# Table of Contents

# ABSTRACT

This study investigates the key determinants influencing the uptake of private coaching among students in India using data from the 75th Round of the National Sample Survey (Schedule 25.2). Employing three supervised machine learning models—Logistic Regression, Random Forest, and Gradient Boosting—we identify the most significant socio-economic, demographic, and educational predictors of private tuition participation. The analysis reveals that **region of residence**, particularly **East and South India**, is a dominant factor, alongside **level of enrolment**, with students in **upper primary and secondary stages** more likely to seek coaching. Additionally, **urban residency**, **higher educational expenditures**, **age**, and **social group identity** also emerge as important drivers. Our findings highlight the disparities in educational support access across regions and social groups and underscore the need for more equitable educational policy interventions to reduce over-reliance on private tutoring.

# 1. Introduction

Private coaching has emerged as a parallel education system in India, increasingly relied upon by students and parents to supplement formal schooling. With growing competition for academic performance and entrance into prestigious higher education institutions, private tuition has become a widely adopted strategy, especially in urban areas and higher school grades. However, this trend raises significant concerns regarding educational equity, affordability, and dependence on out-of-school learning mechanisms.

Understanding the socio-economic and demographic determinants that drive students toward private coaching is essential for evaluating disparities in access to quality education. While prior literature has explored aspects of this phenomenon through surveys or localized case studies, this paper adopts a data-driven, model-based approach using nationally representative microdata.

The objective of this study is to identify and analyze the key factors that influence the likelihood of a student taking private coaching. By employing machine learning models—Logistic Regression, Random Forest, and Gradient Boosting—we aim to uncover patterns and predictors with high explanatory power. The findings will provide policy-relevant insights into the intersection of education, household characteristics, and regional disparities in India's schooling system.

# 2. Data Description and Methodology

## 2.1 Data Source

The study utilizes microdata from the **75th Round of the National Sample Survey (NSS)**, conducted by the **Ministry of Statistics and Programme Implementation (MOSPI)** in 2017–18. Specifically, the data pertains to **Schedule 25.2: "Household Social Consumption: Education"**, which comprehensively captures the educational characteristics of individuals aged 3 to 35 years across Indian households.

To conduct this analysis, we merge information from three relevant blocks:

- **Block 3**: *Household characteristics*
  (`R75252L02`) — provides information on household size, religion, sector, social group, assets like computers, and household-level consumption expenditure.

- **Block 4**: *Demographic and other particulars of household members*
  (`R75252L04`) — includes individual-level demographics such as age, gender,

educational level, disability status, and internet/computer usage.

- **Block 5**: *Educational particulars on basic course of persons aged 3 to 35* (R75252L05) — includes enrolment status, type of institution, course fees, scholarship/stipend status, and whether the individual takes private coaching.

- **Block 6**: *Expenditure on education for current students aged 3 to 35* (R75252L06) — reports educational spending, such as on books, stationery, uniforms, and course-related expenses.

These datasets are merged using household and person-level identifiers,ie HHID and Per_Serial_no. to construct a unified dataset suitable for machine learning classification models.

## 2.2 Key Variables

The dataset used for this analysis contains a total of **30 variables** and **152,558 observations**, capturing a wide range of demographic, educational, and household-level characteristics.

Each observation includes a **household identifier (HHID)**, along with the **state code** (State) and an indicator for whether the household is located in a **rural (1) or urban (2) sector** (Sector). The **household size** (Household_size) reflects the number of members in each household. Religious affiliation is recorded under Religion, while caste or community group is captured using the Social_group variable (e.g., Scheduled Castes, Scheduled Tribes, OBC, Others).

The dataset provides information on physical access to education through three distance-related variables: Distance_primary_school, Distance_upper_school, and Distance_secondary_school, all measured in kilometers. Asset ownership and digital connectivity are represented by binary variables such as HH_Computer and Member_internet, which indicate whether the household owns a computer or has internet access. Economic background is measured by the **monthly per capita consumption expenditure** (HH_Con_exp_rs), and whether any household member aged 3–35 is attending an educational institution is indicated by Any_HH_member_3_35yrs_attnd_edu.

The primary **economic activity of the household** is provided as a categorical variable (HH_Economic_Activity), and Per_serialno is a serial number representing the individual's position in the household roster.

The dataset includes educational details such as the **medium of instruction** (Medium_instruction), the **basic course level** the individual is enrolled in (Enrol_basic_course), and the **type of institution** (Institution_type, e.g., government, private aided, or unaided). Receipt of a **scholarship or stipend** is indicated by Scholarship_stipend. The primary variable of interest, Taking_pvt_coaching, is a binary indicator of whether the individual is taking private tuition (1 = Yes, 0 = No).

Demographic variables such as Age and Gender are recorded for each individual, along with education-related variables like Edu_level_general and Edu_level_technical, which capture the level of general and technical education attained, respectively. Digital literacy is captured through the binary variables Operate_computer_age_5yrs and Operate_internet_age_5yrs, indicating whether individuals aged 5 or above can use a computer or access the internet. The variable Disability_crtificate indicates whether an individual possesses a government-recognized disability certificate.

Lastly, the dataset includes education-related expenditure variables such as Course_fee_amt (annual or term-wise course fee paid) and Books_stationery_uniform_amt (expenditure on books, stationery, and uniforms). The variable Source_funding_basic_exp identifies the **primary source of funding** for basic educational expenses (e.g., household members, scholarship, loan, etc.).

## 2.3 Feature Engineering and Categorical Transformations

To make the data compatible with machine learning models and to improve interpretability, several **categorical variables were re-engineered** using **domain-driven grouping** and **custom functions**. This allowed for better generalization and model performance.

- Household size was categorized into Household_size_category with four groups: Small (1–2 members), Medium (3–5), Large (6–8), and Very Large (9+).
- The primary economic activity of the household was grouped into HH_Economic_Activity_category with categories such as Self-employed, Regular wage/salary, Casual labour, and Others.
- Religion was consolidated into Religion_category distinguishing between Major religions (Hindu, Muslim) and Minor religions (others).
- Caste-based Social_group was recoded as Social_Group_Category comprising SC, ST, OBC, and Others.
- Educational variables were simplified: Edu_level_general was grouped into Low, Basic, Secondary, Higher Education, and Diploma Holders, and Edu_level_technical into Technical Diploma & Degree vs No Technical Education.
- Digital literacy indicators like Operate_computer_age_5yrs and Operate_internet_age_5yrs were binary encoded.
- The medium of instruction was grouped into Medium_instruction_grouped (English, Hindi, Regional, Others).
- course enrollment levels were categorized into Enrol_basic_course_grouped (Pre-primary & Primary, Upper Primary & Secondary, Higher Education, Diploma/Certificate).

- Institution types were reclassified into Institution_category (Government, Private Aided, Private Unaided).
- The source of education funding was grouped under Funding_Category (Household Members, Scholarship, Loan, Student Earnings, Others, Unknown).
- The State variable was grouped into macro-regions (State_Category: North, East, South, West, Central).
- Sector was retained as a binary variable (1 = Rural, 2 = Urban). Finally, the three school distance variables were categorized into Nearby (1–2 km), Moderate (3–4 km), and Far (5+ km) to reflect accessibility to educational institutions.

Following the initial feature engineering process, several variables representing binary categorical outcomes were standardized for consistency across the dataset. Specifically, a group of eight binary variables were recoded to follow a uniform binary scheme. The values were transformed such that **1 indicates "Yes"** and **0 indicates "No"**, by replacing any 2s in the original encoding with 0. This step ensured that all binary variables were aligned and interpretable across modeling pipelines and visualizations.

## 2.4 Final Dataset for Analysis

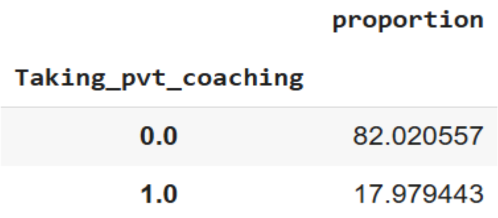
After comprehensive data cleaning, transformation, and feature engineering, the final dataset used for Exploratory Data Analysis (EDA) comprised **152,552 observations** and **30 variables**.

```
<class 'pandas.core.frame.DataFrame'>

Index: 152552 entries, 0 to 181501

Data columns (total 30 columns):

 #   Column                          Non-Null Count   Dtype

---  ------                          --------------   -----

 0   HHID                            152552 non-null  int64

 1   Per_serialno                    152552 non-null  float64

 2   Sector                          152552 non-null  category

 3   HH_Computer                     152552 non-null  category

 4   Member_internet                 152552 non-null  category

 5   HH_Con_exp_rs                   152552 non-null  int64
```

```
6   Any_HH_member_3_35yrs_attnd_edu  152552 non-null  category
7   Scholarship_stipend              152552 non-null  category
8   Age                              122931 non-null  float64
9   Gender                           122931 non-null  category
10  Operate_computer_age_5yrs        120695 non-null  category
11  Operate_internet_age_5yrs        120695 non-null  category
12  Disability_crtificate            122931 non-null  category
13  Course_fee_amt                   130046 non-null  float64
14  Books_stationery_uniform_amt     150577 non-null  float64
15  Household_size_category          152552 non-null  category
16  HH_Economic_Activity_category    152552 non-null  category
17  Religion_category                152547 non-null  category
18  Social_Group_Category            152552 non-null  category
19  Distance_primary_category        152552 non-null  category
20  Distance_upper_category          152552 non-null  category
21  Distance_secondary_category      152552 non-null  category
22  Medium_instruction_grouped       152552 non-null  category
23  Enrol_basic_course_grouped       152552 non-null  category
24  Edu_level_general_category       122931 non-null  category
25  Edu_level_technical_category     122922 non-null  category
26  Institution_category             152552 non-null  category
27  State_Category                   152552 non-null  object
28  Funding_Category                 152552 non-null  object
29  Taking_pvt_coaching              152552 non-null  category
```

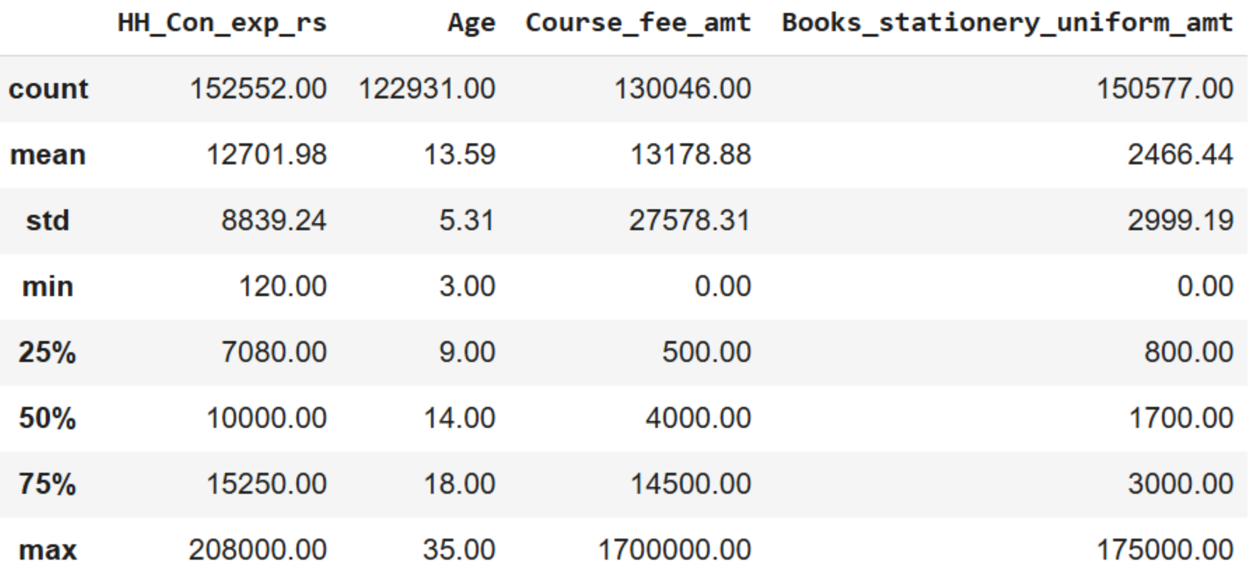
# 3. Exploratory Data Analysis

## 3.1 Class Imbalance:



| | proportion |
|---|---|
| **Taking_pvt_coaching** | |
| **0.0** | 82.020557 |
| **1.0** | 17.979443 |

## 3.2 Descriptive statistics



| | HH_Con_exp_rs | Age | Course_fee_amt | Books_stationery_uniform_amt |
|---|---|---|---|---|
| **count** | 152552.00 | 122931.00 | 130046.00 | 150577.00 |
| **mean** | 12701.98 | 13.59 | 13178.88 | 2466.44 |
| **std** | 8839.24 | 5.31 | 27578.31 | 2999.19 |
| **min** | 120.00 | 3.00 | 0.00 | 0.00 |
| **25%** | 7080.00 | 9.00 | 500.00 | 800.00 |
| **50%** | 10000.00 | 14.00 | 4000.00 | 1700.00 |
| **75%** | 15250.00 | 18.00 | 14500.00 | 3000.00 |
| **max** | 208000.00 | 35.00 | 1700000.00 | 175000.00 |

### Observations from Descriptive Statistics (Numerical Features)

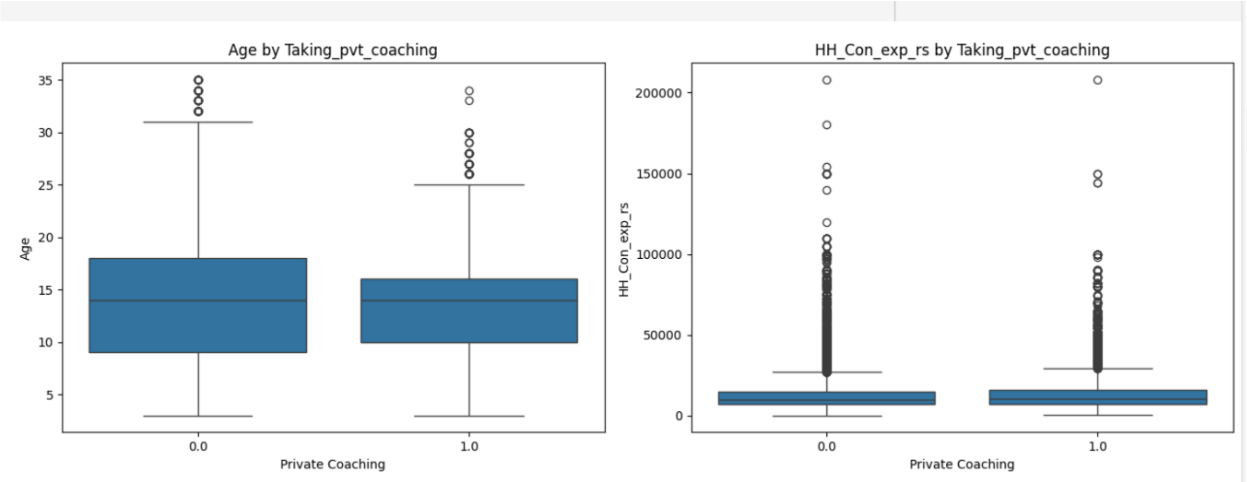
- **Household Consumption Expenditure (`HH_Con_exp_rs`):**
  The average household spends ₹12,702/month with a median of ₹10,000, showing right-skewness and wide variation; 75% of households spend under ₹15,250, indicating concentration in the lower-middle range.

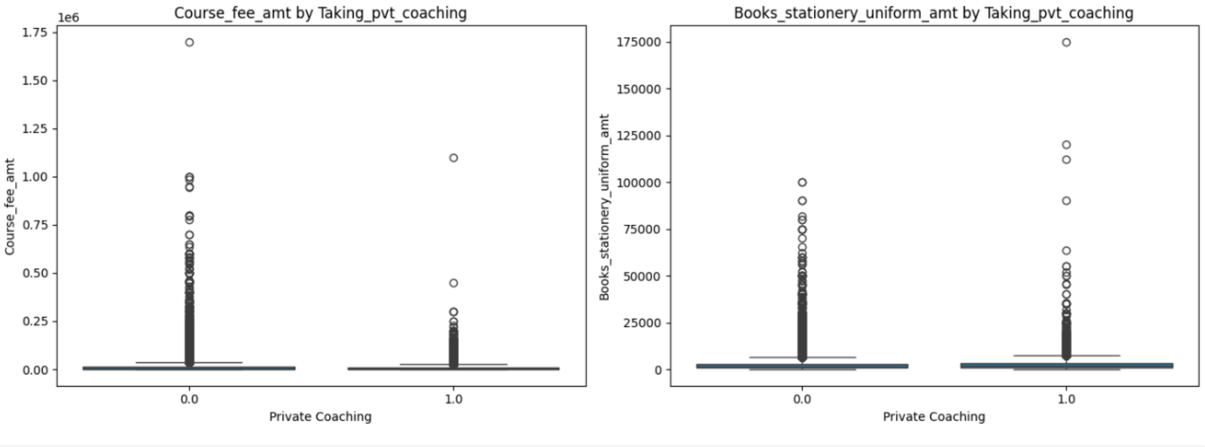
- **Age**:
 The mean age is 13.6 years (median 14), with most students between 9 and 18 years, representing the typical school-going population.

- **Course Fee Amount (`Course_fee_amt`)**:
 With a mean of ₹13,179 and a median of ₹4,000, course fees are highly right-skewed; 75% pay under ₹14,500, but outliers go up to ₹17 lakhs.

- **Books, Stationery & Uniform Expenditure**:
 Average spend is ₹2,466 (median ₹1,700), with 75% of students spending less than ₹3,000; outliers above ₹1.75 lakhs suggest variation in non-fee educational costs.

## 3.3 Bivariate Analysis (Numerical Features vs. Taking Private Coaching)





## Key Takeaways from Bivariate Analysis (Numerical Features)

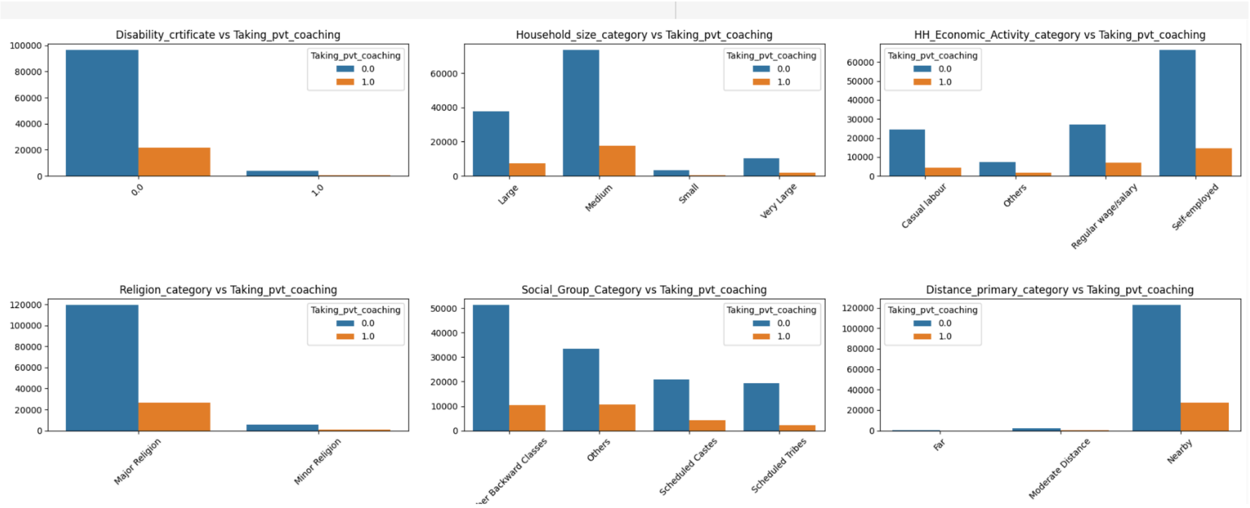
- **Age**:
 Students taking private coaching are typically older, with a median age around 14–15 years. Most fall within the 10–18 age group, which aligns with board exam preparation years. Occasional older outliers (25–35) likely represent higher education students.

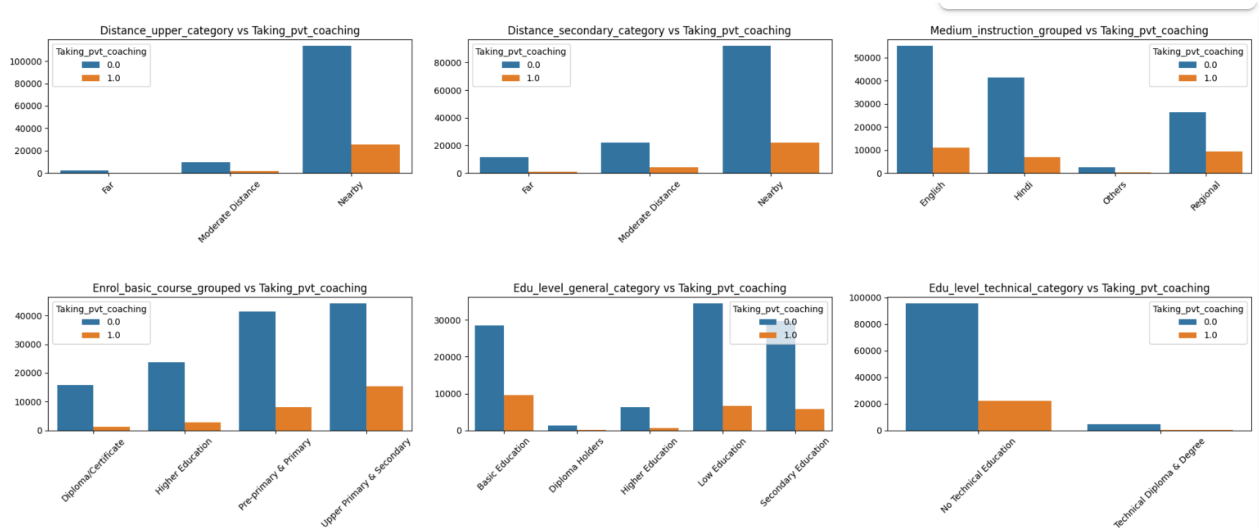
- **Household Consumption Expenditure (`HH_Con_exp_rs`)**:
 Coaching takers come from households with slightly higher average spending. While expenditure is right-skewed with many outliers above ₹1 lakh, the significant overlap suggests household wealth is influential but not exclusive.

- **Course Fee Amount**:
 Students receiving private coaching generally pay higher course fees, with a noticeably elevated median. Some outliers reach up to ₹17 lakhs, pointing to a subset enrolled in elite or high-cost institutions.

- **Books, Stationery & Uniform Expenditure**:
 Educational supply expenses are higher among coaching participants. Though the amounts are modest for most, outliers above ₹1 lakh indicate that some students incur substantial non-tuition educational costs, reflecting broader academic investment.

## 3.4 Bivariate Analysis: Target vs Categorical Features





Based on the count plots, several key patterns emerge regarding the factors associated with students taking private coaching. Students from **medium-sized households** and those whose families are engaged in **regular wage or salaried employment** show higher coaching participation compared to those from large or very large families or from casual labor backgrounds. While students from both **major and minor religions** participate in coaching, the rate is slightly higher among those from **major religions**. In terms of caste, **students from 'Other' and 'OBC' categories** have higher coaching uptake compared to Scheduled Castes and especially Scheduled Tribes. Interestingly, **students living closer to primary, upper, and secondary schools** (categorized as ‘Nearby’) are more likely to take coaching, suggesting that **proximity to educational infrastructure does not substitute for private tutoring**.

When it comes to educational attributes, students studying in **regional-medium institutions** and those enrolled in **upper primary or secondary courses** have significantly higher participation in coaching, indicating a focus on board-exam-relevant stages. Those with **basic general education** and **no technical education** dominate the coaching segment, although diploma holders and higher education students also show non-trivial participation. Private coaching appears more common in **government institutions**, possibly due to a perceived gap in quality or support. Regionally, **urban students** are more likely to attend coaching than rural ones, and both **male and female students** participate at similar rates, suggesting no strong gender gap in access to private tuition.

# 4. Logistic Regression on the Target Variable

## 4.1 Train-Test Split

To evaluate the performance of the logistic regression model effectively, the dataset was split into **training and testing sets** using the `train_test_split()` function from Scikit-learn. The `Taking_pvt_coaching` variable was treated as the binary target, while all other variables formed the feature set.

A **stratified split** was used to ensure that the proportion of students taking and not taking private coaching remained consistent across both the training and testing sets. The data was split using an **80-20 ratio**, where 80% of the data was used for model training and 20% was reserved for validation. The `random_state` was fixed at 42 to ensure reproducibility of results.

## 4.2 Data Preprocessing

To prepare the data for logistic regression, several preprocessing steps were applied to handle skewness, missing values, categorical encodings, and feature scaling.

🔹 **Log-Transformation of Skewed Variables**

Three continuous expenditure-related variables — HH_Con_exp_rs (household consumption expenditure), Course_fee_amt (course fees), and Books_stationery_uniform_amt (expenditure on school supplies) — exhibited substantial right-skewness. To normalize their distributions and reduce the influence of outliers, we applied a natural logarithmic transformation using log1p, which is suitable for handling zero values. The original skewed variables were then dropped from the dataset to avoid redundancy.

🔹 **Categorical Variable Encoding**

The dataset contained both **ordinal** and **nominal** categorical features. These were processed using separate pipelines:

- **Ordinal Features**:
  Variables like Household_size_category, Edu_level_general_category, and distance-to-school variables were encoded using **Ordinal Encoding**, preserving their inherent order. Categories were explicitly specified to maintain consistency in encoding.

- **Nominal Features**:
  Variables such as Religion_category, Social_Group_Category, Medium_instruction_grouped, Institution_category, and others were treated as nominal. These were encoded using **One-Hot Encoding** with the first category dropped to avoid multicollinearity.

🔹 **Numerical Feature Scaling and Imputation**

All remaining numerical features were standardized using **z-score normalization** (StandardScaler), which ensures that variables are centered around zero with unit variance. This step is particularly important for gradient-based models like logistic regression. Missing values in numerical features were imputed using the **median**, which is robust to outliers.

🔹 **Column-wise Pipeline Integration**

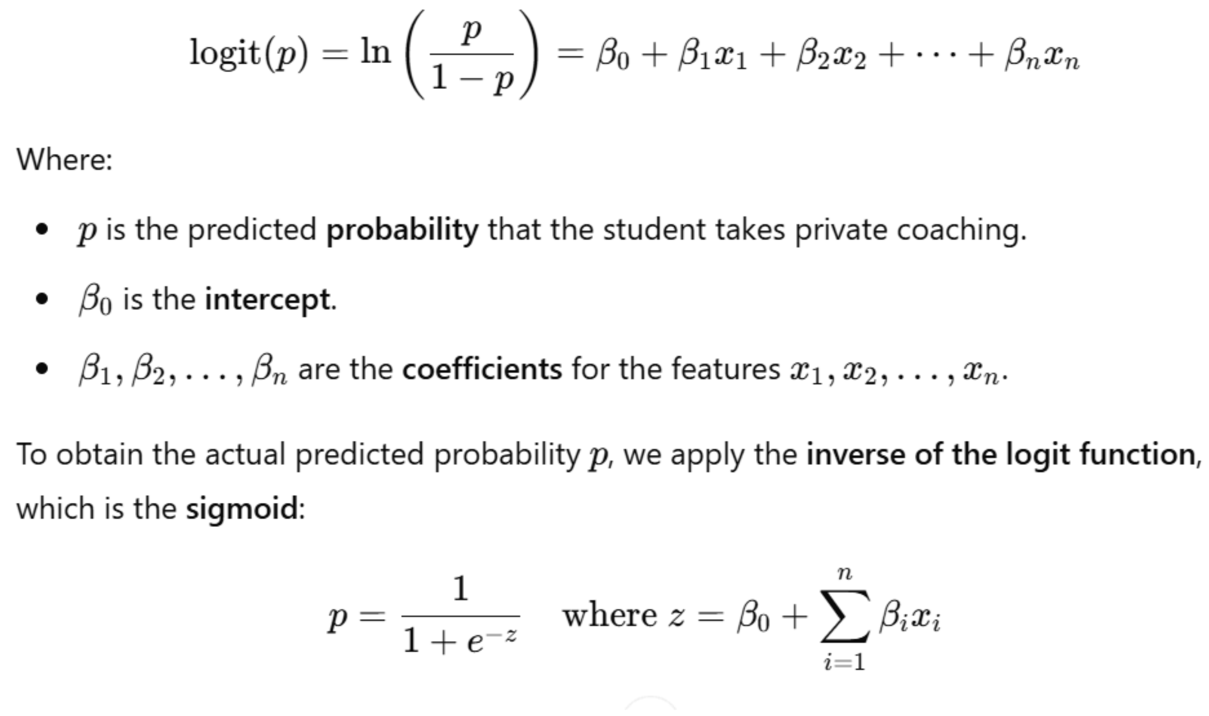
All preprocessing steps were integrated into a **ColumnTransformer**: ordinal_pipeline for ordinal features, nominal_pipeline for nominal features,numeric_pipeline for numerical features. This approach ensures clean, modular, and scalable preprocessing, and serves as a reusable component within the final model pipeline.

## 4.3 Model Building and Hyperparameter Tuning (Logistic Regression)

A logistic regression model was trained using a complete machine learning pipeline consisting of **data preprocessing**, **oversampling**, and **classification** steps.

**How Logistic Regression Works:**

Logistic regression is a statistical model used for **binary classification** problems, where the goal is to estimate the probability that an observation belongs to one of two categories (in this case, whether a student **takes private coaching (1)** or **does not (0)**). Unlike linear regression, logistic regression models the **log-odds** of the probability using the **logit function**:



$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where:

- $p$ is the predicted **probability** that the student takes private coaching.
- $\beta_0$ is the **intercept**.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the **coefficients** for the features $x_1, x_2, \ldots, x_n$.

To obtain the actual predicted probability $p$, we apply the **inverse of the logit function**, which is the **sigmoid**:

$$p = \frac{1}{1+e^{-z}} \quad \text{where } z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

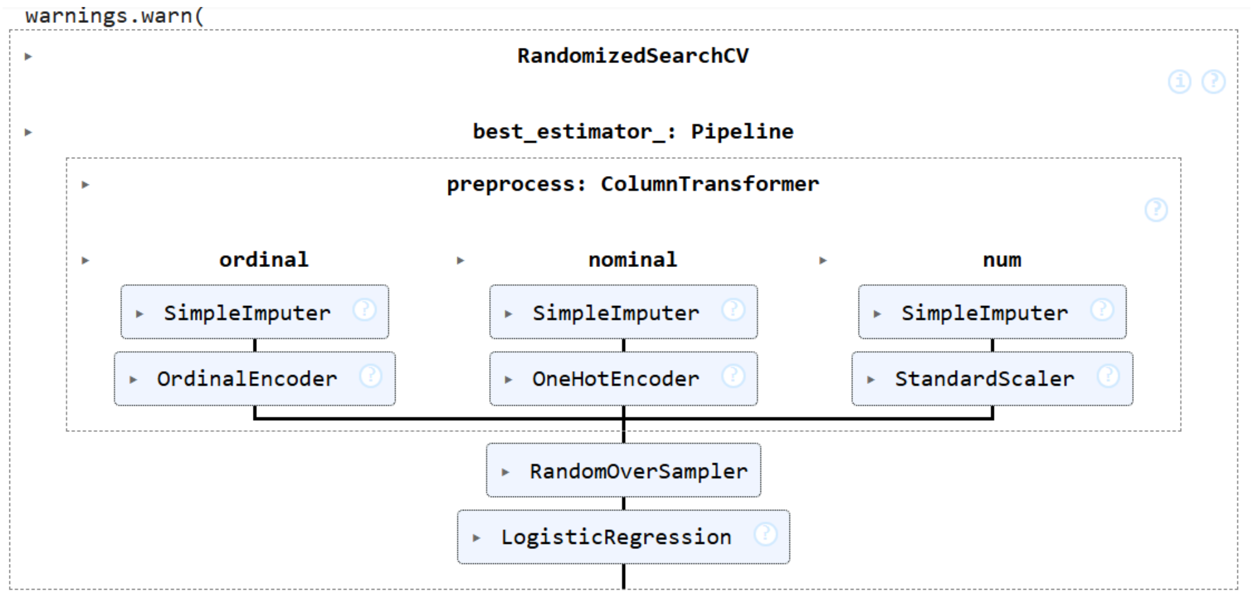
This transformation ensures that the output probability always lies between 0 and 1.

In our case, the model estimates the probability that a student takes private coaching based on features such as household type, state, education level, medium of instruction, and more. A custom threshold (0.6131) was used to convert these predicted probabilities into a binary decision.

**Handling Class Imbalance and Hyperparameter Tuning:**

Since the target variable (Taking_pvt_coaching) was moderately imbalanced, we incorporated **random oversampling** using RandomOverSampler from imblearn to ensure that the minority class (students taking private coaching) was adequately represented during training. The preprocessing pipeline defined earlier was seamlessly integrated using ImbPipeline, ensuring consistent treatment of training data throughout all modeling steps.

The logistic regression model was configured with **class weights balanced**, and **hyperparameter tuning** was performed using RandomizedSearchCV to identify the best combination of regularization strength (C), penalty type (l1, l2, or elasticnet), solver (liblinear, saga, lbfgs), and l1_ratio (for elastic net). A **5-fold cross-validation** strategy was used during tuning, and 50 random combinations were evaluated to optimize performance. This approach helped in systematically finding the best regularized logistic model suited to the structure and imbalances present in the dataset.



```
warnings.warn(
```

## 4.3 Model Evaluation

After training the logistic regression model with optimal hyperparameters, its performance was evaluated on both the **training** and **testing** datasets using key classification metrics. A **custom threshold of 0.6131** was applied to convert predicted probabilities into binary labels, aiming to improve the balance between precision and recall.

The evaluation metrics are as follows:

- **Accuracy**:
    - **Test Accuracy**: **79.89%**
    - **Training Accuracy**: **73.43%**
      These values suggest a good generalization to unseen data, with no significant overfitting.
- **Precision**:
    - **Test Precision**: **45.58%**
    - **Training Precision**: **37.64%**
      This indicates that when the model predicts a student will take coaching, it is

correct roughly 46% of the time on test data. Lower training precision suggests the model is more conservative during training.

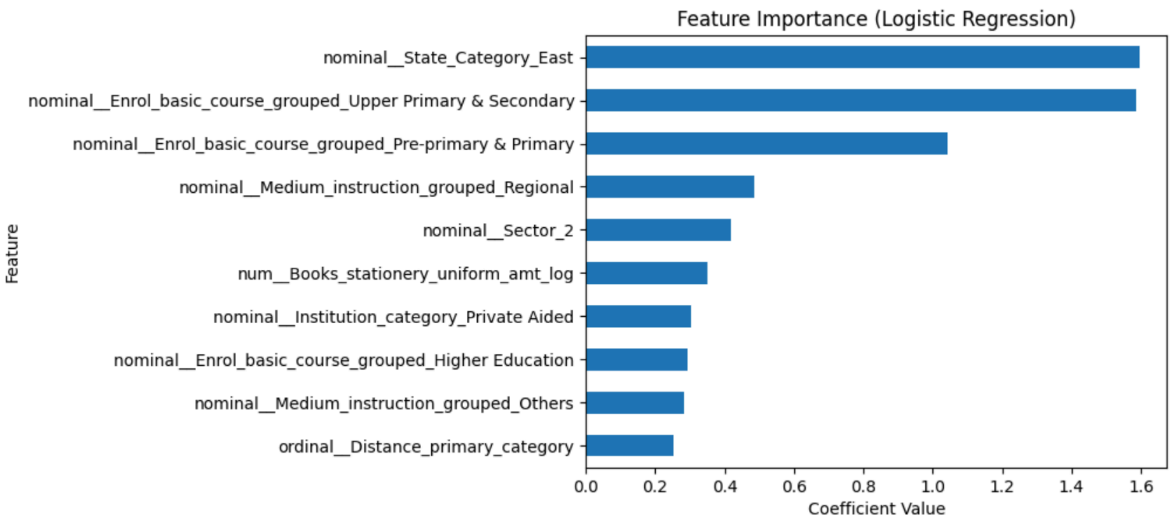
- **Recall (Sensitivity)**:
    - **Test Recall**: **61.05%**
    - **Training Recall**: **72.77%**
      The model correctly captures over 61% of actual coaching takers in the test set, showing strong sensitivity.
- **F1 Score**:
    - **Test F1 Score**: **52.19%**
    - **Training F1 Score**: **49.61%**
      The F1 score, balancing precision and recall, reflects moderate classification performance, with a better test F1 than training.
- **AUC-ROC (Area Under the ROC Curve)**: **Test AUC-ROC**: **0.8063**
  AUC above 0.80 indicates strong discriminative power of the model in distinguishing coaching and non-coaching cases.

Overall, the model achieves a good balance between predictive power and generalization, with strong recall and AUC performance. The slight gap between training and testing metrics suggests some variability but no critical overfitting.

## 4.4 Feature Importance from Logistic Regression

To interpret the drivers of private coaching participation, we examined the **feature importances derived from the logistic regression model**. Since logistic regression is a linear model, the magnitude and sign of the coefficients directly indicate how each feature affects the likelihood of the outcome (`Taking_pvt_coaching = 1`).

The top 10 most influential features, ranked by the absolute value of their coefficients, are visualized in the chart below.



Key Insights:

- **State_Category_East** had the strongest positive influence, suggesting that students in Eastern states are significantly more likely to take private coaching, possibly reflecting regional educational dynamics.

- **Enrollment in Upper Primary & Secondary levels** and **Pre-primary & Primary** were also strong positive predictors, highlighting that coaching is most common during key school years.

- **Medium of instruction** matters: students studying in **Regional language mediums** had a higher likelihood of taking coaching, possibly due to gaps in classroom instruction or aspirations for upward mobility.

- **Urban residence (Sector_2)** also positively influenced coaching participation, aligning with higher access to coaching facilities in cities.

- Higher **spending on books and stationery** was associated with coaching, reinforcing the idea that private coaching is part of a broader pattern of increased educational investment.

- Enrollment in **Private Aided Institutions** and **Higher Education** also appeared among the top features, indicating that coaching needs may extend beyond school years.

These findings provide interpretability and help validate earlier descriptive and bivariate observations with statistically significant model-driven evidence.

# 5. Results

The objective of this study was to examine the factors influencing students' participation in private coaching using a logistic regression model trained on nationally representative educational microdata.

## 🔹 Model Performance

The final logistic regression model, tuned via `RandomizedSearchCV` and evaluated using a custom threshold (0.6131), achieved the following metrics:

- **Test Accuracy**: 79.89% **Training Accuracy**: 73.43%
- **Test Precision**: 45.58% **Training Precision**: 37.64%
- **Test Recall**: 61.05% **Training Recall**: 72.77%
- **Test F1 Score**: 52.19% **Training F1 Score**: 49.61%
- **AUC-ROC (Test)**: 0.8063
-

These results suggest that the model performs well in distinguishing between students who do and do not take private coaching, with strong recall and AUC-ROC scores indicating robust classification power even on unseen data.

## 🔹 Key Predictors of Private Coaching Participation

The most influential variables, as determined by the magnitude of logistic regression coefficients, included:

- **Region**: Students from **Eastern states** were significantly more likely to attend coaching.
- **Education Stage**: Enrolment in **Upper Primary & Secondary** and **Primary** levels were strong predictors, confirming that coaching is most prevalent during school years.
- **Medium of Instruction**: Students in **Regional-language institutions** were more likely to take coaching, possibly to compensate for quality gaps.

- **Urban Residence**: Coaching participation was higher in **urban areas**, where such services are more accessible.

- **Household Educational Spending**: Higher expenditure on **books and supplies** was associated with coaching, reflecting overall academic investment.

- **Institution Type**: Those studying in **Private Aided** institutions and those pursuing **Higher Education** also showed elevated coaching participation.

### 🔹 Descriptive Insights

- Coaching was most common among students aged **10–18**, aligning with key academic stages.

- Households with **moderate income levels** (as measured by monthly consumption expenditure) were more likely to invest in private coaching.

- Spending patterns for course fees and school supplies were highly skewed, indicating significant variation in education-related financial commitments.

These findings provide both statistical and policy-relevant insight into the socio-economic and institutional factors driving private coaching in India. The results suggest that private tutoring is not just an urban elite phenomenon but is shaped by a complex interplay of education stage, location, school type, and household investment capacity.

## 6. Robustness Considerations

To ensure the reliability of our findings, we incorporated several robustness-enhancing practices. A 5-fold cross-validation framework was used during model training to validate performance across different data splits, reducing dependence on a single sample. Randomized hyperparameter tuning further allowed exploration of a wide parameter space to avoid overfitting. We also compared training and testing metrics for accuracy, precision, recall, and AUC-ROC, which were found to be reasonably close, confirming the model's generalizability. A custom decision threshold was selected based on probability scores to improve the balance between sensitivity and precision, reinforcing the model's stability and performance across metrics.

# 7. Discussion

The results of this study reveal important patterns about the determinants of private coaching in India and provide meaningful insight into the evolving dynamics of supplementary education in the country.

### 🔹 Private Coaching as a Response to Institutional Gaps

The strong association between **enrolment in school-level education (primary and secondary)** and private coaching participation highlights a systemic reliance on external academic support during foundational learning years. This reflects widespread concerns about the **quality and uniformity of instruction** in formal schooling, particularly in public institutions, prompting parents to seek reinforcement through tuition.

The finding that students from **Regional-language medium schools** and those enrolled in **Private Aided institutions** are more likely to opt for coaching suggests that private tutoring is often used to **bridge learning gaps or enhance performance**, especially in institutions where infrastructure or pedagogy may be weaker.

### 🔹 Urban-Rural Divide and Educational Access

Urban students were significantly more likely to take private coaching than their rural counterparts, reflecting the **urban concentration of coaching centers**, **greater awareness**, and **higher availability of resources**. However, the presence of coaching even among lower-middle expenditure households (as shown in the descriptive statistics) indicates that families across economic strata are willing to **make financial sacrifices** for better educational outcomes.

This aligns with existing literature on India’s "shadow education" sector — where coaching has become a parallel industry, especially in metros and state capitals, targeting board exams, competitive entrances, and general academic enhancement.

### 🔹 Regional Imbalances

Students from **Eastern India** were more likely to take private coaching, suggesting either a higher cultural emphasis on academic competition or perceived deficiencies in school systems in those states. This points to the **regionally uneven quality of education delivery** and may also reflect the **intense competition for limited quality opportunities** in areas with high population density and fewer reputed institutions.

## 🔹 Socio-Economic Drivers of Coaching

Higher expenditure on **books, stationery, and uniforms**, along with increased **course fees**, were found to be significant predictors of coaching. This reinforces the idea that private coaching is part of a **broader household educational investment strategy**, where coaching is not standalone, but often complements other forms of spending on a child's academic journey.

Interestingly, the model also found that **coaching is not exclusive to the wealthiest households**. Middle-income families form a large part of the coaching demographic, signaling a **widespread perception that success in education — and by extension, social mobility — hinges on private tutoring**, regardless of one's socio-economic class.

## 🔹 Implications for Policy

These findings have several policy implications:

- There is a need to **strengthen foundational schooling**, especially in government and aided institutions, to reduce dependence on external coaching.

- Improved **teacher training**, **curriculum alignment**, and **student feedback mechanisms** can help make classroom learning more effective.

- Regulatory oversight or formal integration of the coaching sector (through accreditation or voluntary registration) may help standardize quality and pricing.

- Education policy must also address the **urban-rural divide** in access to both formal and supplementary learning resources, ensuring equitable academic support.

In sum, private coaching in India emerges not merely as a luxury for the privileged, but as a strategic educational choice made by families navigating a competitive and uneven schooling environment. While it helps fill existing gaps, its growing pervasiveness underscores the **urgent need for systemic improvements** in the country's public education framework.