Education during the pandemic: An evaluation of student performance in Uruguay

Statistical Learning - Final Project

Mathias Cardarello Fierro

7/6/2021

Abstract

The analysis of the impact that the COVID-19 pandemic has across the different sectors in the economy should include the evaluation of educational performances during 2020. This study is focus on the multidimensional factors that impacted on the Maths test scores of the sixth-grade students in Uruguay, including those relating to the health emergency. A set of supervised algorithms were performed to approach a better understanding of the determinants of the educational performance. In the second part, an unsupervised analysis was implemented to detect a set of components that describes the main features of the students (PCA) and classifying them into groups, based on their similarity (k-means).

Statement of the problem

Educational evaluations constitute a fundamental tool to monitor the achievements of the educational systems of the countries, as well as for the design of public policies that improve their quality. For this purpose, the National Institute of Educational Evaluation of Uruguay (INEEd) has been implementing the "Aristas" tests since 2017, in which educational achievements are evaluated from a multidimensional perspective. In addition to measuring the performance of students in Reading and Math tests, "Aristas" collects information about their social emotional skills, opinions and attitudes about the coexistence, among other conjunctural factors. In 2020, the evaluations were performed, despite of the particular background because of the pandemic, with the aim of "generating inputs that allow contributing to develop policies to mitigate the possible effects of the COVID-19 pandemic", according to INEEd (2021).

Therefore, the aim of this study is the analysis of the main factors that could explain the Maths scores of sixth-grade students of Uruguay, in the context of the pandemic. Although there is a debate about the opportunity to take standardized assessments in a year in which learning opportunities have been very diminished, the version 2020 of "Aristas" promote a contextualized evaluation of the results.

Data description

Firstly, we give a brief introduction to the educational system of Uruguay in order to understand the population of the analysis. Public education in Uruguay is free and compulsory for a total of eleven years (2 years of pre-primary, 6 years of primary education, and 3 years of middle education). In primary education, children from age 6 to 11 receive their grounding in academic subjects. The type of public schools that can be found at primary level are Full time (TC), Practice and practice enabled (PR-HP), Common urban (UC), "Aprender" program (APRENDER), Rural (RURAL) and Extended time (Tiempo Extendido). Moreover, although public education is dominant in Uruguay, there is an important quantity of private schools (PRIVADO) that are regulated by the National Public Education Administration (ANEP), but they do not receive public funding.

The variable of study is the skill obtained in the Maths test by sixth-grade students of Uruguay, measured through theta_MAT variable, which was original factorized in 5 levels (Level_MAT). For aesthetic reasons, the complete list of the features that were taking into account in the analysis is in Appendix. In addition to demographic and socio emotional characteristics of the students, the analysis includes variables related to the level of performance in the Reading test, the characteristics of the school, family and home, COVID context, coexistence and participation.

Before starting with the analysis, the original datasets obtained from the web site of INEEd were processed, which includes the translation of the labels from the original language (Spanish), the computation of a subset with the most meaningful features for this study and the elimination of records from individuals that did not answer the socio-emotional or context questionnaires, among the most important tasks. Hence, the dataset is composed by 4.722 observations and 194 variables of which all are factors, with the exception of theta_READ that measures the score in the Reading test, originally standardised as a theta-score.

A. Supervised Analysis

A1. Data Visualization and Exploration

We start the explorative analysis with our dependent variable Level_MAT that is a factor describing the level obtained in the Math test by sixth-grade students of Uruguay in 2020. According to the Maths score, there are five groups of students, but the extreme ones, level 1 and level 5, have a very small share in the sample (6.3% and 8.6% respectively). Therefore, to avoid this unbalanced distribution among categories, level 1 was merged with level 2, while level 4 was merged with level 5.

In the following table there are the most important descriptive statistics of Level_MAT across the three levels. Mean, median, min, max and sd were included thanks to consider the variable theta_MAT of the original dataset, the theta-score version of our dependent variable. As we can see, the factor variable has three levels with different means, each of them with a share between 31% and 35% of the population.

```
  Level_MAT count share    mean  median    min    max    sd
1 1          1649  34.9 -1.02   -0.914  -3.55  -0.419 0.469
2 2          1595  33.8  0.0389  0.0510 -0.416  0.494 0.263
3 3          1478  31.3  1.32    1.13    0.517  5.41  0.731
```

Among the demographic variables, the gender of the student has a balanced distribution and, on average, males (49.7%) have a better score level in Maths (0.10) than females (0.04). After performing a Chi-squared independence test, we do not reject the null hypothesis of independence among the variables Gender and Level_MAT at 95% of confidence, so there is not a significant relationship between them.

```
  Gender count share `mean score`
1 F       2375  50.3       0.0417
2 M       2347  49.7       0.102

Test results:
   X-squared statistic:  3.626
   degrees of freedom:  2
   p-value:  0.163

Other information:
   estimated effect size (Cramer's v):  0.028
```

In the following boxplot, where theta_MAT is a function of Gender, the p-value of the ANOVA test is shown, which also do not reject the independence, in this case with the continuous variable score.

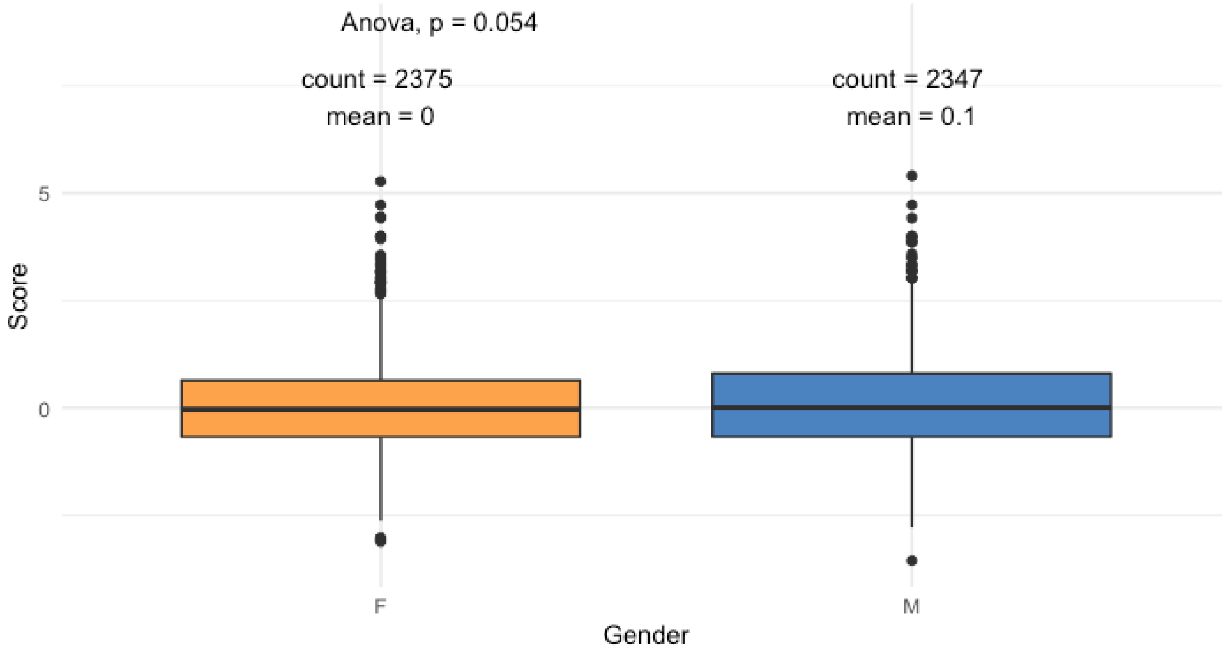

Figure 1: Gender boxplot

Approximately 40% of the students assist to schools in Montevideo, the capital city of Uruguay, while the other group belong to centres of the interior of the country (Interior). This is not surprising since there is a macrocephalism in Montevideo, a disproportionate population growth of a city with respect to the rest, which concentrates almost half of the population of the country (1.3 million inhabitants in 2011). It can be state that, on average, students from Montevideo have a better score level in Maths (0.12) than those who study in the Interior (0.04).

```
  Region     count share `mean score`
1 Interior    2866  60.7       0.0427
2 Montevideo  1856  39.3       0.117
```

In this case, the hypothesis that variables are independent of one another is rejected at 99% of confidence, thus the relationship between Level_MAT and Region is significative. However, the association between the variables is not very strong, according to Cramer's V statistic whose value goes from 0 (no at all association) to 1 (perfect association). In this case, the estimated effect size is 0.07.

```
Test results:
   X-squared statistic:  22.846
   degrees of freedom:  2
   p-value:  <.001

Other information:
   estimated effect size (Cramer's v):  0.07
```

The variable Age was factorised to avoid the extreme values (10 and 15 years) because there is only one student in each category. Therefore, we get a variable with three categories: less or equal than 11y (31.6%),

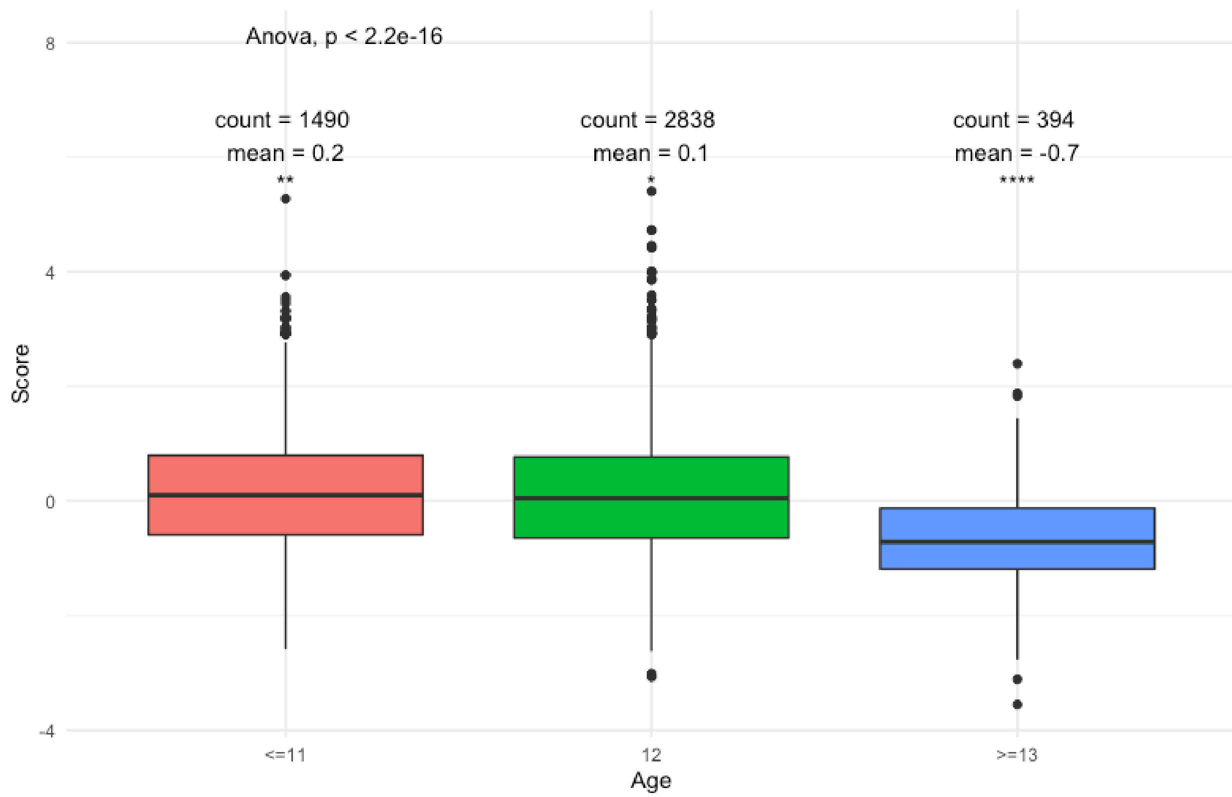

Figure 2: Age boxplot

12y (60.1%), greater or equal than 13y (8.3%). It should be said that usually sixth-grade students are 11 or 12 years old.

The Maths score seems to have a negative relationship with the number of years of students. This was confirmed with the Chi-squared tests, which rejects the null hypothesis of independence at 99% of confidence, with and estimated effect size (Cramer's v) of 0.135.

```
Test results:
   X-squared statistic:  171.051
   degrees of freedom:  4
   p-value:  <.001

Other information:
   estimated effect size (Cramer's v):  0.135
```

Regarding the centre type, in Uruguay there are seven types of schools where the most frequent are Common urban (UC) and "Aprender" (APRENDER), with a share of students of 29.2% and 26.3% respectively. The latter is referred to centres belonging to a special program with the aim of "promoting educational activities that reduce repetition rates, reduce absenteeism and improve learning levels", and they are usually concentrated in the contexts of greatest socio economic vulnerability. It is particularly interesting that students from this schools has the lower score in Maths, with an average of -0.391.

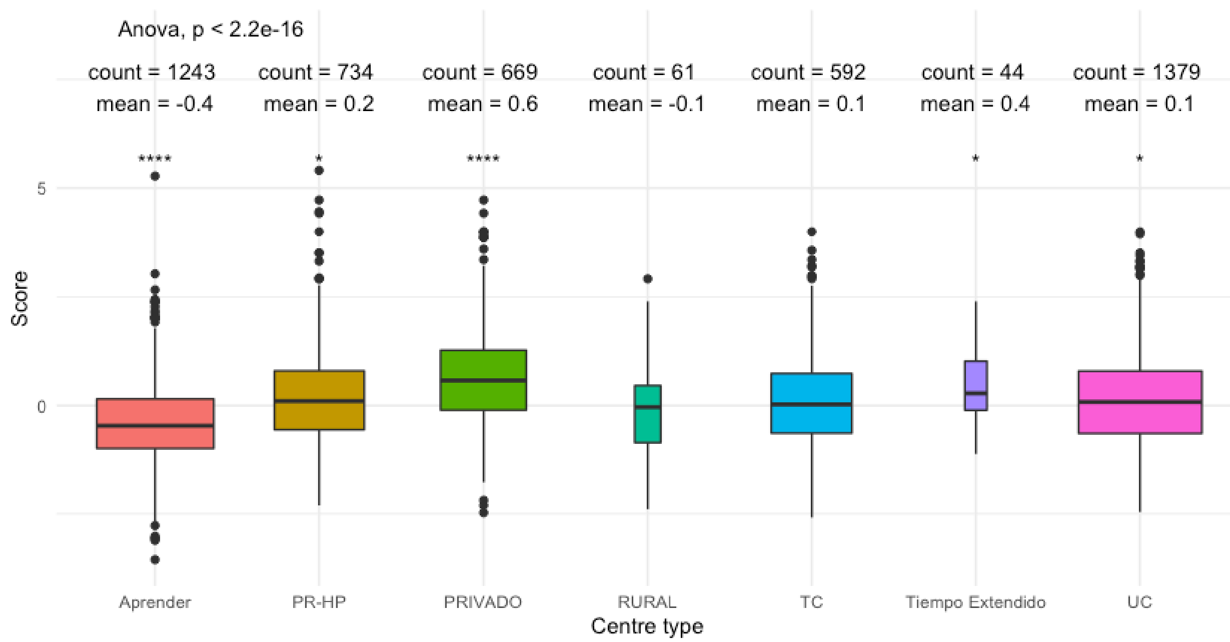

Figure 3: Centre type boxplot

As expected, the hypothesis that the level score in Maths is independent from centre type is rejected at 99% of confidence. Moreover, the association between the variables is stronger in comparison to the age, according to Cramer's V statistic (0.2).

```
Test results:
   X-squared statistic:  398.408
   degrees of freedom:  12
   p-value:  <.001
```

```
Other information:
   estimated effect size (Cramer's v):  0.205
```

The variable Context, that is a measure of the socio economic and cultural context of the centre defined by ANEP, is a factor composed by 11 levels (quintiles 1 to 10 for rural and urban sectors, and private schools). The students from private schools has the highest average score (0.65) while the only two students from Rural Q1 have the lowest one (-1.21).

```
   Context           count   share `mean score`
 1 PRIVADO             669 14.2           0.646
 2 Rural Quintil 1       2  0.0424      -1.21
 3 Rural Quintil 2      15  0.318       -0.660
 4 Rural Quintil 3      15  0.318        0.137
 5 Rural Quintil 4      16  0.339        0.103
 6 Rural Quintil 5      13  0.275        0.368
 7 Urbano Quintil 1    769 16.3         -0.465
 8 Urbano Quintil 2    772 16.3         -0.218
 9 Urbano Quintil 3    585 12.4         -0.103
10 Urbano Quintil 4    750 15.9          0.0734
11 Urbano Quintil 5   1116 23.6          0.397
```

Due to the low frequencies of some categories, a Pearson's Chi-squared test with a simulated p-value was performed to get a more robust approximation. The outcome was the rejection that variables are independent of one another at 99% of confidence.

```
    Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  df$Context and df$Level_MAT
X-squared = 511.63, df = NA, p-value = 0.0004998
```

Then, the same test was performed between Context and Centre_type and the results pointed that there is a contingency between both variables, at 99% of confidence. In fact, the estimated effect size is 0.7, showing a strong association. This results were expected due to the fact that the type of some centres are strongly related to the context, as "Aprender" and privates ("PRIVADO") schools.

```
    Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  df$Context and df$Centre_type
X-squared = 13858, df = NA, p-value = 0.0004998
```

The variable strat has four levels and measures the school size, from sample stratum 1 to 5. The majority of the students belong to the strata 4 (45.2%) and 3 (39.3%), but does not seems that there is a strong relationship with the Maths score. However, according to the Chi-squared test, strat and theta_MAT are not independent of one another at 99% of confidence, with a small effect size (0.07).

```
Test results:
   X-squared statistic:  44.21
   degrees of freedom:  6
   p-value:  <.001

Other information:
   estimated effect size (Cramer's v):  0.068
```

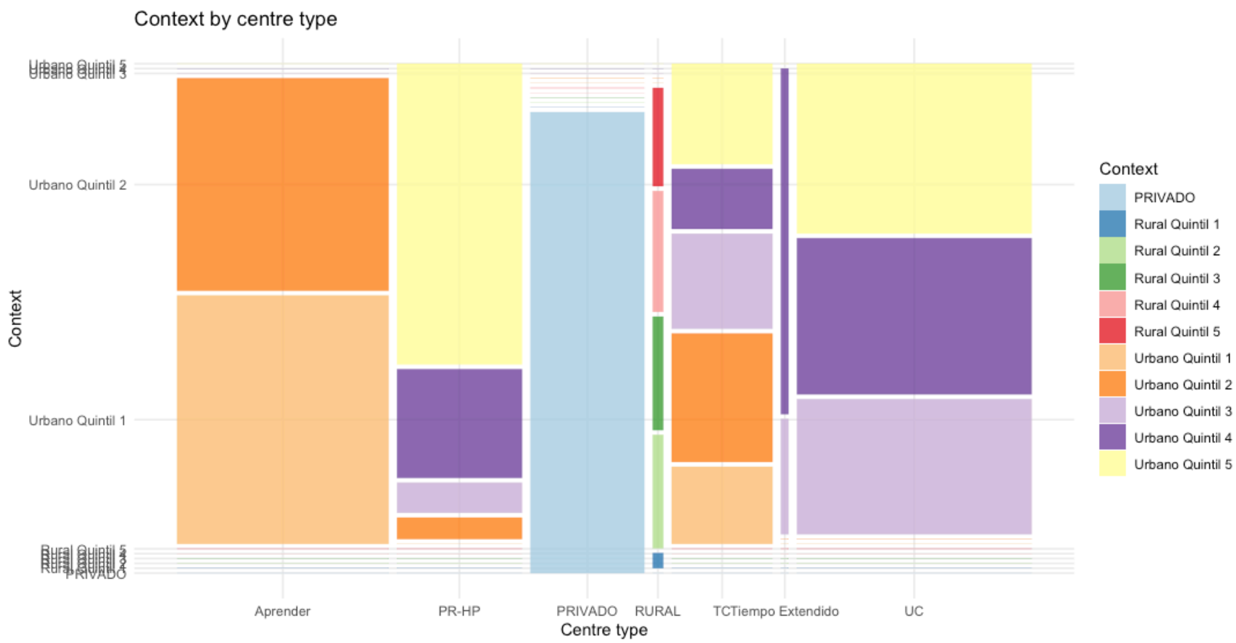

Figure 4: Centre type and Context mosaic

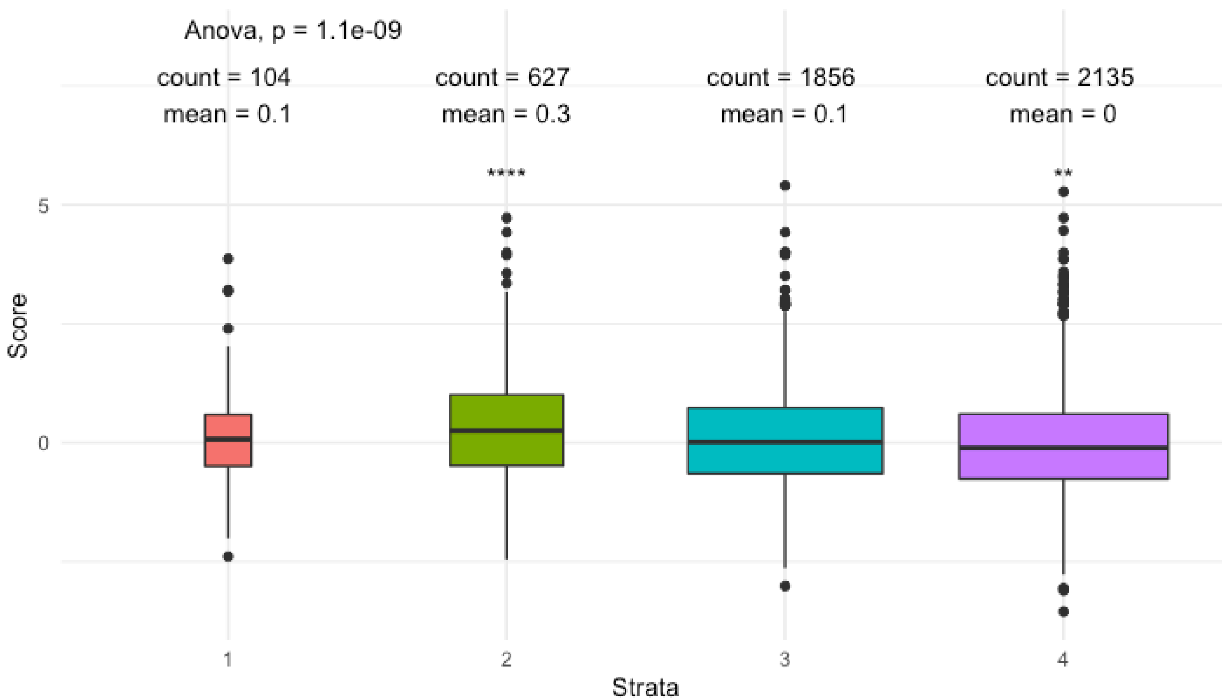

Figure 5: Strata boxplot

The unique continuous variable in the dataset is related to the score of the reading tests (theta_READ), which a priori can be a proxy for the academic level of the students, but not necessarily a good performance in Reading should be related to a good score in Maths. This is what was checked with the ANOVA, which is used to determine whether there are any statistically significant differences between the means of the groups. The p-value of the test, shown in the boxplot, is very low ($p < 2.2e\text{-}16$) so the null hypothesis is rejected at 99% of confidence. Therefore, there are at least two group means that are statistically significantly different from each other. As a robustness check, a Chi-squared test between the factor version of the Reading score (Level_READ) and Level_MAT was performed and we get the same conclusion, with an estimated effect size of 0.438.

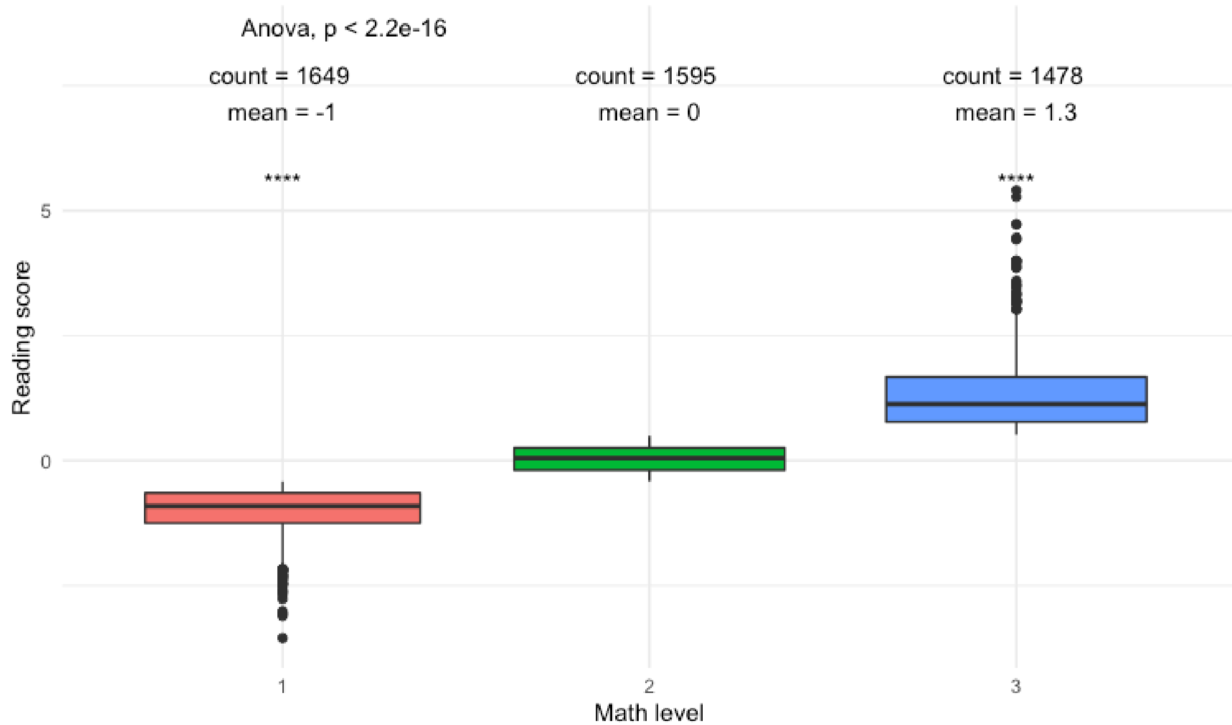

Figure 6: Reading score boxplot

```
Test results:
   X-squared statistic:  1809.914
   degrees of freedom:  10
   p-value:  <.001

Other information:
   estimated effect size (Cramer's v):  0.438
```

Moreover, a scatter plot to analyse the relationship of both theta scores was done, grouped by Gender. In the following figure it can be seen that there is a kind of linear correlation between theta_READ and theta_MAT, where females have a higher score in Reading in relation to the males students with the same score in Maths. The Pearson's product-moment correlation was performed, pointing out that the variables are significantly correlated with a coefficient of 63%. The score distributions for Maths and Reading were also plotted, grouped by Gender.

Furthermore, it would be interesting to analyse the relationship between the outcomes in Maths tests and a variable affected by the COVID-19 pandemic as the attendance of the student in 2020 (Attendance). The

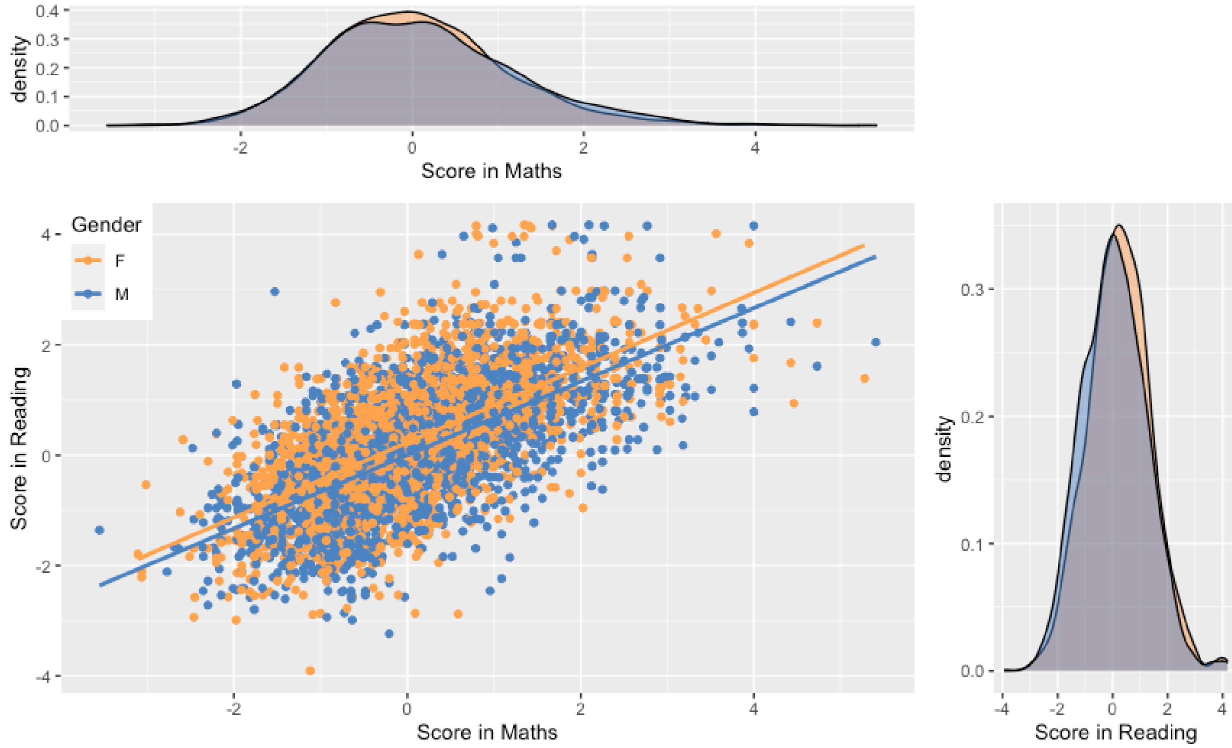

Figure 7: Scores scatter plot

factor is composed by three levels: regularly, little (every week is a few days away), very little (approximately once every 15 days). During the pandemic, the government of Uruguay announced a two-week suspension of classes at public and private schools on 14 March 2020. Then, a plan to let students return to school on a voluntary basis starting in June 2020 was announced on 21 May 2020. However, in this period, students transitioned to online classes using the computers and online tools that had already been set up through the "Ceibal" connectivity project.

```
  Attendance                                 count share `mean Reading score`
1 regularly                                   4266 90.3                  0.123
2 little (every week is a few days away)       326  6.90                -0.376
3 very little (approx. once every 15 days)     130  2.75                -0.497
```

More than 90% of students of the sample had a regularly attendance during 2020, while only 2.75% only assisted once every 15 days. The mean score in the first group (0.12) is much higher than the one among students with a lower level of attendance, which range between -0.38 and -0.5. The Chi-square association test confirm this relationship by not rejecting the null hypothesis of independence at 99% of confidence, with a Cramer's v statistic of 0.095.

```
Test results:
   X-squared statistic:  84.789
   degrees of freedom:  4
   p-value:  <.001

Other information:
   estimated effect size (Cramer's v):  0.095
```

Finally, Chi-squared tests for all the factors in the dataset were run to detect which variables are not significantly associated to Level_MAT. There are only 16 out of 193 variables which are significant independent from Level_MAT at 95% of confidence, among them there are variables related to reading habits and Gender, as it was previously showed.

	statistic	p.value
E6F25a_3_2	4.137746	0.65804177
E6CF25a_2_3	5.166892	0.52259330
E6CF25a_2_4	5.326420	0.50267921
E6F25a_1	6.108798	0.41111362
E6F25a_3_4	6.568629	0.36258633
E6CF25a_2_1	6.636884	0.35573626
E6F25a_3_1	6.779063	0.34176354
BE1_2	6.810747	0.33870446
ES7_5	7.599484	0.26893839
E6F26	5.911256	0.20587448
E6F25a_3_5	8.725279	0.18962785
Gender	3.626119	0.16315424
E6Co12_3	12.761690	0.12031641
BE1_7	10.938375	0.09029979
E6CF25a_2_5	11.096693	0.08543361
BE1_3	12.570369	0.05038912

A2. Linear Discriminant Analysis (LDA)

The first supervised algorithm that is applied correspond to the Linear Discriminant Analysis (LDA), which is a dimensionality reduction technique that helps to find the linear combination of original variables that provide the best possible separation between the groups in a categorical variable. In our analysis, Level_MAT is used as the response variable.

As LDA can be affected by the scale of the variables, we normalise the continuous predictor theta_READ. Then, a pairwise Chi-squared test was performed across all the features to detect multicollinearity in the dataset, taking as a rule of thumb a p-value less or equal than 0.01. The 19 variables that are significantly related between them, at 99% of confidence, are:

- "Gender"
- "E6CF25a_2_3"
- "E6CF25a_2_4"
- "E6Co4"
- "E6Co6_4"
- "E6Co6_5"
- "E6Co6_6"
- "E6Co7_2"
- "E6Co7_4"
- "E6C44m_2"
- "E6F12"
- "E6F20"
- "ES7_2"
- "Region"
- "E6F11"
- "E6C47c_11"
- "Attendance"
- "Age"

- "E6F6a"

We perform this strategy to detect multicollinearity because the standard approach using the Variance Inflation Factors (VIF) is not fully applicable to this model, which include categorical variables. This is because the correlations among these variables are induced by the model structure and depend on how they are coded, specifically which category is considered the reference level. Although there is an alternative solution by using the Generalized Variance Inflation Factors (GVIF), it did not show a good performance due to the dimensions of our dataset.

Finally, we get rid of those which have the lowest Chi-squared statistic (< 70) in the independence test with respect to Level_MAT. Therefore, among those variables, we keep in the model the following features, of which most are related to the COVID-19 context:

- E6Co6_6: "Most of the time I did homework by myself because no one could help me"
- Age: Age of the student, in ranges
- E6Co7_4: "At home I felt that I learned more than when I go to school"
- Attendance: Attendance report provided by the teacher
- E6Co6_5: "Most of the time I did homework just because I didn't need help"
- E6Co6_4: "For doing homework I was helped by another person (for example, a neighbour or someone who takes care of me)"

In order to run the LDA, the final sample with 4.722 observations and 181 variables was split between a training (70%) and test (30%) sets for prediction and testing purposes. Based on the training set, 34.9% belongs to level 1, 34.1% to level 2 and 31% to level 3.

After performing the algorithm, we get two linear discriminants (LD) that explains 83.9% and 16.1% of the between-group variance in the dataset, respectively. We also are interested in recognise the most relevant features for each LD, by looking at the absolute value of their coefficients. In the first linear discriminant, the most relevant features are:

- theta_READ*: Reading test score
- E6F4_5: Repetition (4 times)
- E6Co5_99: Perception about homework (no response)
- Centre_typeTiempo Extendido*: School category ("Extended time")
- E6F4_4: Repetition (3 times)

```
theta_READ                 E6F45                      E6Co599   Centre_typeTiempo Extendido
5.098236                   -2.681083                  -1.881023                    1.129972
E6F44
-1.043102
```

While, for the second LD the features most important are:

- E6Co5_99*: Perception about homework (no response)
- E6F4_4*: Repetition (3 times)
- ContextRural Quintil 2*: Socio economic and cultural context of the centre, defined by ANEP (Rural Quintile 2)
- E6F32_99*: Use a PC for testing in the classroom (no response)
- E6F33_99*: Use PC similar in the classroom (no response)

Where variables with * indicates a positive sign of the coefficient.

```
E6Co599                  E6F44  ContextRural Quintil 2             E6F3299             E6F3399
2.601530              1.943649                1.725869            1.683117            1.683117
```

From the next biplot based on LD1 and LD2, it can be concluded that the first linear discriminant separates the three levels of Maths score while the second one distinguish the extreme levels (1 and 3) from level 2. Having a high Reading test score and belonging to a Extended time school has a positive influence on the student's performance in Maths, while having repeated several times has an expected negative effect. Meanwhile, the most relevant variables for the LD2 are those related to homework, repetition (3 times), context of the school (Rural Quintile 2) and the use of PC in the classroom. They all have a positive coefficient, but there is an important overlap between levels, as shown in the plot.

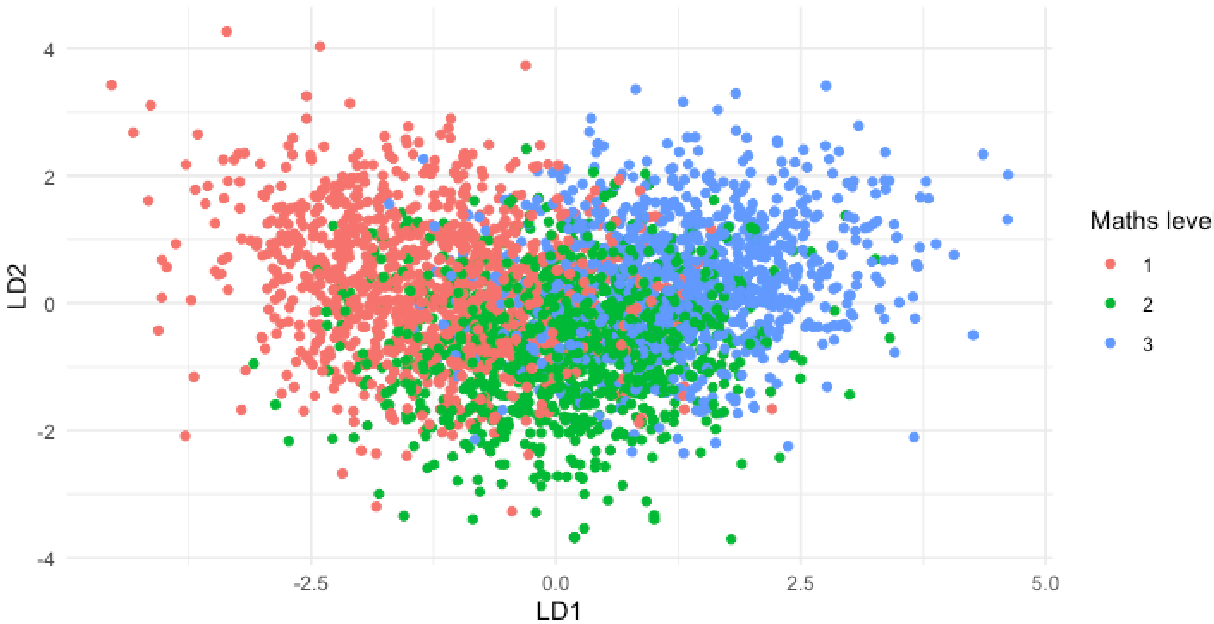

Figure 8: LDA biplot

Then, using the testing dataset we predict the levels of Maths score (Level_MAT) based on the Linear Discriminant model fitted. The accuracy of the model is around 56.2%, so our model correctly classified more than half of the observations, which is a good starting point. Although it is an important measure of the explanatory capacity of the model, the study is not focused on predicting but on finding the most relevant variables that impact on the Math level of the students.

```
         Actual
Predicted   1   2   3
        1 308 127  27
        2 158 188 127
        3  26 149 294

[1] 0.5626781
```

A3. Decision Trees

In this section, we try to classify the students according to the Maths score level of the "Aristas" tests conducted in 2020, by performing Decision trees. This nonparametric algorithm has some advantages for classification problems, as the easy interpretation of the results and the identification of the most important variables.

Firstly, the dataset is separated between training and testing sets and the decision tree model is created with all the 194 features of the dataset, because it is not affected by the presence of multicollinearity. In the following plot, there is a graphical representation of the procedure of how the tree splits the sample. The decision tree obtained with the function rpart() has the following characteristics: a depth equal to 3, 4 decision splits along values of the features theta_READ and Centre_cat, and 5 prediction regions. The decision tree also shows that Context and the socio-emotional variables E6F22 ("Expectations about the school trajectory"), ES2_1 ("If you're smart, you don't need to push yourself to do well in school") and ES2_4 ("A good student does not need to work hard to do well in school") were important features to split the nodes.

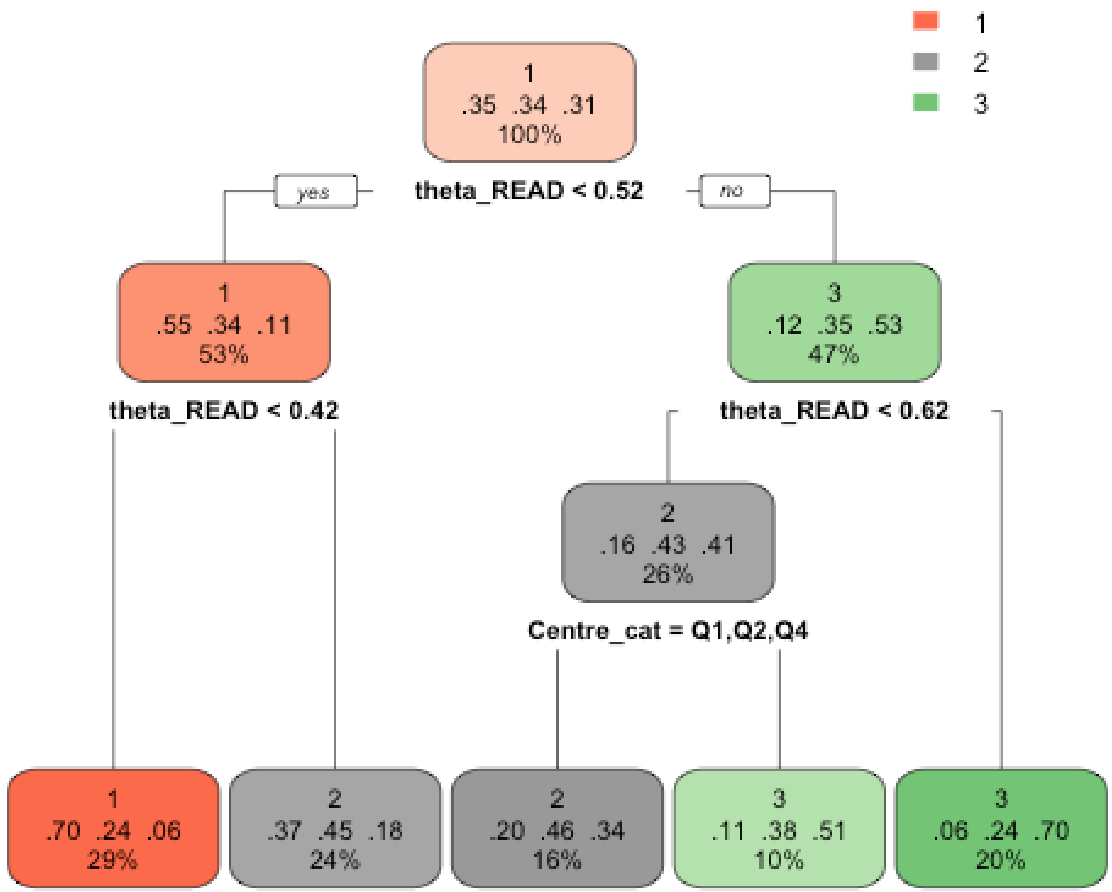

Figure 9: Decision tree (depth 3)

Then, a pruning complexity parameter (cp) plot was done to show the relative cross validation error (y-axis) for various cp values (lower x-axis). We should keep the cp value of the smallest tree (upper x-axis) that has the smallest cross validation error. Looking at the plot, it can be seen that the function rpart() performed an automated tuning and chose the optimal subtree of size 3 and cross validation error of 0.67.

After the model was fitted using the training set, we evaluate it by making predictions against the test set. This tree has an accuracy of approximately 58%, a little improvement with respect to the linear discriminant model. Moreover, the worst prediction performance is related to the observations in level 2, while in the other levels the balanced accuracy is higher than 72%.

```
Confusion Matrix and Statistics
```

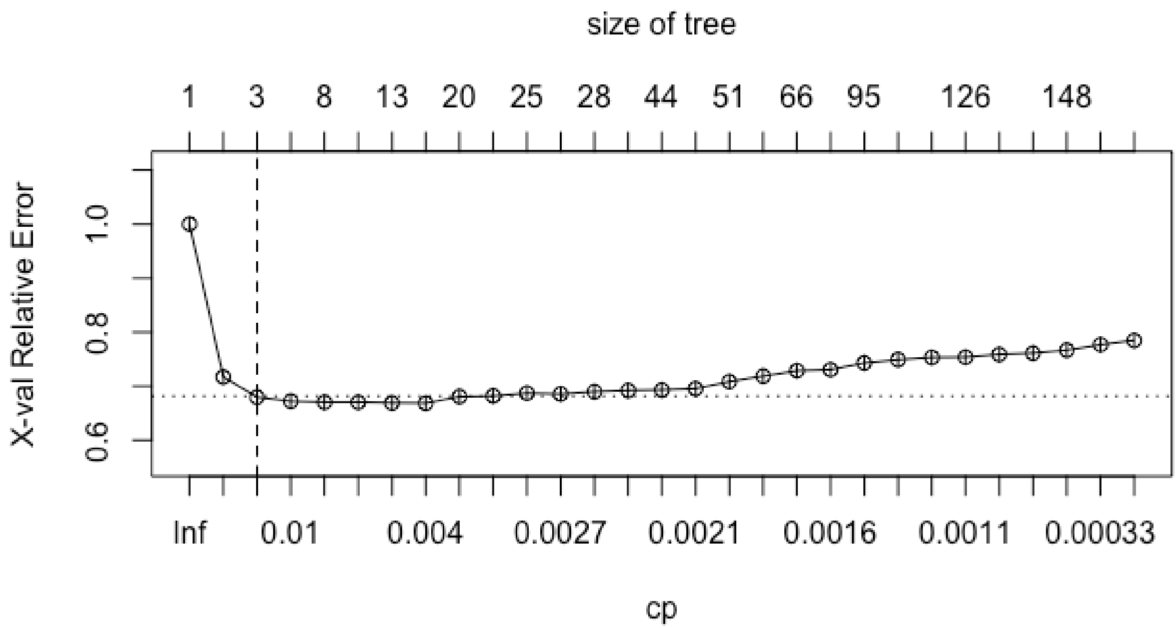

Figure 10: Cp plot

```
          Reference
Prediction   1   2   3
         1 674 232  56
         2 404 604 329
         3  79 295 645

Overall Statistics

               Accuracy : 0.5796
                 95% CI : (0.5626, 0.5964)
    No Information Rate : 0.3487
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3689

 Mcnemar's Test P-Value : 2.602e-11

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.5825   0.5340   0.6262
Specificity            0.8667   0.6648   0.8365
Pos Pred Value         0.7006   0.4518   0.6330
Neg Pred Value         0.7950   0.7340   0.8325
Prevalence             0.3487   0.3409   0.3104
Detection Rate         0.2031   0.1820   0.1944
Detection Prevalence   0.2899   0.4030   0.3071
Balanced Accuracy      0.7246   0.5994   0.7314
```

By incrementing the size of the tree to 7, the cross validation error lowered a bit (66.4%) and the accuracy was improved (62.1%). The variables added to the decision process are ES2_4 ("A good student does not need to work hard to do well in school"), E6Co9_1 ("I was able to talk with my teacher when the tasks raise doubts"), E6C40m_4 ("We all made the class rules"), E6F22a_3 ("Expectations when you are 15 years old: Working"), E6Co6_5 ("Most of the time I did homework just because I didn't need help"), E6C46a_2 ("School support activities participation (face-to-face and/or virtual)"), ES3_3 ("I like hard works because they are a challenge"), ES6_6 ("I tell my friends my feelings") and E6F22 ("Expectations about the school trajectory"). Among some interesting findings across this last decision tree is that students that strongly agree that "A good student does not need to work hard to do well in school" are predicted to belong to the lower level in Maths (level 1), while students that answered that sometimes "We all made the class rules" are predicted to have the highest level score (level 3).

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3
         1 820 303  76
         2 241 538 253
         3  96 290 701

Overall Statistics

               Accuracy : 0.6206
                 95% CI : (0.6038, 0.6371)
    No Information Rate : 0.3487
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4305

 Mcnemar's Test P-Value : 0.007687

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.7087   0.4757   0.6806
Specificity            0.8246   0.7741   0.8313
Pos Pred Value         0.6839   0.5213   0.6449
Neg Pred Value         0.8410   0.7406   0.8525
Prevalence             0.3487   0.3409   0.3104
Detection Rate         0.2471   0.1621   0.2113
Detection Prevalence   0.3614   0.3110   0.3276
Balanced Accuracy      0.7667   0.6249   0.7559
```

A4. Random Forest

Finally, in order to improve the accuracy of the tree classifier, a Random Forest for classification was applied to the dataset. This algorithm is a collection of decision trees that specifies the categories with much higher probability, exceeding the performance of Decision trees.

In first place, as usual, we split the dataset with the 194 features into train and test set, at the ratio 70:30. Next, we will create a Random Forest model with default parameters by running the function randomForest() on the training set. As it can be seen in the output, the model was fitted with 500 trees (ntree) and 13 variables randomly sampled at each stage (mtry), obtaining an error rate of 41.9%. The default value of ntree is 500 and, for classification problems, mtry is approximately the squared root of the number of variables.

```
Call:
 randomForest(formula = Level_MAT ~ ., data = train, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 13

        OOB estimate of  error rate: 41.92%
Confusion matrix:
    1   2   3 class.error
1 834 234  89   0.2791703
2 400 409 322   0.6383731
3  95 251 684   0.3359223
```

In the following plot the most important variables are showed according to two different measures, the Mean Decrease Accuracy and the Mean Decrease Gini, which represents the mean decrease in node impurity. According to the first criterion, theta_READ, Context, Centre_Cat, ES2_4 ("A good student does not need to work hard to do well in school"), E6F5a ("Thoughts about repetition"), and E6F22 ("Expectations about the school trajectory") are the most relevant features. In the case of the Gini index related, theta_READ, Context, Centre_Cat, E6F15 ("Number of people living in the household"), E6F22 and Centre_type are among the most important.

By making predictions over the test set, we get an accuracy of 60.5%, which is lower than the one obtained with the decision tree classifier. Therefore, we try to improve the model by changing the number of trees (ntree) and the number of variables randomly sampled at each stage (mtry). Regarding the first one, the number of trees needs to be sufficiently large to stabilize the error rate and a rule of thumb is to start with 10 times the number of features in the dataset, more than 1.000 in our analysis. In the following figure, the Out Of the Box (OOB) error is plotted as a function of the number of trees, in which is clear that classifying error is much higher in predicting level 2 (green line).

After increasing the number of trees to 1.000, we obtain a model with 13 variables randomly sampled at each stage (mtry) and a lower error rate (40.75%). The accuracy of the model does not change (60.5%) and is still lower than the obtained with the Decision trees algorithm.

```
Call:
 randomForest(formula = Level_MAT ~ ., data = train, importance = TRUE,      ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 13

        OOB estimate of  error rate: 40.75%
Confusion matrix:
    1   2   3 class.error
1 852 224  81   0.2636128
2 383 424 324   0.6251105
3  87 253 690   0.3300971
```

Finally, we tune the hyperparameter mtry by running a Random Forest with 500 trees and mtry=20. The results do not make a big improvement in the accuracy of the model (61.5%).

```
Call:
 randomForest(formula = Level_MAT ~ ., data = train, ntree = 500,      mtry = 20, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 20
```

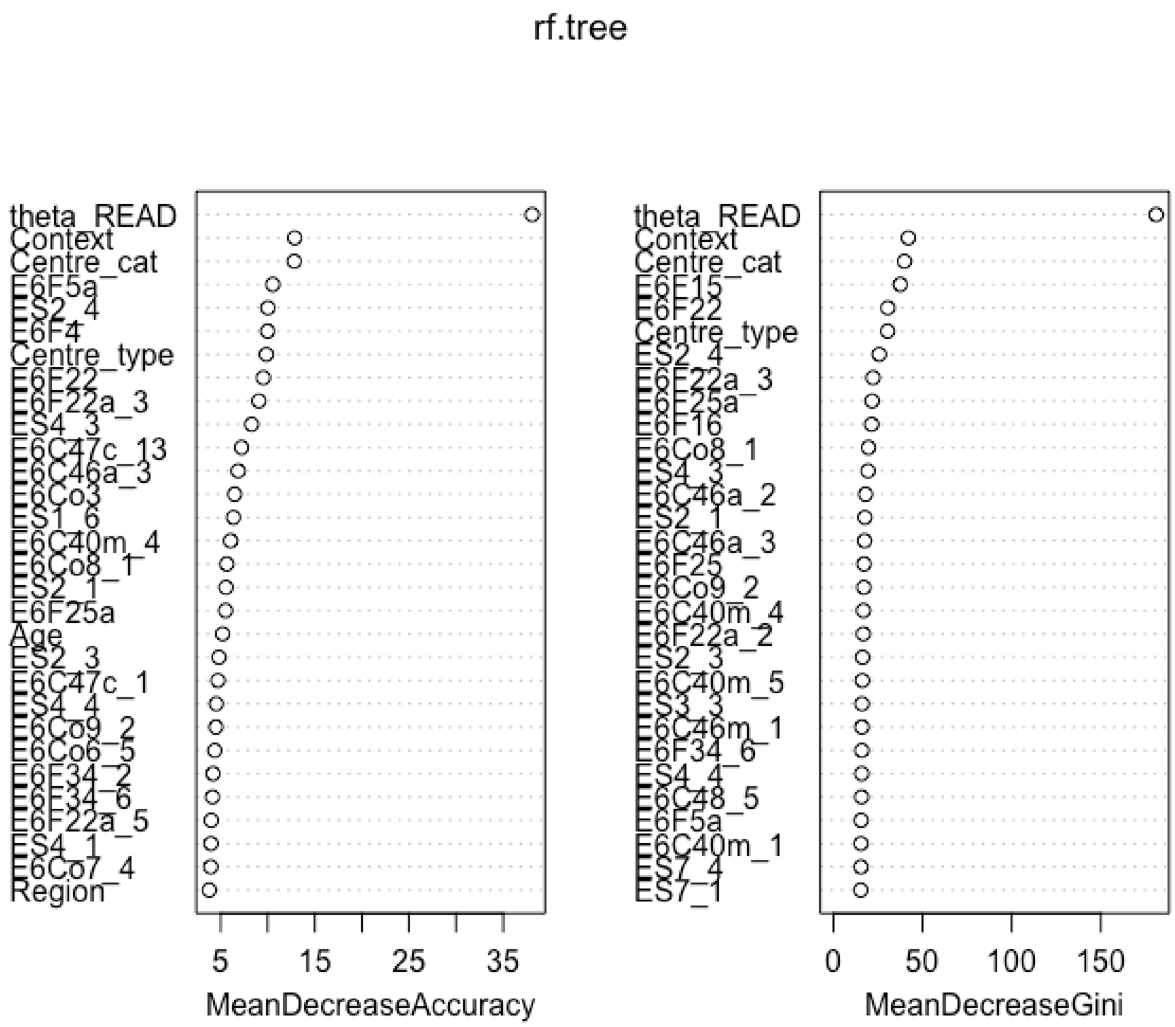

Figure 11: Variable importance plot

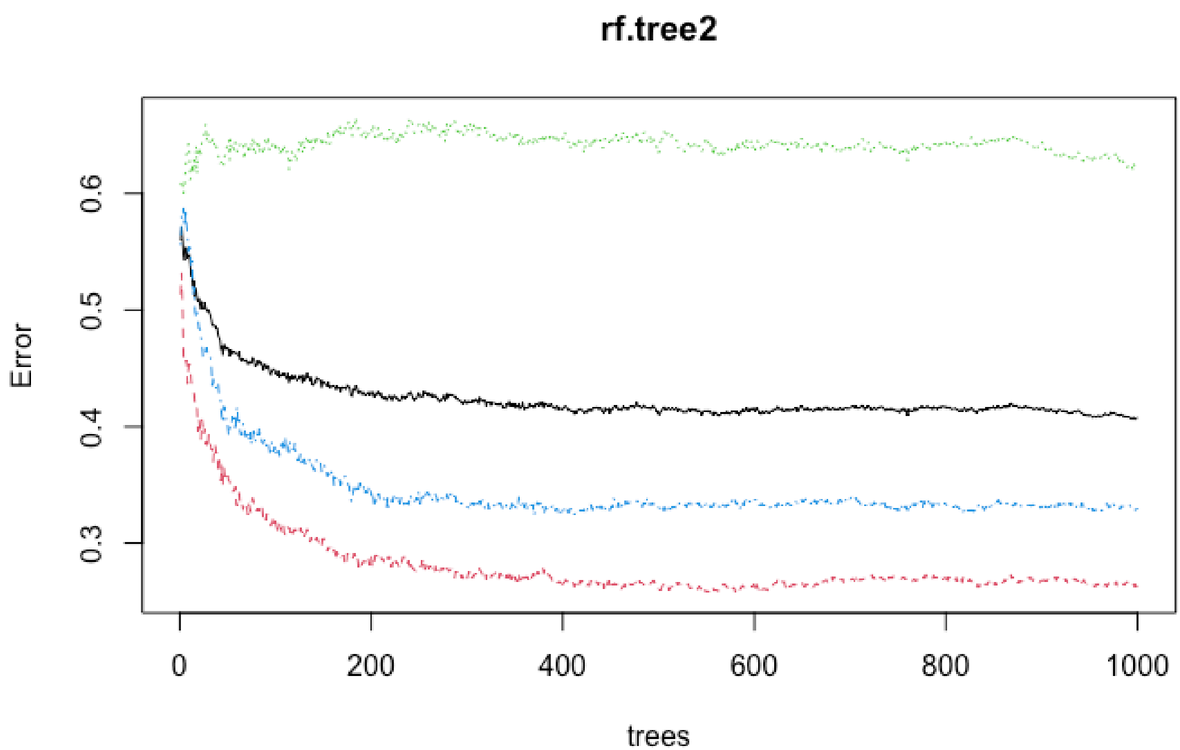

Figure 12: Error by number of trees plot

```
        OOB estimate of  error rate: 41.26%
Confusion matrix:
    1   2   3 class.error
1 840 232  85   0.2739844
2 384 431 316   0.6189213
3  85 267 678   0.3417476
```

The last attempt to obtain a better performing model is by using the Boruta algorithm, which is based on Random Forest. It creates a classification model with shadow and original attributes, and then assess the importance of the attributes. After performing 499 iterations in 1.7 hours, 49 attributes were confirmed as important (Age, Attendance, Centre_cat, Centre_type, Context, and 44 more), while 141 attributes confirmed unimportant (BE1_1, BE1_2, BE1_3, BE1_4, BE1_5, and 136 more), and 3 tentative attributes left (E6C40m_5, E6Co9_1, ES8_6). Among the rejected features, there are variables related to how the student feels at school (socio emotional attributes).

```
Boruta performed 499 iterations in 1.68544 hours.
50 attributes confirmed important: Age, Attendance, Centre_cat, Centre_type, Context and 45 more;
142 attributes confirmed unimportant: BE1_1, BE1_2, BE1_3, BE1_4, BE1_5 and 137 more;
3 tentative attributes left: E6C40m_5, E6Co9_1, ES8_6;
```

Therefore, based on feature selection performed by Boruta algorithm, we run a Random Forest including only the non rejected features. The model with 500 trees and mtry=7 has an OOB estimate of error rate of 41.14% and an accuracy of 60.4%, while the model fitted with 1.000 trees has an error rate of 40.81% and an accuracy of 60.75%. According to the Mean Decrease Accuracy, theta_READ, Centre_Cat, Context, E6F5a ("Thoughts about repetition") and ES2_4 ("A good student does not need to work hard to do well in school") are the most relevant features. In the case of the Mean Decrease Gini index, theta_READ, Context, Centre_Cat, E6F22 ("Expectations about the school trajectory") and E6F25a ("Reading on weekends") are among the most important.

In summary, it seems that it could be difficult to get a better model by tuning the hyperparameters of the Random Forest algorithm. However, we could obtain the most important features that explain the performance of students in the Maths test.

A5. Conclusions and main findings

After running three types of Supervised Machine Learning algorithms for our classification analysis (LDA, Decision Trees and Random Forest), we could detail some of the key findings obtained through the analysis of the results of "Aristas". The main goal was to detect the variables that could explain the Math test outcomes of sixth-grade students in Uruguay during 2020. Moreover, this analysis was also a great opportunity to have a first approximation about the impact that the pandemic had in the educational performance of students during that particular year.

The Reading test score is the most important variable to explain Level_MAT for all the models fitted in this study. Then, the variables related to the socio economic and cultural context of the school (Centre_cat, Context) are also relevant. It could be said that some socio emotional variables that measure the student feelings are related to the level in Maths, as Expectations about the school trajectory and Thoughts about repetition, for example. The repetition is another key factor that explain the student's performance, as it was expected. Finally, a group of variables related to the COVID-19 context, that describes some characteristics as the capacity of the student to do homework, student-teacher relationship and school support activities participation, were found to be relevant in the models.

In summary, from the 193 features that we have in the dataset, it can be concluded that only 21 variables (11%) are considered the most relevant to explain the student's level performance in Maths tests in Uruguay, during the year of the emergence of the pandemic. As it was previously stated, some factors that could be related to the effects of the COVID-19 like doing homework and studying at home, were found relevant to model our dependent variable Level_MAT.

B. Unsupervised Analysis

B1. Statement of the problem

In the second part of the study, we enrich the analysis performing two unsupervised learning methods with the aim of visualize and understand similarities between the sixth-grade uruguaian students, by detecting their common attributes.

In particular, we use Principal Component Analysis (PCA) to identify variables that could be projected into a lower dimension, helping to get a better description of all the features related to the students. Then, a K-means clustering algorithm is performed to classify students into mutually exclusive groups, or clusters, such that individuals within the same cluster are as similar as possible, whereas observations from different clusters are as dissimilar as possible.

For both techniques, we apply the same dataset used in the Supervised analysis, taking into account that Level_MAT is considered as another feature of the sample only in the case of the PCA.

B2. Principal Component Analysis (PCA)

As our main goal is to explain the variance in the data within a small number of components, thus we proceed by extracting the correlation matrix from the dataset, and its eigenvalues. Because of the fact that the dataset contain many categorical variables we use Generalized Low Rank Models (GLRM), an optimization algorithm for dimensionality reduction, to perform the PCA. As the following output shows, there is a first principal component that explains more than 75% of the variance, while the second component explains only 5.2% and the third 3.1%.

```
                              pc1        pc2        pc3        pc4        pc5        pc6
Standard deviation     12.0730515 3.17565839 2.45573716 1.96308089 1.69675341 1.57217668
Proportion of Variance  0.7511746 0.05197258 0.03107925 0.01986018 0.01483694 0.01273824
Cumulative Proportion   0.7511746 0.80314714 0.83422639 0.85408657 0.86892350 0.88166174
                              pc7         pc8         pc9        pc10
Standard deviation     1.332272179 1.300892895 1.144114982 1.047443485
Proportion of Variance 0.009147295 0.008721473 0.006745996 0.005654158
Cumulative Proportion  0.890809037 0.899530510 0.906276507 0.911930665
```

Then, we define how many principal components to keep by applying the three most common approaches in helping to make this decision. The first criteria is based on selecting the number of PC where the sum of eigenvalues is greater than or equal to 1. In our analysis this would have us retain the first 10 components. Secondly, the Proportion of Variance Explained (PVE) identifies the optimal number based on the total variability that we would like to account for, usually above 75%. Thus, in our data we should choose only the first component, which explains 75.1% of the variance. Finally, the scree plot criterion looks for the 'elbow' in the plot that shows the eigenvalues for each individual PC, and keeps all principal components before the line flattens out. In this case, we should choose the first 2 principal components.

In conclusion, different criteria suggest to retain 2 (scree plot criterion), 10 (eigenvalue criterion), or 1 (based on a 75% of variance explained requirement) components. Since in this analysis PCA is used for description and visualization, we select the two first principal components, which explains more than 80% of the variability in our original dataset.

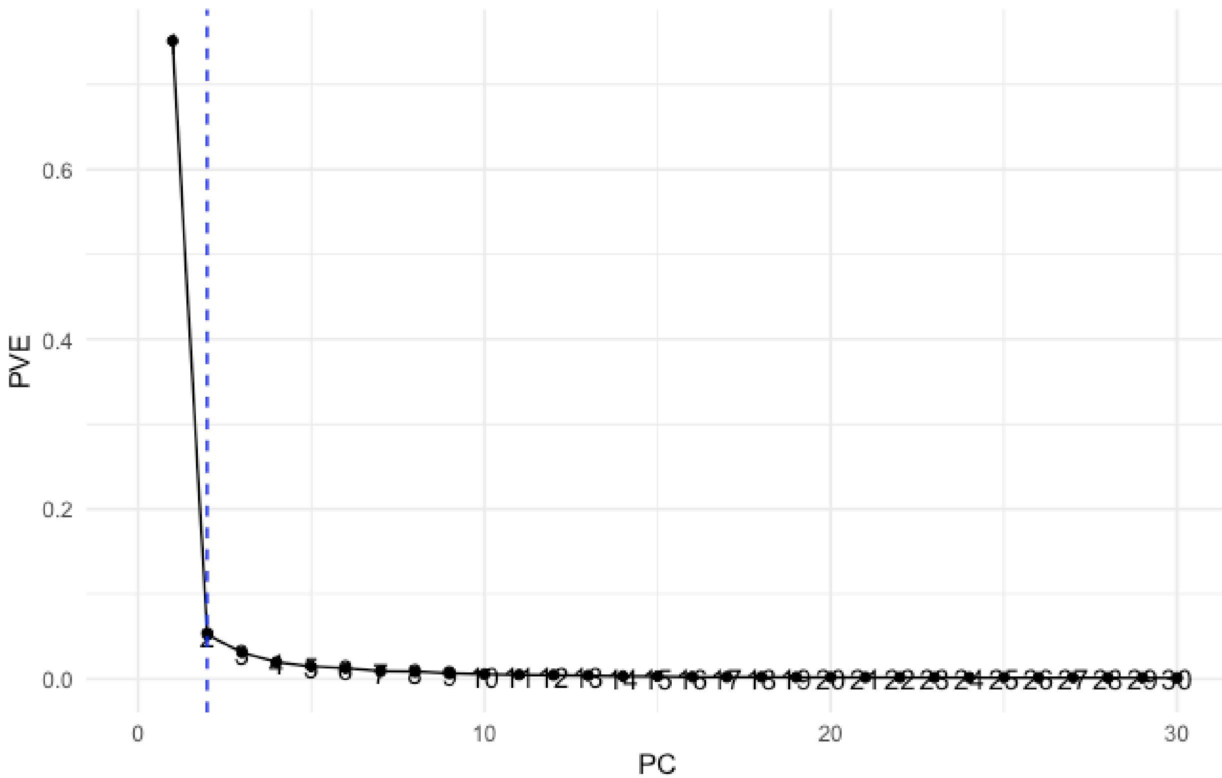

Figure 13: Scree plot

We can identify which of our original features contribute to the PCs by assessing the features influence on the associated PC. If we plot the loadings for PC1 we see that the largest contributing features are:

- ES5_5.1: “I never help my classmates when they have a problem”
- E6C36_1.1: “The teacher is never there to help when the children have a problem”
- E6C48_2.1: “The teacher never takes into account the opinion of the students”
- BE1_1.1: “I never have a good time at school”
- ES1_6.1: “When I realize that a work is going wrong, I never correct it”
- E6F37_1.1: “The teacher never explains to me again when I don’t understand something”

They are related to negative aspects about the experience of the students at school, and all of them contribute positively to the first PC. According to the second principal component, the main features are:

- E6F25a_1.1: “I talk about what I read”
- E6CF25a_2_1.0: “I recommend books”
- E6F25a_3_4.0: “I talk about what I read with the teacher”
- E6F25a_3_5.0: “I talk about what I read with other people”
- E6Co10_3.4: “Back to school feelings: I really wanted to see my teacher”
- E6Co1.1: “I read during the pandemic”
- E6Co10_1.4: “Back to school feelings: I was really excited to go back to school”
- E6F25a_3_2.0: “Do not talk about what I read with non-school friends”
- E6C42_2.4: “I go to school willingly”
- E6Co10_2.4: “Back to school feelings: I was very happy when I knew that I would go back to class”

It is clear that having reading habits of the students is a crucial factor to explain the second PC, while back

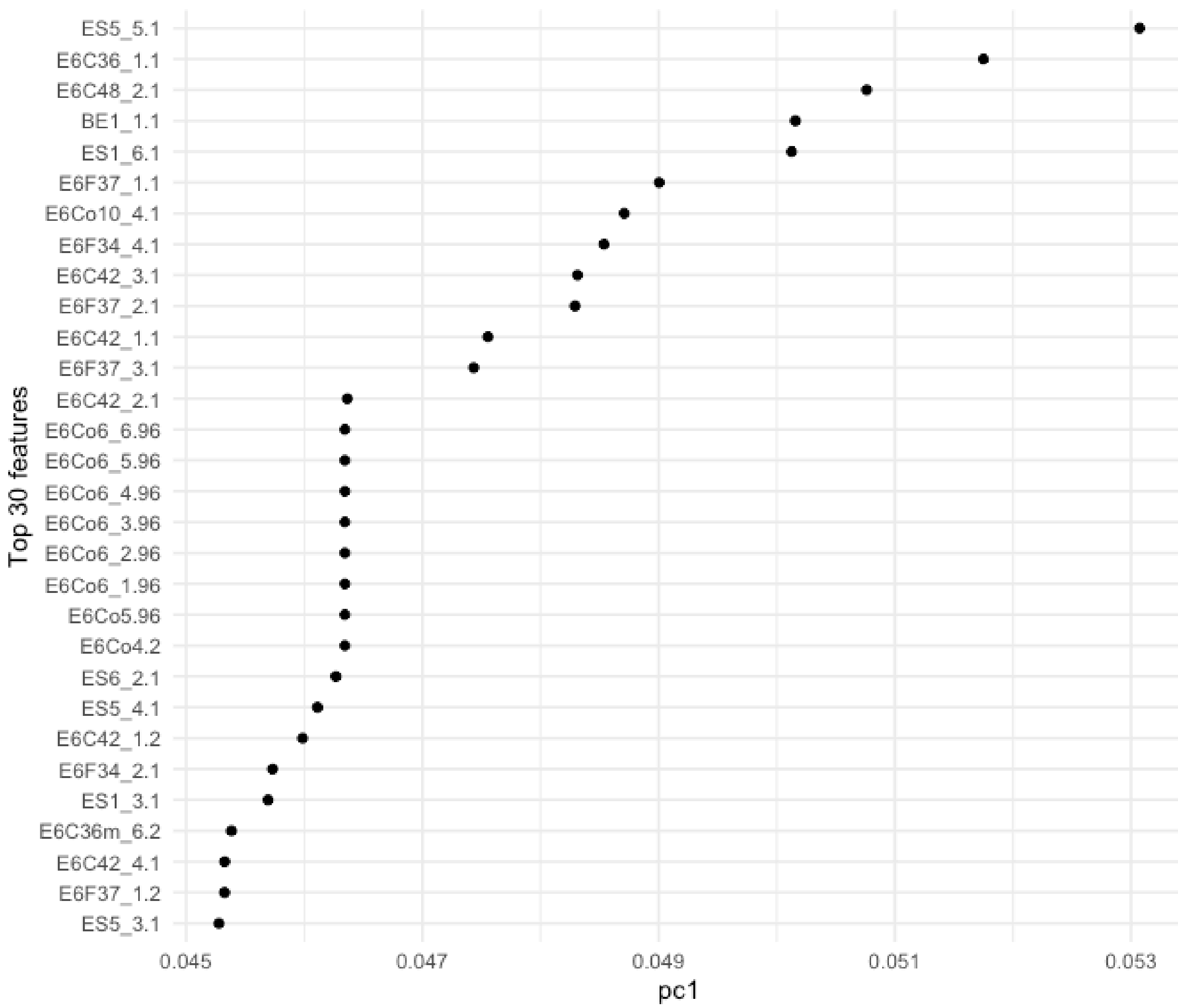

Figure 14: Variable importance for PC1

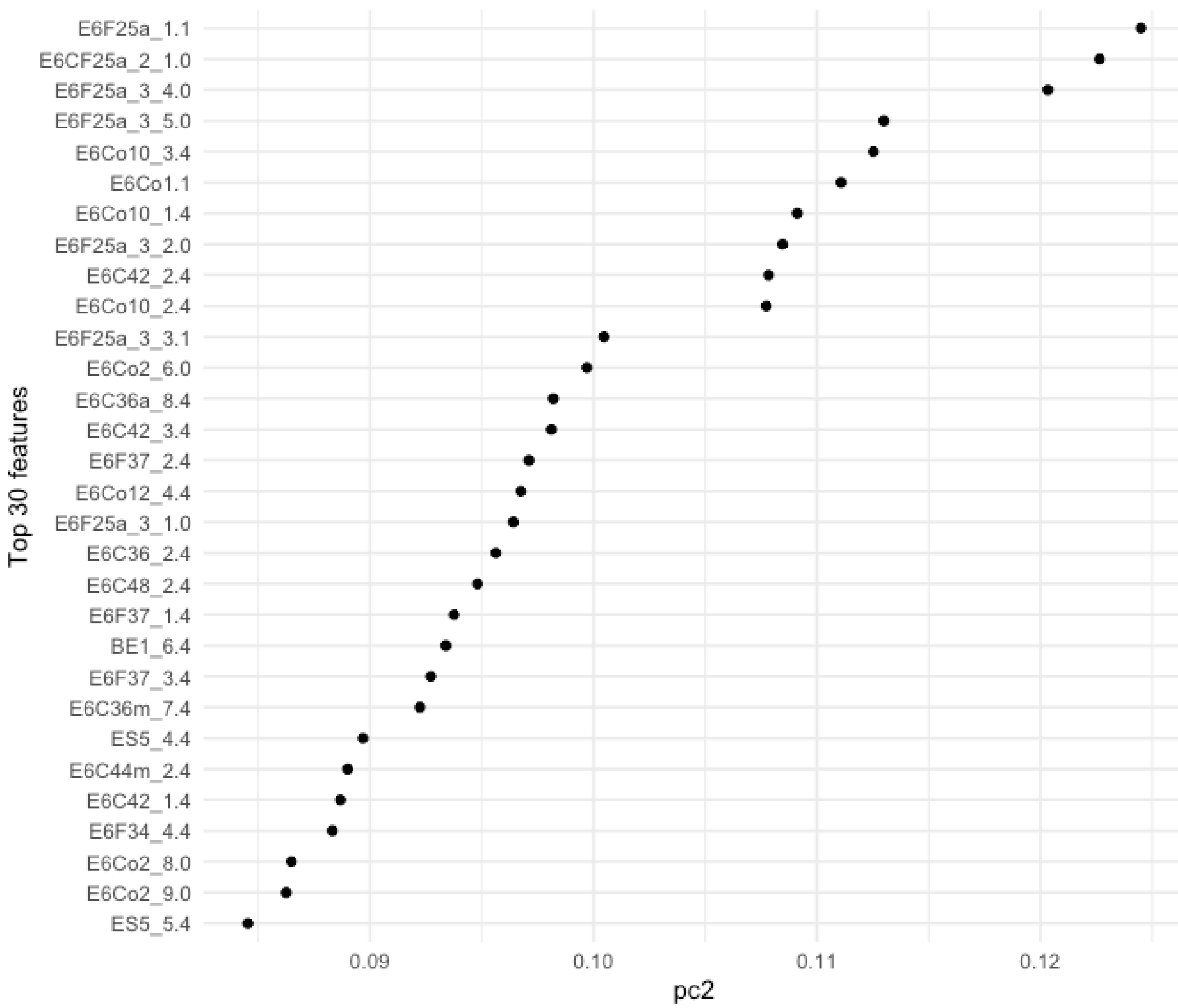

Figure 15: Variable importance for PC2

to school feelings after lockdown have a great impact too. Also in this case, all this features have a positive influence on the second PC.

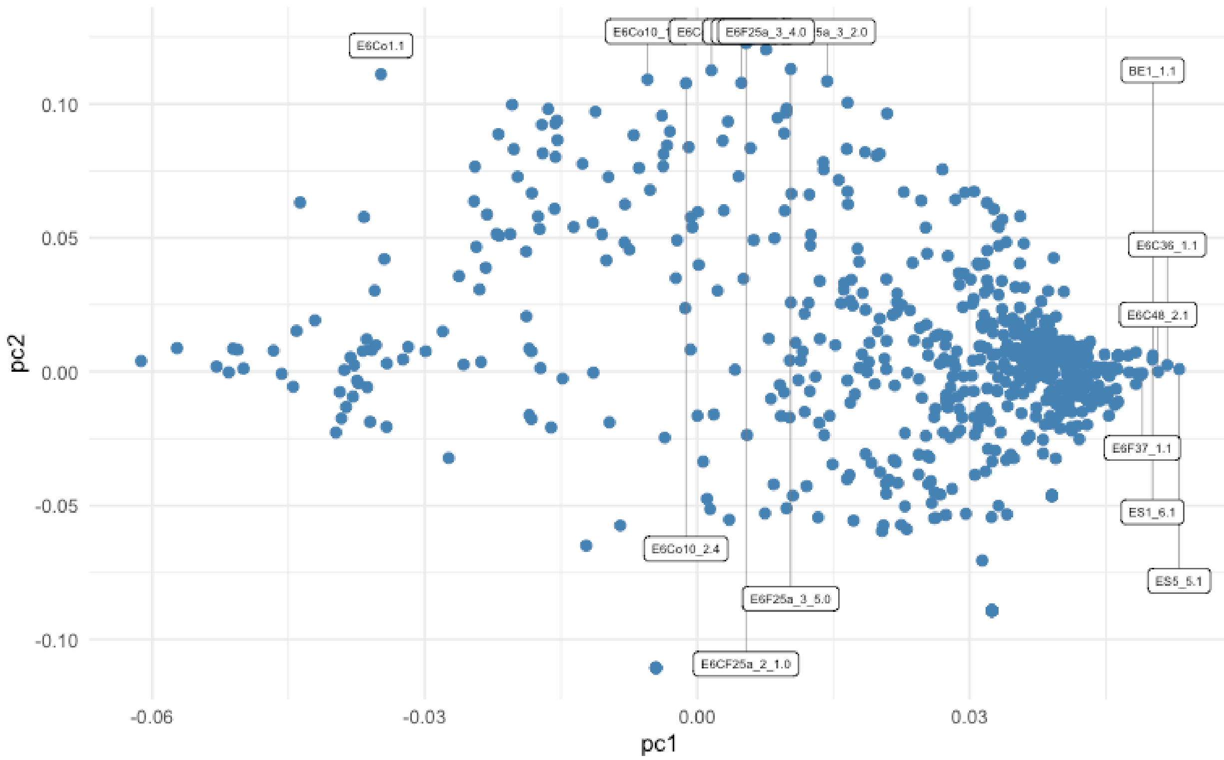

Figure 16: Biplot of PC1 and PC2

B3. K-means clustering

The main goal of performing clustering is, firstly, grouping sixth-grade students which share common features, and then, evaluate whether the clusters obtained resemble the level of Maths score for the individuals in the dataset. Thus, we get rid of the variable Level_MAT when running the clustering algorithms.

The choice of distance measure is a critical step in clustering because defines how similarity is calculated and influence on the shape and size of the clusters. We decided to chose the Gower distance measure, which is commonly used for data sets containing categorical and ordinal features. For this purpose, the K-medoids algorithm PAM was applied, which consist on a robust alternative to k-means for partitioning a data set into clusters of observations.

Firstly, to perform k-means on mixed data we should one-hot encode the nominal categorical variables. Although the number of clusters we specified could be based on the quantity of levels of the Level_MAT factor (3), we perform this cluster analysis to identify what natural distinct groups may exist in our dataset. Therefore, we start applying one of the most popular methods called “Elbow”, in which the location of a bend in the plot of total within-cluster sum of squares for each number of clusters (k) is generally considered as an indicator of the optimal k. In our case, the plot shows the “elbow” appears to happen when k = 5.

By computing the algorithm with k=5 we get five balanced clusters of sizes 883, 1045, 1010, 1249 and 535. As a measure of the goodness of the classification k-means, we look at the average silhouette width for each cluster and the whole dataset to check if data is well or poorly structured. A possible interpretation is made by Kaufman and Rousseeuw (1990), in which values higher than 0.5 imply that data is well structured. In our case, the dataset is poorly structured, with an overall average silhouette of about 0.013. The average value is also small for each cluster, as it is shown in the following output.

```
Numerical information per cluster:
     size  max_diss   av_diss  diameter separation
[1,]  883 0.2496148 0.1727093 0.3235747 0.08320493
[2,] 1045 0.2681048 0.1855929 0.3451464 0.08936826
[3,] 1010 0.3328197 0.1660208 0.4314330 0.08320493
[4,] 1249 0.3929122 0.2005216 0.5084746 0.12480740
[5,]  535 0.2773498 0.2081477 0.3482280 0.14329738


Average silhouette width per cluster:
[1]  0.020290007 -0.004414756  0.034252297  0.003381489  0.013651174
Average silhouette width of total data set:
[1] 0.01258457
```

Then, a 3-means clustering was fitted to be able to compare the similarity between the level in Maths test (Level_MAT) and the cluster that each student belongs to. We get three balanced clusters with size 1588, 1547, and 1587, respectively, and an overall average silhouette equal to 0.027. The value for each cluster is also very low, as it is reported in the next output.

```
Numerical information per cluster:
     size  max_diss   av_diss  diameter separation
[1,] 1588 0.2989214 0.1918947 0.3605547 0.08936826
[2,] 1547 0.3328197 0.1724676 0.4314330 0.08936826
[3,] 1587 0.3929122 0.2070521 0.5084746 0.12480740


Average silhouette width per cluster:
[1] 0.001394824 0.078113883 0.002665057
Average silhouette width of total data set:
[1] 0.02695608
```

Despite the poor results obtained when clustering individuals, we proceed to evaluate how the k-means algorithm categorized the students compared to the variable Level_MAT. Thus, we check whether the clusters resemble the level of Maths test scores of the individuals. By performing a Chi-squared independence test, we do not reject the null hypothesis of independence among the variables cluster and Level_MAT, at 99% of confidence, so there is a significant relationship between them. However, the estimated effect size is 0.08, showing a weak association.

```
Test results:
   X-squared statistic:  62.518
   degrees of freedom:  4
   p-value:  <.001

Other information:
   estimated effect size (Cramer's v):  0.081
```

By looking at the mosaic plot of the two variables considered, it can be seen that there is an inverse relationship between the level of the Maths score and the cluster that each individual belongs to. A possible interpretation could be that in the first cluster there is a trend towards students with the highest score in Maths, while in the last one there is a majority of students with the worst performance.

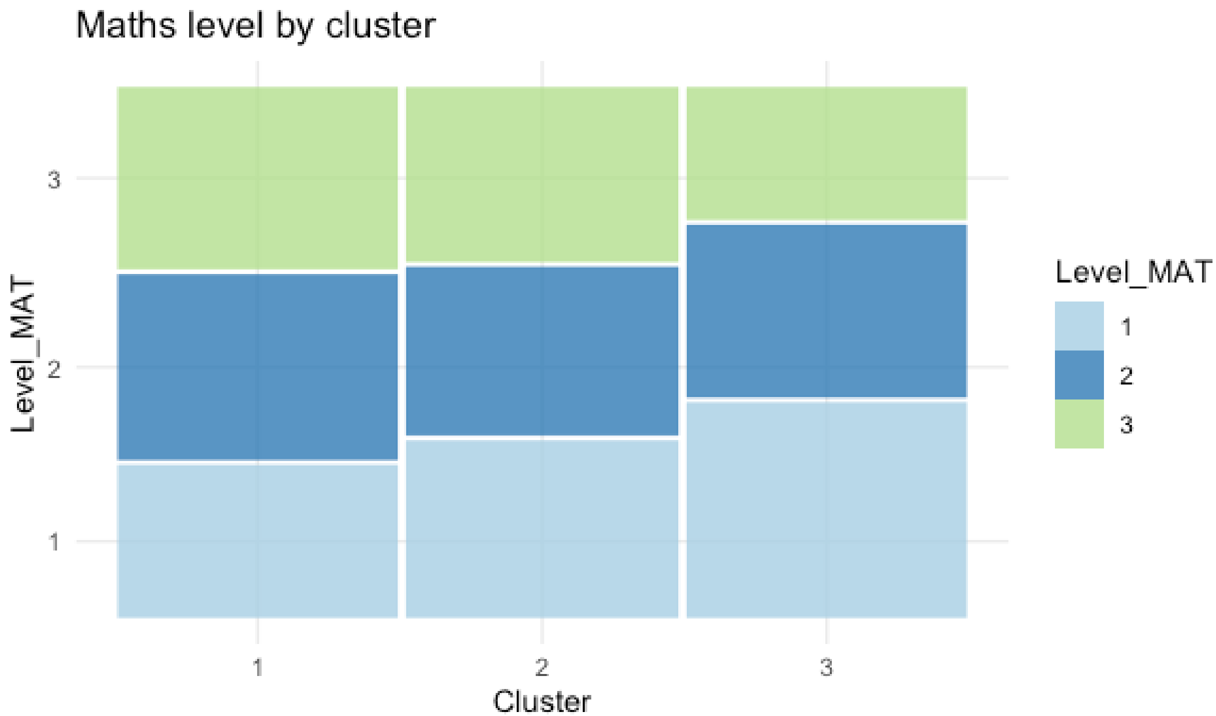

Figure 17: Level in Maths by Cluster mosaic

B4. Conclusions and main findings

In summary, in this second part of the study two Unsupervised learning algorithms were performed to analyse similarities between the sixth-grade students of Uruguay, by detecting a set of components that describes the features (PCA) and classifying observations into k groups, based on their similarity (k-means). The key findings can be summary as follows:

- The first two principal components explains a great proportion of the variability in the data, more than 80%.
- The group of variables with the greatest impact on PC1 are related to negative aspects about the experience of the students at school during 2020.
- Having reading habits and the feelings of the children about back to school after lockdown are important factors to explain PC2.
- Some variables related to the socio emotional effects of the COVID-19 on students were also found relevant in the Unsupervised analysis.
- Although the clustering did not generate well structured groups, we got three clusters of students that have a statistically significant association to the variable Level_MAT. For example, more than 40% of the students with the worst performance in Maths were grouped in the first cluster.

These results highlights the importance of the student socio emotional characteristics, in addition to providing insights on how the impact of the COVID-19 pandemic is relevant to the achievements of the educational system.

References

- INEEd (2021), *Aristas 2020. Primer informe de resultados de tercero y sexto de educación primaria.* Extracted from https://www.ineed.edu.uy/images/Aristas/Publicaciones/Aristas2020/Aristas-2020-Primer-informe-de-resultados-de-tercero-y-sexto-de-educacion-primaria.pdf
- Santiago, P. et al. (2016), *OECD Reviews of School Resources: Uruguay 2016*, OCDE Publishing, Paris. Extracted from https://read.oecd-ilibrary.org/education/oecd-reviews-of-school-resources-uruguay-2016_9789264265530-en#page4

Appendix

Data dictionary

Variable name	Label
Centre_cat	Socio-economic and cultural context of the centre (coded)
theta_MAT	Mathematics test score
Nivel_MAT	Mathematics performance
theta_LEC	Lecture test score
Nivel_LEC	Lecture test performance
Region	Region
Age	Age
Genre	Sex
Centre_type	School category
Context	Socio-economic and cultural context of the centre (defined by ANEP)
strat	School size
ES1_1	I check my works to make sure I did it right
ES1_2	When I am studying, I ask myself questions to know if I am understanding well
ES1_3	When I have to do a job, I manage to organize myself (for example: I look for materials)
ES1_4	I read the test questions carefully before I start answering
ES1_5	If I don't understand something, I reread it more carefully until it is clear to me.
ES1_6	When I realize that a work is going wrong, I correct it
ES3_1	I study to learn
ES3_2	I like studying
ES3_3	I like hard works because they are a challenge
ES3_4	If I get excited about a topic, I want to continue learning beyond class
ES3_5	I like learning new things
ES2_1	If you're smart, you don't need to push yourself to do well in school.
ES2_2	You can get better marks by studying more
ES2_3	Some kids are just born smart and do better in school
ES2_4	A good student does not need to work hard to do well in school.
ES4_1	If a task is hard, I leave it undone
ES4_2	I abandon tasks before finish them
ES4_3	If a task is long, I only do the easiest parts
ES4_4	I forget to do the homework
ES4_5	I lose borrowed things
ES5_1	I can recognize when a classmate feels sad
ES5_2	I know when a classmate is angry, even when he does not say anything
ES5_3	It bothers me when a classmate is bullied
ES5_4	I am happy when others are happy
ES5_5	I help my classmates when they have a problem

Variable name	Label
ES6_1	I make friends easily
ES6_2	I feel part of a group because we do things together
ES6_3	At break, I invite other children to do things together
ES6_4	When something happens to me that I can't solve, I ask for help
ES6_5	When someone gets mad at me, I speak up to fix things
ES6_6	I tell my friends my feelings
ES7_1	When I am angry I think of something else
ES7_2	I can calm down when I am nervous
ES7_3	When I'm sad about something, I try to think of something happy
ES7_4	When something makes me wrong, I try to think that it is not so bad
ES7_5	When I'm sad I try to distract myself
ES8_1	I am easily distracted during lesson
ES8_2	It is hard for me to wait my turn
ES8_3	I get angry easily
ES8_4	I do things without thinking (such as hitting, insulting, assaulting) but later I regret
ES8_5	In some situations I act without thinking
ES8_6	If I don't get a job, I get angry and stop doing it
BE1_1	I have a good time at school
BE1_2	I start the day thinking that good things are going to happen
BE1_3	I like how I am
BE1_4	The teacher notices when I do something right
BE1_5	When I am sad, my colleagues ask me what is wrong with me
BE1_6	I like going to school
BE1_7	When I have a problem I tell a friend
BE1_8	I am satisfied with myself
E6F6a	Number of houses
E6F15	Number of people living in the household
E6F16	Number of siblings you live with
E6F7	Live with mother
E6F8	Live with father
E6F9	Live with aunt
E6F10	Live with uncle
E6F11	Live with grandmother
E6F12	Live with grandfather
E6F13	Live adult brother
E6F14	Live other adult
E6F17	Has a place of study
E6F18	Has a table and a chair to study
E6F18a	Have a computer just for you to do your homework
E6F19a	Have a shared computer that you share with other people in your house, to do homework
E6F19	Good lighting for reading and studying
E6F20	A place to store school supplies
E6F21	They ask you about homework
E6F3	Change of school
E6F4	Repetition
E6F5a	About repetition
E6F22	Expectations
E6F22a_5	Expectations when you are 15 years old: Still at primary school
E6F22a_1	Expectations when you are 15 years old: Studying at the high school
E6F22a_2	Expectations when you are 15 years old: Studying at UTU (technological high school)
E6F22a_3	Expectations when you are 15 years old: Working

Variable name	Label
E6F22a_4	Expectations when you are 15 years old: still at primary school
E6F26a	Liking to read
E6F26	Liking to read outside school
E6F25	Reading time
E6F25a	Reading on weekends
E6F25a_1	Talk about what you read
E6F25a_3_1	Talk about what you read with school friends
E6F25a_3_2	Talk about what you read with non-school friends
E6F25a_3_3	Talk about what you read with family
E6F25a_3_4	Talk about what you read with the teacher
E6F25a_3_5	Talk about what you read with other people
E6CF25a_2_1	I do not recommend books
E6CF25a_2_2	I recommend books to my school friends
E6CF25a_2_3	I recommend books to my non-school friends
E6CF25a_2_4	I recommend books to my family
E6CF25a_2_5	I recommend books on social media (for example, YouTube, Instagram, Facebook, etc.)
E6Co1	Read in pandemic
E6Co2_1	Read in pandemic: Poems
E6Co2_2	Read in pandemic: Song lyrics
E6Co2_3	Read in pandemic: Stories
E6Co2_4	Read in pandemic: Written novels
E6Co2_5	Read in pandemic: Comic books
E6Co2_6	Read in pandemic: Science Articles
E6Co2_7	Read in pandemic: Study books
E6Co2_8	Read in pandemic: News
E6Co2_9	Read in pandemic: Social media posts (for example, Facebook, Instagram, etc.)
E6Co3	Had classes through platforms
E6Co4	Homework submissions
E6Co5	Perception about homework
E6Co6_1	I was helped by someone in my family
E6Co6_2	I was helped by my teacher
E6Co6_3	I was helped by my classmates
E6Co6_4	I was helped by another person (for example, a neighbor or someone who takes care of me)
E6Co6_5	Most of the time I did homework just because I didn't need help
E6Co6_6	Most of the time I did it by myself because no one could help me
E6Co7_1	I was able to study and do homework for school quietly, without distractions
E6Co7_2	If I did not understand, someone in my house tried to explain
E6Co7_3	I enjoyed studying from home without going to school
E6Co7_4	At home I felt that I learned more than when I go to school
E6Co8_1	I communicated with my classmates (by phone, Zoom, Conference, WhatsApp)
E6Co8_2	Between classmates we help each other with the tasks that they send us
E6Co8_3	I kept abreast of the family situation of my colleagues.
E6Co8_4	Although I did not see everyone, I felt accompanied by my classmates during lockdown
E6Co9_1	I was able to talk with my teacher when the tasks raise doubts
E6Co9_2	I was in contact with my teacher (either by WhatsApp, Zoom, CREA, email, etc.)
E6Co9_3	When I needed it, I was able to share with my teacher how I was feeling
E6Co9_4	I felt supported by my teacher during the time I worked from home
E6Co9_5	My teacher asked me if I understood the tasks I had to do
E6Co9_6	My teacher corrected the homework that I handed in
E6C47c_1	Communication channels: Notice board on the school door

Variable name	Label
E6C47c_3	Communication channels: Through WhatsApp with the teacher (messages, personal audios, a class chat group)
E6C47c_6	Communication channels: Through WhatsApp with my classmates (messages, personal audios, a class group)
E6C47c_7	Communication channels: Through my parents
E6C47c_8	Communication channels: The school website / Facebook page
E6C47c_11	Communication channels: A specific application of the center
E6C47c_12	Communication channels: Through your teachers in interviews in the press (radio, newspapers, television)
E6C47c_9	Communication channels: Through video calls (Conference, Zoom, Meet, Webex, etc.)
E6C47c_13	Communication channels: Through the teachers when we were going to pick up the meal at school
E6Co10_1	Back to school feelings: I was excited to go back to school
E6Co10_2	Back to school feelings: I was happy when I knew that I would go back to class
E6Co10_3	Back to school feelings: I wanted to see my teacher
E6Co10_4	Back to school feelings: I wanted to see my school friends again
E6Co11	Virtual lessons
E6Co12_1	It was difficult to get used to wearing the mask
E6Co12_2	It was hard going back to school and keeping a distance at school break
E6Co12_3	I missed not being able to play what I wanted at school break
E6Co12_4	It was hard going back to school and not being able to hug my classmates
E6F34_1	My classmates take care of each other
E6F34_2	In my class we are good mates
E6F34_3	When I have a problem, my classmates help me
E6F34_4	I have a good time with my classmates
E6F34_5	Children in this class worry about the rest of their classmates
E6F34_6	Classmates in this class treat each other with respect
E6C36_1	The teacher is there to help when the children have a problem
E6C36_2	The teacher tells me when I do a good job
E6C36m_6	The teacher treats me well
E6C36m_7	I get along with the teacher
E6C36a_8	I am confident to talk to my teacher
E6F37_1	The teacher explains to me again when I don't understand something
E6F37_2	The teacher listens to what I have to say
E6F37_3	The teacher values my effort when I do a work
E6C42a	Written rules
E6C40m_1	We talk about why it is necessary to have rules
E6C40m_2	We did activities to work on how we children feel about sanitary measures
E6C40m_3	We talked about why it is necessary to put the health protocol into practice
E6C40m_4	We all made the class rules
E6C40m_5	Between the children we agree on what to do at recess
E6C44m_1	Rules help us get along better
E6C44m_2	Penalties apply to all children equally
E6C44m_3	The rules are the same for everyone
E6C46a_2	School support activities (face-to-face and/or virtual)
E6C46m_1	Activities with the class (for example, debates, thematic workshops, solidarity campaigns, etc.)
E6C46a_3	Virtual educational activities (for example, visits to shows, theaters, etc.)
E6C46m_4	Activities to which relatives and neighbors are invited (for example, festivals, open workshops, national events, etc.)
E6C43a	Class delegate existence

Variable name	Label
E6C43	Class delegate election
E6C48_1	The teacher asks the opinion of the students before carrying out any activity (virtual or face-to-face)
E6C48_2	The teacher takes into account the opinion of the students
E6C48_3	The students decide on things at school
E6C48_4	The students propose activities that we would like to do
E6C48_5	Sixth graders vote to decide something
E6C42_1	I like my school
E6C42_2	I go to school willingly
E6C42_3	I feel part of this school
E6C42_4	It would make me sad to have to change schools
E6C42_5	I would be upset if someone spoke ill of my school
E6F32	Use PC test in the classroom
E6F33	Use PC similar in the classroom
Attendance	Attendance report provided by the teacher