Lyn Detwiler and Lucille Barnett

Global Education Data Analysis

DS-160

The data set we chose to focus on for our final project explores the relationships between educational proficiency and societal outcomes. Using a dataset that captures various educational and socio-economic indicators across global regions. Below we will have a list of all of the variables in our dataset, but a few key variables we plan to focus on include proficiency in reading and math at different educational stages, out-of-school rates, literacy rates, and unemployment rates.

### World Education Data Variables and Definitions

* **Countries and Areas**: Name of the countries and areas.
* **Latitude**: Latitude coordinates of the geographical location.
* **Longitude**: Longitude coordinates of the geographical location.
* **OOSR\_PrePrimary\_Age\_Male**: Out-of-school rate for pre-primary age males.
* **OOSR\_PrePrimary\_Age\_Female**: Out-of-school rate for pre-primary age females.
* **OOSR\_Lower\_Secondary\_Age\_Male**: Out-of-school rate for lower secondary age males.
* **OOSR\_Lower\_Secondary\_Age\_Female**: Out-of-school rate for lower secondary age females.
* **OOSR\_Upper\_Secondary\_Age\_Male**: Out-of-school rate for upper secondary age males.
* **OOSR\_Upper\_Secondary\_Age\_Female**: Out-of-school rate for upper secondary age females.
* **Completion\_Rate\_Primary\_Male**: Completion rate for primary education among males.
* **Completion\_Rate\_Primary\_Female**: Completion rate for primary education among females.
* **Completion\_Rate\_Lower\_Secondary\_Male**: Completion rate for lower secondary education among males.
* **Completion\_Rate\_Lower\_Secondary\_Female**: Completion rate for lower secondary education among females.
* **Completion\_Rate\_Upper\_Secondary\_Male**: Completion rate for upper secondary education among males.
* **Completion\_Rate\_Upper\_Secondary\_Female**: Completion rate for upper secondary education among females.
* **Grade\_2\_3\_Proficiency\_Reading**: Proficiency in reading for grades 2-3 students.
* **Grade\_2\_3\_Proficiency\_Math**: Proficiency in math for grades 2-3 students.
* **Primary\_End\_Proficiency\_Reading**: Proficiency in reading at the end of primary education.
* **Primary\_End\_Proficiency\_Math**: Proficiency in math at the end of primary education.
* **Lower\_Secondary\_End\_Proficiency\_Reading**: Proficiency in reading at the end of lower secondary education.
* **Lower\_Secondary\_End\_Proficiency\_Math**: Proficiency in math at the end of lower secondary education.
* **Youth\_15\_24\_Literacy\_Rate\_Male**: Literacy rate among male youths aged 15-24.
* **Youth\_15\_24\_Literacy\_Rate\_Female**: Literacy rate among female youths aged 15-24.
* **Birth\_Rate**: Birth rate in the respective countries/areas.
* **Gross\_Primary\_Enrollment**: Gross enrollment in primary education.
* **Gross\_Secondary\_Enrollment**: Gross enrollment in secondary education.
* **Unemployment\_Rate**: Unemployment rate in the respective countries/areas. respective countries/areas.

In starting our exploratory analysis of this robust dataset, the initial research objectives were twofold: first, to analyze how proficiency in reading and math at the end of primary education influences unemployment rates, and second, to determine whether an individual’s literacy can be classified based of a threshold in reading proficiency. By addressing these questions, this study aimed to uncover critical patterns in education and its socio-economic impacts.

First, with the help of a multiple linear regression analysis, we explored the relationship between reading and math proficiency and unemployment rates. 1Using Jupyter Notebook, we trained a model with “Primary End Proficiency Reading” and “Primary End Proficiency Math” as the independent variables, and “Unemployment Rate” as the dependent variable. The model was evaluated using mean squared error (MSE), root mean squared error (RMSE), and R-squared value (R²). And, despite our expectations, there was a significant deviation from actual values with a MSE of 36.2264 and RMSE of 6.0188. Furthermore, the model failed to explain unemployment rate variability due to a negative R2 value of -0.0064 and a correlation matrix revealed a near-zero correlation between the variables. Consequently, it was concluded that there is no significant linear relationship between primary end proficiency in reading and math and the unemployment rate in this dataset. However, further investigation of the data revealed a stronger correlation between reading and math proficiency and an enrollment in tertiary education.

In the next section of our analysis, we focused on our second research question: “Can we classify a person as literate if their proficiency in reading is above a certain number at the end of primary or lower secondary education? If so, what is that number?” Unfortunately, while beginning work on the coding for this section of our analysis, we ran into significant problems with the structure of the data. Given this significant roadblock, we decided to pivot our focus towards another topic of interest.

Using logistic regression, we wanted to discover the main indicators that classified a country as developed or undeveloped. First, we began by classifying countries based on their average literacy rates into two categories: "Developed" and "Not Developed." This classification was determined by setting a threshold of 95; countries with literacy rates above this threshold were considered developed. The next step involved preparing the data for analysis, which included converting relevant columns (such as birth rate, primary education enrollment, and tertiary education enrollment) to integer types and applying one-hot encoding to categorical variables to facilitate the logistic regression modeling. The dataset was then split into training and testing sets, with an 80-20 split, ensuring a robust evaluation of the model's performance.

Upon training the logistic regression model, we evaluated its accuracy and generated a classification report. The model achieved an overall accuracy of approximately 70.73%, with notable performance metrics for the "Not Developed" class: a precision of 0.73, recall of 0.93, and an F1-score of 0.82. The confusion matrix further illustrated the model's performance by showing the number of true positives, true negatives, false positives, and false negatives. Although the model performed well for the "Not Developed" class, there was room for improvement in accurately predicting the "Developed" class. This analysis demonstrated the effectiveness of logistic regression in classifying countries based on development status, providing valuable insights into the factors influencing this classification.

In conclusion, our analysis of the World Education dataset provided valuable insights into the relationships between educational proficiency and societal outcomes, though it also highlighted the complexities and limitations of these relationships.

Overall, this project reinforced the importance of data-driven analysis in addressing global educational and socio-economic challenges and emphasized the need for more nuanced approaches to understanding development dynamics. The insights gained lay a foundation for future research, emphasizing the need for robust datasets and more sophisticated modeling techniques to better capture the complexities of education’s role in shaping societal outcomes.