#### **2. Data Collection and Preprocessing**

The dataset was collected from internally displaced persons (IDP) sites across northern Nigeria, including states such as Adamawa, Borno, and Yobe. The dataset comprises 74 features, including demographic, educational, health, and socioeconomic variables. Key preprocessing steps include:

1. **Target Variable Engineering**:
   1. The target variable, school\_attendance, was derived from the feature Children at the site attending school. Attendance percentages were mapped to binary values:
      1. 0 (Low attendance: "none", "<25%", "25%-50%").
      2. 1 (High attendance: "51%-75%", ">75%").
2. **Handling Missing Data**:
   1. Numeric features were imputed using the median strategy.
   2. Categorical features were imputed using the most frequent value or a constant placeholder (e.g., "missing").
3. **Encoding Categorical Variables**:
   1. Ordinal encoding was applied to distance-related features (e.g., Distance to nearest education facility) to preserve logical order.
   2. One-hot encoding was used for high-cardinality categorical features (e.g., State, Occupation trade).

#### **3. Model Development**

A machine learning pipeline was developed to train and evaluate multiple models, ensuring robustness and interpretability. The pipeline consists of the following steps:

1. **Preprocessing**:
   1. Numeric features were scaled using StandardScaler.
   2. Categorical features were encoded using OneHotEncoder or OrdinalEncoder.
   3. A ColumnTransformer was used to apply different preprocessing steps to numeric and categorical features.
2. **Model Selection**:
   1. Six models were evaluated:
      1. **Logistic Regression**: For interpretability and baseline performance.
      2. **Random Forest**: For handling non-linear relationships and feature importance.
      3. **Decision Tree**: For simplicity and interpretability.
      4. **Support Vector Machine (SVM)**: For high-dimensional data.
      5. **LightGBM**: For efficiency and performance on large datasets.
      6. **CatBoost**: For handling categorical features natively.
3. **Cross-Validation**:
   1. A 5-fold stratified cross-validation was used to evaluate model performance, ensuring robustness and reducing overfitting. Metrics such as accuracy, precision, recall, and F1-score were computed.
4. **Model Evaluation**:
   1. Models were evaluated on a held-out test set (20% of the data). Classification reports and accuracy scores were generated to compare performance.

#### **4. Explainability and Ethical Considerations**

To ensure transparency and ethical AI practices, SHAP (SHapley Additive exPlanations) was used to interpret model predictions. Key steps include:

1. **Global Interpretability**:
   1. SHAP summary plots were generated to rank features by their impact on predictions. This helps identify key drivers of school attendance, such as distance to schools, teacher availability, and gender ratios.
2. **Gender-Sensitive Analysis**:
   1. SHAP dependence plots were used to analyze interactions between gender-sensitive features (e.g., Female hygiene and Security is provided on-site). This highlights how gender disparities influence attendance.
3. **Regional Disparities**:
   1. SHAP values were analyzed at the state level to identify systemic barriers in specific regions (e.g., Borno vs. Adamawa).
4. **Bias Mitigation**:
   1. Models were evaluated for fairness across gender and location. For example, SHAP values were used to detect if predictions disproportionately favored certain states or genders.

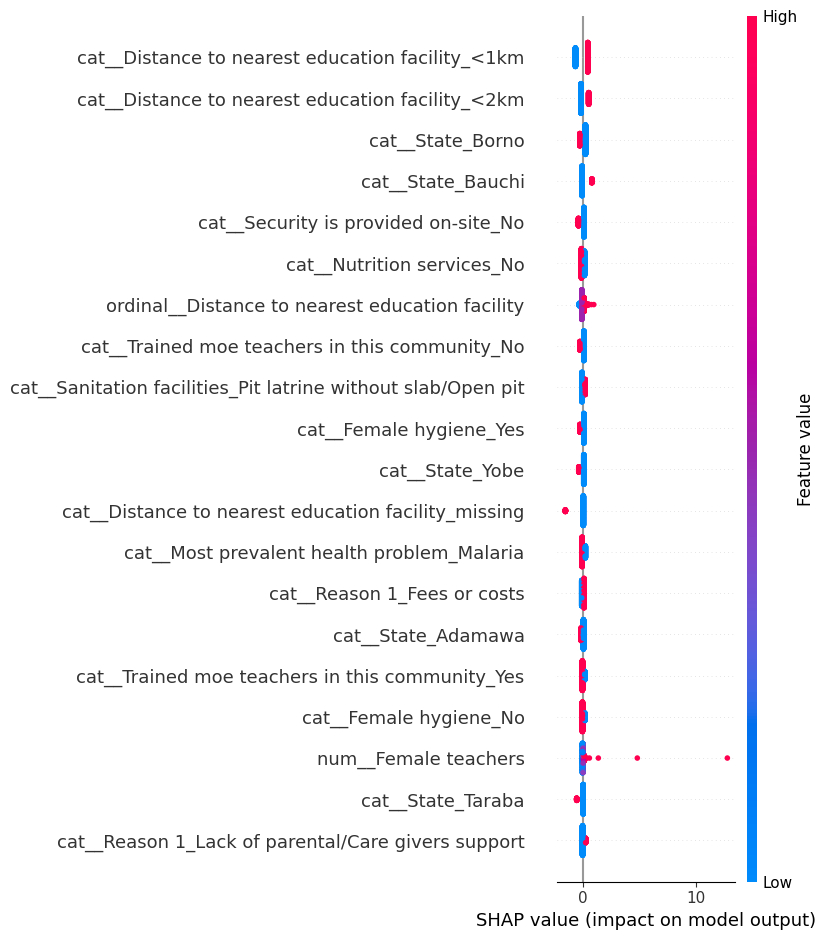
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression | 0.697 | 0.608 | 0.324 | 0.421 |
| Random Forest | 0.792 | 0.747 | 0.587 | 0.658 |
| Decision Tree | 0.713 | 0.575 | 0.597 | 0.586 |
| Support Vector Machine | 0.659 | 0.350 | 0.006 | 0.012 |
| LightGBM | 0.781 | 0.706 | 0.615 | 0.657 |
| CatBoost | 0.783 | 0.733 | 0.573 | 0.643 |

#### **Key Findings**

1. **Best Performing Model**:
   1. **Random Forest** achieved the highest accuracy (0.792) and F1-score (0.658), indicating strong overall performance.
   2. It also demonstrated the highest precision (0.747) and recall (0.587), suggesting a good balance between identifying true positives and minimizing false positives.
2. **Interpretable Models**:
   1. **Logistic Regression** provided moderate accuracy (0.697) but had the lowest recall (0.324), indicating challenges in correctly identifying positive cases (high attendance).
   2. **Decision Tree** performed reasonably well (accuracy: 0.713, F1-score: 0.586) and offered interpretability, making it suitable for stakeholder discussions.
3. **Ensemble Models**:
   1. **LightGBM** and **CatBoost** performed similarly, with LightGBM slightly outperforming CatBoost in recall (0.615 vs. 0.573) and F1-score (0.657 vs. 0.643).
   2. Both models demonstrated robust performance, making them suitable for deployment in resource-constrained settings.
4. **Support Vector Machine (SVM)**:
   1. SVM performed poorly, with the lowest recall (0.006) and F1-score (0.012). This suggests that SVM is not well-suited for this dataset, likely due to its sensitivity to imbalanced data and high dimensionality.

Logistic regression

SHAP analysis revealed that proximity to educational facilities (<1km and <2km) emerged as the strongest predictors in our model. Geographic location, particularly in states such as Borno and Bauchi, also demonstrated significant influence. Infrastructure factors, including security provision and nutrition services, showed moderate effects, while socioeconomic factors such as fees and parental support demonstrated measurable but comparatively lower impact on the model's predictions.



Lightgbm

Key findings from the analysis:

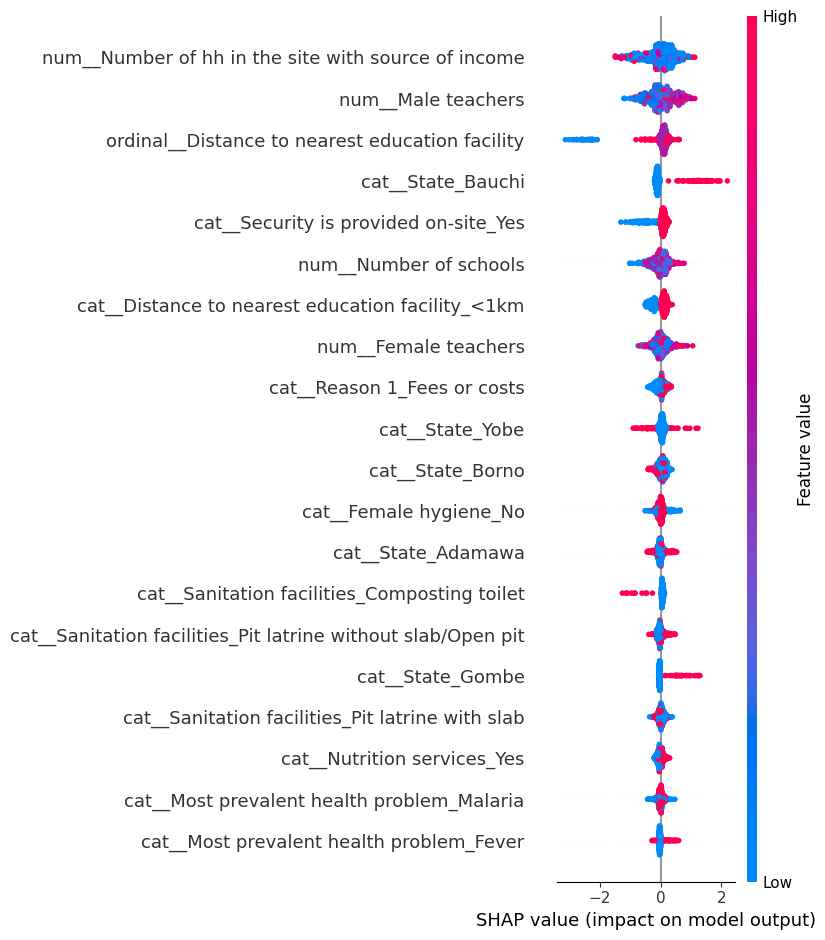
The most influential features (in order of impact magnitude) are:

1. Number of households with source of income
2. Number of male teachers
3. Distance to nearest education facility
4. State\_Bauchi (categorical)
5. On-site security provision

Notable patterns:

* Socioeconomic factors: The number of households with income sources shows a complex relationship, with both positive and negative impacts depending on the value
* Educational resources: Male teacher count demonstrates substantial bidirectional influence
* Geographic factors: Distance to educational facilities shows clear negative correlation with the outcome when distances are greater
* Regional variations: Certain states (particularly Bauchi and Gombe) show strong positive associations
* Infrastructure: Security provision and sanitation facilities demonstrate moderate but consistent effects

For your research, this analysis suggests that both socioeconomic and educational infrastructure factors play crucial roles in the model's predictions, with household income sources and teacher availability being particularly significant determinants.



Key Findings:

1. Teacher Gender Distribution:

* The strongest correlation appears between male and female teachers (0.96), indicating a highly positive relationship
* This suggests schools tend to maintain balanced gender ratios in their teaching staff

1. School Resource Independence:

* The number of schools shows very weak correlations with other variables (all correlations ≤ 0.02)
* This suggests school quantity operates independently of other educational factors

1. Income Source Relationships:

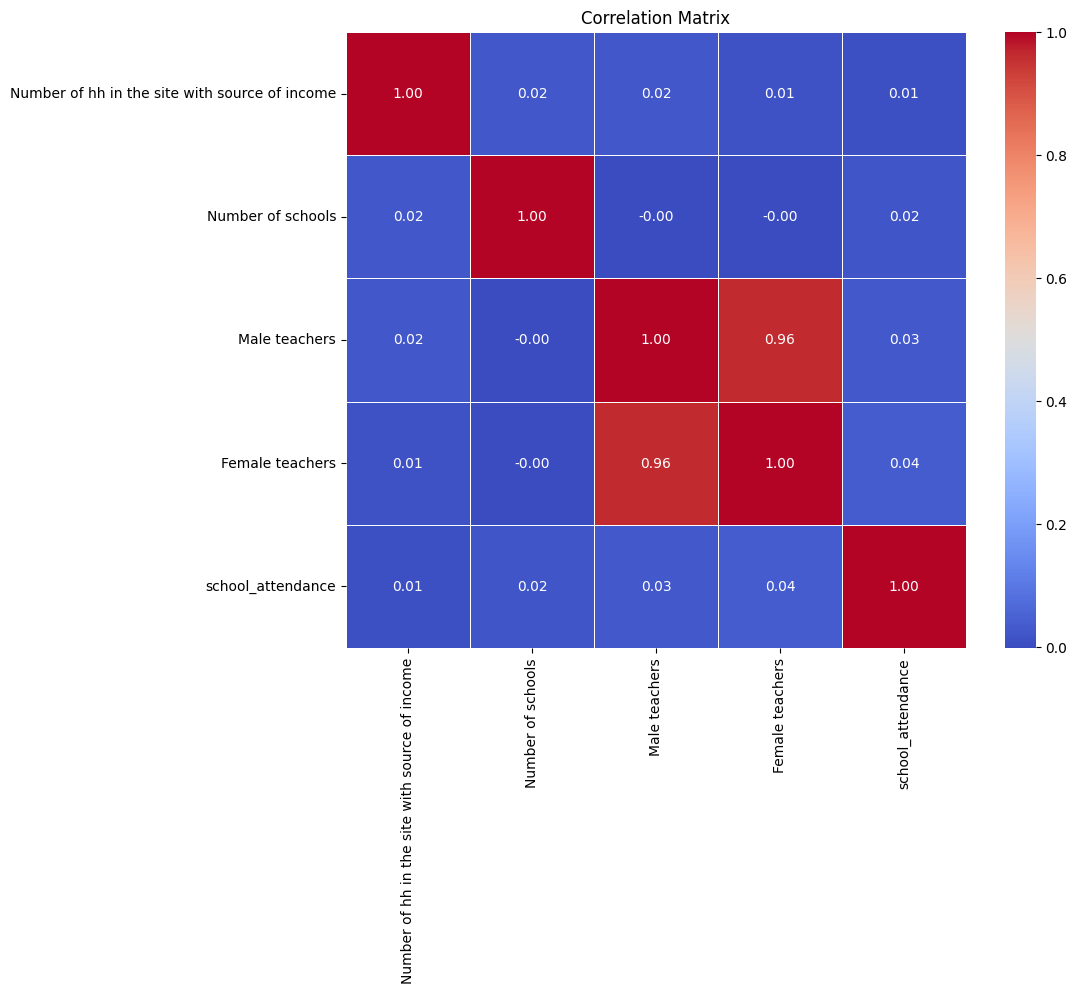
* Household income sources show minimal correlation with other variables (all correlations ≤ 0.02)
* This indicates that household economic status has limited direct relationship with educational staffing patterns

1. School Attendance Patterns:

* School attendance shows weak positive correlations with teacher variables (0.03-0.04)
* The relationship with household income sources is negligible (0.01)

For your research, you might want to emphasize that:

"The correlation analysis reveals interesting patterns in educational resource distribution. Most notably, there is a strong positive correlation (r = 0.96) between male and female teacher numbers, suggesting systematic staffing patterns that maintain gender balance in teaching staff. However, other variables such as household income sources and school quantities show minimal correlations (r ≤ 0.02) with other factors, indicating that these aspects operate relatively independently within the educational system."



1. **Distance to Nearest Education Facility**:
   1. Features like <1km, <2km, and missing have significant impacts. Shorter distances (e.g., <1km) are associated with higher school attendance, while missing data or longer distances negatively impact attendance.
2. **State**:
   1. The state of residence (e.g., Borno, Bauchi, Yobe, Adamawa, Taraba) plays a critical role. For example:
      1. Borno and Yobe have negative SHAP values, indicating lower attendance rates, likely due to higher insecurity and resource constraints.
      2. Adamawa and Taraba show positive impacts, suggesting better educational infrastructure or security.
3. **Security**:
   1. Security is provided on-site\_No has a strong negative impact, highlighting the importance of safety in ensuring school attendance.
4. **Nutrition Services**:
   1. The absence of nutrition services (Nutrition services\_No) negatively impacts attendance, emphasizing the link between health and education.
5. **Sanitation Facilities**:
   1. Poor sanitation (Pit latrine without slab/Open pit) is associated with lower attendance, reflecting the role of hygiene in educational outcomes.
6. **Female Hygiene**:
   1. The presence of female hygiene facilities (Female hygiene\_Yes) positively impacts attendance, underscoring the importance of gender-sensitive infrastructure.
7. **Teacher Availability**:
   1. The absence of trained teachers (Trained moe teachers in this community\_No) negatively impacts attendance, while their presence (Trained moe teachers in this community\_Yes) has a positive effect.
8. **Health Issues**:
   1. The prevalence of malaria (Most prevalent health problem\_Malaria) negatively impacts attendance, highlighting the intersection of health and education.
9. **Reasons for Non-Attendance**:
   1. Financial barriers (Reason 1\_Fees or costs) and lack of parental support (Reason 1\_Lack of parental/Care givers support) are significant predictors of low attendance.

#### **2. Direction of Impact**

* **Positive Impact**: Features like Distance to nearest education facility\_<1km, Female hygiene\_Yes, and Trained moe teachers in this community\_Yes are associated with higher attendance.
* **Negative Impact**: Features like Security is provided on-site\_No, Nutrition services\_No, and Most prevalent health problem\_Malaria are associated with lower attendance.

#### **3. Gender-Sensitive Insights**

* The presence of female hygiene facilities (Female hygiene\_Yes) and female teachers (num\_Female teachers) positively impacts attendance, highlighting the importance of gender-sensitive interventions.
* The absence of these features disproportionately affects girls' attendance, reflecting systemic gender disparities.

#### **4. Regional Disparities**

* States like Borno and Yobe show negative impacts, likely due to ongoing conflict and resource limitations.
* States like Adamawa and Taraba show positive impacts, suggesting better infrastructure and security.

### **Key Analysis for Your Paper**

#### **1. Policy Recommendations**

1. **Infrastructure Development**:
   1. Prioritize building schools within a 1-2 km radius of IDP sites to reduce travel barriers.
   2. Improve sanitation facilities, especially gender-sensitive hygiene infrastructure.
2. **Security Measures**:
   1. Enhance on-site security to create a safe learning environment.
3. **Health and Nutrition**:
   1. Provide nutrition services and malaria prevention programs to address health-related barriers.
4. **Teacher Training**:
   1. Increase the availability of trained teachers, particularly female teachers, to improve attendance.
5. **Financial Support**:
   1. Address financial barriers (e.g., school fees) through scholarships or subsidies.

#### **2. Ethical and Gender-Sensitive Interventions**

* Ensure that interventions are tailored to address gender disparities, such as providing female hygiene facilities and increasing the number of female teachers.
* Monitor for regional biases and ensure equitable resource allocation across states.