Explainable Policy Decision Model: Linking Education Spending and Outcomes Using XAI

Apoorva Hegde
School of Computer Science and Engineering
RV University, Bengaluru, India
Email: apoorvah.bsc22@rvu.edu.in

***Abstract*—This paper presents a framework based on Explainable Artificial Intelligence (XAI) to understand how government spending affects learning outcomes across the globe. By using a Random Forest regression model along with SHAP(SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) tools, the study identifies the key social and policy factors that affect literacy and enrollment levels. In addition to predicting academic success, the model also points out the main variables responsible for these outcomes. The results suggest that explainable AI can make policymaking more transparent and grounded in evidence, helping governments use educational funds more wisely and effectively. Overall, this project supports accountability and encourages the use of data to design better, more focused education policies.**

***Index Terms*—XAI, SHAP, LIME, Education Policy, Governance, Random Forest**

I. Introduction

Economic progress, social justice, and national development are all based on education. To improve literacy, enrollment, and learning results, governments all around the world devote large amounts of their GDP to education. Nonetheless, it is still difficult to determine how expenditure results in actual benefits in education. Economic structure, gender parity, access disparities, and demographics all influence how education policies are implemented.

Although helpful, traditional econometric techniques frequently fall short of capturing nonlinear relationships or provide comprehensible insights into the efficacy of policies. In order to close this gap, this work presents a **Explainable AI (XAI)** framework that combines interpretability methods with machine learning. This method helps policymakers understand "why" and "how" specific factors influence educational success by fusing predictive power and transparency.

In particular, predictions made by a Random Forest model trained on global education metrics are explained by the model using SHAP and LIME. These explainability methods assist more informed and open government by identifying the factors that most affect results, such as education spending, gender gaps, and tertiary enrollment.

II. Motivation and Significance

To achieve equitable and sustainable development goals, public investment in education plays a crucial role. However, the relationship between spending and educational outcomes is not always straightforward or linear. Because of social, infrastructural, or governance-related factors, countries with similar education budgets often display varying levels of literacy and enrollment.

This project aims to support governments by:

1) Presenting a data-driven, explainable model that identifies the key factors influencing educational outcomes.
2) Promoting clear and transparent communication of the model's reasoning to policymakers and stakeholders.
3) Fostering accountability and informed budget planning through interpretable data analysis.

By incorporating XAI into education policy studies, decision-makers can move beyond opaque, black-box models and better understand the underlying causes that drive educational progress.

III. Dataset and Methodology

A. Dataset Description

The dataset was obtained from the Kaggle Global Education Statistics repository, covering data from 1820–2020 for multiple countries. It includes key indicators such as literacy rates, education spending (% of GDP), enrollment ratios, gender-based education gaps, and the share of population by education level.

B. Data Preprocessing

The data was cleaned and processed using Python (Pandas, NumPy). Missing values were handled via forward-fill imputation, numerical features were normalized, and categorical variables were encoded. An 80/20 split was applied for model training and testing.

C. Exploratory Data Visualization

Exploratory analysis was conducted to understand relationships and distributions among education indicators.

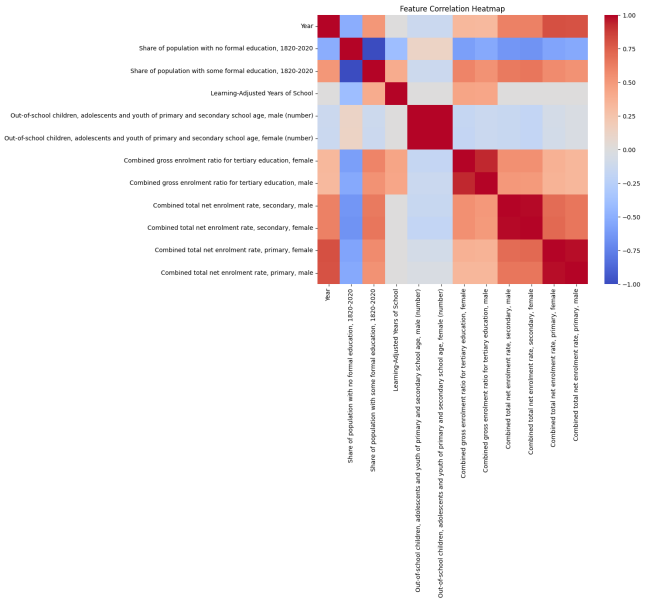

Fig. 1: Correlation heatmap showing relationships among key variables.

Over the long period of 1820-2020, the correlation heatmap (Fig. 1) visually interprets the statistical correlations between numerous demographic and educational measures, including enrollment rates, years of schooling, and population share with/without formal education. It employs a color scale in which dark blue denotes a high negative correlation (one variable increases while the other falls) and brilliant red denotes a strong positive correlation (variables increase together). The heatmap clearly illustrates the overall historical trend of increasing educational access and attainment by showing a strong negative correlation between the Year and the Share of the population without formal education and a strong positive correlation between the Year and metrics like Learning-Adjusted Years of School and total net enrollment rates.

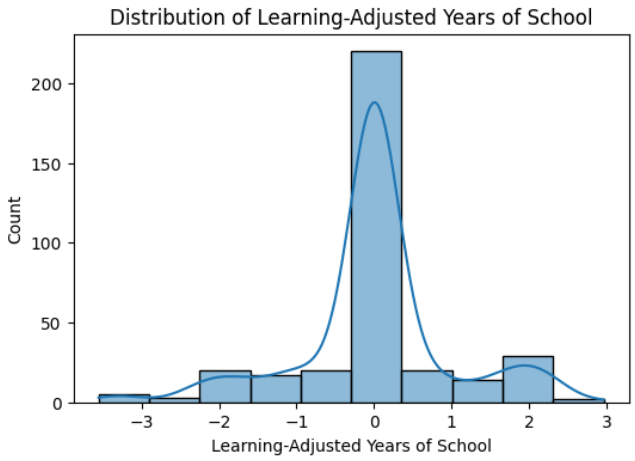

Fig. 2: Distribution plots of key education indicators.

The distribution of Learning-Adjusted Years of School is shown in this distribution plot (Fig. 2), which is a histogram with a kernel density estimate (KDE). Since the data is highly concentrated and peaks at zero (0), this value is the dataset's mode, or most frequent occurrence. A leptokurtic distribution is characterized by a very high central peak and a generally symmetrical shape. The values, which range roughly from -3 to +3, steadily drop as they spread out toward the tails, indicating that the variable was probably normalized before plotting (where 0 denotes the mean).

D. Model Architecture

A Random Forest Regressor was employed to predict a composite education performance index based on spending and demographic factors. The model averages outputs from multiple decision trees, each trained on random subsets of the data and features, to reduce overfitting and improve accuracy.

Mathematically, the Random Forest prediction for input $\mathbf{x}$ is:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}) \tag{1}$$

where T is the total number of trees, and $f_t(\mathbf{x})$ represents the prediction of the t^{th} tree. The objective minimizes the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

E. Explainable AI Techniques

SHAP (SHapley Additive exPlanations) quantifies each feature's contribution to a model's prediction using game theory. It distributes feature importance fairly across all possible feature combinations:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{i\}) - f(S)] \tag{3}$$

SHAP provides both global interpretability (feature importance across the dataset) and local interpretability (explanation for single predictions).

LIME (Local Interpretable Model-Agnostic Explanations) builds a simple local surrogate model (e.g., linear regression) that approximates the behavior of the complex model around a specific instance:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{4}$$

This helps understand why the model predicted a certain outcome for an individual data point.

F. Visualization Pipeline

Visual analytics were created using Python to help interpret the data:

- Correlation heatmaps and feature importance plots.
- SHAP summary and dependence plots (global explanations).
- LIME local explanations for selected countries and years.

IV. Results and Discussion

A. Feature Importance

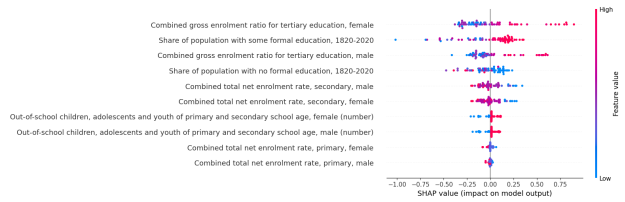

Fig. 3: Feature importance derived from the Random Forest model.

This diagram is a SHAP (SHapley Additive exPlanations) summary plot, which explains the global feature importance and impact of the educational and demographic variables on the model's output across the entire dataset. Each horizontal row represents a feature, and each dot is an observation from the data. The horizontal position of a dot demonstrates that feature's impact on the model output (SHAP value), wherein values to the right increase the prediction, and values to the left decrease it. The color of the dot indicates the actual feature value: red indicates a high feature value, and blue indicates a low feature value. This plot reveals that high tertiary education enrollment ratios (female and male) and also a high share of the population with formal education (red dots) strongly push the model output to the right (positive impact), making them the most influential features. Alternately, a high share of the population with no formal education (red dots) and high out-of-school numbers generally pushes the model output to the left (negative impact), confirming their strong influence over the overall model behavior.

B. Global SHAP Interpretation

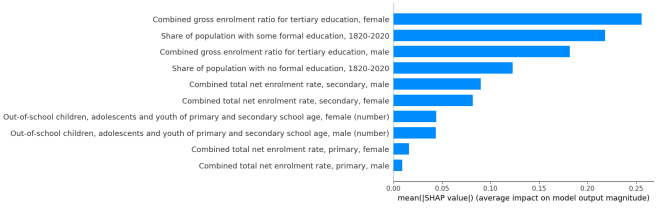

Fig. 4: SHAP summary plot showing feature impact distribution.

This is a SHAP (SHapley Additive exPlanations) Global Feature Importance summary plot. It ranks the educational and the demographic features based on the average magnitude of their impact on the model's output across the entire dataset. The horizontal axis represents the average absolute SHAP value, i.e features are ordered by how strongly they generally influence the prediction, regardless of the direction (positive or negative).The plot shows that the combined gross enrollment ratio for tertiary education is the key indicator. Among the factors, females have the strongest influence. This is followed closely by the share of the population with some formal education from 1820 to 2020, and then by the combined gross enrollment ratio for tertiary education among males. Alternatively, primary education enrollment rates and out-of-school numbers have the lowest average impact, clearly demonstrating that the model relies more on higher education and general attainment metrics when making its global predictions.

C. Local LIME Explanation

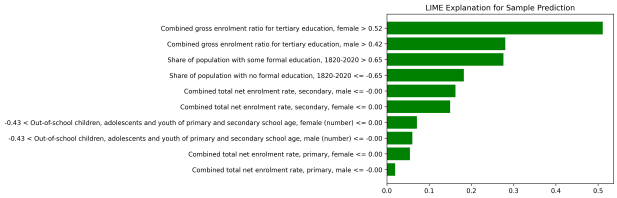

Fig. 5: LIME explanation for a specific country-year instance.

This figure is a LIME (Local Interpretable Model-agnostic Explanations) explanation which explains the factors that contributed to a machine learning model's specific prediction for a single data sample. The plot seen in the diagram, is a horizontal bar chart where the length of the green bars indicates the magnitude of influence each feature had on the final prediction, with all features pushing the prediction in the same direction. The analysis clearly shows that for this particular sample, the prediction was most heavily driven by high enrollment rates in tertiary education for both females (with a ratio above 0.52) and males (above 0.42), followed by a high share of population with some formal education (above 0.65).Factors like, lower-level enrolment rates (primary and secondary) and out-of-school metrics were less critical to the outcome .

V. Conclusion

The study demonstrates how Explainable AI (XAI) can help revolutionize evidence-based policymaking in the field of education. The use of SHAP and LIME explanations with predictive modeling, would help governments identify the most influential variables driving learning outcomes, further improving budget allocation, and ensuring transparency in policy design.

Combining both , global and local interpretability will help bridge the gap between complex analytics and actionable governance. Future works can extend this framework to incorporate causal inferences, longitudinal analysis, and regional education quality indices.

Acknowledgment

The author expresses sincere gratitude to the School of Computer Science and Engineering, RV University, for their academic guidance and infrastructure support.

REFERENCES

[1] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *ACM SIGKDD*, 2016.

[2] World Bank Open Data, "Education Statistics," 2020.

[3] Kaggle Dataset, "Global Education Data," *https://www.kaggle.com/datasets/imtkaggleteam/global-education?select=1-+share-of-the-world-population-with-at-least-basic-education.csv*, Accessed 2025.