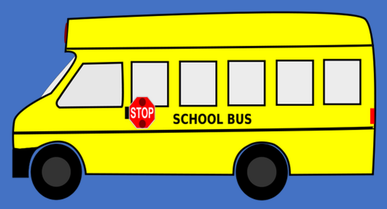


Access to Education Globally

DS 3000 Final Project

Keily Hernandez, Luke Jianu, Claire Mahon, David Alade

Abstract

Our project explores how education is a significant indicator of a country's status. We utilized the Human Development World Index [1] created from Human Development Reports by the United Nations Development Program [4], which contains data on different metrics of human development across countries and territories worldwide. The dataset spans from 1990 to 2021. Multiple algorithmic models including linear regressions, a knn algorithm analysis, a random forest regressor, as well as an SVM model were used to predict the main factors that contributed to educational access (measured mainly in a country's Mean Years of Schooling). The performance of the different models also contributed to our final conclusions.

Introduction

Defining the Problem

Especially since 2019 when the pandemic began, what defines a proficient education has changed for many students. Looking at news sources and websites such as UNICEF and The Red Cross, we realized there were many factors that played into educational access by region. UNICEF's website states, "Over 600 million children and adolescents worldwide are unable to attain minimum proficiency levels in reading and mathematics, even though two thirds of them are in school. For out-of-school children, foundational skills in literacy and numeracy are further from grasp." Overall, there are inequalities globally that control whether or not students can receive a proficient education.

In general, educated populations perform significantly better than poorly educated populations in various metrics including life expectancy, GDP, mortality rate, etc. We believe that working towards more educated populations will allow underdeveloped countries to become more self-sufficient and will cause these countries to improve in various metrics.

Motivation For the Topic

We choose to pursue this topic for three reasons: education has been a big part of our lives, education is something we value highly, and we're interested in comparing education quality/development in countries around the world. To elaborate, we are college students who have been in school for almost our entire lives. For most of the year, for around 6 hours a day, for almost 13 years, we've been in a classroom. When we're not in a classroom, there's time spent doing HW and pursuing other activities relevant to education. So much of our lives has been dedicated to education, which is one of the reasons why we see it as highly valuable. We think that the privilege of having easy access to education has given us the opportunity to live good lives. Education allows us to become better people, as we learn about the world and become more informed global citizens. It gives us the skills to succeed in the workforce. It also exposes us to interesting concepts/ideas, which can lead us to discover our passions and hobbies. Lastly, we are genuinely interested in learning more about education quality/development in other countries, and seeing if we can make any meaningful observations that give insight to how access to education can be improved globally.

Goals and Objectives

With this project, we used the Human Development World Index dataset to gather information about how different factors affect access to education among a wide range of countries. We planned to gain insight on how education varies from country to country and identify trends in our data that indicate which attributes (if any) affect a country's access to education. By pinpointing the certain commonalities between contributing factors, we can draw answers as to why some countries have education more readily available to citizens than others. After the data processing, we hope to be able to use our results and conclusions to form predictions and for the future about educational access globally.

Related Work

We found two projects related to our work, either seeking to find similar metrics, or using the dataset for exploring different global impact hypotheses. We took inspiration and notes from both, but ultimately were answering different questions for our analysis.

Our World in Data - Global Education is an online data website looking at global data sets for how access to education affects different economic and societal factors. This closely related to our work, as it sought to predict many similar factors. However, this is done on a much larger scale than our project was looking to draw conclusions from. [2]

Kaggle Sourced Plotly Comparison between Rich and Poor was a Kaggle python code project that analyzed global differences between rich and poor countries. They used the same dataset we used for our project (Human Development World Index), but drew more conclusions about economic impacts. Less of their project related to differences in access to education, although they did draw conclusions about a country's per capita GDP and how it affected a citizen's Mean Years of Schooling. [3]

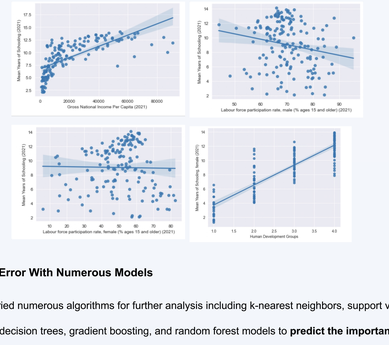
Methodology

Data Preparation: Cleaning and EDA

Our dataset was obtained from Kaggle in the form of a csv file. The dataset initially had 195 rows and 879 columns. We used deletion and imputation in the preparation stage. We first removed all columns that had more than 34 null values. We then removed all rows that had more than 5 null values. At this point, we were left with 145 rows and 612 columns. We still had a few null values remaining, which we imputed using the mean of the column.

Feature Selection

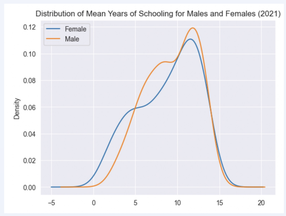
Once our data was cleaned, we had to choose the best features for our model. In this process, we first decided what features we didn't want to include in our model, as some features would give us uninteresting or inaccurate results. For example, we didn't want to include "Expected Years of Schooling" as a feature for predicting "Mean Years of Schooling", since these these two observations are inherently correlated.. After removing some features, we took what was left and picked the 300 features that had the highest correlation with our target variable. We hypothesized that features such as **Gross National Income per Capita, Labour Force Participation Rate**, and **HD Groups** would factor highly into a country's **Mean Years of Schooling** (which we used as the main feature that would reflect access to education). These original hypotheses were tested with linear regressions (shown below).



Trial and Error With Numerous Models

We tried numerous algorithms for further analysis including k-nearest neighbors, support vector machines, decision trees, gradient boosting, and random forest models to **predict the importance of different features** in our dataset. After running extensive tests and drawing conclusions from multiple models, we ended up selecting a **Random Forest Model**, as it performed the best on our dataset. Random forests are good for making regressions from large datasets, and are more accurate than decision trees alone. Although they can be less accurate with sparse data, we cleaned null and inconclusive values in our dataset to allow for clearer results. It is also a good machine learning algorithm to avoid overfitting datasets. Since our target variable was continuous, we used a regression model.

Further Selecting Our Target Feature:
Our dataset had features for mean years of schooling male and female. As seen from our plots, the distribution for the mean years of schooling for both males and females have similar shapes. The standard deviation for males is less than the standard deviation for females and the mean for males is greater than the mean for females. We assume that the cause of this is the result of gender inequalities in different countries. However, because the shapes of their distributions are very similar, we can generalize the results of our model to males with a high level of confidence. We selected data from the year 2021 in order to gain a better view of how our access to education has been affected most recently.



Hyperparameter Tuning

We tuned the hyperparameters using GridSearchCV. The first hyperparameter we tuned was **n_estimators,** which is the number of trees in our random forest. The second hyperparameter we tuned was **max_features**, which is the maximum number of features to consider at each split in a tree. The third hyperparameter we tuned was **max_depth,** which is the maximum height of any tree in our forest. We also tuned the **min_samples_split**, **min_samples_leaf**, and **bootstrap** hyperparameters, which have to do with how our data samples interact with our model. We used 4 folds for our cross-validation.

We evaluated our model's using mean squared error and the r squared value. To see what features were important in our forest, we used the feature_importances_ attribute of our model. To analyze the presence of overfitting in our dataset, we looked at the cv_results_ attribute of our GridSearchCV, which included training scores and testing scores that we compared.

Results and Evaluation

Our GridSearchCV discovered that the following values were ideal: bootstrap as True, max_depth as 50, max_features as 1.0, min_samples_leaf as 2, min_samples_split: 2, and n_estimators as 20.

N_estimators set to 20 means our regression used 20 decision trees in the model in order to produce the best mean squared error.

When our hyperparameters were finalized to achieve the best results, we managed to reach a **mean squared error of 1.08**, indicating our model performed well, as this is much lower than the standard deviation of our target, **3.58**.

Our model also had a corresponding **R-squared of 0.92**, indicating that over 92% of the variation in our model is explained and that our regression model was a very good fit for our data.

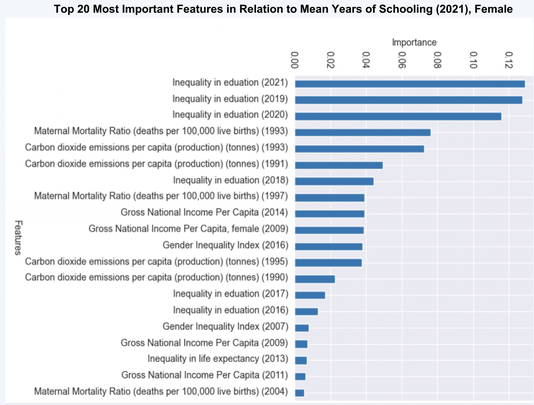
	mean_train_score	std_train_score	mean_test_score	std_test_score
481	0.975767	0.001438	0.863038	0.016659
485	0.975767	0.001438	0.863038	0.016659
483	0.975767	0.001438	0.863038	0.016659
241	0.975767	0.001438	0.863038	0.016659
85	0.975767	0.001438	0.863038	0.016659

The differences in our Mean Train Score and Mean Test Score (at 0.976 and 0.863) as well as the differences in our Std Train Score and Std Test Score (0.0014 and 0.0167) indicate a slight overfitting and potential bias in the model.

Explaining Overfitting and Bias

Despite the slight discrepancies between test and training data, the discrepancies were small enough for us to conclude that our model performed very well. Especially considering our data came from a variety of geographic locations and sources, it can be assumed that not all countries would be able to provide exact data for every feature involved, while we did our best to clean the data, there is almost always room for improvement. We were happy with the performance of the Random Forest Regression, especially considering our other algorithms were operating with Mean Squared Errors more than double the error of our model (the KNN algorithm was 2.84 and the SVM Model was 2.04).

Top 20 Most Important Features in Relation to Mean Years of Schooling (2021), Female



Interpreting Our Results

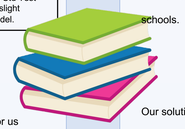
The most important features in our dataset were **Inequality in Education** (2021, 2019, and 2020) at around 0.12 importance followed by **Maternal Mortality Ratio** (1993) and **Carbon Dioxide Emissions per Capita** (1993 and 1991) around 0.07 and 0.05 importance. These results were surprising to us because we had not predicted they would have such large importances on Mean Years of Schooling. **Gross National Income per Capita** and **Gender Inequality Index** were also prominent features in the top 20 we had not necessarily expected.

Inequality in Education is a measure in our dataset that evaluates countries on adult literacy rates as well as combined primary, secondary, and tertiary gross enrollment ratio. Countries that have large discrepancies in these metrics across their populations receive higher Inequality scores. This feature importance implies that countries where there are large differences in education between rich and poor populations, as well as urban and rural populations, are also likely to have more significance on their average years of schooling for females.

Maternal Mortality Ratio is a measure of the number of maternal deaths during a given time period in comparison to 100,000 live births. It calculates how many mothers are dying during the period of pregnancy as a result of complications. Our regression suggests that maternal mortality is closely tied to mean years of schooling. An article from BMC Public Health states, "The more socially and economically advantaged people are, the better their health. Years of formal education are a well-recognised indicator of social position … studies show that people with progressively more advanced levels of education have better health and longer lives than those without … women's educational levels (relative to those of men) have been found to be associated with maternal death." [6] In general, the connection between education and maternal death/risk plays a prominent role in our findings.

Carbon Dioxide Emissions per Capita measures how much carbon dioxide a country is producing. Since the Industrial Revolution, carbon dioxide emissions have grown rapidly. Although they are a large indicator of climate change, they are also a large contributor to a country's technological innovation and development [5]. From our model, we can assume that larger versus smaller levels of carbon dioxide emissions are closely connected to a country's mean years of schooling average. This is likely because countries that have more technological innovation and development, also have more means of educating their populations, which in turn leads to more development (and more CO2 emissions).

Gross National Income per Capita is a measure of a nation's income as a whole per year. It is likely that nations with higher Gross National Incomes have more money that can be funneled into public and private education for their populations. At the same time, it is likely that countries with lower GNPs have less money to fund their schools.



The Impacts

Our solution can impact the real world and anyone trying to learn more about increasing years of schooling in a nation and in turn increasing a population's access to education. The results gained from our model can be used to enact change in the real world because it allows people to gain a better understanding of which factors most strongly contribute to changes in the mean years of schooling. While our results specifically looked at Mean Years of Schooling for women, you could also take these findings and generalize them to how populations as a whole could be benefitted by enacting change in some of the other connecting factors. The results could be a tool for any researcher who is curious about what features impact years of schooling for females across all nations. If a researcher were also looking at the Human Development World Index dataset, and trying to draw general conclusions about factors that contribute to a populations wellbeing they could easily see the connections between education, maternal health, carbon dioxide production, and more. Also, it can be useful to any activist who is seeking to aid women's access to education, the general populations access to education, increase the years of schooling for people in their nation, improving literacy rates, as well as improving development and growth. The list of people affected by our solution would be those across the nations whose years of schooling and/or access to education increases because of insights or changes our solution provoked.

Conclusion

Through our exploration of this dataset, we gained a lot of insight into the relationships between education and its contributing factors across the globe. Using our dataset, we wanted to identify different trends that could alert us to patterns that affect access to education in different countries . Our goal was to discover these patterns so they could be used as a basis for measures/further studies to improve global access to education. For example, our target column was data from 2021 but columns ranging from the years 1993 to 2019 were important features when accurately predict this target. This means that there potentially exists confounding events from 1993 that affect access to education in 2021 almost as much as events in 2019 did. We could use trends like these to help improve access to education worldwide by launching deeper studies into columns that create trends such as these.

Potential Improvements and/or Future Work

If we had the chance to expand this project further, we would have likely tried to access more education-specific datasets. The Human Development World Index dataset was large and provided a lot of insights for global differences in education, but lacked in its range of features directly related to education. While this allowed us to draw broader conclusions for factors that contribute to global educational access, we would have liked to do further research into factors such as government spending on public education, quality of education, percentages of population enrollments in primary versus secondary education, graduation rates, and more. The problem we chose was broad and despite our decision to focus on average years of schooling as the main feature of educational access, this problem could have been interpreted in numerous other ways.

References

[1] Human Development World Index https://www.kaggle.com/datasets/iamsouravbanerjee/human-development-index-dataset
[2] Our World In Data - Global Education Analysis https://ourworldindata.org/global-education
[3] Plotly Comparison Between Rich and Poor https://www.kaggle.com/code/mikevanderklis/plotly-comparison-between-rich-and-poor
[4] Human Development Reports https://hdr.undp.org
[5] A History of Carbon Dioxide Emissions https://www.wri.org/insights/history-carbon-dioxide-emissions
[6] Maternal Mortality Ratio and Education https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-11-606