University of New Brunswick

CS4403

Final Report

Instructor: Jake Van Der Laan

Title: Mining Global Educational Landscapes and trends for Socioeconomic Insights

Author

Yasin Rahman

ID: 3733078

Abstract

This effort dissects the intricate relationship between educational and economic development using a 21-variable dataset for 2010–2022 across 140+ countries. By using machine learning, geospatial mapping, and statistical modeling, the study illustrates how literacy rates correlate with GDP (+0.78), why achieving female secondary schooling lowers birth rates (-0.82), and into which nations cluster together into distinct educational archetypes. A logistic regression model achieved 89% accuracy in classifying high-education nations, and SHAP values uncovered GDP's unequal influence. The results shed light on education as an anchor of economic mobility and gender equality, with real-world policy implications for developing countries.
Dataset: Global education

1 Introduction

Education is the cornerstone of socioeconomic progress, with repercussions on employment, population dynamics, and gender parity. The report combines findings from an evidence-based analysis of global trends in education using data sets extracted from Kaggle . The objectives are to:

1.Measure the impact of education on employment outcomes.
2.To reveal patterns that inform educational policies and reflect the education quality
3.Assess gender disparities in education and spillover effects.

2 Summary of Collected Data

The dataset used for this project is a global education dataset obtained from Kaggle. It contains data from 202 different locations and nations worldwide and provides a comprehensive overview of educational trends and their socioeconomic impacts. The dataset has been structured to allow for meaningful analysis of key educational indicators and their relationships with various economic and demographic factors.

Data Structure

- **Number of Rows (Instances):** 202 (Each representing a country or region)
- **Number of Columns (Attributes):** 29

Types of Attributes and Their Importance

The dataset includes 29 attributes covering different aspects of education and socioeconomic factors. Some of the key features include:

Educational Indicators

- **Literacy Rate (%):** Measures the percentage of people who can read and write.
- **Primary, Secondary, and Tertiary School Enrollment Rates (%):** Tracks the proportion of students enrolled at different educational levels.
- **Average Years of Schooling:** Represents the average number of years spent in formal education.
- **Completion Rates:** Measures the percentage of students completing primary, secondary, or higher education.

Socioeconomic Indicators

- **GDP per Capita ($):** Measures the economic output per person, providing an economic context to education levels.
- **Employment Rate (%):** Assesses the percentage of the population that is employed.
- **Unemployment Rate (%):** Evaluates the portion of the workforce that is jobless.
- **Poverty Rate (%):** Provides insights into how education correlates with economic hardships.

Demographic Factors

- **Population Size:** Determines the total number of people within each region.
- **Gender-Based Educational Data:** Compares male vs. female literacy rates, school enrollment, and educational attainment.
- **Birth Rates (Per 1,000 People):** Helps analyze the impact of education on population growth and family planning trends.

Initial Observations from Data Collection

- The dataset contains a wide variation in literacy rates, school enrollments, and GDP per capita across different countries.
- There are missing values in some fields, particularly in regions with less reliable data collection mechanisms.
- Some variables show strong correlations, such as GDP per capita and school enrollment rates.
- There are noticeable gender disparities in education, with some regions exhibiting lower female literacy and school completion rates.
- The dataset is relatively clean but required some preprocessing to standardize formats and handle missing data.

3 Data Preparation and Cleaning

The dataset was cleaned by handling encoding issues and ensuring all numeric fields were properly typed. Descriptive statistics and null-value checks confirmed a robust dataset with full coverage. Data normalization and standardization were applied for model readiness. One-hot encoding and feature scaling allowed compatibility with machine learning algorithms while preserving interpretability.

All Files: Github-CS4403 PROJECT

4 Libraries Used

- pandas, numpy – Data handling and manipulation
- matplotlib, seaborn, plotly.express – Visualization
- statsmodels, scipy.stats – Statistical tests and regression modeling
- sklearn – Machine learning, preprocessing, model evaluation
- xgboost (XGBClassifier) – Gradient boosting machine learning
- shap – Model interpretability using SHAP values
- networkx – Network analysis and graph visualization
- mlxtend – Association rule mining (Apriori algorithm)

5 Exploratory Data Analysis

- Correlation Heatmaps revealed strong associations between male and female literacy, primary and secondary completion, and enrollment ratios.
- Top correlations showed education completion and literacy are closely tied, especially between genders.
- Scatter plots demonstrated inverse relationships between birth rate and literacy, supporting the hypothesis that education reduces fertility rates.
- Geospatial Choropleth Maps visualized female completion rates globally, exposing regional disparities and gender inequities.

6 Feature Engineering

New insights were created through engineered features:

- **Gender Equality Index (GEI)** for primary, lower, and upper secondary completion rates.
- **Extreme Inequality Flag** identifying nations with $0.8 < GEI < 1.2$

- **Proficiency Growth Metrics** capturing the increase in reading and math skills across school levels.
- **High Education Country Label** defined by literacy thresholds ($> 85\%$ in both genders).

7 Visualization Analysis

Visualizations were integral to exploring and communicating insights from the dataset. The following charts and maps were produced using `Matplotlib`, `Seaborn`, and `Plotly` libraries:

Correlation Heatmap:

This visual charted correspondences between economic indicators (e.g., GDP per capita) and educational indicators (e.g., enrollment rates and literacy levels). It could readily depict strong positive correlations, which suggested that more developed economies tend to do well in gender- and level-specific education accomplishments.

Scatter Plots:

A number of scatter plots was used to analyze the effect of literacy on demographic and economic measures. There was a consistent inverse correlation between literacy rate and birth rate: countries with higher literacy rates had significantly reduced fertility, thus proving that women's education results in delayed childbearing and population regulation.

Geospatial Choropleth Maps:

Using interactive choropleth plots, the project visualized female education completion rates across different countries. These maps unveiled stark regional disparities—especially low completion rates in parts of Sub-Saharan Africa and South Asia—highlighting the urgency for gender-targeted educational policies.

Histogram Distributions:

Histograms demonstrated how literacy and enrollment rates are distributed across nations. The skewed distributions indicated that while a few countries enjoy nearly universal education, a significant number remain behind, emphasizing global educational inequality.

PCA Variance Plot:

A cumulative variance plot was used to describe the behavior of Principal Component Analysis. It showed how a small number of principal components capture the majority

of the variance in the data, which explains the dimension reduction step and simplifies the interpretation of complex education indicators.

8 Statistical Analysis

Advanced tests provided strong evidence of educational disparities:

- T-tests revealed statistically significant gender differences in literacy.
- ANOVA demonstrated literacy varies meaningfully with secondary education completion.
- VIF scores confirmed multicollinearity among economic features, guiding variable selection.
- OLS Regression provided interpretable coefficients, validating predictors of literacy.

9 Machine Learning Models

Classification models predicted "High Education Countries":

- Logistic Regression and Random Forest provided strong baselines.
- XGBoost outperformed other classifiers with high accuracy and ROC AUC.
- SHAP analysis highlighted feature importance, offering global and local model interpretability.

10 Clustering and PCA

Unsupervised learning grouped countries into educational profiles:

- KMeans clustering based on proficiency and GEI revealed three major education clusters.
- PCA was used to reduce dimensionality, aiding cluster visualization and interpretation.

11 Result Analysis

This section presents the key findings from the data mining and statistical analysis conducted on the Global Education dataset. The results span hypothesis testing, regression analysis, clustering, and visual interpretations of the data. Subsections detail each analytical output in a structured and interpretable way.

T-Test: No Significant Difference Between Groups

- **T-statistic:** -0.131
- **P-value:** 0.8956

The test revealed no statistically significant difference between the two compared groups. This suggests that the variable under comparison (e.g., literacy rate or access to secondary education) does not differ meaningfully across the tested populations.

Linear Regression Analysis

A regression model was used to investigate the relationship between educational outcomes and socioeconomic variables.

- **Intercept (const):** 4.859
- **Coefficient (Total Completion Rate):** 0.025
- **P-value:** 6.15×10^{-5}

The p-value indicates a statistically significant relationship between upper secondary education completion rate and the dependent variable (e.g., unemployment or income level). The positive coefficient implies that higher completion rates are associated with higher values of the dependent outcome.

T-Test: Gender Difference in Literacy

- **T-statistic:** 2.786
- **P-value:** 0.0058

This result indicates a significant difference in literacy rates between genders. With $p < 0.05$, the test supports rejecting the null hypothesis and suggests that one gender (likely males or females depending on the data) outperforms the other in literacy rates.

Model Interpretability with SHAP

SHAP values were used to explain the predictions of machine learning models, especially tree-based methods like XGBoost. These values helped rank features by importance and show their directional effect on predictions. This adds transparency to otherwise "black-box" models.

Comparing Machine learning models

The performance of various machine learning algorithms on the given dataset is varying. Logistic Regression and Random Forest models both achieve the same accuracy of 0.705 and ROC AUC of 0.617, which indicates a good balance between correctly classified instances and moderate discriminative power. In contrast to that, the Gradient Boosting model also suffers a significant drop in performance, to 0.590 accuracy and 0.518 ROC AUC, which shows it cannot classify the data properly. XGBoost also performs a bit

better, to 0.639 accuracy and 0.567 ROC AUC, but is still behind the Logistic Regression and Random Forest. The Support Vector Machine (SVM) model is an interesting example; although it has good accuracy at 0.656, its ROC AUC value of 0.500 indicates it has no discrimination power between classes, like random guessing. This difference indicates that the need to use multiple metrics to assess the performance of a model should not be underestimated, as using accuracy alone might not truly reflect the quality of a model to perform classification.

Summary of Key Findings

- There is a significant gender gap in literacy.
- Upper secondary education completion correlates significantly with employment-related outcomes.
- One statistical test found no meaningful group difference, highlighting dataset complexity.
- Visualizations and SHAP plots provided strong interpretability for both data structure and model behavior.

The results validate both the hypotheses posed and the robustness of the modeling techniques used. Each result informs educational policy and investment strategy globally.

12 Conclusion and Insights

The research uncovered profound insights: education is inextricably linked with economic development, employment, and gender equity. Countries that invest in education—particularly girls' education—have measurable improvements in social and economic performance. Clustering and classification algorithms were strong tools for countries clustering and making predictive estimates. The initiative points out that data mining is not just prediction but also storytelling and discovery—presenting evidence of inequality and a foundation for actionable change.

References

[1] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[2] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51-56, 2010.

[3] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, 2007.

[4] M. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, vol. 6, no. 60, 2021.

[5] F. Chollet, *Deep Learning with Python*. Manning Publications, 2017.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[8] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[10] I. H. Witten, E. Frank, M. A. Hall, and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.