Final Project

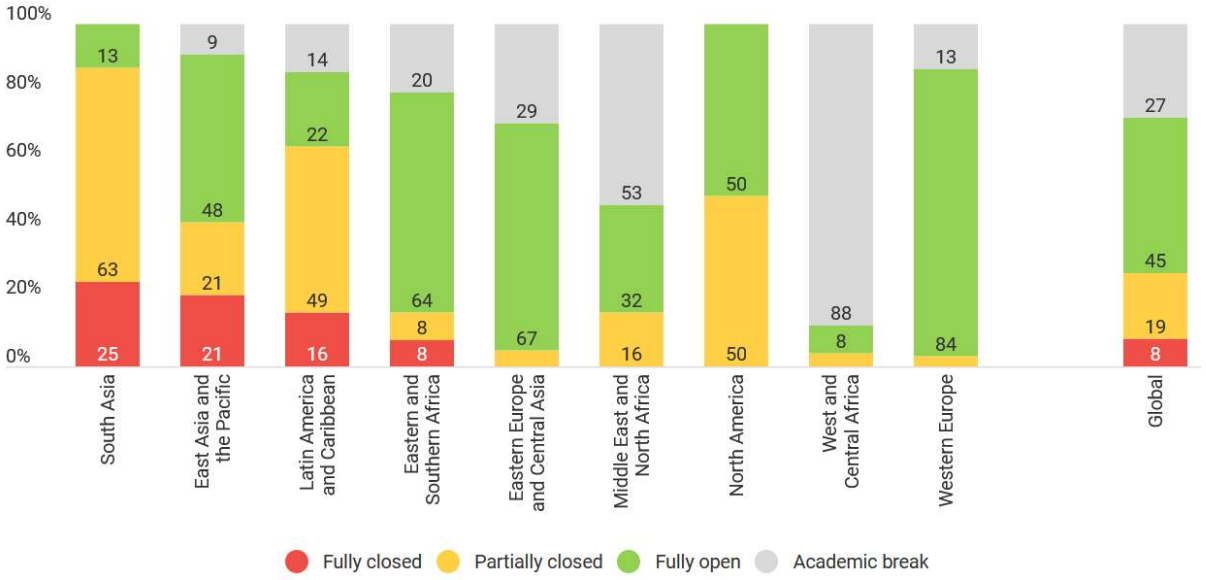
Impact of COVID-19 on Primary Education Completion Rates Worldwide

CSCI-109A Course – Fall 2021, Harvard University Extension School

I. Problem Statement and Objective of the Project

The COVID-19 crisis has severely affected the economy and most areas of life, including the educational sector. The new norms and policies for socially distancing to reduce health risks involved the closing of schools. This interrupted the normal function of the educational system with a growing evidence of a reduction of student learning achievements worldwide (link). In 2020, schools around the world were fully closed for 79 instruction days on average (from 53 days in high-income countries to 115 days in lower-middle-income countries) (link). As of September 2021 (almost a year and a half since the pandemic was declared), 27 per cent of the countries continued to have fully or partially closed schools. There is no region across the world where all countries have fully reopened schools (link).

Figure 1: Share of countries by school closure status, by region, as of September 2021



Source: (link)

The responses to the COVID-19 school closures mainly included remote learning solutions such as:
- paper-based take home materials
- broadcast media (radio and TV)
- digital platforms.

However, there are differences in the responses by the income level of the respective country. Responses focusing on broadcast media (specifically radio) were more popular among low-income countries (92 percent) than high-income countries (25 percent). Conversely, responses relying on-line platforms were more common in high-income countries (96 percent) than in low-income countries (58 percent) (link).

Despite these locally adapted mitigation measures, research suggests that students affected by school closures experienced significant learning losses. In general, COVID-19 could result in a loss of between 0.3 and 0.9 years of schooling (adjusted for quality), which will reduce the effective years of lifetime schooling from 7.9 years to between 7.0 and 7.6 years (link). As such, these students may earn $49,000 to $61,000 less over their lifetime due to the impact of the COVID-19 on their schooling (link).

Drop-outs and an overall reduction in primary and secondary completion rates are also expected. At global level, the International Labour Organization (ILO) estimated that four out of every five workers were unable to work due to lockdowns. In poor households, the associated decline in income will limit their potential investments in education (link). Close to 7 million students from primary and secondary education may drop out due to the income shock of the pandemic alone (link).

During these trying times and (upcoming) learning crisis, countries need reliable information to plan not only immediate response efforts but also to prepare recovery strategies in the educational sector (link). Therefore, the main objective of this final project is to use publicly available datasets to contribute to the understanding of the scope of the learning crisis by determining the impact of COVID-19 on primary education (although still with usual time limitations for working on this final project and only with our restricted current knowledge of data science).

II. General Background on the Data Used in the Project

The original final project topic H was "United Nations International Children's Emergency Fund (UNICEF) Data Set: Impact of COVID-19 on Secondary And Primary School Education". However, the UNICEF dataset (link) provided for the analysis only includes data pre-COVID (from 2019 or before) and does not include data post-COVID (from 2020), making it not suitable for an evaluation of the impacts of the pandemic[1]. Therefore, we identified and used a comprehensive United Nations Educational, Scientific, and Cultural Organization (UNESCO) dataset on educational indicators (link), which we complemented with a World Bank dataset on economic indicators (link). The main characteristics of both datasets are provided in the table below:

[1] The World Health Organization declared the Covid-19 outbreak as pandemic on March 11, 2020 (link), with the first official case going back to November 17, 2019 (link).

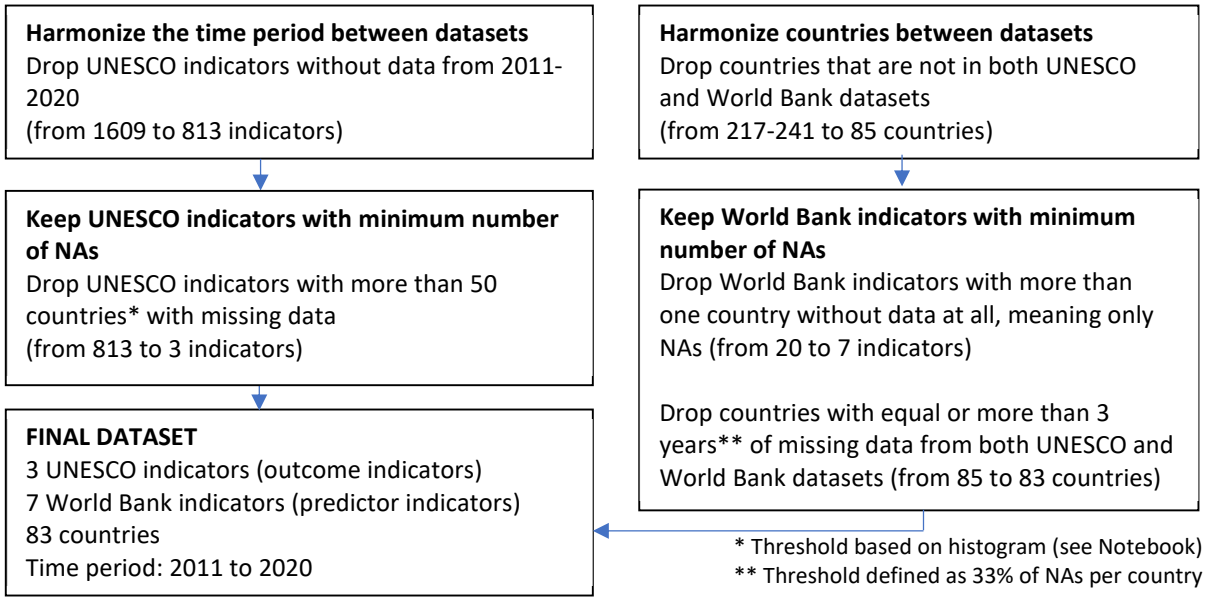
Table 1: Main characteristics of datasets used in the project

Characteristic	UNESCO dataset	World Bank dataset
Number of indicators	1609 SDG-related[2] educational indicators	20 economic indicators
Number of countries[3]	241	217
Years of data	1970-2020	2011-2020

III. Data Cleaning and Reconciliation

The data cleaning and reconciliation was extensive as both datasets include an unequal number of countries and years of data per indicator. Besides, they include multiple NAs unequally distributed among indicators, countries, and years of data. In general, our data cleaning and reconciliation procedure focused on identifying those indicators with the minimum number of NAs and followed the steps in the (simplified) flow chart below:

Figure 2: Flow chart summarizing the main steps for data cleaning and reconciliation

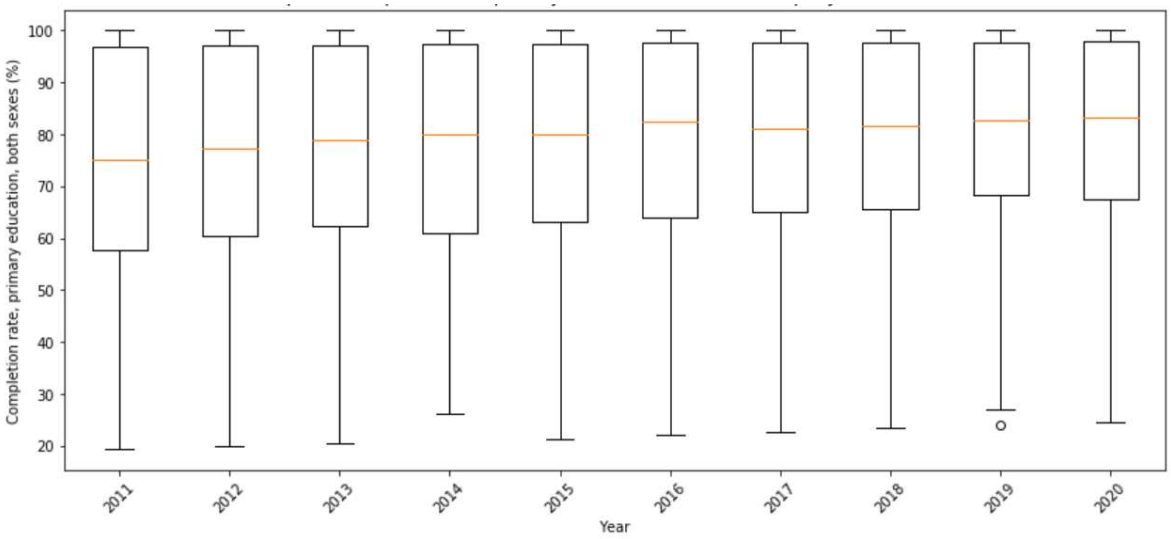


[2] SDG stands for "Sustainable Development Goal". The UN has identified 17 SDGs, which humanity needs to achieve to guarantee sustainable progress for all (link). UNESCO is the custodian agency of SDG 4 or SDG on "Quality Education". UNESCO uses this data to measure progress towards the achievement of that specific goal.
[3] The number of sovereign countries, which are recognized by UN and Members States of UN, is 193 (link). However, UNESCO and World Bank datasets also include as "countries" currently not existing countries and some territorial dependencies of other countries.

IV. Findings of the Exploratory Data Analysis (EDA)[4] and Final Variable Selection

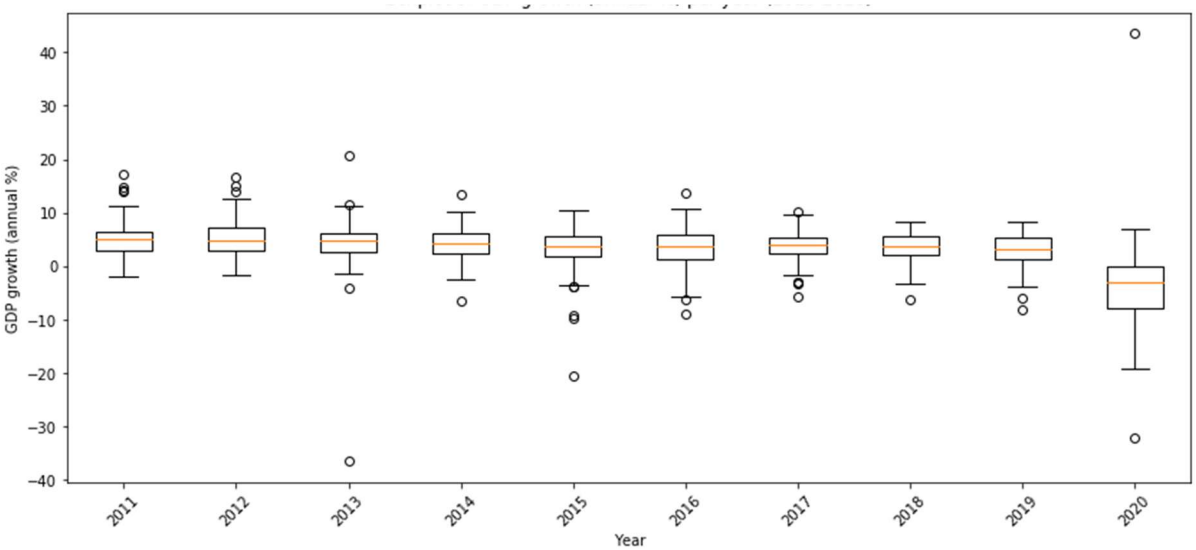
Our EDA included boxplots per year to observe the trend over time per indicator. From the boxplot of the outcome variable, we observe that primary education completion rate has increased from 2011 to 2020, but there was not a significant drop in 2020 in comparison to 2019. This suggests that the impact of COVID is still not strong or evident in (some) educational indicators.

Figure 2: Boxplot of completion rate, primary education, both sexes (%) per year (2011-2020)



Conversely, there has been a drop in economic indicators from 2019 to 2020, such as in annual GDP growth.

Figure 3: Boxplot of GPD growth (annual %) per year (2011-2020)



[4] Our EDA includes a histogram, boxplots, scatter plots, a correlation heatmap, and Pearson correlations, but only a couple of boxplots are shown here due to space constraints (complete results in Notebook).

After cleaning and reconciling our educational outcomes and economic predictors data, the following number of variables remained for further analysis (see also Section III for the flow chart summarizing the cleaning and reconciliation of our data in this document and the Notebook for further details and explanations):

- Educational outcomes: 3 educational outcomes (with 99 countries and data from 2011 to 2020 each of them)
- Economic predictors: 7 economic predictors (with 99 countries and data from 2011 to 2020 each of them)

Based on the EDA results, we selected only 1 educational outcome variable:

- Completion rate, primary education, both sexes (%) in 2019

We dropped the other two educational variables, which were also primary education completion %, but disaggregated by sex (male and female), as the trends were similar to primary education completion both sexes % (the selected educational outcome indicator).

In addition, we selected 5 predictor variables (economic indicators):

- GDP growth (annual %) in 2011, 2015 and 2019
- GDP per capita (current US$) in 2011, 2015 and 2019
- Inflation, GDP deflator (annual %) in 2011, 2015 and 2019
- Mobile cellular subscriptions (per 100 people) in 2011, 2015 and 2019
- Population growth (annual %) in 2011, 2015 and 2019

We dropped two economic indicators (current GDP in US$ and total population), as they were already accounted for by the selected economic indicators (and were visually similar to them).

Regarding the time dimension of our data, we are not using time series models for this final project (as they are not covered by the CS-109A course). However, to account for the effect of time in our models, we have included time lag variables as predictors (as the reference year is 2019, we included 2015 to account for the situation 5 years ago, and 2011 to account for the situation 10 year ago).

The intuition behind incorporating time lag variables is that the effects of some predictor variables on the others require time to become evident. For example, the effects of an increase on the GDP growth (annual %) may take five or even a decade before they can be associated with a reduction on population growth (annual %).

The inclusion of time lag variables is in line with approaches followed by similar type of research. For example, Kane et al. (2014) compared the performance of a time series model (ARIMA) and a random forest model using lag time variables. They found that the random forest model outperformed the ARIMA model in predicting a disease outbreaks (link).

Beyond the predictor variables indicated above, our models included two types of categorical/dummy variables:

- Geographic location (per continent or sub-continent: Africa-Northern, Africa-Sub-Saharan, Asia-Central and Southern, Asia-Eastern and South-Eastern, Asia-Western,

Latin America and the Caribbean, Northern America and Europe, and Oceania-Excl. Australia/New Zealand[5]; as defined by UNESCO)
- Development status of the country (low-income, middle-income and high-income, as classified by the World Bank as of July 2021)

In this case, as the categorical variables provide too many subgroups and our number of observations is small (83 countries), we decided to include only one dummy per type (for location, Africa-Sub Saharan=1; 0 otherwise; and for development status, low-income=1; 0 otherwise).

We expect that the dummies (partially) capture the effect of variables not available for inclusion in our models and that vary according to geographic location and development status of the country. For example, broadcasting in radio was more common as remote learning response to COVID-19 in low-income country than in other countries. In relation to geographic location, Africa Sub-Saharan is the poorest region in the world, and likely very different in responses and approaches than the rest of the countries.

V. (Refined) Research Question

Based on the selected educational outcome variables, our (refined) research question is: What is the impact of COVID-19 on primary education completion rates (%) worldwide?

VI. Methodology and Assumptions

The main steps we are following for answering our research question include:
- using pre COVID-19 data (2011-2019) to build model(s) to predict 2020 (we will consider this prediction as the equivalent to what would have happened without COVID-19).
- calculating the gap between our 2020 model predictions and what actually happened with COVID-19 (this means the difference between our 2020 prediction and the actual/real 2020 data). We will consider this gap as the impact of COVID-19.

Our main (strong) assumptions are:
- Only the selected economic predictors influence our outcome variable (primary completion rates).
- The observed gap between the predictions of our model for 2020 and the actual/real data for 2020 is only/mainly due to COVID-19.
- The effects of COVID-19 on education are already evident one year after the pandemic started.
- The different local measures in place to mitigate the impact of school closures and student dropouts in 2020 had similar effects across the globe (and do not need to be explicitly included in the models[6]).

[5] Australia and New Zealand are not part of the 83 countries analyzed in this project

[6] We are aware that this is a very strong assumption, but there is not information available of measures taken per country. However, to (partially) account for these differences (and other related differences), we included dummy variables for geographical region and development status of the country in our models.

VII. Regression Model Results

We ran different types of regression models for comparing their performance and the results of their predictions[7]:

- Linear
- Lasso linear
- k-NN
- Single decision tree
- Bagging
- Random forest
- AdaBoost

The hyperparameters of all the regression models (with the exception of the standard linear) were tuned using cross-validation (please see Section V in our Notebook for all the detailed results and graphs). Also, to work with a balanced data, the train – test split was stratified by development status of the country (low-income country =1; 0 otherwise).

The results of the regression models are summarized in the table below:

Table 2: Summary of main results of the regression models

Regression	Actual value	Prediction	Mean difference %	Train accuracy	Test accuracy
(Standard) Linear	79.7745	80.1050	2.94	0.8103	0.7190
Lasso linear	79.7745	80.1026	3.53	0.7850	0.7494
k-NN	79.7745	79.4850	2.74	0.7647	0.6982
Single decision tree	79.7745	80.8430	5.39	0.7587	0.6317
Bagging regression	79.7745	79.6700	2.62	0.9441	0.7189
Random forest	79.7745	79.4707	2.26	0.9516	0.7185
Adaboost	79.7745	77.8177	-0.27	0.8847	0.8080

Actual value: mean primary completion % rate in 2020 worldwide (83 countries included in the project)
Prediction: mean primary completion % rate predicted by the regression models for 2020
Mean difference %: difference between the actual value and the prediction in %
Number of observations: 83 (countries)

In general, despite the small sample size after cleaning the data (83 countries) the test accuracy of all the models was medium to relatively high (between 0.63 to 0.81), while the difference between the predictions of the models for 2020 and the actual value of primary completion rate for 2020 was between -0.27 and 5.39 percent.

As expected, the AdaBoost regression showed the best test accuracy (0.81) (followed by the random forest, bagging, k-NN, and single decision tree). Also, the (standard and lasso) linear

[7] We followed the same procedures as in class to fit the models, tune the models, estimate accuracies, calculate feature importance, and predict using the models. The main difference was that in class we focused on classification models, while we used regression models for this project.

models provided high test accuracies but violated the homoskedasticity assumption (detailed results in Section 5.2 in our Notebook) and the multicollinearity assumption, as predictors are correlated (as shown in Section 4.7 in the EDA in our Notebook). Therefore, these models were not considered suitable for prediction (but we included them here as reference).

The AdaBoost regression was tuned at 300 iterations, as the test accuracy results were stable at that number of iterations (see Figure 4); and at a maximum depth of 2, as the train and test accuracy results were not so distant between them and the test accuracy results were also stable after 200-300 iterations for this maximum depth (see Figure 5).

Figure 4: Accuracy of AdaBoost regression as training progresses

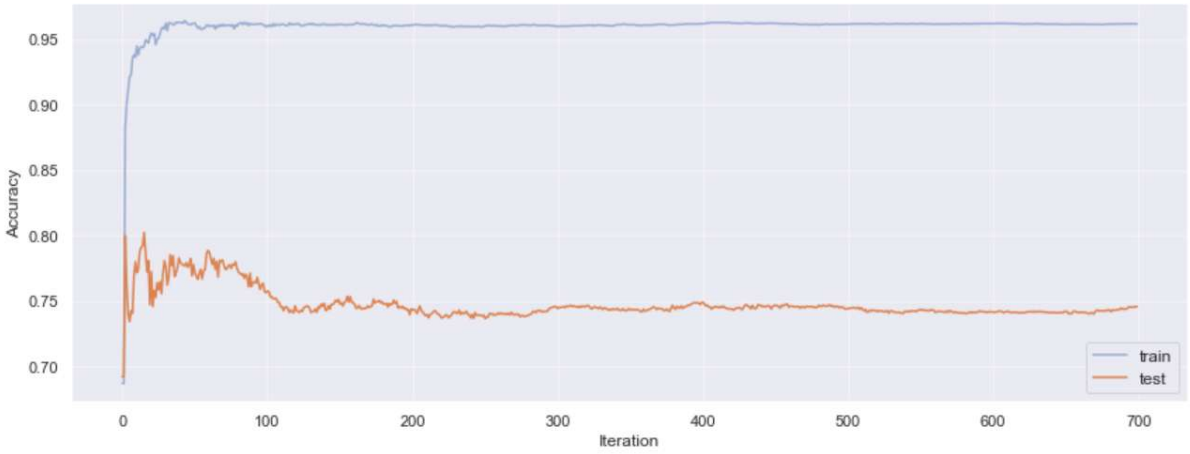
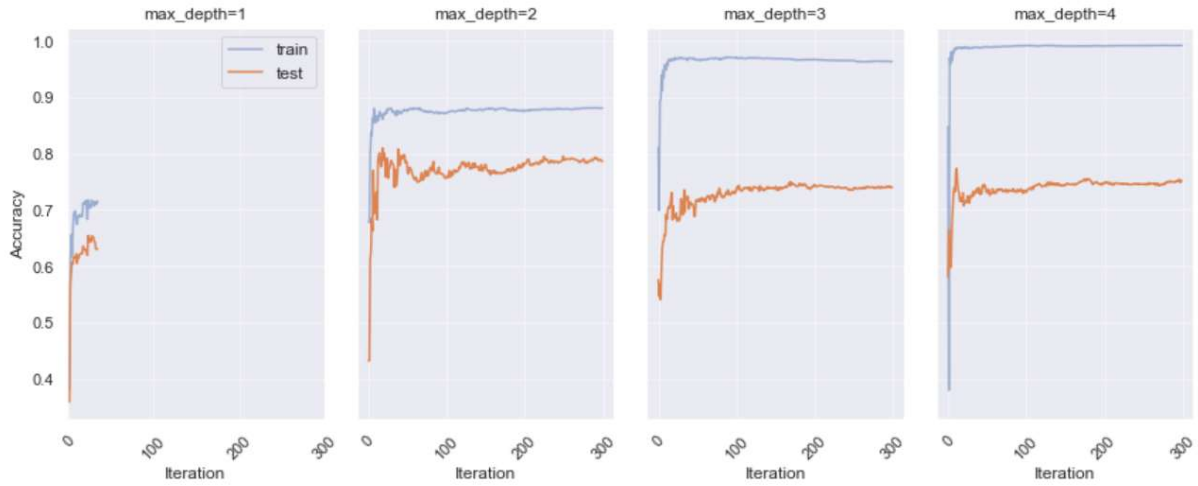


Figure 5: Accuracy of AdaBoost regression by maximum depth as training progresses



We also evaluated if all the predictors were important for the AdaBoost model. The analysis of the feature importance suggested that all the predictors had influence in the accuracy of the model (Figure 6), while the permutation feature importance[8] considered only three

[8] The feature importance tells us about the influence of the predictors in the accuracy of the model; while the permutation feature importance randomly shuffles the feature values and evaluate the decrease in accuracy of the model. If the decrease is very small, then the feature was not very important for the model predictions.

predictors as not having influence in the accuracy of the model and none of them was negative (Figure 7). Therefore, we decided not to prune (drop) any of the predictors from the AdaBoost model.

Figure 6: Feature importance for the AdaBoost Regression

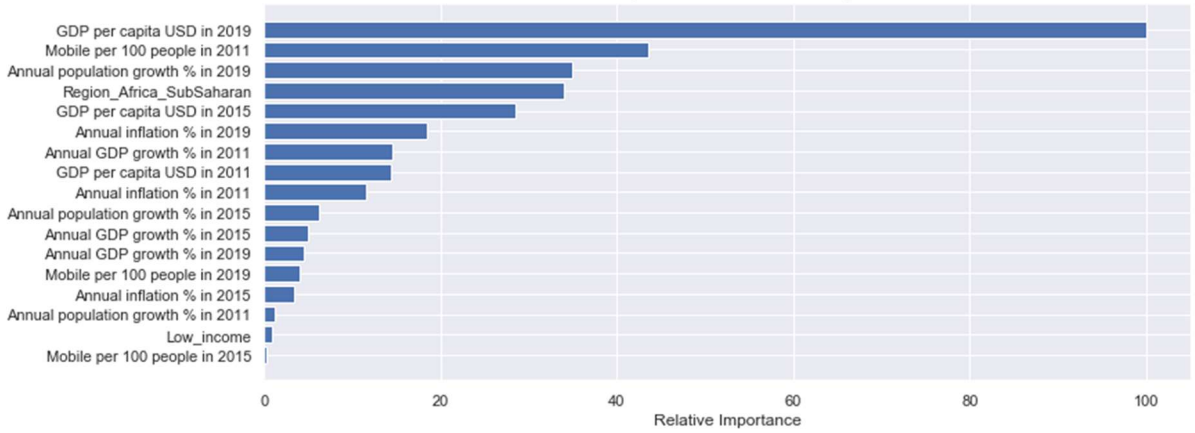
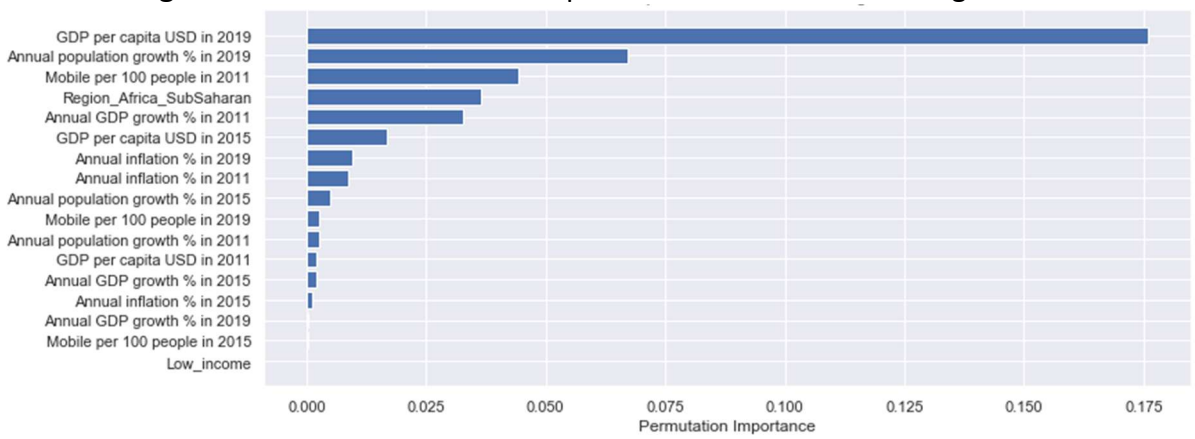


Figure 7: Permutation feature importance for the AdaBoost Regression



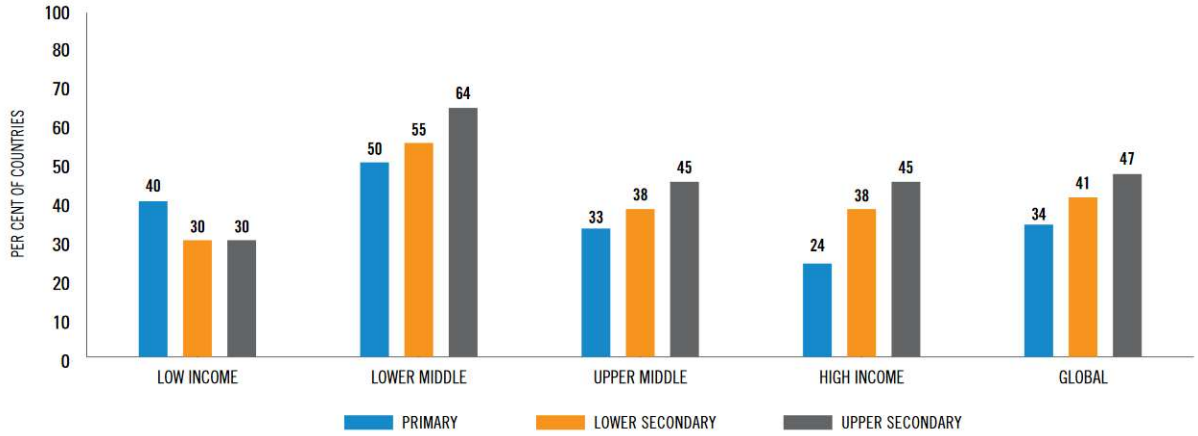
Finally, the difference between the prediction for 2020 of the AdaBoost model and the actual 2020 value for primary completion rate was less than 1 percent (-0.27 percent), suggesting that the impact of COVID-19 is (still) not evident for this educational indicator.

VIII. Conclusions and Way Forward

COVID-19 severely impacted all areas of life. In relation to school education, countries around the world ordered school closures and when feasible, put measures in place to mitigate their impact, focusing on facilitating the continuity of education for all through remote learning (link). However, countries also responded to the pandemic with other measures specifically oriented to mitigate potential decreases in primary and secondary completion rates. For example, 41 per cent of countries extended the academic year and 42 per cent prioritized certain curriculum areas or skills (link). As well, 34 percent of countries included plans to

adjust graduation criteria for the school year 2019/2020 (end of 2020) for the primary level (link).

Figure 8: Share of responding countries that introduced adjustment to graduation criteria (end of school year 2019/2020), by level of education and income group



Source: (link)

In general, our results suggest that these targeted measures worked in mitigating the impact of COVID-19 on primary completion (%) rates. There was not a decrease in this educational indicator in 2020 in comparison to 2019. However, countries may not be able to cope using the same strategies in long term, as school closures make more difficult for some students to go back to school, especially if their households are under economic stress (link). Evidence from other crises indicates that the longer vulnerable children are out of school, the less likely they are to return (link). Unfortunately, it could be expected that dropout rate would increase (and primary completion rates would decrease) if school closures continue worldwide.

Importantly, the pandemic has forced us to pay more attention on learning quality, equity and inclusion. The most challenging issue in education under the current crisis is to not simply the school completion rates, but that all children could indeed receive quality education (link). Furthermore, schools do more than offer formal education (meaning teaching children how to read, write and count). In many cases, they provide meals and nutrition, health and hygiene services; and psychosocial support that many children lack at home (link). As such, future research in this area may need to evaluate the impact of COVID-19 in learning quality and other child development indicators beyond school completion rates (obtaining a primary high school degree does not necessarily entails the knowledge and skills children may need to success later in life). Ideally, the data on this type of indicators will be publicly available and volunteers like us may be able to help to get more insights on this relevant topic.