Improving Education Through Data

Final Report

Authors: Josh Grace and Ziwei Wu
Advisors: Dr. Franz Kurfess, Dr. Brian Stacy
CSC 580 Artificial Intelligence
Cal Poly San Luis Obispo, Computer Science and Software Engineering Department
Submitted December 7, 2021

Abstract

The World Bank has collected data from schools in Jordan, Peru, Rwanda, and Ethiopia. The data focuses on administrative characteristics, student academic performance, teaching ability, and access to resources in the classroom. With the collected data, the World Bank wants to find possible insights about how each factor in a school may affect student performance through machine learning (ML). Their ultimate goal is to use these factors to inform public policy to help improve student outcomes. The World Bank is also interested in finding techniques not commonly used in economics that might help future projects analyze data.

To achieve these goals, we applied clustering and regression techniques. Using these methods allowed us to use standard methods in Machine Learning to find insights that had sufficient evidence to be applied. These also allowed us to use well-known metrics to validate our models. We found possible factors influencing student outcomes using various regression and clustering techniques. These include the involvement of parents in schools and whether the internet and textbooks are available in schools. We also demonstrate that clustering and regression techniques could find relations in high-dimensional education data.

Overview

We want to understand the many factors around a school and how they influence students' learning ability. These include factors such as the behavior of public administration officials, economic activity around the school, and teachers' access to resources. Currently, the World Bank has data on schools that we will use to find insights related to how each of these factors may affect school outcomes/student learning. At the start of the project, this included Jordan, Peru, and Rwanda, with data from Ethiopia getting added at the end of the project. We are also interested in finding factors that influence student performance globally. These would help create general policies regarding school governance and ways to improve underperforming school systems. Student outcomes are measured through standardized tests and other activities to find their levels of knowledge, similar to previous efforts to quantify student performance.

The World Bank has significant resources to help governments in disadvantaged regions improve their countries. There are groups at the World Bank interested in improving students' educational outcomes in these countries. They will use our insights to inform policy decisions on what aspects of school systems to invest in and add requirements to other loans to improve education in those regions. To support the team at the World Bank, we detail our methods and insights into the data. These will help the World Bank economists and bankers understand the trends we found to apply our insights to their policies.

Difficulty

The dataset we are using from the World Bank was only recently cleaned and had minimal analysis performed. So, we may encounter difficulty working with the data. Additionally, there are significant data from the four countries the World Bank studied, which may cause issues when processing the data. Furthermore, neither of the project team members had experience in Economics or Data Science. So, we first needed to learn techniques to find trends in the data before starting significant work on the project. Accordingly, we believe the project will have a difficulty of 9.

Relevance

Our project matches well with the class topic, as our project is applying decision-making procedures in complex environments. We will use Machine Learning (a branch of AI) to find correlations between data and how it will affect school outcomes. To achieve this, we will experiment with different ML methods to find the best match for the data. The project will give us great opportunities to improve our machine learning skills and produce research directly applicable to real-world problems. Furthermore, our team is confident that our project will satisfy all the learning outcomes as it will give us opportunities to carry out the research topic in a very applied environment. Finally, we will evaluate how current trends in machine learning can be applied in a non-CS related field, which is one of the main parts of the course description. Based on this evaluation of our project, we are confident that the project matches very well with the class topics.

Requirements

We will deliver this report and our code to the World Bank. The report will detail our insights into the data and our methods to generate the insights. This will clearly document work that the World Bank can use to verify our methods and reproduce our project. We will also detail the methodology we used to gain these insights and why the approaches we tried best fit the data we had. This will give the World Bank sufficient justification for our project outcomes to implement the findings into real-world policy.

Background

The World Bank has been performing studies of schools to determine what factors make schools more effective at teaching students. They focused on Practices (School Resources, School Management, Teachers, Student Learning), Policies (Teacher attraction, Clarity of Functions, Nutrition Programs, School Monitoring), and Politics & Bureaucratic Capacity (Financial, Impartial Decision Making, Quality of Bureaucracy). This data was collected in Jordan, Peru, Rwanda, and Ethiopia. The World Bank collected data by performing on-the-ground surveys and observations. For example, they performed school surveys by testing students in literacy and math using a standardized test, surveying teachers about their teaching methods, testing teacher knowledge, observing classes, etc. The World Bank also inspected the schools to determine the availability of resources like textbooks, pens, etc. and the state of the school infrastructure like if there is running water, disabled student access, etc., They studied Politics and Bureaucracy by surveying teachers about hiring practices, the opportunity for advancement, and the overall running of the schools. They also surveyed principals about their knowledge of the schools, teacher evaluations, and interactions with the school. Furthermore, they surveyed higher-level officials to understand the overarching bureaucratic decision-making. The answers to these questions were then quantified and added to CSV files for each country. Using the data, the World Bank hopes to understand the factors that enable schools to educate students successfully.

Related Work

This project was previously worked on by Babar Ayan [1], a student at a university in Germany. They were responsible for cleaning the data and writing a program characterizing the relation between school district success and the distance to the central office. The previous student was able to find insights about the relationship of school characteristics to the academic outcomes of 4th graders.

There have also been papers published exploring the relationships of many factors to students' academic performance. Park et al. [6] studied the relationship of student academic performance and the school's learning environment to three types of parent involvement (PI): involvement in improving the school (public-good PI), involvement to help the parents' own childrens' schooling (private-good PI), and networking among parents (networking). The authors found that public-good PI and networking had more benefits (more students having better scores than the national average) for schools with high socioeconomic status. For schools with low socioeconomic status, private-good PI and networking helped students more. Bierman et al. [2] studied an intervention method for children identified to have aggressive-disruptive problems. They concluded that the intervention did not significantly improve long-term school outcomes. Buyse et al. [3] studied the impact of first-year teacher-child relationship quality and their impact on the children's psychosocial adjustment and academic achievements. The authors found that the relationship quality and classroom relational variables were associated with the students' psychosocial adjustment in their first few years of school. On the other hand, these variables had no significant effect on the students' academic achievement.

System Design

Our project takes the information gathered by the World Bank on school education systems using questionnaires and research to find trends/ insights between the education system and the administration in schools. To achieve this, we explored the use of machine learning (ML) techniques to find which factors influence the outcomes of student academics. To ensure that our methods are generalizable to other data gathered in different countries, we used automated data processing methods and added documentation in the code to make future modifications and expansion easier. With the knowledge gained from our research, the World Bank will be able to make better decisions to help schools and students. We will also be publishing papers about our methods and the insights from the data given. In the future, anyone interested in continuing this project will be able to recreate our results and improve upon our methodology.

Evaluation Criteria

We generated requirements to define good insights to ensure our insights are accurate and applicable to the real-world. First, our insights must have sufficient statistical backing. For supervised machine learning tasks, we split the data into training and testing sets. The ML models are trained with the training data and are tested with the testing data. This way, we will better understand how the ML methods perform when faced with different data points and ensure they are not overfitting. We use the mean squared error (MSE) and R^2 scores to evaluate the models. We believe that the combination of the metrics will give us confidence in the quality and accuracy of our models and provide sufficient support for our report to the World Bank.

We used standard correctness measurements for unsupervised tasks to ensure that the models generating our insights have sufficient backing. Furthermore, we ensured that any groupings our models generate would balance school countries across clusters. As we know that certain countries have stronger school systems than others, insights that relate to the country a school is from are not valuable. Instead, we want to find insights related to the relative performance of schools of similar resources. These insights have much broader applicability and will be much more helpful to the World Bank.

Functional Components

Our project first reads data from the World Bank and third parties as needed. Then, it combines the varied datasets into a single Dataframe for easy manipulation and application of statistical methods. We will then clean the data as needed to ensure missing data does not cause issues with any methods we call. Once the data is ready for analysis, it will be passed to the insight component.

The insight component applies machine learning and statistical methods to extract insights and relations from the data. This finds relations between variables and creates predictions about the data using extracted trends. Then, it creates visualizations to help our customers at the World Bank understand our insights.

Finally, these insights and visualizations were incorporated into this paper, describing what we learned from our project. These describe the insights we gained into the data and help the World Bank implement what we learned.

Model, Data Structures

During processing, the data is held in Pandas [5] Dataframes. This is a standard structure that almost all Python Machine Learning and Data Science programs can read and write from. Furthermore, it has native support for CSVs (the filetype of our data). So, this makes reading/ writing files much more straightforward. Dataframes also make data cleaning and selecting data very simple, so this saved lots of development time by streamlining common data processing steps.

We used machine learning algorithms included in sklearn [7] for models. In addition, we relied on MatPlotLib [4] and Seaborn [8] for visualization.

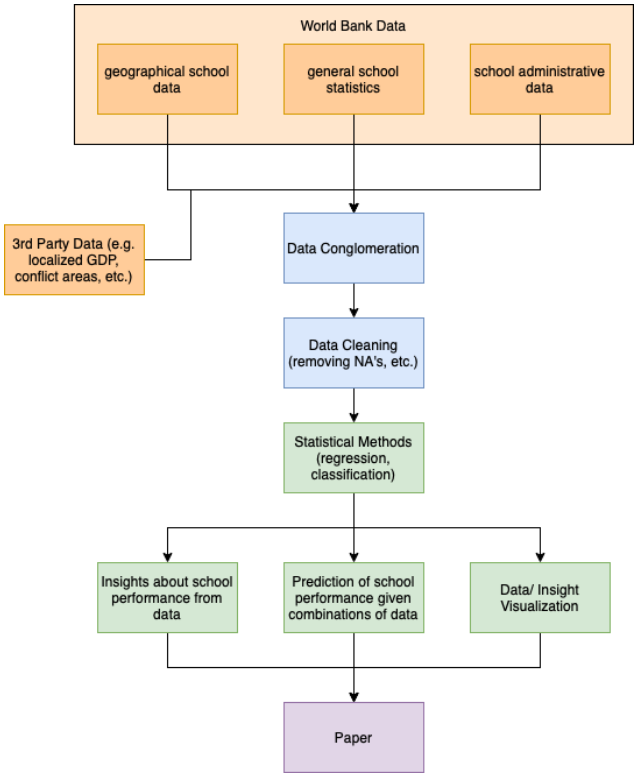

Figure 1: Project software block diagram

Technologies, Tools, Languages, Development Environments

We wrote our code in Python, and as such, we used Python libraries to help us
Our data was read using Pandas, which made interfacing with the data from our CSV easier. For the clustering algorithm and multiple regression methods, we used sklearn to generate the models.

Implementation

We chose two techniques to find insights in the data: regression and clustering. We were concerned that, given our lack of familiarity with data mining techniques, we would not choose a method with a high likelihood of successfully finding insights. So, we split our efforts to implement two different techniques, which increased the chance of successfully finding a valid method.

Clustering Implementation

Several data pre-processing steps were performed to correctly cluster the data to ensure any insights found applied to all countries. First, the individual data CSVs were combined using the school hash value. Then, any schools missing students_knowledge data were removed as insights in schools that did not have this data would be less interesting as conclusions about student performance could not be made.

Once the data was preprocessed, clusters could be generated. First, the variables to be used in the clusters were selected to gain insights into specific group-level relations. Next, each of these groupings is evaluated with both the overall and country-specific data. This ensured that relationships could be found in the aggregate and peculiarities within specific countries. The following steps are performed for the overall data and each country's data.

- Missing or NA values were replaced with the variable's mean across the whole group. This ensured that a missing variable did not require the whole school to be excluded from the cluster.
- Variables were demeaned by country. This project was interested in the relative performance of schools. As certain countries may perform better overall, when evaluating data in aggregate, we needed to ensure we found trends that affected the relative performance of the schools. We removed the country mean from each school's variable to achieve this.
- Each variable was scaled to be on the interval 0 to 1. As clusters are evaluated using euclidean distance, variables need to be on the same scale to ensure that variables with a large scale do not skew the clusterings.

Once the data was standardized, a k-means cluster model from sklearn was fit to the data. The labels from this model were added to the data so insights could be gained using the cluster groupings from the model. Next, graphics were generated to help understand the data visually. As graphics are an easy way to understand complex relations, we were especially interested in graphs that described our insights. We used spider graphs to visualize the relationships between clusters. They visualize each cluster as a different color. The spider graphs are polar graphs where the variables used in the clustering are listed outside the graphs. Starting at each variable and extending to the center are the mean lines. The variable mean of each cluster is shown as the cluster line's interaction with the variables' mean line. We evaluated the shape of the clusters to determine the relative positions of each cluster. We were particularly interested in which clusters were subsumed by other clusters as this would represent a worse performing school according to the data. We were also interested in partially subsumed clusters as this

would represent different classes of schools where higher scores in one area may cause lower scores in other areas. Spider graphs were beneficial in quickly understanding the outcomes of our clusters as they provide a quick visual way to evaluate clusters. An example spider graph is shown in Figure 2. In addition to spider graphs, we also graphed the cluster distribution of each country. We needed to ensure that specific countries were not over-represented in clusters as this indicated the clusters were based on country quirks instead of global trends. We also generated correlation matrices and heatmaps to try and find strong relations between variables. However, no insights were found with these graphics.

We also wanted to ensure we had a quantitative understanding of the data. So, tables of the cluster means were used to back up our insights into the clusters. We visualize the cluster labels as columns and variables as rows. Each cell was the mean of the variable in each cluster. The labels were sorted by the sum of the means. As all variables had a higher is better scoring system, this allowed us to see which clusters were performing better quickly. To sort the rows, the variance across the label means was used as this allowed us to quantify how far each cluster was in that dimension. This was important as it let us prioritize variables that the cluster identified as having a higher impact than others. These tables helped to verify insights seen in the spider graphs and helped us understand where the significant differences in the clusters occurred. An example of this table is shown in Table 1.

We also printed evaluation statistics as described in the Clustering Evaluation section for each cluster to ensure that the insights from our graphs had sufficient backing for them to be valid and useful.

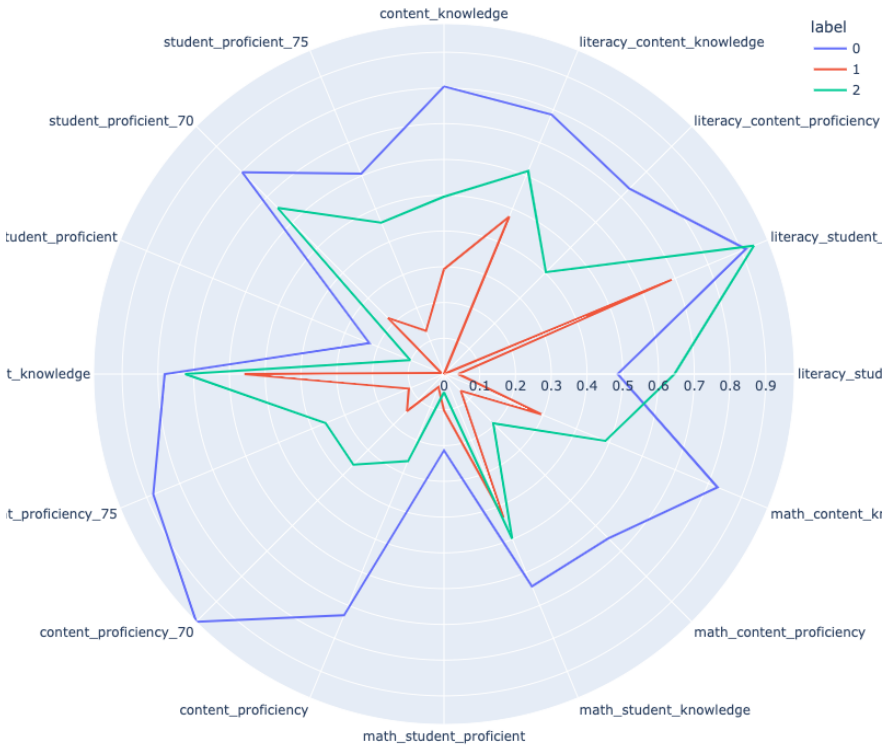

Figure 2: Example spider graph

	1	2	0
sch_goals_exist	0.961968	0.973408	0.069709
sch_goals_relevant	0.961968	0.973408	0.069709
sch_goals_measured	0.828335	0.813020	0.057826
sch_goals_clear	0.741605	0.750069	0.049695
principal_used_skills	0.842512	0.146334	0.516397
prinicipal_trained	0.848415	0.187687	0.524387
principal_management	0.794714	0.778536	0.309251
sch_support	0.758414	0.241307	0.496955
principal_training	0.443968	0.119872	0.273137
principal_offered	0.707150	0.441849	0.544112
drinking_water	0.667599	0.602449	0.489403
monitoring_infrastructure	0.613572	0.543358	0.452180

Table 1: Table of most different features and cluster means in overall data

Regression Implementation

Initial Regression Analysis: Linear Regression
Before we could apply models, the data had to be pre-processed. So, first, we read the CSV files containing the World Bank data into Dataframes using the Python pandas library. After that, we took out all student outcome columns from the Dataframes as we looked at administrative factors of features.

For regression, we wanted to find the regression method that was the best fit for the given data. In order to do so, we used a brute force method. For each regression method, we first split the data in an 80-20 ratio for training and testing, respectively. Second, any NAs left in the training or testing data were replaced by either the mean or the median of their respective columns. Next, we trained the model, used it to predict the outcomes, and evaluated the model using mean squared error (MSE) and R^2 metrics. Finally, we repeated the first and second steps up to 10 times and acquired the average of each metric. As the training and testing splits are randomized each time, this minimized the impact of randomness on the output.

Unfortunately, most models were bad fits, with many of the r-squared scores being negative. Linear regression, however, seemed to perform the best, especially for student knowledge and literacy student knowledge scores with R^2 scores of 0.59 and 0.61, respectively. Looking at the previous work, the highest R^2 score achieved was roughly 0.69. Accordingly, we thought linear regression might be worth looking into.

In order to find possible features for linear regression, we turned to stepwise feature selection, a feature selection algorithm made in mind for linear regression. It has two main steps in a loop: adding and removing features based on p-value thresholds. Once a step is reached where nothing has been added or removed, the names of the chosen features are returned. After choosing the features, they are validated by training a linear regression model (80-20 train-test split). The trained model is then evaluated by MSE and R^2 scores.

As training and testing splits are randomized, we used the feature selection algorithm up to 50 times. The chosen features were then counted for how many times they were chosen out of the total number of runs. Finally, the averages of the MSE and R^2 scores were also recorded.

Peru and Jordan vs. Rwanda and Ethiopia
Following the success of using regression methods to find possible insights, the World Bank wanted us to experiment with splitting up the data based on countries with similar economic backgrounds. Data from Peru and Jordan would be combined and compared against Rwanda and Ethiopia for any possible insights. The steps to find a potential model for this data split were done the same way as the initial regression analysis.

Unfortunately, none of the regression models could get good results from this data split. Therefore, it was concluded that this was a dead-end for us because there was not enough data

to perform a data split in this way. For example, linear regression had a -1019490453554.5273 R^2 score for the Peru and Jordan dataset for predicting student knowledge. However, we did find that the random forest regression method had the highest R^2 score. We thought this result suggested that it was a more flexible model than the others.

Random Forest Regression

Because of its flexibility, we also explored using random forest regression to find feature importances for the World Bank. The random forest regressor implemented in the Python *sklearn* library contains a feature importance attribute, where each feature is ranked on how much it influences the final outcome based on the training data. The ranks are normalized so that all of them add up to 1.

Unlike the previous methods, all of the data would be subtracted by their country mean as the first preprocessing step. After that, the rest of the steps are done the same way: train-test 80-20 split done 50 times with the results being an average of all the runs. And instead of using any feature selection algorithm, the World Bank gave us a list of 12 outcomes, where each outcome had its own list of features for us to investigate with.

Obstacles and Implementation Issues

Clustering

One challenge with clustering was in standardizing data. If the scale of variables is too different, then the Euclidean distance-based k-means clustering algorithm will overrepresent the larger variables in the cluster. So, determining the best way to balance the scaling was challenging as this had a significant impact on the insights of our clusters' output. While there were features that we believed would be more impactful to the overall school performance, we determined that keeping all variable scales (thus balancing variable weights) was the best method. Changing variable weights could taint the data, so we decided equal weighting was the optimal strategy. Another implementation issue encountered was the order of operations for applying the clustering operations. For a few of the steps, we used a different order initially, like applying the NA replacement using the overall data instead of the overall data and the country data separately. While this did not significantly impact our insights, we made a few similar mistakes throughout the development process. So, we had to spend extra time checking our work to ensure we did not make further mistakes.

Regression

Regression methods have a few issues. The main one is that most regression methods were not good fits for the data. In fact, a lot of them produced a negative R^2 score. Moreover, even though linear and random forest regression performed well, they only performed well as predictors for a few outcomes. In addition, while the R^2 score may indicate that a model performs well, the MSE values are high.

The second main issue is that the results change for the feature selection algorithm when the training set changes. There is not a good way to determine how many runs would be necessary to prevent randomness from being an issue. We chose to use 50 runs because it would minimize the effects of randomness while having decently fast performance.

Testing and Validation

Clustering Validation

For clustering, we did not use labeled data to define what our clusters should be (there were no incorrect clusters), so we were more interested in how differentiated the clusters were and their relative sizes. So, we used industry-standard techniques to estimate how effective our clusters are. Specifically, we used Silhouette, Calinski-Harabasz, and David-Bouldin scores to evaluate our clusters. Silhouette Score measures the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). This determines how similar a point is to its cluster relative to other clusters. Calinski-Harabasz score is a measure of the ratio between the within-cluster dispersion and the between-cluster dispersion. The David-Bouldin score measures the inter-cluster distance and the between cluster distance and helps to ensure that clusters are more spread out and the data is tighter.

We first used the scores described above to validate that our cluster models correctly modeled our data. These were primarily used to ensure that the cluster had sufficient separation from other clusters and actually captured differences in the data. If clusters are too close together, they do not have strong support. Once the scores were sufficiently high, the percent of the data in each cluster was visualized using a pie chart. Ideally, the clusters should be reasonably balanced to ensure that the data is distributed well. For clusters with data from all countries, we also evaluated the country level percentage for each cluster. We found that for clusters for the total data, the clusters were essentially capturing the country-level variation. So, to ensure that our methods captured differences due to administration/ resources and not just the differences from the countries, we ensured that the country-level cluster distributions were roughly the same. We used elbow graphs of the Calinski-Harabasz scores to ensure that our k value was at the optimal value. Overall, our evaluation criteria were used to ensure that the data we were running clustering on was able to be successfully clustered and that any insights we gained from the data had sufficiently strong backing. Furthermore, to ensure that we captured global insights instead of country-specific trends, we graphed the distribution of each country's data across the cluster.

Regression Validation

For regression methods in general, we used the metrics that Babar, the previous student who worked on this project, used. Namely, the mean squared error (MSE) and R^2 scores. The MSE value tells us the average of the squared errors. The R^2 value tells us how much of the data can

be explained by the regression model. R^2 scores can be negative, which means that the model's predictions are worse than just using the mean of the true values each time. Before each model evaluation, the data is split into training and testing sets, with an 80-20 split, respectively. Because the training/testing split is randomized each time, several runs of each regression method are done to minimize the randomness. At the end, the average of the MSE and R^2 scores are used to evaluate the models.

In addition, we also used the scores to validate the features chosen by the stepwise feature selection algorithm. The chosen features and the outcome they predict are fitted into a linear regression model, which is evaluated by the two scores. By doing this, we are able to validate that the chosen features are good predictors of the outcome.

Mappings of Features to Requirements

As our requirements are all based on the paper with our results, our requirements do not map directly to our features. The features of our project are to find interesting insights, which, based on our evaluation metrics, we have achieved. Accordingly, we will integrate these insights into our paper to satisfy the requirements.

Insights

Insights from Clustering:

Using clustering, we found several interesting relationships. Generally, the clusters can separate the data into a high-performing group, a low-performing group, and a 3rd group which was either a medium performing group or another high-performing group. The relative performance of these groups helped us understand what combinations of variables were interrelated and what impact these relations had. The Spider Graphs proved especially useful in understanding the relationships as they represented the clusters in an easy-to-understand manner. While the tables of ranked means were also used occasionally, finding clear insights from them was much more difficult due to the volume of data present.

Figure 3 is a spider graph of all the data with all variables in all countries. The clusters were able to cluster group 0 as schools with low principal and administrative scores. These clusters had low scores on both school goals and principal training metrics. Interestingly, this cluster also had lower infrastructure scores, including drinkable water, pens, access to the internet, and electrification. This is believed to be a correct insight as incompetent principals and administrators likely do not have the skills necessary to ensure their schools get the required resources. Another fascinating insight from this cluster is that cluster 0 also has low infrastructure monitoring scores. This means that no one checked to ensure that schools had access to the needed resources. So, administrators may be less likely to be aware of the lack of resources in cluster 0.

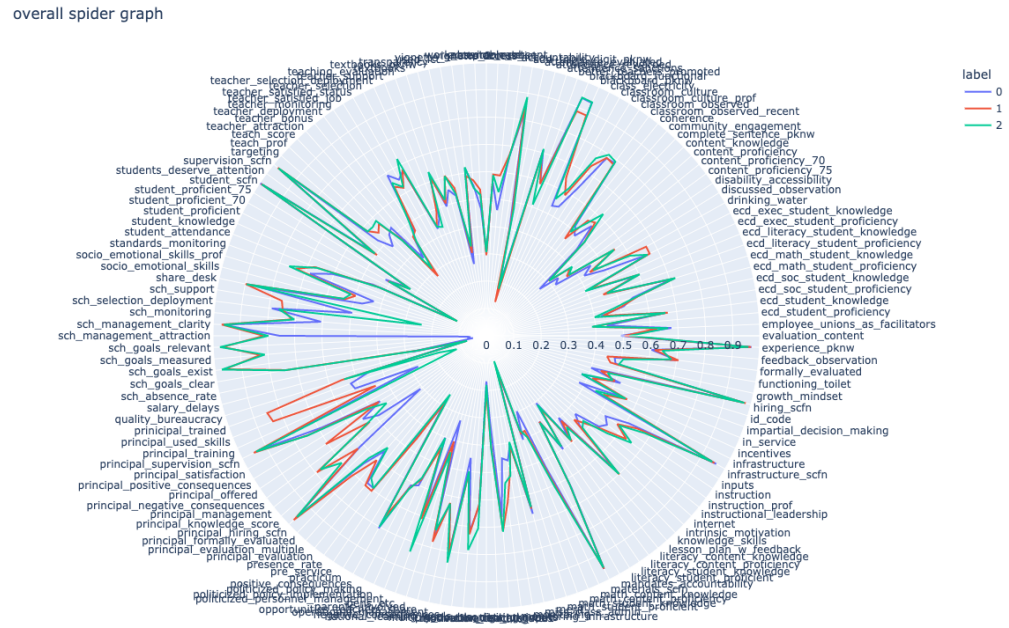

Figure 3: Spider Graph of clusters for all variables in all schools

Another interesting insight was found in Figure 4. This was created by clustering schools from all countries using the student knowledge scores *student_proficient* and *student_knowledge* and the teacher knowledge scores *content_proficiency* and *content_knowledge*. We expected that effective teachers would result in higher student scores. We see in cluster 1 that ineffective teachers are associated with underperforming students. But, we found that cluster 0, which featured high student scores, had low teacher scores. In contrast, cluster 2 with high-performing teachers had low student knowledge scores. This is an unexpected insight as these variables would be expected to depend on each other. So, further analysis is needed.

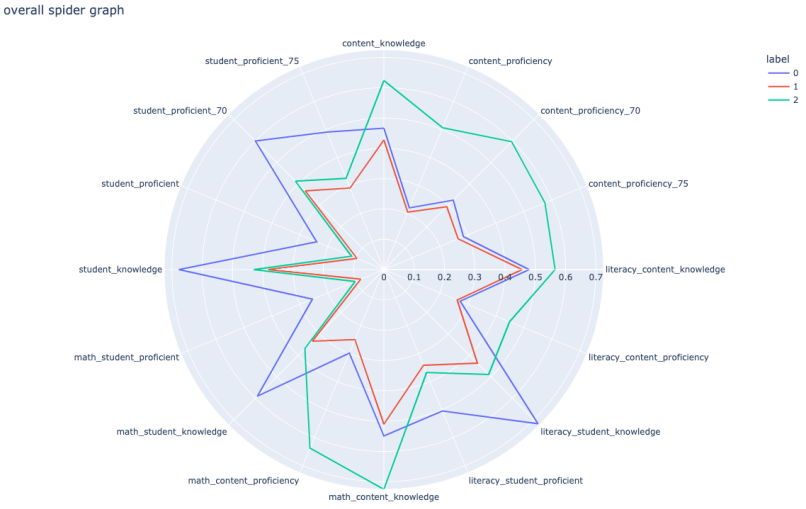

Figure 4: Spider Graph of clusters for student and teacher knowledge in all schools

When comparing this output to the graph including Peruvian schools shown in Figure 5, we see the expected differentiation of clusters. Cluster 0 contains schools where both students and teachers perform poorly, while cluster 2 contains schools where students and teachers perform in the middle of most categories. Finally, cluster 1 contained schools where both groups were performing well. This cluster is much more logical as it aligns with the intuitive understanding of the student-teacher relationship. So, further analysis should be performed on the previous graph to determine the underlying causes within the schools of the unexpected groupings.

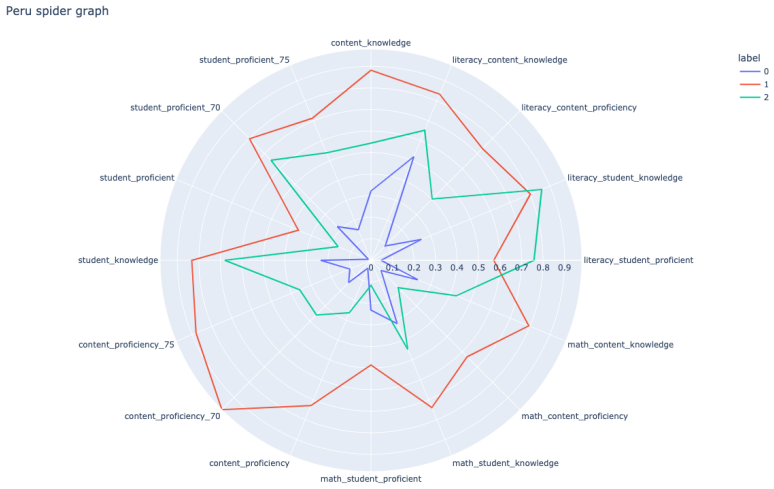

Figure 5: Spider Graph of clusters for student and teacher knowledge in Peruvian schools

We also evaluated clusters of school access to resources in Figure 6. We can see that the schools in clusters 1 and 0 had low access to resources, while schools in clusters 2 had better access to resources. This graph shows that schools with low access to resources generally have low access across all metrics.

Figure 6: Spider Graph of clusters for school resources in all countries

We also evaluated teacher administrative factors using clusters in Figure 7. These included how fairly promotions for teachers were given and how teacher continuing education is run. Again, we see high, medium, and low cluster groups (clusters 0, 1, and 2, respectively). These clusterings show that schools with good administration tend to have high scores in these categories across the board, while low-scoring schools have overall poor administration.

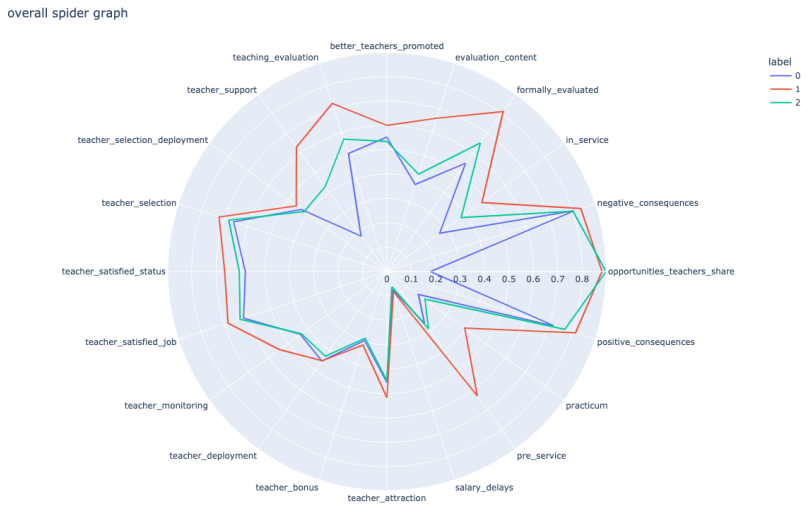

Figure 7: Spider Graph of clusters with teacher administration in all countries

We also found that clustering could successfully differentiate between the countries when running on data without removing country-level means. Ethiopia was in one cluster, Peru and Jordan were in another, while Rwanda was in a final cluster. While this was not the ideal behavior as we were aware of that similarity in the data, this was still an interesting insight that

our method could find in the data. This also demonstrated early on in the project that our methods were able to find insights without knowledge of the underlying data characteristics.

We should also note that many more clusters were generated, and there are likely more insights that have not been found yet. So, additional insights will likely be found after future work.

Insights from Regression:

Using a brute force method of testing each regression method, we found two methods that worked well for the data: linear regression and random forest regression.

Using stepwise feature selection, an automatic feature selection algorithm based on p-values, we found possible features that the World Bank could explore. The chosen features were validated by training a linear regression model with them and having the model's output be evaluated by the mean squared error (MSE) and R^2 metrics. Some of the chosen features for predicting student knowledge and literary student knowledge were: miss_class_admin, parents_involved, vignette_2, students_deserve_attention, and much more. The chosen features were then sent to the World Bank representative for a closer look at how each feature may be an influence on the outcome.

Table 2 below contains results for predicting the 4th-grade literary knowledge score. As mentioned before in the implementation section, we use the stepwise feature selection on an 80-20 train-test split. As the train-test splits are random each time, we perform up to 50 runs to minimize the randomness.

The number after each feature is the number of times they were chosen out of 50 runs, sorted by the number of times chosen from highest to lowest. At the bottom are the average MSE and R^2 values of the 50 runs.

miss_class_admin : 50 parents_involved : 50 vignette_2 : 50 students_deserve_attention : 50 used_ict : 49 practicum : 49 textbooks : 44 principal_positive_consequences : 40 class_electricity : 39 teacher_deployment : 37 classroom_culture_prof : 36 teacher_selection : 30 better_teachers_promoted : 28 sch_monitoring : 24 community_engagement : 23 presence_rate : 20 socio_emotional_skills : 20 teacher_selection_deployment : 18 monitoring_inputs : 15 national_learning_goals : 14

absence_rate : 13
GDP : 12
in_service : 8
sch_selection_deployment : 7
quality_bureaucracy : 6
sch_absence_rate : 6
teacher_bonus : 5
discussed_observation : 5
drinking_water : 5
principal_negative_consequences : 4
knowledge_skills : 4
socio_emotional_skills_prof : 4
teacher_monitoring : 4
attendance_rewarded : 4
sch_management_clarity : 3
pens_etc : 3
inputs : 2
instructional_leadership : 2
monitoring_infrastructure : 2
evaluation_content : 2
principal_training : 2
pre_service : 2
teacher_satisfied_status : 2
principal_management : 2
teaching_evaluation : 1
id_code : 1
employee_unions_as_facilitators : 1
merit : 1
infrastructure_scfn : 1
multiply_double_digit_pknw : 1
content_proficiency_75 : 1
politicized_policy_implementation : 1
complete_sentence_pknw : 1
teacher_attraction : 1
textbooks_pknw : 1

Average MSE: 274.6450595560181
Average R-squared: 0.6329421416693779

Table 2: Results of 80-20 train-test split-run 50 times for predicting 4th-grade literary knowledge

As seen in the table, some features are more prominent than others. However, where the cutoff point should be when choosing the features to examine or focus on economically is still unknown. In addition, 50 runs may not be enough to determine which features are more useful.

One of the next steps was to perform regression by splitting the data by countries of a similar economic background: Jordan and Peru vs. Rwanda and Ethiopia. Unfortunately, most regression methods performed horribly here as there may not have been enough data, which became a dead-end for us. However, we did find that the random forest regression method was a flexible model and had a built-in feature importance attribute.

The World Bank gave us a list of 12 outcomes and hand-picked features for each of the outcomes. Using random forest regression, we calculated the feature importances. From this, we found out that GDP and the distance from a school to its administrative office may be important features in determining outcomes.

Results are shown in Table 3 below. For each of the 12 outcomes the World Bank gave us, we found the five most important features of each outcome. From that, we also looked at which five features were the most important across all outcomes, which is what the table below shows. Each feature has a fraction to its right, showing how many times it was in the top 5 most important features out of 12 outcomes. Some features are also outcomes themselves, and since an outcome should not be its own feature, they are counted with 11 columns instead.

Top 5 important features:
Distance to office (8-9/12)
Student Knowledge (6-7/11)
ECD Student Knowledge (7/11)
GDP (5-6/12)
Infrastructure (5/11)

Table 3: Top 5 features that influenced the most outcomes.

Relevance for AI

One of the core uses of AI is making predictions and understanding high-dimensional data. The information from the World Bank included a significant number of questions asked of each school. While these questions were highly valuable, the relation between questions is less clear, especially for humans. So, more advanced methods were necessary to understand the data and understand how the data can be applied to decision-making at the World Bank. So, we used machine learning techniques, including clustering and regression models, to understand the relationships between the variables and how they might affect school performance. Accordingly, our project is directly applicable to AI because our project involved applying machine learning techniques.

Lessons Learned

One of the biggest lessons learned is how complex data science is. We made several minor errors, such as not applying de-meaning operations as we thought we were or including variables that should not be present. While these were minor errors, they significantly impacted the insights we took from the data. Given the number of operations required in a successful data science project, it is easy to overlook a missing or misapplied step. So, we learned that significant care is required to ensure that the code is doing what we expect it to.

Another lesson learned is how much data is necessary to predict outcomes. Unfortunately, we did not have access to data from Ethiopia during most of our project, as it had not been processed yet. Once the data from Ethiopia became available, many of our insights and feature importances changed, which impacted the insights we saw in our data. This showed us the importance of increasing the amount of data, as a smaller subset will not be fully representative of the overall trends in the data.

We also learned that real-world data contains missing areas that need to be handled carefully. While the World Bank provided data on a significant number of schools, many of these schools

were missing variables from the data. Initially, we just dropped rows with missing data. But, this resulted in very few rows being available for analysis. So, we had to spend time thinking about handling these missing variables so we could use all the rows in our data. So, we had to learn how to handle the missing data that is prevalent in the real world.

Future Work

As the World Bank continues the project, they will collect more data from more schools in different countries. The additional data will likely have further insights that can be extracted as the school and administrative factors will be different in these different countries. So, additional analysis will be required when this new data is collected. Furthermore, once more data is collected, feature importance will likely change. As additional countries that will have data collected will use different organizational systems, these organizational changes will have an impact on how variables are interrelated. So, further work will be required to understand these changes and the impact that feature importance changes will have.

There is also interest in creating predictive models. Currently, the World Bank is working to understand all the data. Once this data is fully understood, a model that can relate principal factors or school resources to student performance would be highly valuable. It would help the World Bank identify schools that need help in certain areas to improve their student outcomes. The World Bank's final goal is to identify schools that require additional resources to ensure students are successfully educated. So, a model that can predict the performance of a school's students given administrative factors would help the World Bank achieve its goals.

Finally, additional analysis should be performed on the World Bank data to look for more insights into the data. While we have accomplished a significant analysis of the data using a few different methods, further insights may be present in the data that our methods have been unable to surface. So, more work could be done on the data to apply more machine learning methods to find insights that our methods have been unable to discover. We also want to spend additional time evaluating our clusters to find further insights into the analysis we have already performed.

Conclusion

We show that applying machine learning techniques to high-dimensional education data is able to find interesting insights into the data. In particular, clustering and random forest analysis yielded valid and applicable insights into the education dataset provided by the World Bank. Furthermore, we showed that ineffective principals resulted in schools with lower access to vital resources and that schools with knowledgeable teachers were not necessarily schools with highly knowledgeable students. Additionally, we found that teachers who had to miss class due to administrative issues, the involvement of parents in ensuring that classes have access to resources, and the distance from the school to the district office were highly impactful features.

References

[1] Ayan, B. “Personal Communication”, email, October 2021

[2] Bierman, K.L., Coie, J., Dodge, K., Greenberg, M., Lochman, J., McMohan, R., Pinderhughes, E. Conduct Problems Prevention Research Group, 2013. School outcomes of aggressive disruptive children: Prediction from kindergarten risk factors and impact of the Fast Track prevention program. Aggressive behavior, 39(2), pp.114-130. Accessed October 2021

[3] Buyse, E., Verschueren, K., Verachtert, P. and Damme, J.V., 2009. Predicting school adjustment in early elementary school: Impact of teacher-child relationship quality and relational classroom climate. The Elementary School Journal, 110(2), pp.119-141. Accessed October 2021

[4] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007. Accessed October 2021

[5] McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Reds), Proceedings of the 9th Python in Science Conference (bll 56–61). doi:10.25080/Majora-92bf1922-00a. Accessed December 2021

[6] Park, S., Stone, S.I. and Holloway, S.D., 2017. School-based parental involvement as a predictor of achievement and school learning environment: An elementary school-level analysis. Children and Youth Services Review, 82, pp.195-206. Accessed October 2021

[7] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011. Accessed December 2021

[8] Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. doi:10.21105/joss.03021. Accessed December 2021