Group Project

SOC476: CONTEMPORARY APPLICATIONS OF SOCIAL DEMOGRAPHY

Identifying the Factors Affecting School Dropout Rates in India

- Anisha Srivastava: 220149
- Diyansh Mittal: 210358
- Nishant Shanmukham: 210674
- Rashi Sharma: 220872
- Shreya Gupta: 221027

Introduction – Importance of analyzing student dropout rates

Even after 75 years of independence, India continues to experience significant socioeconomic disparities. While a 5% percentage of the population enjoys a standard of living comparable to developed nations, 95% of the population faces challenges in access to basic necessities such as food, healthcare, and shelter. Despite economic growth and an increase in the number of Indian billionaires, India ranks 134th on the Human Development Index. For many families, an unexpected medical emergency can create lifelong financial debt.

In this context, education, though essential, may become a lower priority as children contribute to household income. High dropout rates, particularly in rural areas, can limit opportunities for social and economic mobility. Understanding the factors behind student dropouts and exploring solutions can help improve access to education and promote a fairer society.

Literature review

How Significant is India's Problem of Child Dropouts?

Child dropout is a critical issue in India, impacting the country's efforts to achieve universal education. High dropout rates affect individual students' personal growth and have broader implications for the nation's development, contributing to social and economic disparities[1][2]. Studies highlight that dropout rates peak at the secondary level, with significant regional disparities[3]. Due to the scarcity of data at the secondary level, we focus on school dropouts at the primary school level, showing regional disparities. As seen in Fig.04, the **strong correlation between the predicted dropout rate for classes 1-3** and the predicted dropout rate for classes 4-5, highlights that the findings from this study can be used to understand factors behind school dropout across classes.

Major Factors

1. Poverty and Financial Constraints: Financial difficulties often force children to withdraw from schools to support their families, especially in low-income households[1][4].
2. Lack of Quality Education: Inadequate number and quality of teaching staff, outdated methods, and lack of engaging education also contribute to high dropout rates[1][2].
3. Infrastructure Deficiencies: Poor school infrastructure, including a lack of basic amenities and/or academic facilities, discourages students from attending school[5][1].
4. Health Issues: Malnutrition and chronic illnesses are significant factors leading to dropouts[2].
5. Social Factors: Child labor, child marriage, and gender bias are prevalent reasons for dropout, particularly among girls[2][4].

Metrics Used in Data Analysis

The metrics used in data analysis are crucial for understanding and addressing educational challenges. These include:

- Enrollment and Attendance: These metrics help identify gaps in student participation and potential dropout indicators. Studies have shown that factors like infrastructure and teacher quality significantly influence attendance and enrollment rates[5].
- Teacher Coverage: This includes appointments and attendance of principals, regular government teachers, and para-teachers. The presence of qualified and regular teachers is crucial for maintaining student interest and reducing dropout rates6.
- Infrastructure Score: This assesses the availability of facilities like computers, electricity, libraries, and playgrounds. Research indicates that better infrastructure can improve learning outcomes and reduce dropout rates[5][6].
- Food and Water Score: This evaluates the provision of mid-day meals and access to clean water. Access to basic amenities like these can enhance student attendance and retention5.

The usage of these metrics is supported by research highlighting the importance of infrastructure, teacher quality, and basic amenities in reducing dropout rates[5][1][6].

Our Study

This project examines the key factors contributing to school dropout rates in rural India, with a focus on school infrastructure, teacher attendance, and mid-day meal availability. Using data from the **Annual Status of Education Report (ASER)** for the years 2010, 2014, and 2018, we analyze trends and correlations between the aforementioned metrics and the probability of student dropout, to further the current understanding of reasons behind student dropouts which can aid future researchers in developing solutions.

About the Dataset

The dataset is from the Annual Status of Education Report (ASER) and was provided by the Pratham NGO. It contains school-level, cross-sectional data across 26 states and 3 Union Territories, covering rural districts. Urban areas are excluded from the survey.
Datasets from the years 2010, 2014, and 2018 were selected for this project due to the maximum availability of overlapping data across these years and to conduct a trend analysis at regular intervals of four years. It includes information on student enrollment and attendance from Class 1 to Class 8, principal, teacher, and para-teacher appointments and attendance, details of school infrastructure (such as classrooms, blackboards, and playgrounds), availability of basic amenities (drinking water, electricity, toilets), and mid-day meal services.
Due to scarce data for Classes 6, 7, and 8, our analysis **uses data from Class 1 to Class 5**. The sample size for the total dataset across the three years is about 49,000 points.
The dataset provides insights into the overall quality of schooling in rural India, allowing for analysis of infrastructure, teaching conditions, and student participation.

Link to the Python Notebooks

1. Notebook 1: Link to the Data Cleaning and EDA Python code
2. Notebook 2: Link to the Python code for data analysis

Data Preprocessing

This Python script **loads, cleans, merges, and processes the ASER school data** from multiple years (2010, 2014, 2018), ensuring consistency in column names and handling missing data.

The common columns across the dataset of years 2010, 2014, and 2018 were grouped into four categories as shown below:

```python
common_columns = [
    'state_name', 'state_code', 'district_code', 'district_name',
    # Enrollment & Attendance (Classes 1-8)
    'childenrollment_class_1', 'childenrollment_class_2', 'childenrollment_class_3',
    'childenrollment_class_4', 'childenrollment_class_5',
    'childattendence_class_1', 'childattendence_class_2', 'childattendence_class_3',
    'childattendence_class_4', 'childattendence_class_5',
    # Teacher & Principal
    'principal_appointment', 'principal_attendence',
    'regulargovtteacher_appointment', 'regulargovtteacher_attendence',
    'parateacher_appointment', 'parateacher_attendence',
    'pe_class_teacher', 'pe_other_teacher',
    # Infrastructure
    'computer_in_school', 'sch_electricity_conn', 'library_books__in_school_usable',
    'playground_in_school', 'blackboard_class_2',
    # Food & Water
    'midday_meal_in_school', 'kitchen_shed_in_school', 'tap_in_school_usable',
    'meal_other_evidence_in_school'
]
```

Code Snippet 01

The load_and_clean_data does the following steps:

- Reads the CSV/Excel files.
- Prints the Raw data shape.
- Converts column names to lowercase for uniformity.
- Select only the necessary columns (those in common_columns that exist in the file) from classes 1 to 5.
- Renames them to a standardized format.
- Standardizes state names (capitalizing first letters).
- Adds a year column to track the dataset's source year.

The following other steps were done to pre-process the datasets.

- Combines data from different years into one data frame.
- Missing numeric values (enrollment, attendance) are filled with 0, while missing binary indicators (teacher presence, electricity, infrastructure) are also set to 0 and converted to integers.
- The shape of the Final dataset is printed.

The following is the size of the dataset, before and after the processing:

```
Raw 2018 data shape: (17171, 83)
Cleaned 2018 data shape: (17171, 63)
Raw 2014 data shape: (16215, 101)
Cleaned 2014 data shape: (16215, 63)
Raw 2010 data shape: (15756, 117)
Cleaned 2010 data shape: (15756, 63)
Final dataset shape after merging: (49142, 63)
Final cleaned data shape: (49142, 63)
```

Code Snippet 02

Dropout Rate Calculation Explained:

→ The next part of the code calculates dropout rates for **Classes 1-3** and **Classes 4-5**. The dropout rate is computed as the percentage of students who enrolled but did not attend.

→ Creation of four new columns: "enrollment_1to3", "attendance_1to3", "dropout_rate_1to3", "enrollment_4to5", "attendance_4to5", and "dropout_rate_4to5".

Teacher Coverage Ratio Calculation Explained:

The calculation for the teacher coverage ratio was done as followed using Python:

- Calculates the total number of appointed teachers, including regular government teachers, para-teachers, and principals.
- Counts the total number of teachers present on a given day, summing attendance from the same categories.
- Computes teacher coverage, which is the ratio of total appointed teachers to the total enrolled students (Classes 1-5). If no students are enrolled, coverage is set to 0.
- Determines the teacher attendance ratio, which is the fraction of teachers present out of the total appointed. If no teachers are appointed, the ratio is set to 0.

Infrastructure Score Calculation Explained:

The **infrastructure score** is the sum of all infrastructure indicators for each school, where higher values signify better facilities, and the presence or absence of a facility is represented by 1 or 0, respectively.

Food and Water Score Calculation Explained:

The **food and water score** is the sum of all such availability indicators for each school, with higher values representing better meal and water provisions. The presence or absence of a facility is represented by 1 or 0, respectively.

Summary of the newly added columns:

- **Dropout**: (Child Enrollment - Child Attendance)/ Child Enrollment %
- **Teacher Coverage**: No. of Appointed Teachers per Enrolled Student Ratio
- **Teacher Attendance Ratio:** (Teachers Present/ Teachers Appointed) %
- **Infra Score:** A 1-5 scaled score of infrastructure based on the blackboards, computers, and playground columns.
- **Food and Water Score:** A 1-4 scaled score of meals per day and water availability.

Note: *The attendance is recorded on a random day (day of the survey), which could make the data prone to outliers.*

Observations

Our data analysis indicates that dropout rates across India range from approximately 25% to 40% for Classes 1 to 3, with a higher standard deviation observed for Classes 4 to 5. However, this represents the national average, and a more detailed state-wise analysis is provided below.

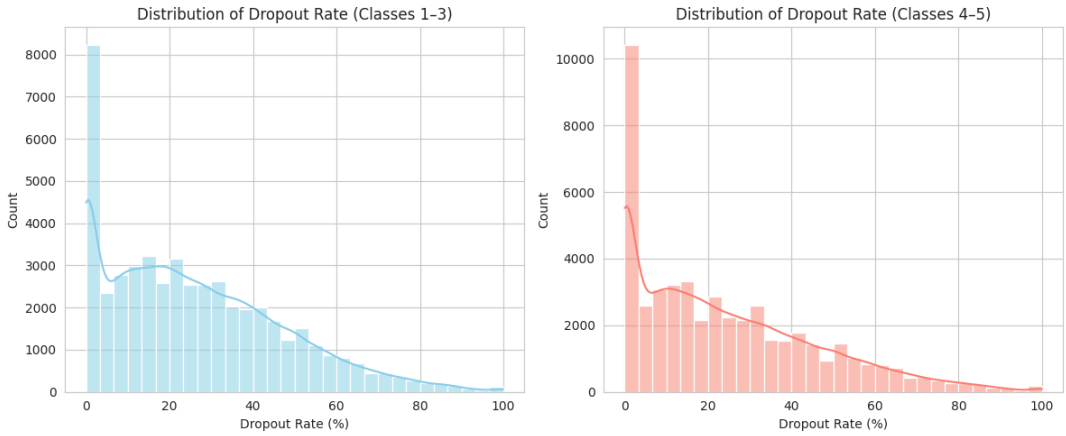

Fig 01 – Distribution of Dropout rates over India

We conducted a detailed state-wise analysis of dropout rates and found that Bihar, Uttar Pradesh, and Jharkhand exhibit the highest dropout rates, whereas Puducherry, Goa, and Tamil Nadu have the lowest.

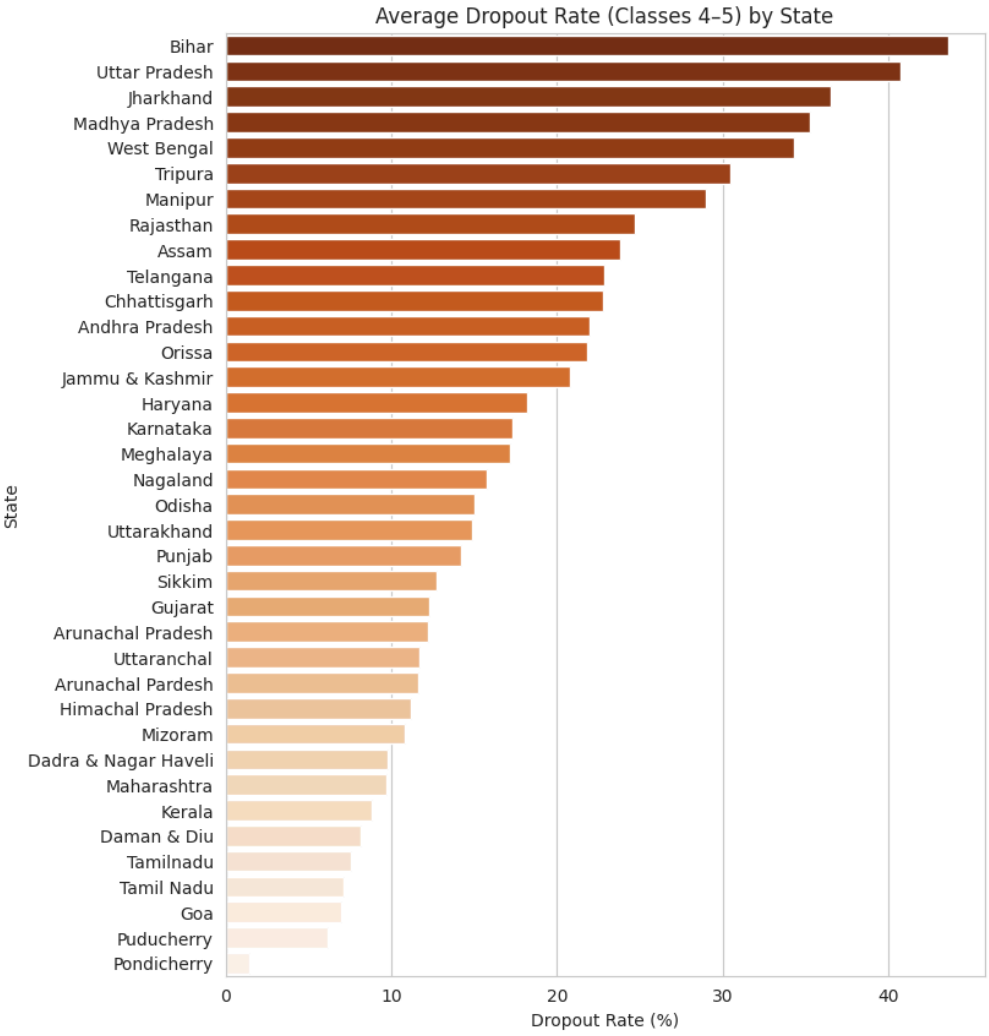

Fig 02 – State-wise dropout rates for Class 4-5

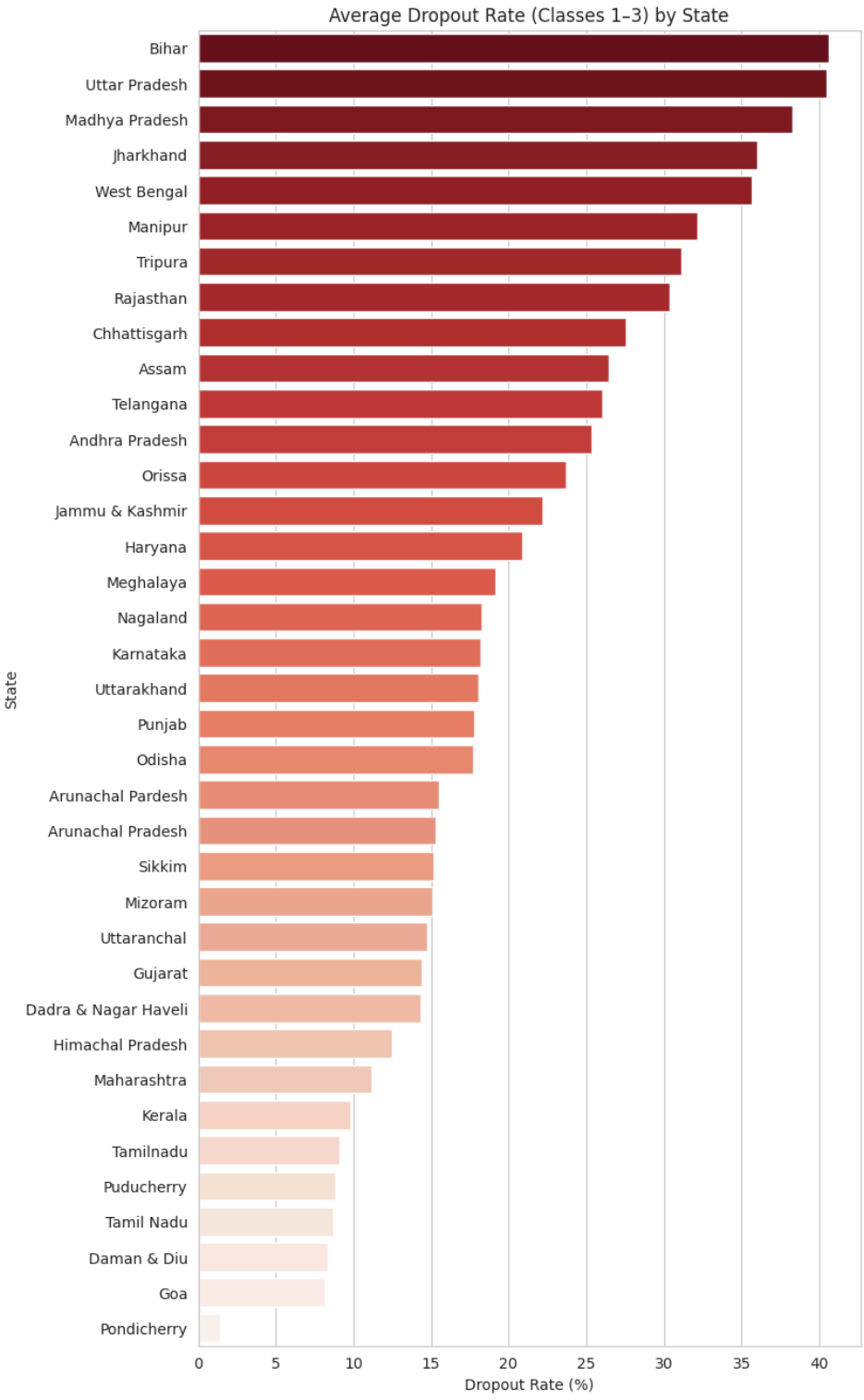

Fig 03 – State-wise dropout rates for Classes 1-3

After analyzing the dropout landscape, we examined the key contributing factors, validating our literature review and hypothesis.

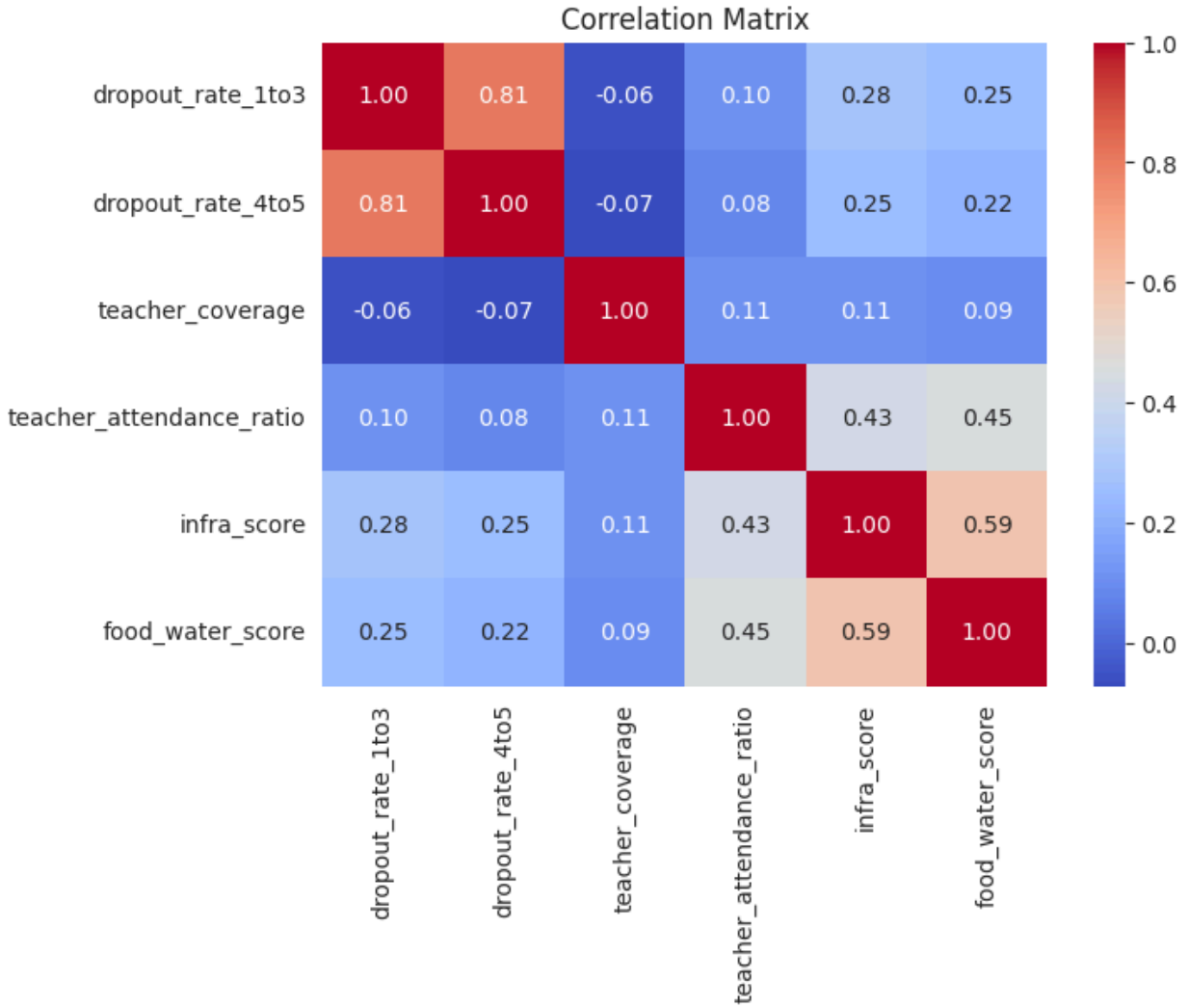

Fig 04 – Confusion Matrix

The above Confusion matrix highlights that Dropout rates are influenced by "infra_score" and "food_water_score"(0.59). We also noticed there was a significant correlation with "teacher_attendance_ratio", "infra_score",(0.43) and "food_water_score"(0.45), as inferred from Fig 04

After understanding the strongest correlations present in the dataset, we have inferred that for high teacher attendance, we can observe a lower dropout rate, as mentioned in Fig 05 (1 being highest attendance and 0 being lowest teacher attendance).

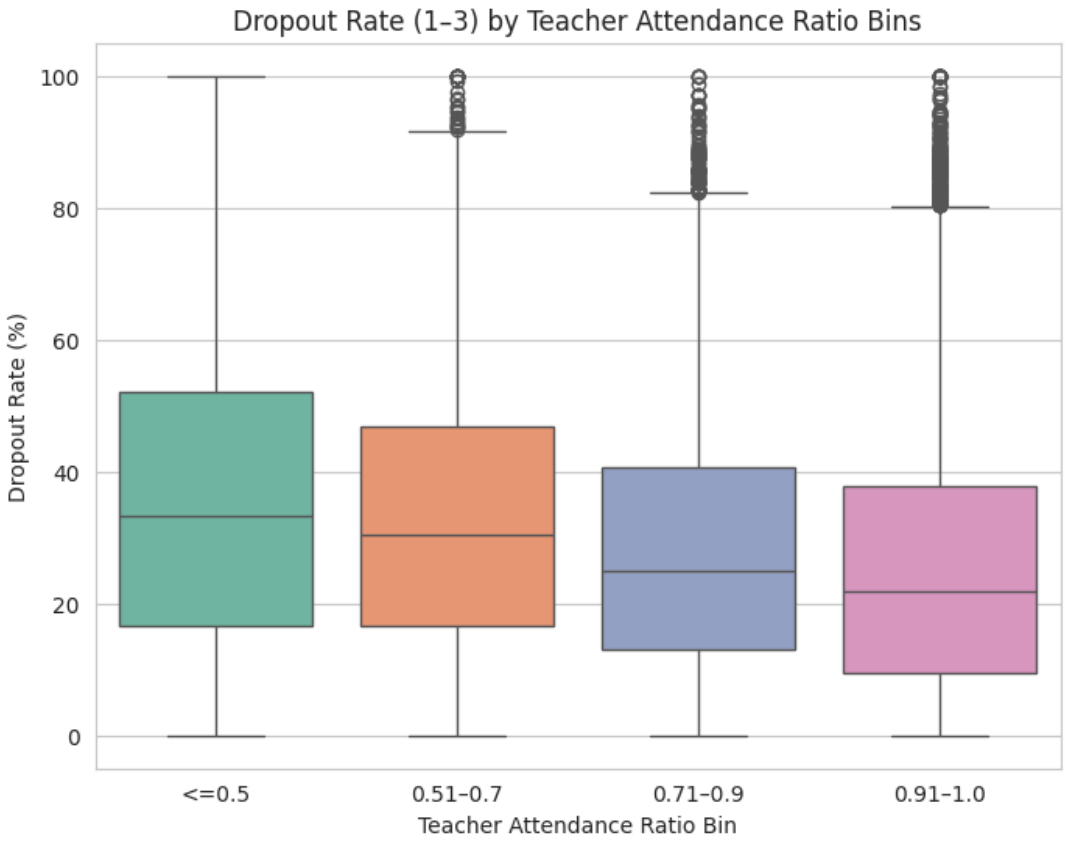

Fig 05 – Dropout rate correlation with Teacher Attendance for classes (1-3)

Four categories of the Teacher attendance ratio were made <=0.5, 0.51-0.7, 0.71-0.9 and 0.91-1.0.

From our next analysis, we infer that the infra_score increased over the years 2010-2018 (in the sampling of 4), and dropout rates fell inversely, as displayed in Fig. 06. However, the data must be taken with additional scrutiny since we observe significant outliers, as depicted in Fig. 05

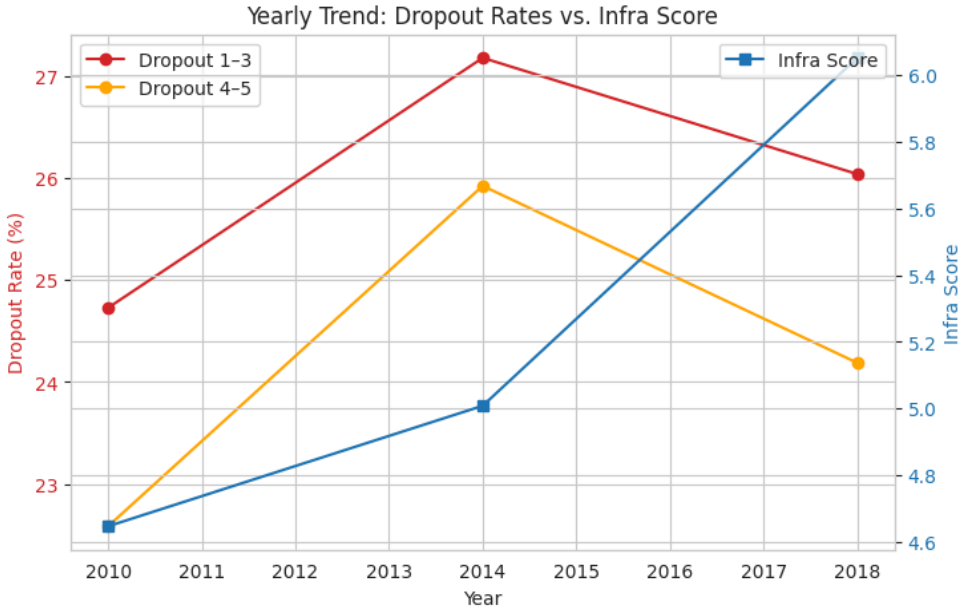

Fig 06 – Infrastructure correlation with Dropout rates

To provide a comprehensive analysis of the factors correlating with dropout rates and their extent of influence, we selected four states representing different regions across India: Assam, Bihar, Maharashtra, and Uttar Pradesh.

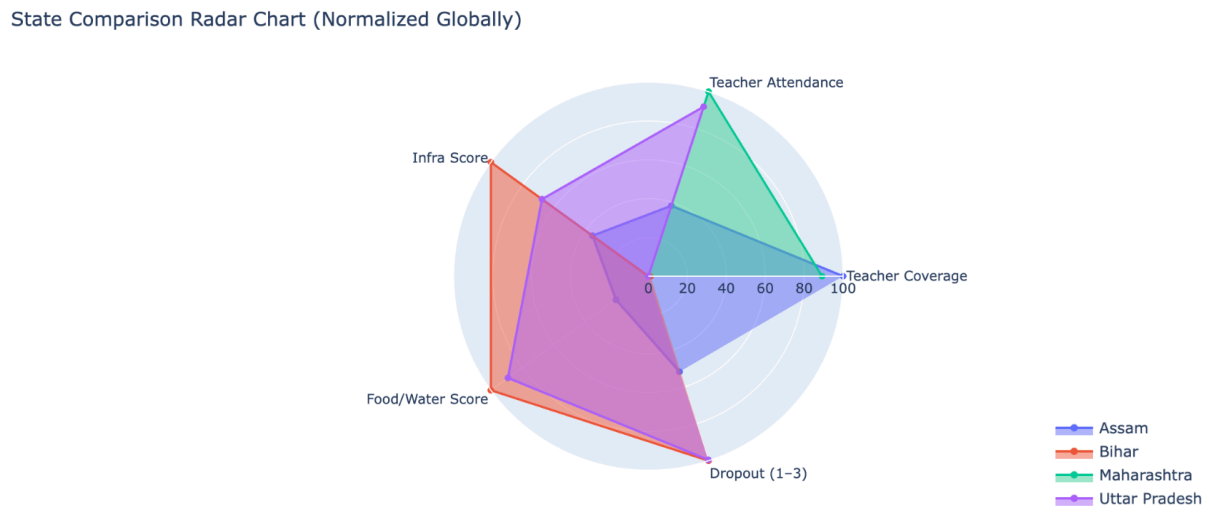

Fig 07 – Infrastructure correlation with Dropout rates

Maharashtra exhibits the lowest dropout rates in classes 1–3, which appears to correlate directly with the highest teacher attendance and coverage. In contrast, Uttar Pradesh has the highest dropout rates in classes 1–3, coinciding with the lowest teacher coverage.

Precaution

This dataset was manually collected by volunteers organized by ASER and Pratham, which may introduce outliers and inaccuracies. Additionally, data collection was not conducted uniformly across all states, resulting in varying sample sizes. Consequently, discrepancies in the dropout rates may partly stem from differences in data availability rather than actual variations in student attrition.

Conclusion

High dropout rates in India remain a critical issue, as highlighted in the literature, which points to financial hardships, inadequate school infrastructure, and socio-cultural barriers as key drivers7. Our study reinforces these findings, demonstrating that dropout rates are strongly linked to school infrastructure, teacher attendance, and mid-day meal availability8. States like Bihar and Uttar Pradesh, with poor facilities and teacher shortages, experience the highest dropout rates, whereas Maharashtra and Tamil Nadu, with better educational resources, see significantly lower dropout rates. Although dataset limitations such as non-uniform data collection must be considered, the results emphasize the urgency of intervention10. Strengthening school infrastructure, ensuring teacher accountability, and enhancing student welfare programs are vital steps in addressing this issue11.

Policy initiatives in India have aimed to tackle these issues. Programs like *Sarva Shiksha Abhiyan* and *Samagra Shiksha Abhiyan* focus on improving school infrastructure, teacher training, and providing support to disadvantaged students12. The *Mid-Day Meal Scheme*, now known as PM POSHAN, has been

effective in increasing attendance and reducing dropout rates by addressing nutritional needs and providing incentives for students to remain in school13. Strengthening these initiatives and ensuring uniform data collection are vital steps in addressing the dropout issue effectively14.

Work Distribution

- Literature Review: Anisha
- Data Procurement and Approval: Nishanth, Rashi
- Data Pre-Processing: Nishanth, Shreya, Divyansh
- Data Analysis: Divyansh, Rashi, Shreya
- Editorial Work: Shreya, Rashi, Nishanth, Anisha

References

1. Dropout Rates in Schools in India
2. School drop-out rate is a serious concern | Matters India
3. TRENDS IN SCHOOL DROPOUT RATE IN INDIA Dr. KV Sridevi
4. Determinant of school dropouts in India: A study through survival analysis approach
5. https://documents1.worldbank.org/curated/en/099308007102434482/pdf/IDU1bac1b6fa165e014df81930512e0d1336463c.pdf
6. https://jhr.uwpress.org/content/early/2019/07/02/jhr.56.2.1216-8430R3
7. https://guides.library.unr.edu/mlacitation/intextcitation
8. https://www.scribbr.com/apa-style/in-text-citation/
9. Sarva Shiksha Abhiyan
10. https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/in_text_citations_the_basics.html
11. https://library.nd.edu.au/instruction/referencing/ama/in_text
12. https://www.scribbr.com/category/apa-style/6th-edition/
13. https://libraryguides.vu.edu.au/ieeereferencing/CitingSources
14. https://apastyle.apa.org/style-grammar-guidelines/citations/basic-principles/multiple-works