Colombian Education Case Study Report

Author:	Luis González
Email:	luisf.gonzalezv@yahoo.com
Date:	March 18, 2023

Table of Contents

Introduction

The following report describes a case study I (the author) completed on Colombian Education, specifically on the performance of students on the state-mandated test at the end of their secondary education (Pruebas Saber 11). The test is administered by *ICFES* (Colombian Institute for the Promotion of Higher Education), and it is meant to be a reliable indicator of how students are faring in the country´s educational system. ICFES also asks students presenting the test to answer a series of questions regarding their socio-economic background for study purposes, which is the data employed by this study.

Figure 1. ICFES Logo

Colombia, as other countries in South America, struggles with economic inequality and lack of opportunities for less-wealthy families, and this study looks to answer the following questions:

- Which of the variables within the data collected by ICFES are most significant to explain student scores on the test?

- How well can student scores wit be predicted using the data collected by ICFES and machine learning models?

Having understood the context of this exercise, we can move on to the data analysis methodology.

Note: This case study was originally developed in Spanish using python and its dependent libraries, and this report is meant to be a short summary of the data analysis process undertaken, as well as its main findings. A full description of the project is available in an *.ipynb* file, containing the full-detailed data wrangling and analysis process, including a functioning dashboard app using *panel* and *plotly .*

Methodology

1. Extract-Transform-Load (ETL)

The original datasets were collected from a public ICFES repository, located at https://www.icfes.gov.co/data-icfes. The data collected corresponded to the last four published results to date, ranging from the second semester of 2021 to the first semester of 2023 (the test occurs once every semester). After downloading locally the .txt files, they were loaded into a *pandas* dataframe.

The data originally contained over 1 million rows, with each row being one student and their data. The dataset contained originally 82 columns, which were reduced to only the 25 most indicative of the student´s background. Also, a 5% random subset of the data was selected, in order to reduce the computational burden required to process such a lengthy data structure (An SQL RDBMS should be considered for future applications).

Data was also cleaned of nulls by deleting them, which in total left the database with 39660 rows and 25 columns. The most indicative columns selected were:

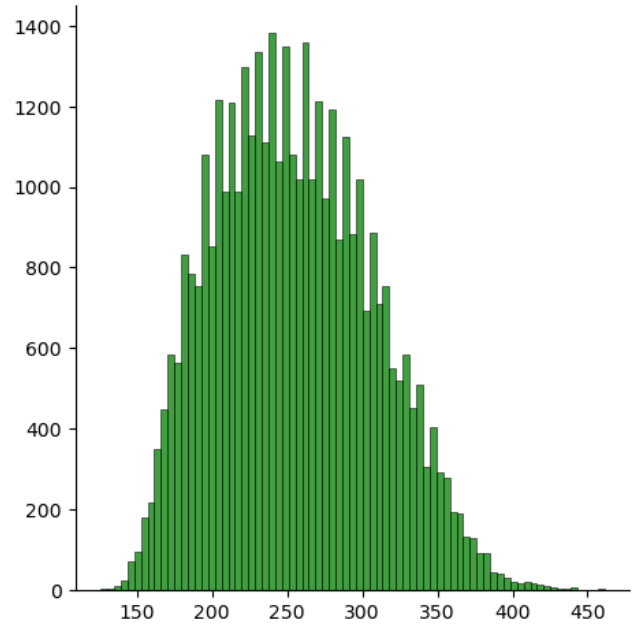
- Family housing level (from 1 to 6, 1 being near poverty and 6 being very wealthy)
- Data relating to the student´s school (bilingual, mixed gender, calendar type, type of school)
- Data relating to the student´s performance on the test (score and percentile by subject and in total)
- Data relating to the student´s use of their personal time (dedication to reading, internet and to work)

2. Exploratory Data Analysis (EDA)

In this phase, several interesting findings were extracted from the data. Below are some of the most interesting:

- Total score

The total score of the students on the test goes from 0 to 500, and resembles a normal distribution with a slight skew to the left. Lower values are more common than higher values, and the mean is around half the total possible score.

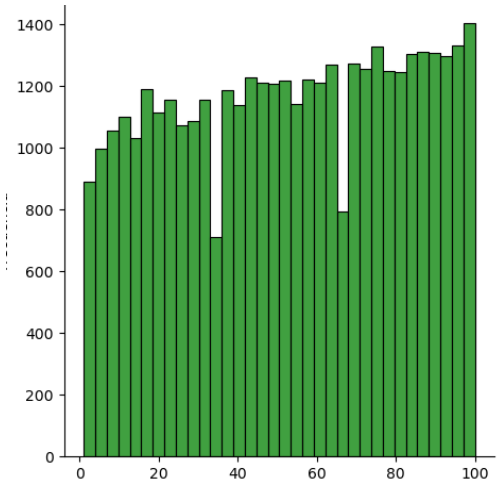


Mean	254.39
St. Dev	52.13
Min	126
Max	461
25%	214
50%	251
75%	291

Figure 2. Distribution of Total Score

- Overall Percentile

This metric ranks students by their score, and it reflects what percentage of students their score was above from (ICFES adds 1 more point to this metric, so it goes from 1 to 100). The graph resembles a uniform distribution, with mean and standard deviation very close to its theoretical values.



Mean	52.86
St. Dev	28.67
Min	1
Max	100
25%	28.0
50%	54.0
75%	78.0

Figure 3. Distribution of Overall Percentile

As for the predictive variables, in general, some correlations were identified, but it is best to discuss these in the following section, concerning machine learning models.

3. Machine Learning Model Selection and Development

In developing this section, the preference was to settle for a highly-interpretable model, even at the expense of predictive power, as a goal of the study was to identify the most important variables in determining student performance on a test. Prior to that, variable transformation was realized, including encoding, in order to use python´s *sklearn* libraries for machine learning, which mostly accept only numerical variables. Additionally, feature selection was considered by filter, using an F-Statistic lower bound of 50. All features passed this threshold. Filter methods have the advantage of being computationally more efficient than wrapping methods, but they have the disadvantage of being somewhat subjective (50 is a pretty arbitrary threshold).

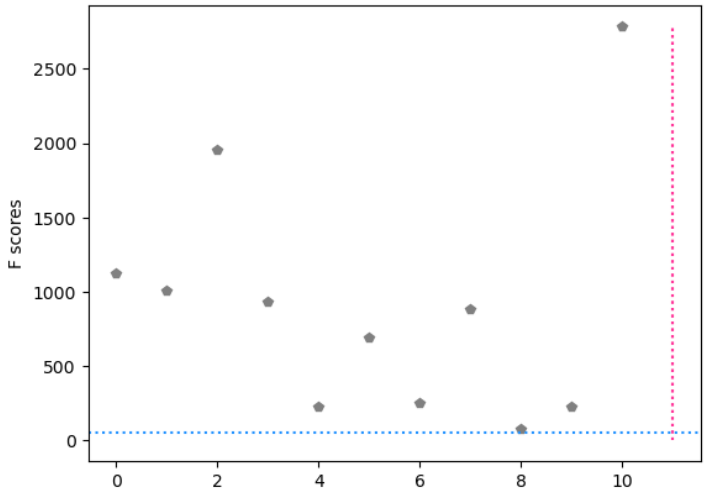


Figure 4. F-score results by feature number, feature selection

After performing the above- mentioned tests and transformations, the following are the resulting predictive variables for the model:

Variable	Description
ESTRATO (Housing Level)	The level of housing in which the student resides, ranging from 1 to 6. 1 means near poverty, and 6 means very wealthy.
LECTURA_DIARIA (Daily Reading Time)	The approximated time dedicated by the student every day to reading.
INTERNET_DIARIO (Daily Internet Time)	The approximated time dedicated by the student every day to the internet.
TRABAJO_DIARIO (Daily Working Time)	The approximated time dedicated by the student every day to working.
PRIMER_SEMESTRE (First Semester)	A dummy variable, indicating 1 if the test was presented on the first semester of the year, and 0 if otherwise.
COLE_MIXTO (Mixed School)	A dummy variable, indicating 1 if the school is mixed (i.e., allowing both male and female students), and 0 if otherwise.
GENERO_M (Masculine gender)	A dummy variable, indicating 1 if the student is male and 0 if they are female (no other gender identities are surveyed by ICFES).
ETNIA_Si (Ethnicity)	A dummy variable, indicating 1 if the student belongs to an ethnicity, and 0 if otherwise.
BILINGUE_S (Bilingual School)	A dummy variable, indicating 1 if the school is bilingual, and 0 if otherwise.
COLE_ACADEMICO (Academic School)	A dummy variable, indicating 1 if the school is of an academic character, and 0 if otherwise (other options include technical, or technical/academical).
JORNADA_COMPLETA (Complete Day)	A dummy variable, indicating 1 if the school operates during the whole day (morning to afternoon), and 0 if otherwise (morning, at night, on weekends, etc.).

Table 1. Model Feature description

Linear regression was considered on a first instance for model selection, looking to predict for overall percentile. However, this wasn´t a good idea, as most of the predictive variables were categorical and one-hot encoding was needed for several variables, making mlr a somewhat forced model. Checking for the assumptions regarding linear regression, no significant collinearity was found for the predictive variables.

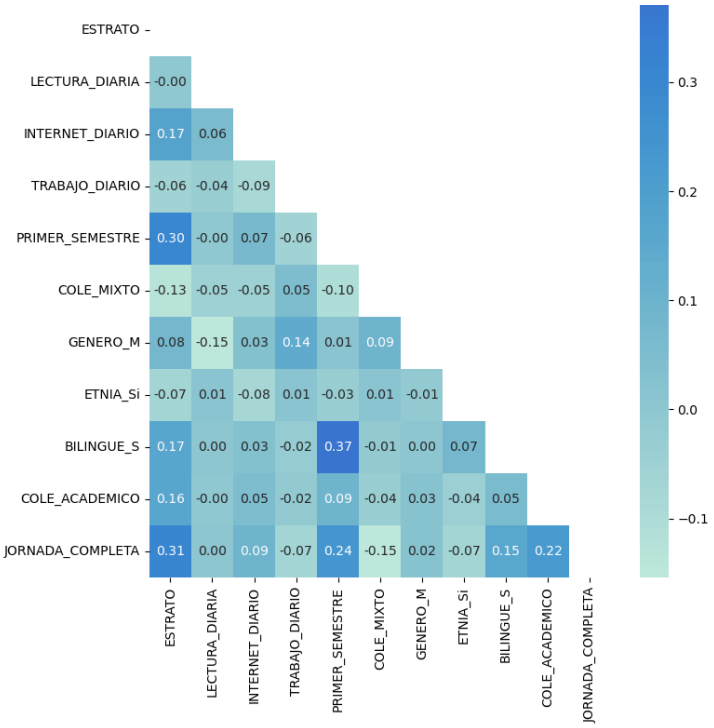


Figure 5. Pairwise correlation matrix, predictive variables

However, the model failed the White Test for heteroscedasticity, meaning the variance of the error wasn´t constant along its range, and the assumptions for linear regression were violated. To solve for this, a Lasso Model was considered, which penalizes high error terms, even reducing them to 0, and possibly fixing the heteroscedasticity problem. However, the results obtained on the test data were less than acceptable.

```
Penalization parameter: 0.0026126752255633263
R- squared: 0.1721372489429388
```

The penalized mlr model managed to explain only 17% of the total variability in overall percentile, which suggest regression wasn´t the right approach to the problem.

Going forward, a new approach to the problem was considered, involving classification, as opposed to regression. The goal was to identify students performing above 75% percent or more of their classmates, using the same predictive variables.

Response Variable (Y)	
1	The student placed in the 75th percentile or above
0	The student placed below the 75th percentile

Table 2. Response variable description

Decision trees were selected, as they are robust to problems in the data, like multicollinearity. However, decision trees on their own are prone to both bias and overfitting of the data, so **Random Forests** were chosen as a solution, an elegant alternative that averages out several trees in which the predictive variable is chosen at random for each split. Below are the results obtained:

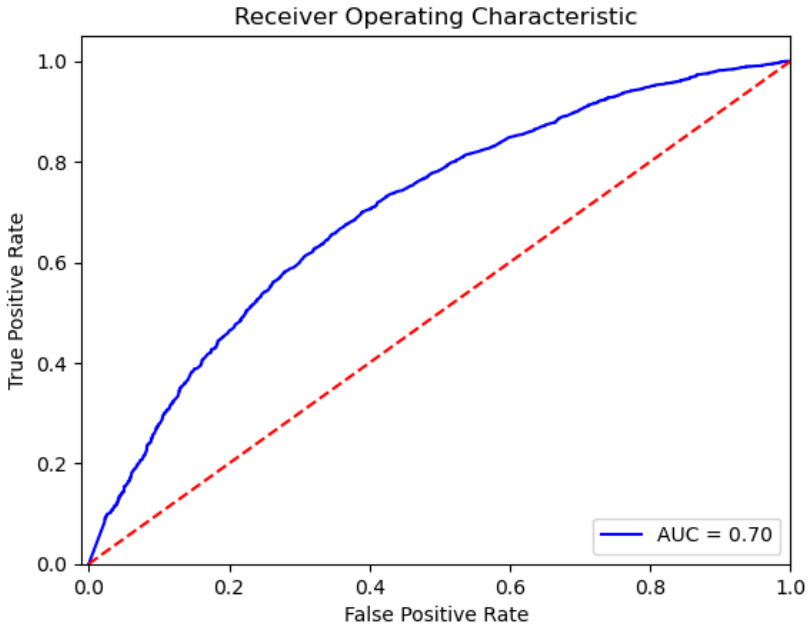


Figure 6. AUC for Random Forest Model

```
Accuracy: 0.7405
Precision: 0.5705
Recall: 0.2665
```

The results of the model can be interpreted as follows:

- 74% of the time, the model classifies correctly between students on the upper quartile of the test scores.
- 57% of the predictions made for the positive class (Y=1) were correct.
- 26% of the actual positive class instances were correctly classified.
- Overall, the model scores a 7/10 in performance.

Additionally, the most relevant predictive variables can be found by computing the Mean Decrease in Impurity (MDI) from the model:

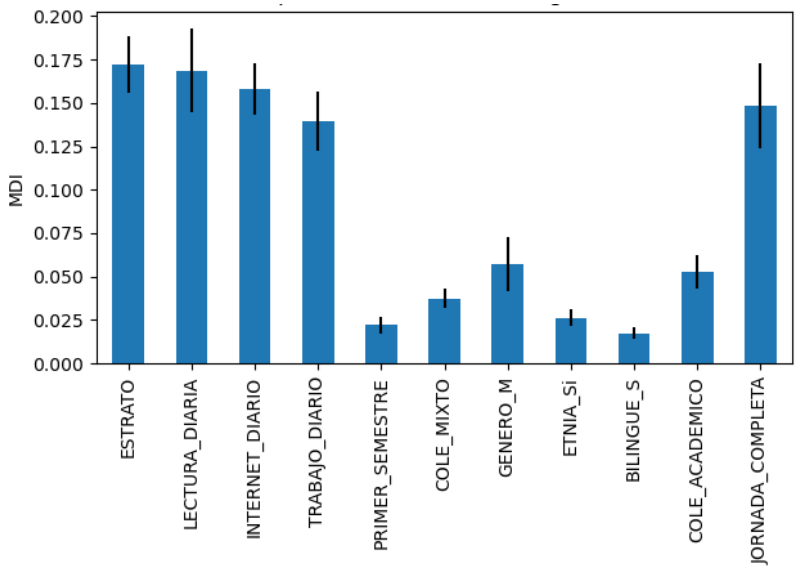


Figure 7. Model Feature Importance by MDI

The results show that the most influential variables within the dataset to determine whether a student has a good score (upper quartile) on the test are:

- Their housing level (the higher the better)
- Their daily dedication to reading (the more, the better)
- Their daily dedication to the internet (the more, the better)
- Their school being full-time (it is preferable to other options
- Their dedication to work (it is preferable that you not work)

The other variables included in the model showed little relative importance. Also, the direction of the relation to the target variable was obtained during EDA.

It is important to highlight that, although the development of models was not exhaustive, it was possible to correctly identify the variables that have the most effect on a student performing well on the tests. Also, it is reasonable that very high performances are not obtained for the models, since the predictive variables are not sufficient to capture the great complexity of variables that ultimately determine student performance. For example, having information on IQ tests for each student would probably be a useful metric in predictive models, but in reality it is highly unfeasible to achieve on such a large scale.

4. Interactive Dashboard App

As part of the process, although it isn´t necessary for the data analysis methodology, a dashboard app was developed using panel and plotly. The objective was to present to the user the possibility of viewing by themselves the behavior of the data, going beyond what was discovered during the exploratory and modelling phases.

The tool allows the user to:

- Choose a target variable (overall or subject-specific score or percentile)
- Choose a predictive variable about the student's socioeconomic status (school characteristics, housing level, etc.)

With these input values, the user will be able to view:

- A bar graph for the distribution of values according to the grade and subject chosen.
- A pie chart with the distribution of students by level of the variable.
- A histogram with the distribution of the target variable, its mean, standard deviation and other descriptive statistics.

Below are some screenshots of the output (source code is available on the jupyter notebook):

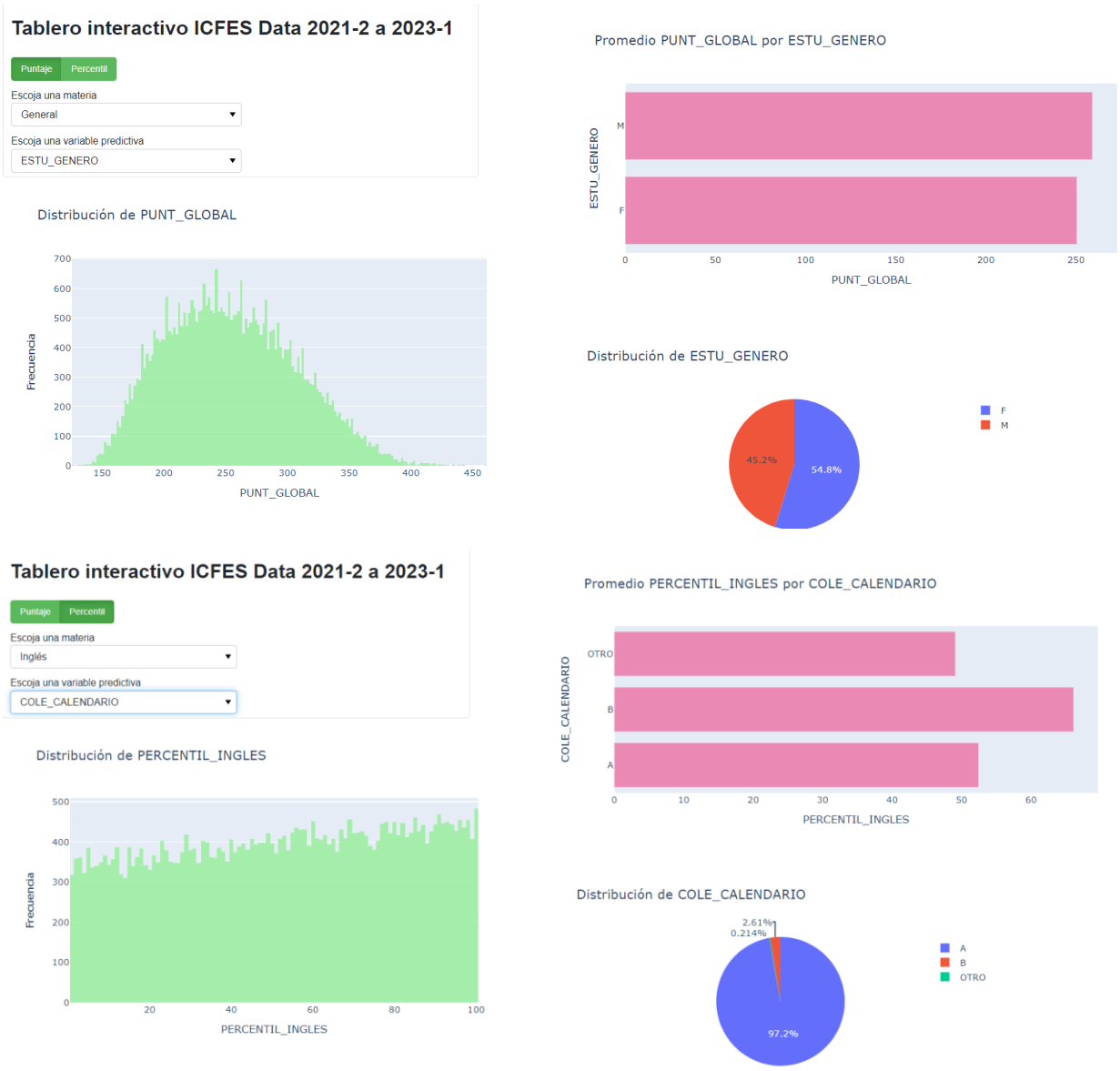


Figure 8. Dashboard app using panel and plotly

Unfortunately, the dataset used in the project can´t be shared publicly by me, as it is property of ICFES. However, I would be delighted to share it privately upon request by email, to any interested reader.

Conclusions

In conclusion, the present study successfully identified influencing variables to determine a student's academic performance in the national end-of-secondary tests in Colombia. Of course, the socioeconomic status of the student's family plays an important role, and it cannot be denied that students from families with more economic means have more opportunities for a better education than other students.

However, the study also shows that not all of the variability of the problem can be explained by conditions external to the student, such as economic status. It was found that, among other things, those students who read for longer tended to have better test scores. The same with the student's school, in which regardless of wealth, those belonging to full-day and academic schools also proved to be better prepared for the tests.

The findings found about schools and good time use practices can be a good starting point for parents or government agents who are interested in improving the state of education for their children or young people in Colombia or elsewhere in the world, and helping better education lead to the progress of society as a whole.

Regarding the data analysis methodology and process, I admit this scope was severely limited and can be improved in several ways, including studying all available ICFES datasets (spanning over 20 years) and storing the data on an RDBMS. Also, more exhaustive model selection can be employed, in order to validate results obtained previously. However, part of the objectives of this study was to rely mostly on pandas and intuitive models, even if at the expense of a more limited working scope.

Lastly, I thank you for your time and interest in reading this report.