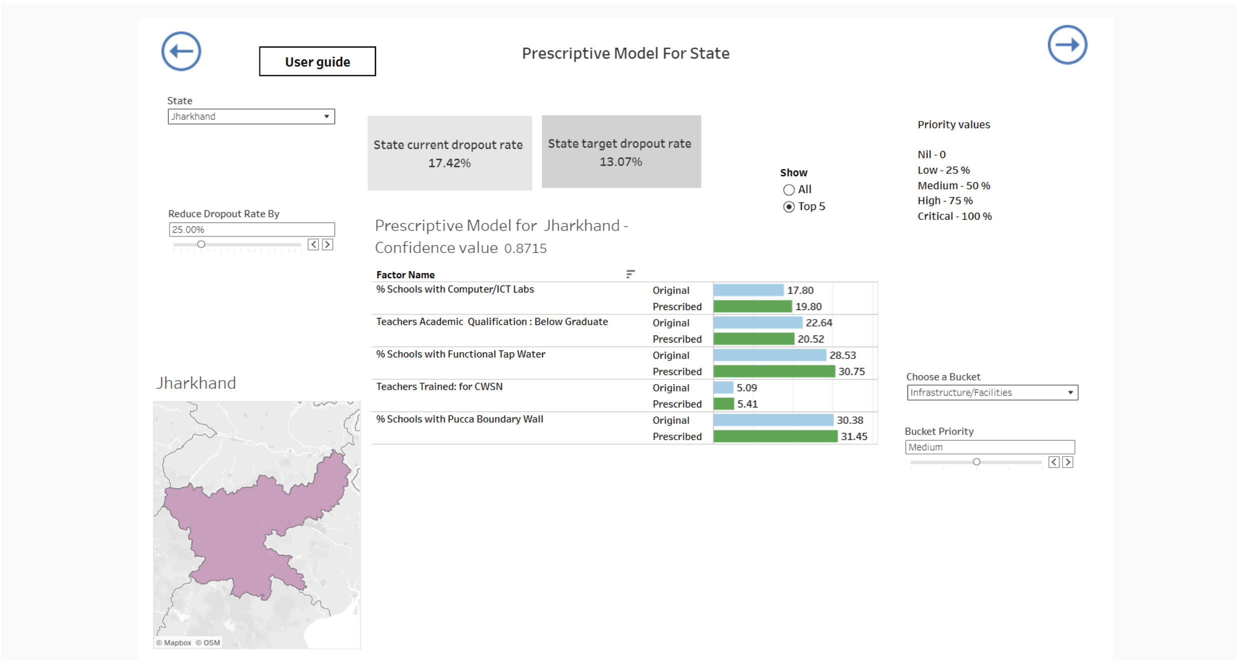
- Darker color → More intervention required
- Lighter color → Less intervention required

This visualization enables the rapid identification of states requiring prioritized administrative attention.

The table on the top left displays the top five factors for a selected state that require the most intervention. When a factor is selected from the table, the bottom table shows the list of other states where that factor also appears.

2.7 Prescriptive Model for State

Figure 2.9: Tableau - Prescriptive Model for State Dashboard - Jharkhand



Step-by-Step Workflow

Step 1: Select the District

Users select the relevant district to load current metrics and establish reduction targets.

Step 2: Set the Dropout Reduction Target

Adjusting the reduction slider sets the model's target dropout rate. The system subsequently calculates prescribed values for each factor that collectively achieve the desired target.

Step 3: Review Prescribed vs. Original Values

The dashboard displays a comparative view for each factor:

- Original Value: Represents the current factor value.
- Prescribed Value: Represents the computed factor value after target reduction

Factors are categorized into thematic buckets, such as Infrastructure / Facilities, Digital/ICT, and Teacher Profile. This view assists in identifying which areas need the most improvement and where efforts should be focused.

The system assigns a default priority to each factor bucket, generating recommended levels for each factor. These values work together to reach the dropout target by guiding where change is most needed.

Step 4: Adjust Bucket Priorities

Users can adjust the priority level for specific buckets on the right side of the dashboard. The remaining effort is then distributed proportionally among the other buckets as follows:

- Nil (0%)
- Low (25%)
- Medium (50%)
- High (75%)
- Critical (100%)

Users may adjust these settings based on policy focus. For example, if infrastructure is a priority, setting the Infrastructure / Facilities bucket to *High* results in the remaining 25% being distributed proportionally among the other buckets. Each bucket's share is then distributed proportionally among its individual factors.

The model utilizes these weighted priorities to refine its prescriptions, giving greater weight to higher-priority buckets and proportionally less to lower-priority ones.

Step 5: View Confidence value

The dashboard also displays the following:

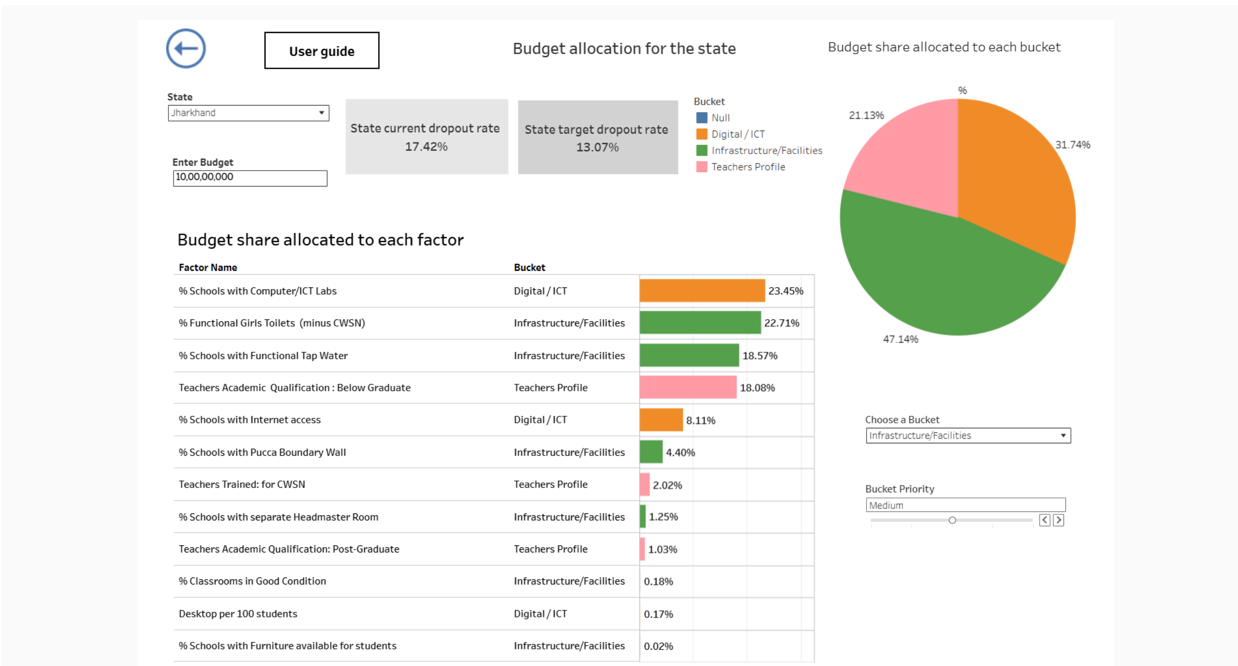
- **Confidence Value**: 0.8715 — this indicates a strong fit between the model's predictions and the current dropout rate.

This metric serves as an indicator of model reliability for the selected state. Higher confidence values represent a closer alignment between model projections and empirical data.

2.8 Budget Allocation

This dashboard provides a framework for allocating budgetary resources across factors influencing dropout rates, utilizing model-driven prescriptions and thematic priorities to optimize resource distribution.

Figure 2.10: Tableau - Budget Allocation for State Dashboard - Jharkhand



Step-by-Step Workflow

Step 1: Enter the Budget Amount

The dashboard automatically imports current dropout rates and reduction targets from the prescriptive modeling interface.

Upon entering the total available budget, the system calculates the optimal distribution across each individual factor based on sensitivity and required change.

Step 2: Review Budget Share by Factor

The dashboard displays a bar chart showing how the budget is distributed across individual factors. Each factor is grouped under a thematic bucket (e.g., Infrastructure / Facilities, Socio-Economic), and the chart shows the percentage share of the total budget.

Factors receive budget shares that are inversely proportional to their sensitivity and the magnitude of change required to meet targets. The portion of the total budget allocated to each factor is determined by the required magnitude of change (Current vs. Prescribed) and its assigned bucket priority. Highly sensitive factors may require less intensive budgetary shifts to achieve significant results compared to other factors.

Step 3: Adjust Bucket Priorities

On the right side of the dashboard, users can adjust the priority level for a chosen bucket using the slider:

- Nil (0%)
- Low (25%)
- Medium (50%)
- High (75%)
- Critical (100%)

The model incorporates these priorities to determine how budget shares are calculated, assigning greater weight to higher-priority buckets and proportionally less to lower-priority ones.

3 Appendix: Formulas and Calculations

3.1 Univariate Regression Model for Sensitivity Analysis

$$O = m * f + c$$

- O: Represents the targeted outcome (e.g., secondary school dropout rate)
- f: Refers to the specific individual factor under analysis
- m: The slope coefficient, quantifying the sensitivity of the outcome to changes in the factor
- c: The regression constant (intercept)

Purpose

This univariate model evaluates the isolated impact of individual factors on the targeted outcome. It provides the statistical foundation for generating both factor profile and district profile.

a. Factor Profile
   The factor profile calculates the proportional contribution of district *i* relative to the aggregate impact of factor f_i across all districts within the state.

b. District Profile
   The district profile identifies the percentage contribution of a specific factor f_i to the cumulative impact of all factors influencing the district level outcome. Contributions are derived using the sensitivity ratio (normalized), where the percentage contribution = S_i * 100.

$$S_i = \frac{\dfrac{m_i f_i d_i}{m_i f_i + c_i}}{\sum_{j \in I} \dfrac{m_j f_j d_j}{m_j f_j + c_j}} = \frac{\dfrac{\Delta O_i}{O_i}}{\sum_{j \in I} \dfrac{\Delta O_j}{O_j}}$$

where,

$$I = \begin{cases} \mathcal{F}, & \text{for factor-wise profiling} \\ \mathcal{G}, & \text{for geography-wise profiling} \end{cases}$$

3.2 Multivariate Regression for Prescriptive Modeling

$$O^g = \beta_0 + \beta_1 f_1^g + \beta_2 f_2^g + ... + \beta_k f_k^g$$

The multivariate regression model is employed to determine the integrated impact of all contributing factors on the primary outcome variable, O.

a. Prescriptive Factor Update Rule

To reach the user-defined proportional reduction targets, prescribed factor values are computed using the following formula:

$$f_i^{\text{new}} = f_i^{\text{old}} + \alpha_i \cdot \frac{O_{\text{current}}(1-r) - O_{\text{current}}}{\mathrm{k_i}\,\beta_i}$$

Where:

r: Target proportional reduction in the outcome (dropout rate)

$O_{current}$: Current value of the outcome variable

β_i: Statistical model coefficient for f_i

α_i: Designated priority level for the factor

k_i: Total number of factors within the specific bucket

f_i^{old} : Original value of the factor

f_i^{new} : Revised prescriptive value for the factor

In instances where a single simultaneous fit for all factors is statistically unfeasible, a secondary model integrates the remaining factors. This ensures the retention of the maximum number of indicators for analysis. The dashboard defaults to the dominant model that maximizes both factor inclusion and the cumulative factor sensitivity ratio to provide the most comprehensive prescriptive coverage.

b. Confidence Value

The calculation follows the formula below:

$$\text{confidence}_s = 1 - \text{error}_s$$

$$\text{error}_s = \frac{|O_s^{\text{model}} - O_s^{\text{actual}}|}{O_s^{\text{model}}}$$

Where:

O^{model} : Outcome value as predicted by the regression model

O^{actual}: Actual outcome value derived from the dataset

c. Intervention Index

The Intervention Index assigns a quantitative score to each state, reflecting the required level of intervention as determined by the prescriptive model's recommendations.

$$InterventionIndex_s = \frac{P_s \cdot confidence_s}{\sum_{s'} (P_{s'} \cdot confidence_{s'})}$$

Where:

$$P_s = \frac{1}{n} \sum_{i=1}^{n} c_{s,i} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\left| f_{s,i}^{\text{new}} - f_{s,i}^{\text{old}} \right|}{f_{s,i}^{\text{old}}} \right)$$

d. Budget Allocation

Once priorities are established, resources are allocated to align with identified intervention needs. The proportional weight of factors, normalized across the state based on their established priorities and statistical sensitivities, is provided below:

$$prop_{s,i} = \frac{c_{s,i}\alpha_i/(k_i\,\beta_i)}{\sum_{j=1}^{n}(c_{s,j}\alpha_j/(k_i\,\beta_i))}$$

This proportion is applied to the state's total budget to derive the recommended expenditure for specific factors:

$$Budget_{s,i} = TotalBudget_s \times \text{prop}_{s,i}$$

4 Glossary

- **Case File:** The refined set of most relevant factors generated from the correlation analysis, which is used as input for the runtime modeling and visualization
- **Confidence Value:** A reliability indicator for the model's predictions in a given state; higher values denote stronger alignment between predicted and actual dropout rates.
- **Correlation Analysis:** A statistical process to assess the relationship between individual factors and the target outcome (dropout rate). It assesses statistically significant factors (p-value < 0.05) and uses the Pearson correlation coefficient (r) to establish the direction and strength of the correlation, thereby validating which factors are carried forward into subsequent domain checks and modeling.
- **District Sensitivity Ratio:** A proportional measure representing each factor's relative contribution to dropout rates within a specific district, normalized to a total of 100%.
- **Do-Board:** An innovative prescriptive dashboard that utilizes statistical analysis and Machine Learning (ML) models to provide actionable insights, moving beyond standard dashboards.
- **DoSEL:** Department of School Education & Literacy, Ministry of Education, Government of India.
- **Factor Sensitivity Index:** A metric quantifying the projected impact of a 10% incremental change in a selected factor on district level dropout rate reductions
- **IIITB:** International Institute of Information Technology Bangalore, serving as the Lead Knowledge Institution (LKI) for NITI Aayog's State Support Mission.
- **Intervention Index:** A composite score assigned to each state, reflecting the required level of intervention as determined by the prescriptive model's recommendations.
- **LKI:** Lead Knowledge Institution designated under NITI Aayog's State Support Mission
- **Multivariate Regression Model:** A statistical model employed to determine the combined impact of all contributing factors on the targeted outcome variable (dropout rate)
- **NFHS-5:** National Family Health Survey (Round 5), a comprehensive dataset providing key socio-economic and health indicators.
- **Prescriptive Modeling:** An analytical approach that identifies specific adjustments required in key factors to achieve targeted reductions in dropout rates, guiding policy recommendations.
- **Sensitivity Analysis:** Examination of factor importance, ranking the primary factors to determine their relative influence on dropout rates within a specific district.
- **SSM:** State Support Mission, an initiative by NITI Aayog
- **UDISE:** Unified District Information System for Education, the Ministry of Education (Government of India)'s primary dataset for school level indicators
- **Univariate Regression Model:** A statistical model used to evaluate the isolated impact of a single factor on dropout rates, serving as the basis for factor and district sensitivity profiles

Lead Knowledge Institution (LKI)
under NITI Aayog's State Support Mission

The **State Support Mission (SSM)** is a flagship initiative by NITI Aayog, designed to foster a structured and institutionalized partnership between the Central Government and States/Union Territories. As the designated **Lead Knowledge Institution (LKI)**, IIITB serves as a high-level academic and technical catalyst, bridging the gap between cutting-edge computational research and practical public policy.

Explore more: lki-ssm.iiitb.ac.in