Gender and Socioeconomic Disparities in Global Education

Sean Mulherin
Muzi
Pei
Rachel
University of California, Los Angeles
Department of Statistics

June 5, 2024

Abstract

This report investigates global educational trends and disparities, focusing on gender and socioeconomic factors. Utilizing data from UNESCO, UNICEF, WorldBank, and the United Nations Statistics Division, we analyzed out-of-school rates, literacy rates, and academic proficiency across genders, economic classes, and geographical regions. Our findings reveal significant gender disparities in countries with low and medium Human Development Index (HDI), with females generally exhibiting higher out-of-school rates and lower literacy rates compared to males. Additionally, we observed a strong positive correlation between HDI and academic proficiency. The report concludes with a discussion of the implications of these findings and recommendations for policy interventions to address educational inequalities.

List of Figures

1 Introduction

Education is a critical determinant of both social and economic development. Globally, access to quality education remains uneven, with significant disparities based on gender, socioeconomic status, and geographical location. This report aims to uncover such disparities and provide insight into the global educational landscape. To accomplish this, we utilize data from prominent international organizations including: UNESCO, UNICEF, WorldBank, and the United Nations Statistics Division. Our analysis focuses on out-of-school rates, literacy rates, and academic proficiency grouped by gender and Human Development Index (HDI) class. Additionally, we examine the geographical distribution of educational outcomes to identify patterns and trends that may inform policy decisions.

We have two main objectives. First, we set forth to identify and quantify the extent to which gender disparities exists within education and across different HDI classes. Second, we explore the relationship between socioeconomic factors, particularly HDI and Gross Domestic Product (GDP), and academic proficiency. Achieving these objectives will provide detailed understanding of the factors contributing to educational inequities, equipping educators with evidence-based recommendations for addressing these challenges.

In the following sections, we introduce the data, detail our methodology, present the results of our analyses, and discuss the implications of our findings. We conclude with a call to action aimed at policymakers and educational stakeholders to minimize the educational divide and promote equitable access to high quality education.

2 Data

2.1 Overview

The Global Education Database is an amalgamation of several data sets including: UNESCO, UNICEF, WorldBank, and the United Nations Statistics Division. This data provides a comprehensive global perspective on education, offering vital insights into diverse education systems worldwide. It covers essential metrics such as out-of-school rates, completion rates, proficiency levels, literacy rates, HDI, and GDP. It is a valuable resource for researchers, educators, and policymakers looking to assess and improve education systems.

Country Gross Domestic Product data describes the GDP per capita in US dollars for various countries and regions over a range of years from 1960 to 2022. GDP signifies how wealthy a country is with respect to other countries. A high GDP indicates a wealthy country. Human Development Index data quantifies human development within a range from 0 to 1. HDI incorporates life expectancy, education, and income to determine the overall quality of life for a county's respective citizens. A high HDI indicates a high quality of life.

2.2 Collection, Wrangling, and Feature Engineering

Global Education and Human Development Index data is accessed from *Kaggle.com*. Country GDP data is accessed from *data.worldbank.org*. Since our analysis is based on global education in the year 2021, we select only the year 2021 to get Country GDP when merging data.

Country GDP data included four rows of information not relevant to our objective (i.e. date of pre-

vious update) and so they were removed. All the variables of use are of numeric type. Global Education data is primarily separated into two parts: geographic location and educational indicators. Geographic location is expressed using longitude and latitude. Educational indicators are expressed as percentages. *HDI Class* classifies *HDI Value* into three groups: low, medium, high. *HDI Class* was calculated based on *HDI Value*, partitioning the real-valued *HDI Value* variable into three equal proportions using the `cut()` function in R.

Ultimately, five datasets were merged based on the country to which variables belonged, consisting of a total of 161 observations and 37 variables.

3 Exploratory Data Analysis

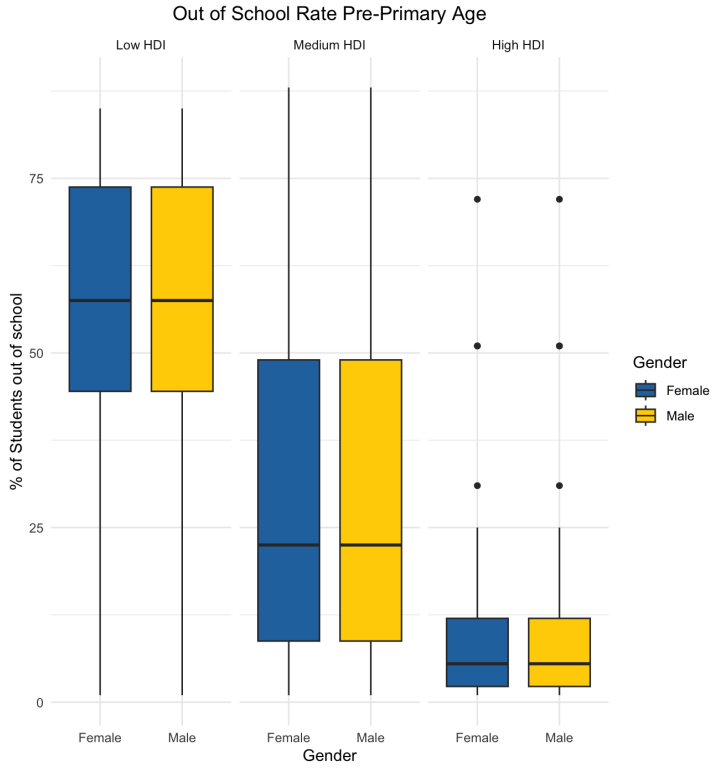


Figure 1: Out of School Rate of Pre-Primary Age Students

Figure 1 portrays a layered box plot comparing out of school rates for pre-primary aged students by gender and HDI class. Pre-primary age is defined as ages 2 - 5 years. Notice the similarities across genders, as there is no discernible difference in mean and variance between males and females within each HDI class. However, there is a noticeable difference when compared across HDI classes. As HDI class increases from low to high, out of school rate decreases substantially among pre-primary aged students.

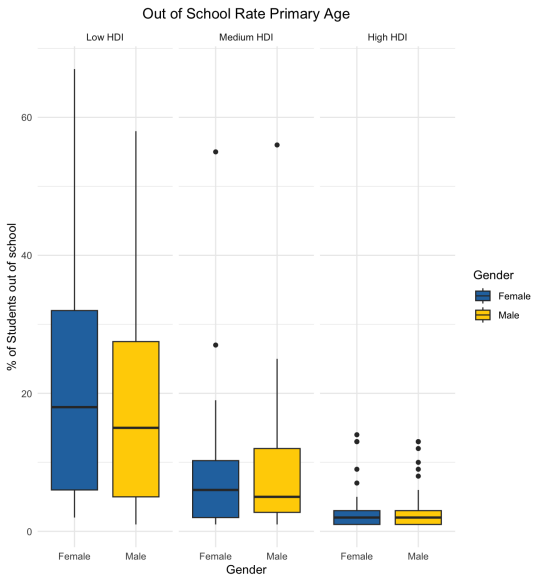


Figure 2: Out of School Rate of Post-Primary Age Students

Figure 2 portrays a layered box plot comparing out of school rates for primary aged students by gender and HDI class. Primary age is defined as ages 6 - 11 years. You will notice similar overall trends described in Figure 1. However, there does appear to be gender disparities within low and medium HDI classes, wherein primary age females have a higher out of school rate than males of the same age. There does not appear to be such gender disparity in high HDI countries. Notice the overall lower out of school rate compared to pre-primary age students shown in Figure 1. This indicates that primary aged children are more likely to be in school than pre-primary aged children.

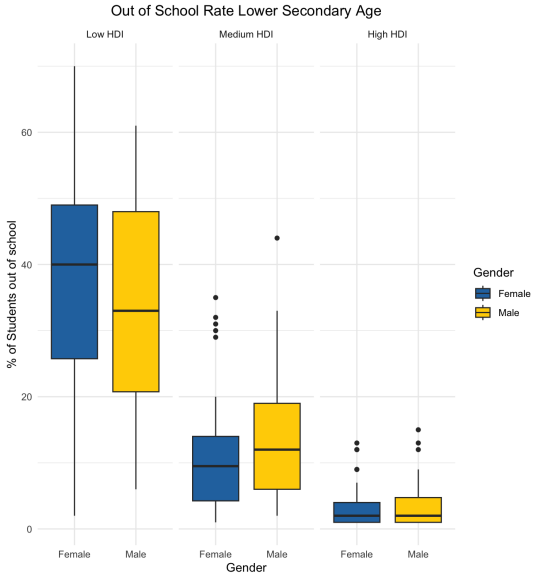


Figure 3: Out of School Rates Lower Secondary Aged Students

Figure 3 portrays a layered box plot comparing out of school rates for lower secondary students aged 12 - 14 years. Data is grouped by gender and HDI class. The overall decreasing out of school rate as HDI class increases is similar to that found in Figure 1 and Figure 2. Notice the differences is out of school rates between males and females. In low HDI countries, males have a lower out of school rates compared to females. Interestingly, this gender dynamic shifts in medium HDI countries where females

have a lower out of school rate compared to males. In high HDI countries, no notable difference exists between genders.

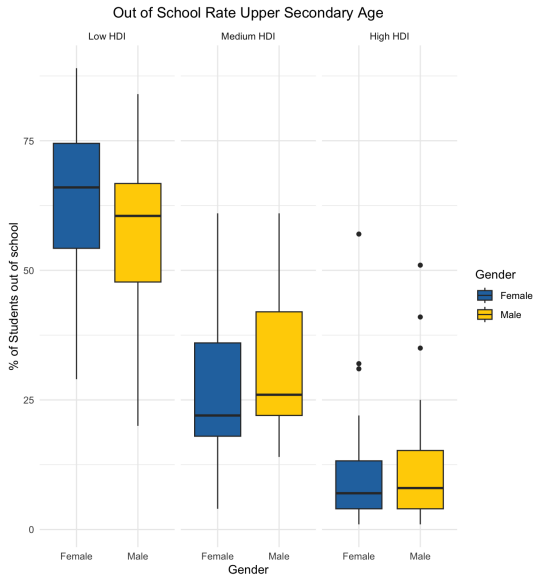


Figure 4: Out of School Rates Upper Secondary Aged Students

Figure 4 portrays a layered box plot comparing out of school rates for upper secondary students aged 15 - 17 years. Notice the higher out of school rate compared to lower age students shown in previous figures. This indicates that upper secondary aged children are less likely to be in school than their younger counterparts. Similar to Figure 3, gender dynamics switch from low to medium HDI countries, with females having a higher out of school rate than males in low HDI countries but a lower out of school rate in medium HDI countries. As was true in all the previous figures, there seems to be no discernible difference in out of school rate between genders, both having low rates.

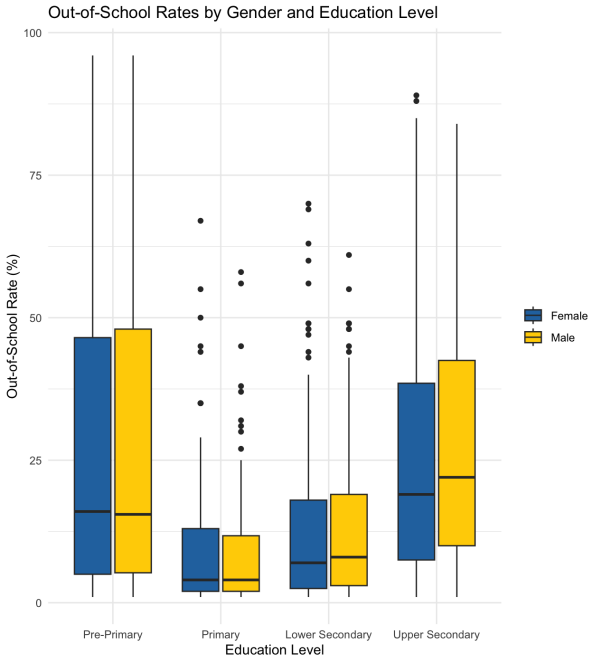


Figure 5: Out of School Rates By Grade Level

Figure 5 displays the distribution of out-of-school rates (OOSR) for males and females across four edu-

cation levels: Pre-Primary, Primary, Lower Secondary, and Upper Secondary. It shows that females have higher median out-of-school rates at the pre-primary and upper secondary levels compared to males. Conversely, males have slightly higher median out-of-school rates at the primary level. Both genders exhibit increased variability in out-of-school rates at the pre-primary and upper secondary levels. The colors used in the graph represent UCLA blue for males and UCLA gold for females.

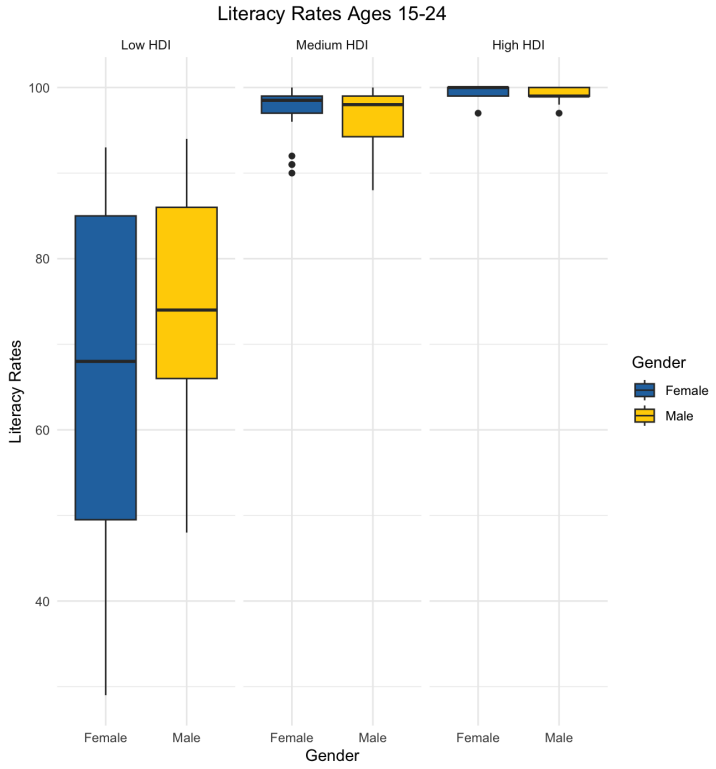


Figure 6: Literacy Rates Age 15-24 Years

Above is a layered box plot portraying literacy rates for citizens aged 15 - 24 years, grouped by gender and HDI class. Notice the overall increasing trend in literacy rates as HDI increases. Also notice the disparities between genders in low HDI countries where females have a lower average literacy rate than males, but a higher variance. In medium HDI countries, the average literacy rate between genders is comparable, but males have a much higher variance than females. High HDI countries have extremely high literacy rates in both males and females.

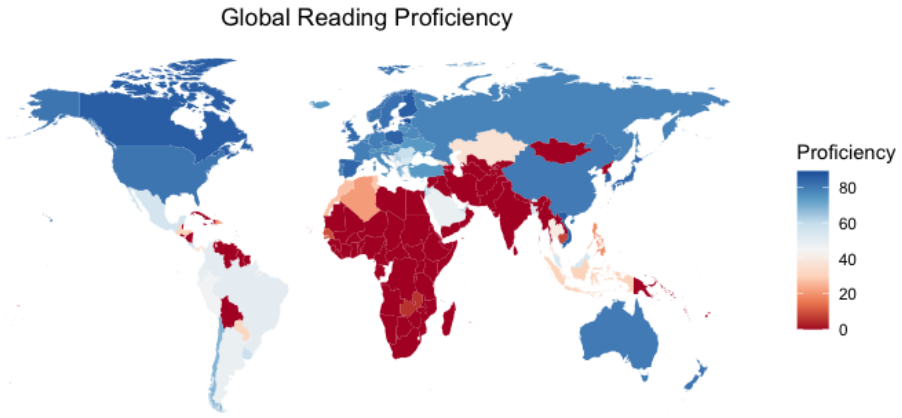


Figure 7: Global Reading Proficiency

Figure 7 portrays a choropleth map of global reading proficiency ratings. The proficiency scale is [0 - 100]. Notice the regions with low proficiency ratings (red), predominantly: Africa, Southeast Asia, West Asia, and some in South America.

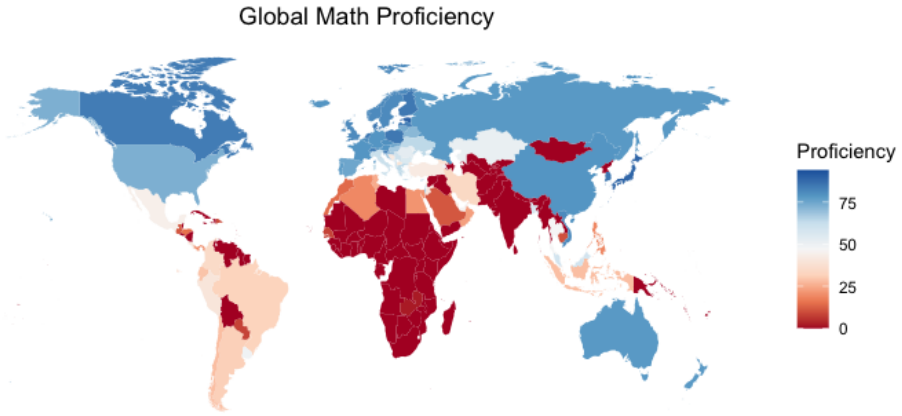


Figure 8: Global Math Proficiency

Similar to Figure 7, Figure 8 portrays a choropleth map of global math proficiency ratings. The proficiency scale is [0 - 100]. Notice the regions with low proficiency ratings (red), predominantly: Africa, Southeast Asia, West Asia, and South America. These match up with the reading proficiency findings described in Figure 7.

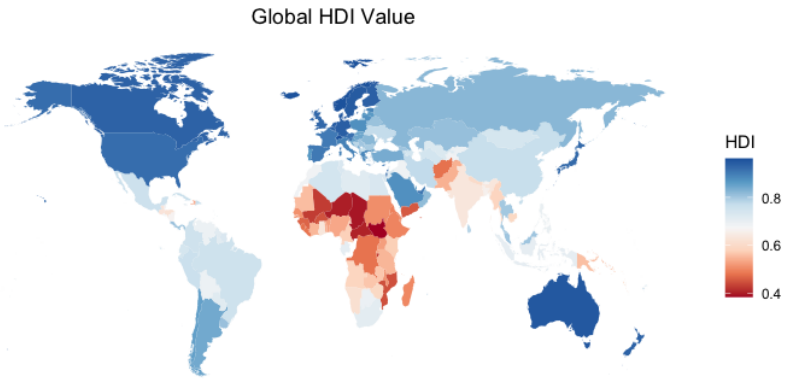
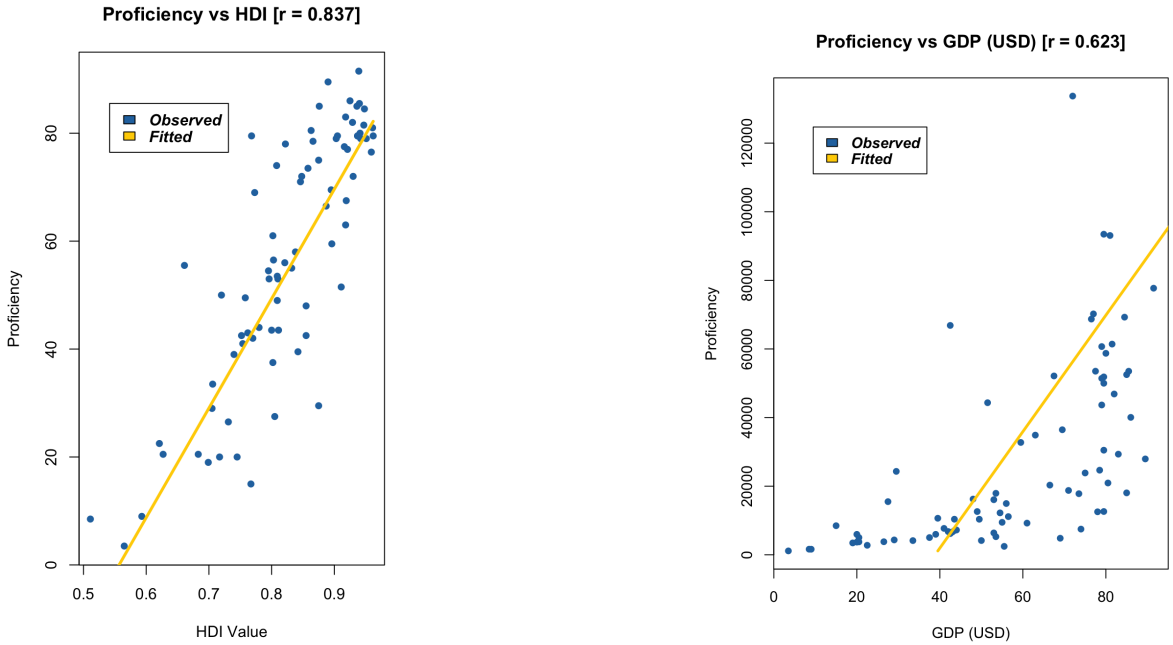


Figure 9: Global HDI Value

Figure 9 portrays a choropleth map of global HDI values. The HDI scale is [0 - 1]. Notice the regions with low proficiency ratings (red), predominantly: Africa, West Asia, and some in Indonesia.



(a): Proficiency vs HDI

(b): Proficiency vs GDP

Figure 10: Socioeconomic Factors on Academic Proficiency

Figure 9(a) portrays a scatter plot comparing proficiency with HDI value. Note, proficiency is the average math and reading proficiencies for each country. Also displayed is a fitted linear regression model (yellow). The correlation coefficient r = 0.756, suggesting a strong positive correlation between proficiency and HDI value. This validates the findings portrayed in previous figures. Figure 9(b) portrays as scatter plot comparing proficiency with GDP (USD). The correlation coefficient r = 0.623, suggesting a moderately strong positive correlation between proficiency and GDP.

Figure 11: Correlation Plot

Figure 11 is a heat map that illustrates the correlations between various education metrics and the unemployment rate. Positive correlations are represented by darker shades of blue, while negative correlations are shown in lighter shades. The strongest positive correlation is observed between lower secondary proficiency in reading and math (0.96). Primary proficiency metrics also correlate positively with each other (0.77). However, the unemployment rate shows weak correlations with most education metrics, indicating that factors other than educational attainment might play a more significant role in influencing unemployment rates.

3.1 Discovery

This section will further elaborate on the results previously displayed. Note, all claims made are based solely on the data described in Section 2. As such, claims and inferences shall not be extrapolated to other years or domains. We first discuss the results pertaining to gender.

As shown in Figures 1 - 5, gender disparities vary across HDI classes. In low HDI countries and across all ages studied, females have lower literacy rates and higher out of school rates compared to their male counterparts. Interestingly, for out of school rates of both upper and lower secondary students, this disparity shifts from low to medium HDI countries where males have higher out of school rates than females. This indicates that, for secondary students, as HDI class increases from low to medium, males are less likely to complete secondary schooling than females. In high HDI countries, gender disparities do not seem to exist within the out of school rate nor literacy rate parameters. Measures of central tendencies are comparable in both parameters and across all ages for countries with high HDI.

We now move to discuss results pertaining to socioeconomic factors. There is strong evidence for the effect of socioeconomic status on academic proficiency. As shown in Figures 1 - 4, as HDI increases, out of school rates decrease. Shown in Figure 6, literacy rates increase rapidly as HDI class transitions from low to medium to high. As an alternative view, choropleth maps were displayed in Figures 8 - 9; all of which indicating a directly proportional relationship between HDI and academic proficiency. In other words, high HDI values are shown to correspond with high academic proficiency. Lastly, shown in Figure 9(a) and 9(b), there exists strong positive correlations between academic proficiency, HDI, and GDP. Academic proficiency increases as HDI increases, and similarly as GDP increases.

4 Methods

4.1 Method and Female Youth Literacy Rate

We fit models to analyze disparities in education between genders and across grade levels. By calculating the differences in out-of-school rates and literacy rates between males and females, the study is able to gain insights into the variability of gender disparities. We use *Female Youth Literacy Rates* as the outcome variable to assess such relationships.

Three models are created to predict *Female Youth Literacy Rates*: random forest, gradient boosting machine (GBM), and support vector machine (SVM). These models are chosen for their ability to handle non-linear relationships and interaction effects among features. Random forest models leverage multiple decision trees during the training process and output the average prediction of the individual trees. It is capable of handling multiple features at the same time. GBM models construct additive models and optimizes the differential loss function with high efficiency. It is known to be useful when analyzing complex non-linear data patterns. SVM models focus on finding the optimal boundaries that best differentiate classes. It is robust to outliers and effective with multiple features as well.

It is difficult to identify the best of these three models solely depending on their characteristics as they are all suitable for a high dimensionality of numerous features. Therefore, we will fit all three models and compare their performances using evaluation metrics. All three models were trained after partitioning datasets into training and testing sets, a process often referred to as *cross-validation*. To mitigate the risk of overfitting, we employed *K-fold cross-validation* to further split the data into five

folds and evaluate their performances based on mean absolute error (MAE), root mean squared error (RMSE), and the R^2 score. The results of this are portrayed in Table 1.

4.2 Method and Geographic Location

The study uses interaction models to further explore the relationships between geographic location, HDI, and gender disparities in completion rates at various grade levels. Based on the previous exploratory data analysis, HDI has an indirectly proportional relationship with gender disparities. In other words, as HDI increases, the disparity between genders in educational outcomes decreases. Therefore, the study decides to only focus on the relationship between HDI, latitude, and disparities.

To fully explore the gender disparities in education, three linear models were constructed based on the categorization of education: primary, lower secondary, and upper secondary. Each model assesses how the interaction between latitude and HDI impacts the disparity in completion rates between males and females. The inclusion of an interaction term examines if the effect of HDI on gender disparities varies across different latitudes. One thing to note is that the latitude has been standardized to ensure scale uniformity even though in the original data latitudes are recorded as non negative values only. This step makes sure further addition or adjustment of data will not impact on the accuracy of the model.

We will use linear regression models with interaction terms to analyse how multiple predictor variables influence the outcome variable, as well as how predictor variables could influence each other. Linear models are useful in estimating relationships between numerical predictors and a numerical outcome variable. The interaction term is important because it examines whether the effect of one predictor on the outcome variable changes at different levels of another predictor. Collectively, these three models offer a general view of how geographic and socioeconomic factors could influence academic performance. Moreover, such influence could vary as education moves into different stages.

5 Result

5.1 Result for Female Youth Literacy Rate

	Random Forest	Gradient Boosting Machine	Support Vector Machine
MAE	43.167	40.185	41.747
RMSE	46.379	45.489	51.639
R^2	0.034	0.064	0.041

Table 1: Evaluation Metrics of Fitted Models

Table 1 presents a summary of the performance of three models. The evaluation focuses on three key metrics: mean absolute error (MAE), root mean squared error (RMSE), and R^2 score. Overall, gradient boosting machine (GBM) outperforms the other two models across all listed metrics except for R^2, suggesting it to be the most effective model for predicting female youth literacy rates.

GBM is effective in managing non-linear relationships in the data and capturing complex patterns more accurately than models such as random forest. While they both build trees, random forest builds in parallel without iterative error correction. GBM also optimizes the loss function and minimizes errors by using *gradient descent*. The lower MAE and RMSE of GBM is evidence that its predictions are closer to the actual data.

5.2 Result for Geographic

Variable	Coefficient	Std. Error	t-value	p-value
Primary Education Completion Differences				
Intercept	4.9459	2.8507	1.735	0.0847
Latitude_std	-0.4455	2.8623	-0.156	0.8765
HDI_val	-7.8968	3.7222	-2.122	0.0354 *
Latitude_std:HDI_val	2.0122	3.4741	0.579	0.5633
Lower Secondary Education Completion Differences				
Intercept	7.6430	3.0970	2.468	0.01466 *
Latitude_std	-6.9140	3.1090	-2.224	0.02759 *
HDI_val	-12.790	4.0430	-3.163	0.00187 **
Latitude_std:HDI_val	10.379	3.7740	2.750	0.00666 **
Upper Secondary Education Completion Differences				
Intercept	7.9050	3.1140	2.539	0.01211 *
Latitude_std	-7.9060	3.1270	-2.528	0.01245 *
HDI_val	-13.032	4.0660	-3.205	0.00164 **
Latitude_std:HDI_val	11.260	3.7950	2.967	0.00348 **

Table 2: Regression Analysis of Gender Differences in Educational Completion Rates

Table 2 presents a summary of the interaction models of the influence of latitude and HDI on gender disparities across various educational levels.

At the primary education level, the analysis reveals that latitude alone does not have a significant direct impact on gender differences in completion rates. In contrast, HDI shows a significant negative effect on gender disparities, indicating that higher HDI values contribute to reducing gender disparities in primary education. The interaction value between latitude and HDI was also not significant, implying that the influence of HDI on reducing gender disparities does not vary with latitude at the primary level.

In lower secondary education, latitude is revealed as a more significant factor on the gender disparity, indicating that a higher latitude tends to be associated with a more reduced disparity. HDI continues to hold a strong negative effect on disparities. The interaction between HDI and latitude is more significant, suggesting that the effect of higher latitudes on reducing gender disparities is more prominent in regions with higher HDI scores.

The trend for upper secondary education closely mirrors those of lower secondary, indicating that higher latitudes combined with higher HDI scores effectively reduce gender disparities.

The difference of the impact of latitude on the gender disparity at different educational levels could be rooted in several factors. At the primary level, educational policies in the globe typically emphasize universal access, leading to relatively high completion rates regardless of geographic differences. However, as students progress to secondary education, the influence of geographic factors, including climate, economic opportunities, and cultural norms, becomes more significant.

Regions at higher latitudes often correspond to regions with more developed economies, and therefore have more resources in education. These investments can lead to the reduction of reduced gender disparities. For example, regions at higher latitudes might have more policies and programs supporting gender equality in education.

The impact of latitude is also more pronounced in the upper secondary education than it is in the lower. Some potential causes could be the result of social expectations and adolescent developments. As students progress from the lower to the upper secondary levels, they are also transitioning from compulsory education requirements to a phase of self discovery, at which latitude-related factors might begin to show their effects. For instance, regions with extreme latitudes often experience more extreme weather, and this factor could potentially affect school attendance for students at greater ages as they could potentially get more involved in outside school activities.

The significant interaction between HDI and latitude in reducing gender disparities suggests that socioeconomic development plays a crucial role in the implementation of effective educational policies at high latitudes. This interaction shows that in more developed regions, the advantages of higher latitudes in educational resources and cultural norms towards gender equality become more pronounced.These two factors together help to minimize gender disparities in education.

5.3 Limitations and Future Areas of Research

Several limitations should be considered in this study. One significant limitation is the presence of missing values within the dataset. Values like literacy and out of school rate are zero-valued for many countries, especially in Africa. These missing values are interpreted as large outliers and were carefully wrangled to prevent misleading conclusions. The linear model tries to reduce the effect of zero-valued entries by calculating the difference rather than including the raw numbers directly. Another significant limitation is the narrowed time span of our study. As discussed previously, we only studied disparities based on data for 2021. This, coupled with the large number zero-valued entries, limits the extrapolation power of our study.

There are many opportunities to engage in this data for future research. It would be interesting to look at the effect of academic performance caused by race or educational policies such as length of school day, number of school days per year, teacher salaries, etc. Furthermore, looking at these social, economic, and academic metrics across decades would provide a much more robust understanding of the trend of education and disparities that exist within.

6 Conclusion

We conclude this paper by providing a brief summary of the findings discussed more thoroughly in previous sections. Recall, the objective of this paper is to unearth global trends, patterns, and disparities that exist within the education sector. To accomplish this, we focused our efforts on both gender and socioeconomic classifiers.

We began with rigorous exploratory data analysis, using box plots, scatter plots, choropleth maps, and correlation coefficients to do so. This helped portray trends unearthing academic disparities respective to both gender and socioeconomic factors. Specifically, it is clear that gender disparities within education are more prominent in low HDI countries compared to high HDI counties. This was revealed by looking at out of school rates, literacy rates, and proficiency rates. Furthermore, we showed socioeconomic status to be a strong factor in academic performance with a directly proportional relationship. This was shown by looking at both GDP and HDI, enabling us to study their correlation with similar academic performance metrics as previously mentioned (i.e. out of school rates, literacy rates, and proficiency

rates).

After performing thorough exploratory data analysis, we employed more advanced statistical methods to delve more deeply into the inner workings of gender and socioeconomic disparities within education. Three models were fit to the data: random forest, gradient boosting machine (GBM), and support vector machine (SVM). These models were tasked to predict social disparities using geographic location and HDI value as predictors (see Section 5.2 for information on feature selection). After evaluating each model's performance using cross-validation, GBM was shown to provide the most accurate prediction.

Lastly, we would like to express gratitude to UNESCO, UNICEF, World Bank, and the United Nations Statistics Division for generously providing data. We also want to take this opportunity to thank Professor Dave Zes for his engaging, inspiring, and informative lectures. As noted in the introduction, education is a sector in which all people come together to grow and develop as individuals. Because of this, it is critically important to understand the patterns, trends, and factors that play a role in people's educational experience. We hope the statistical analysis performed in this paper is used for the advancement of social equity within education.

Social Disparities Within Education

Muzi, Pei, Sean, Rachel

June 2024

Introduction

The objective of this paper is to unearth global trends, patterns, and disparities that exist within the education sector. To accomplish this, we focused our efforts on both gender and socioeconomic classifiers.

Data from UNESCO, UNICEF, WorldBank, and the United Nations Statistics Division, was merged to analyze out-of-school rates, literacy rates, and academic proficiency across different Human Development Index (HDI) classes, Gross Domestic Product (GDP) values, and geographic regions.

Key Findings

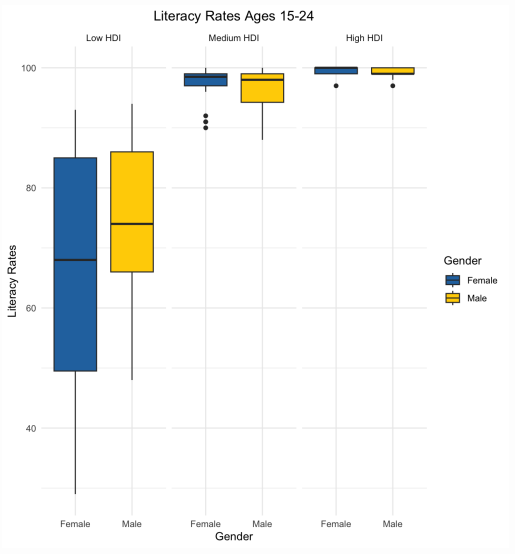
Gender Disparities exist in low and medium HDI countries, but predominantly for secondary-aged students.

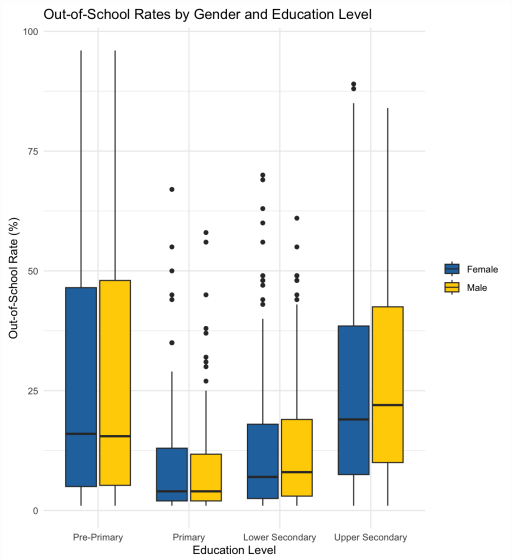
Socioeconomic Disparities exist such that

- HDI $\propto$ academic proficiency
- GDP $\propto$ academic proficiency

Gender Disparities

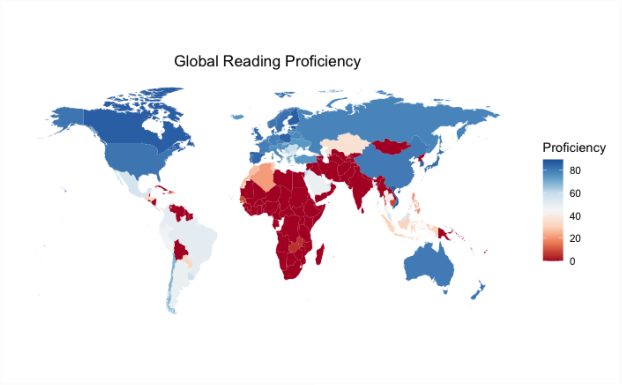
Literacy rates increase as HDI increases. Gender disparities found in literacy rates exist in low HDI countries, but not in medium and high HDI countries. Gender disparities found in out of school rates exist only in older, upper secondary students.

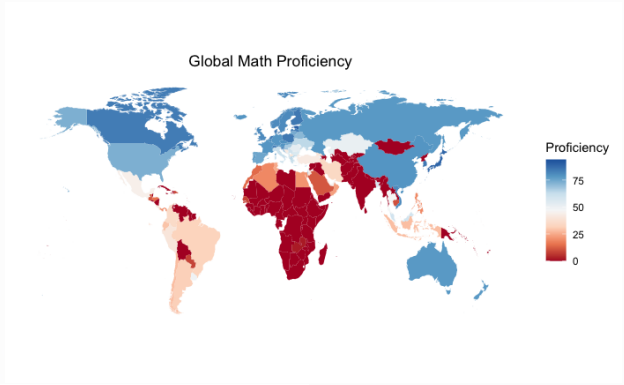


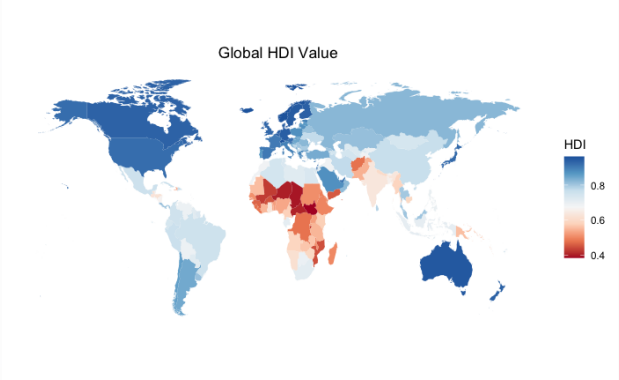


Socioeconomic Disparities

There exists a strong relationship between HDI and academic proficiency, defined by reading and math proficiency scores. HDI measures the quality of life for a country's citizens and is calculated using life expectancy, education, and income. Hence, a higher HDI indicates a higher quality of life.







Conclusion

It is clear that gender disparities within education are more prominent in lower HDI countries compared to high HDI counties. This is revealed by looking at out of school rates, literacy rates, and proficiency rates across HDI classes. Furthermore, socioeconomic status is a strong factor in academic performance with a directly proportional relationship. This is shown by looking at both GDP and HDI with respect to academic proficiency.

We hope the statistical analysis performed is used for the advancement of social equity within education.