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

— 01

Statistical Learning Final Project

Mathias Cardarello Fierro
MSc Data Science and Economics
Università degli Studi di Milano

8th July 2021

Aim of the study

This study is focused on the analysis of the main factors that could explain the Maths scores of sixth-grade students of Uruguay, during the first year of the pandemic.

In 2020, the evaluations "Aristas" were performed in primary schools of Uruguay by the National Institute of Educational Evaluation (INEEd), with the aim of “generating inputs that allow contributing to develop policies to mitigate the possible effects of the COVID-19 pandemic”.

Dataset

The dataset is composed by 4.722 observations and 194 variables of which all are factors, with the exception of *theta_READ* and *theta_MAT* that measures the student score in the Reading and Maths test respectively, standardised as theta-scores.

"Aristas" also collects information about social emotional skills, opinions and attitudes about the coexistence of students, among other conjunctural factors.

Aristas
Evaluación Nacional
de Logros Educativos

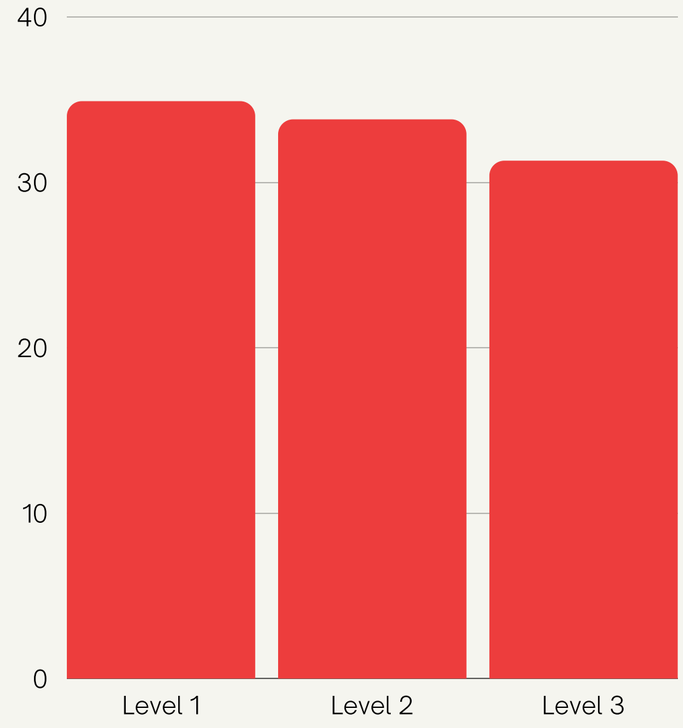
Source: Aristas 2020 (INEEd)

Descriptive analysis

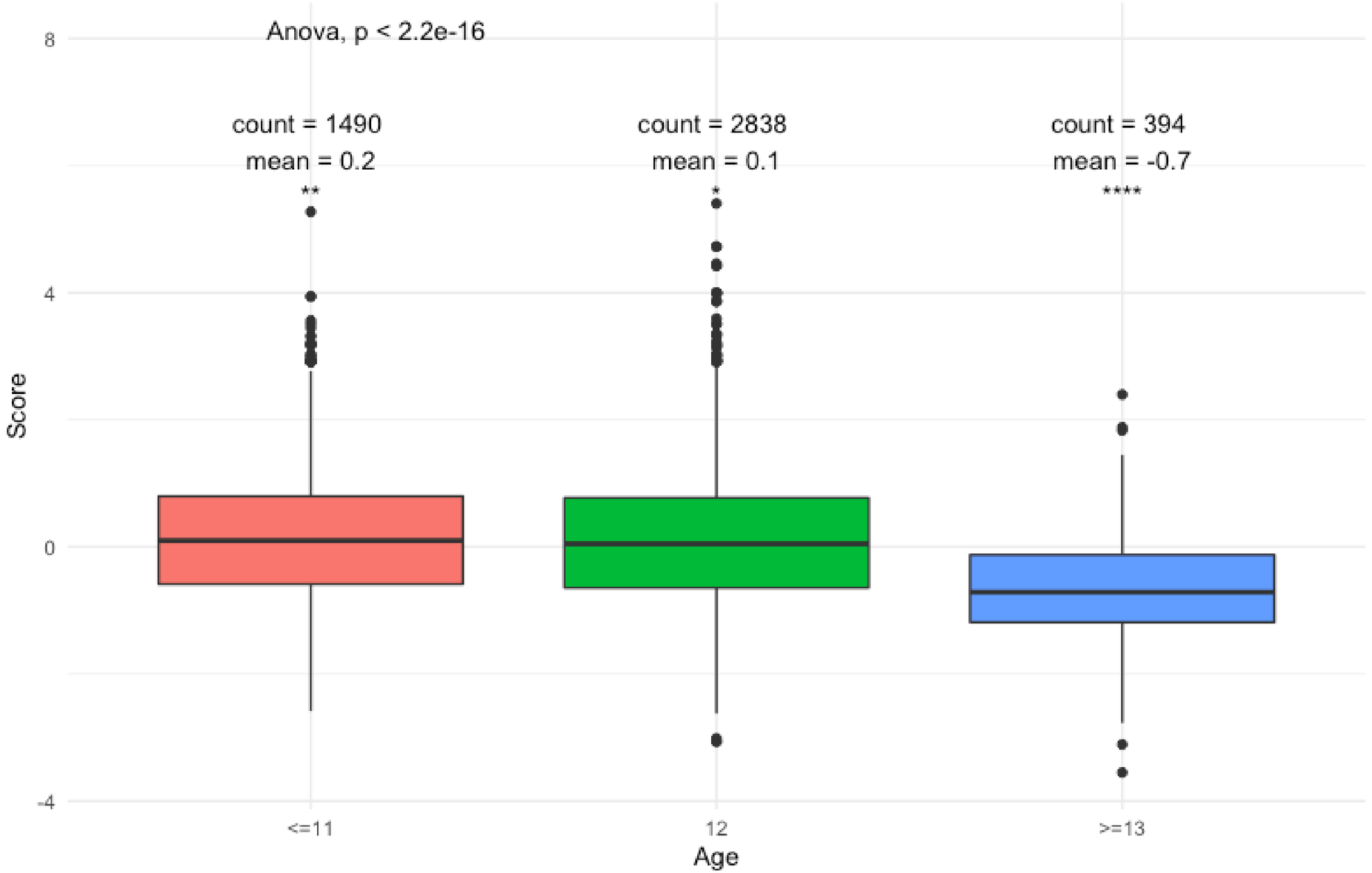
Main findings

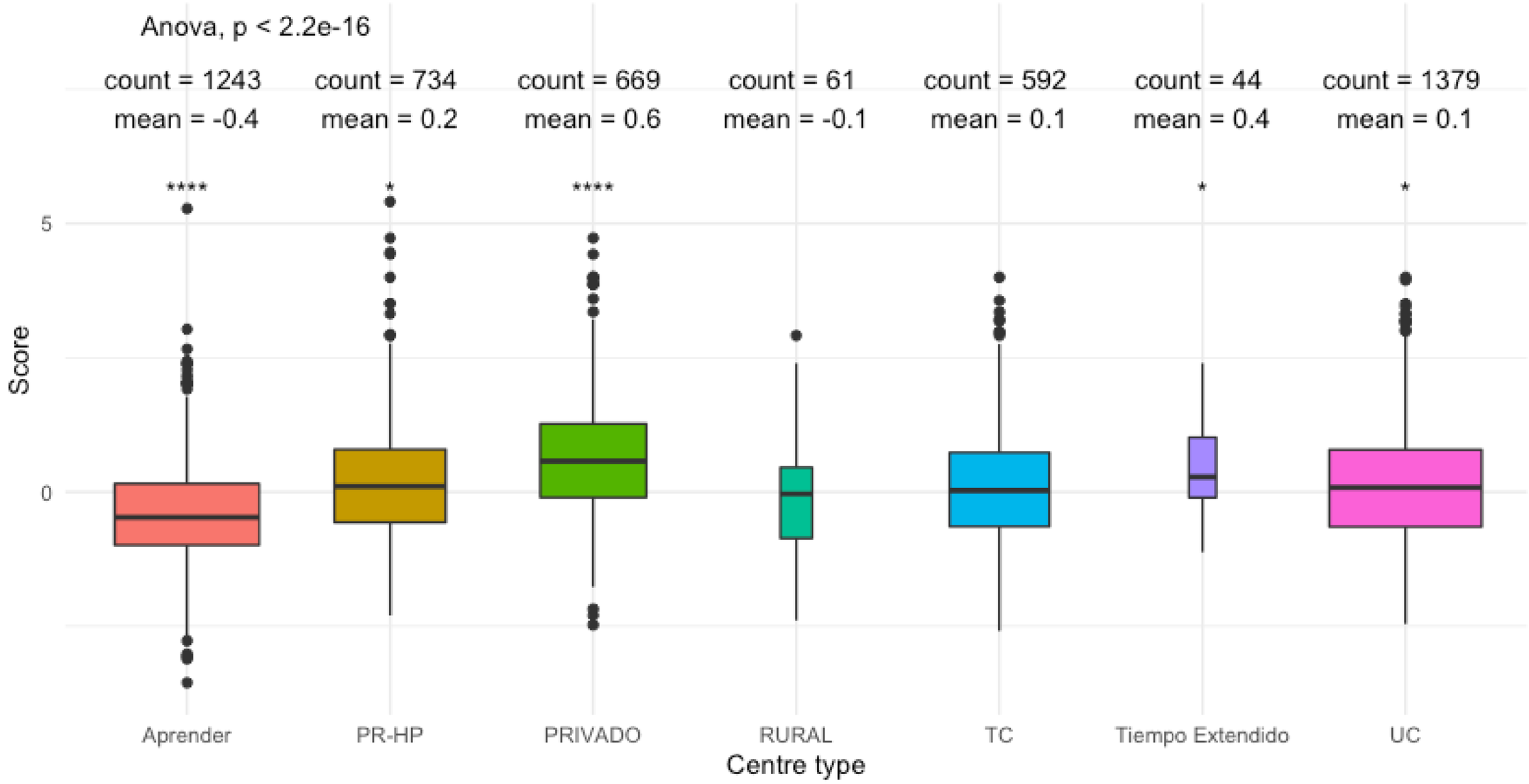
- The variable *Level_MAT* has 3 levels with a share between 31% and 35%.
- Students from Montevideo have a **better score** level in Maths (0.12) than those who study outside the capital city (0.04).
- The Maths score have a **significant negative relationship** with the age of students.
- *Context*, *Centre_type* and *theta_READ* have also a **significative correlation** with Level_MAT.
- The level of attendance during 2020 has a **positive association** with the level in Maths.
- **Only 16** out of 193 variables are significant independent from *Level_MAT*. They are related to reading habits and the gender of students.

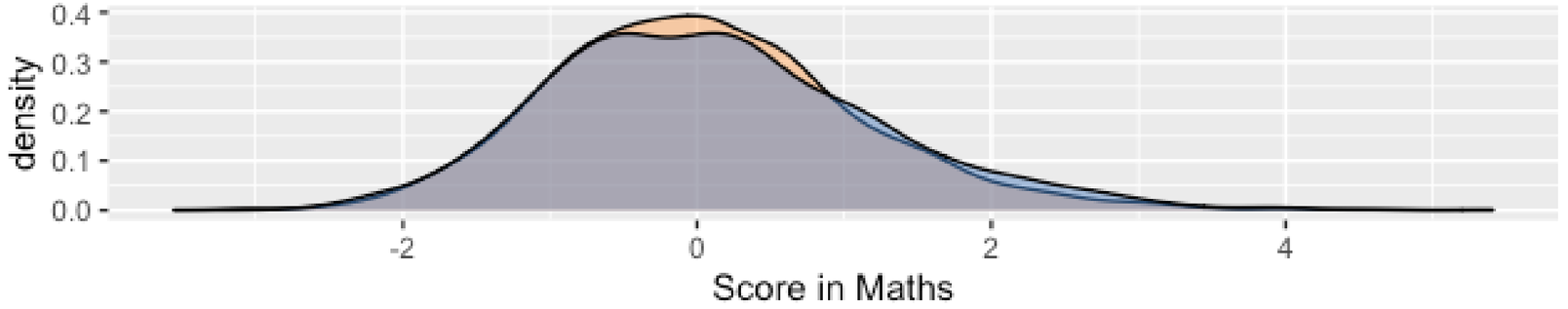
Share of students by Level in Maths (%)



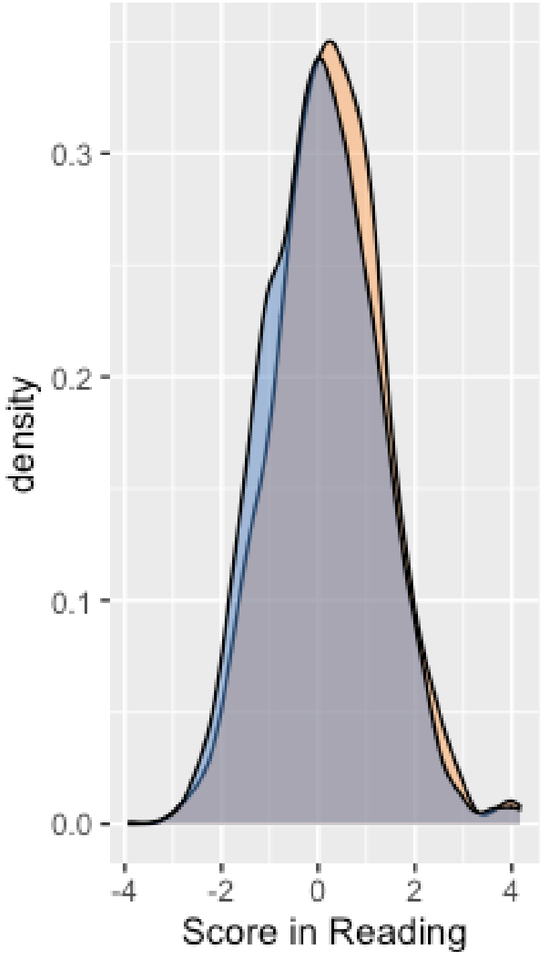
34.9% 33.8% 31.3%

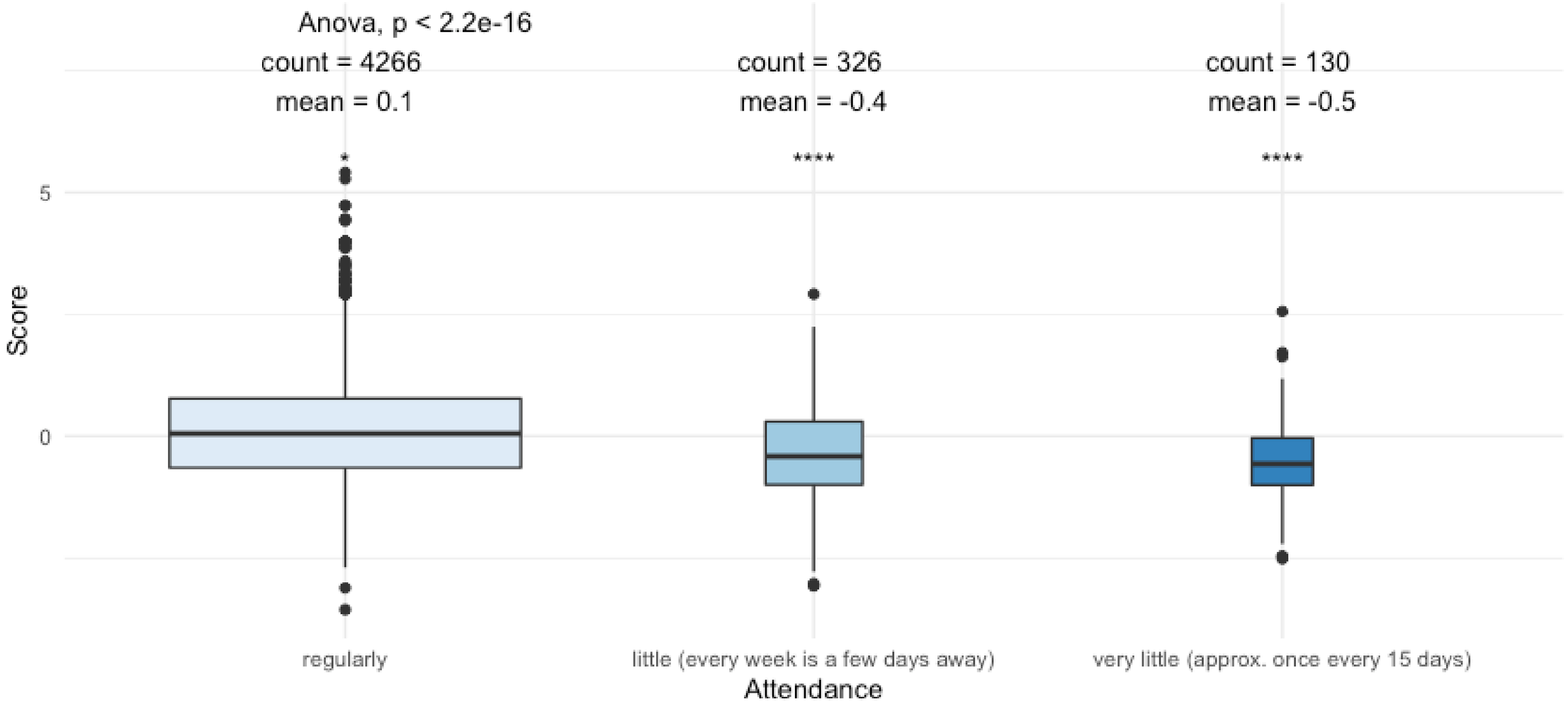

Anova, p < 2.2e-16
count = 1490
mean = 0.2
**
count = 2838
mean = 0.1
*
count = 394
mean = -0.7
****
8
4
0
-4
Score
<=11
12
>=13
Age


Anova, p < 2.2e-16
count = 1243
mean = -0.4
count = 734
mean = 0.2
count = 669
mean = 0.6
count = 61
mean = -0.1
count = 592
mean = 0.1
count = 44
mean = 0.4
count = 1379
mean = 0.1
****
*
****
*
*
5
0
Score
Aprender
PR-HP
PRIVADO
RURAL
TC
Tiempo Extendido
UC
Centre type


density
0.4
0.3
0.2
0.1
0.0
-2
0
2
4
Score in Maths


Gender
F
M
Score in Reading
4
2
0
-2
-4
-2
0
2
4
Score in Maths


density
0.3
0.2
0.1
0.0
-4
-2
0
2
4
Score in Reading


Anova, p < 2.2e-16
count = 4266
mean = 0.1
count = 326
mean = -0.4
****
count = 130
mean = -0.5
****
Score
5
0
regularly
little (every week is a few days away)
very little (approx. once every 15 days)
Attendance

— 08

Part A: Supervised analysis

LDA

- Response variable: *Level_MAT.*
- 4.722 observations and 181 variables, after multicollinearity analysis.

Decision trees

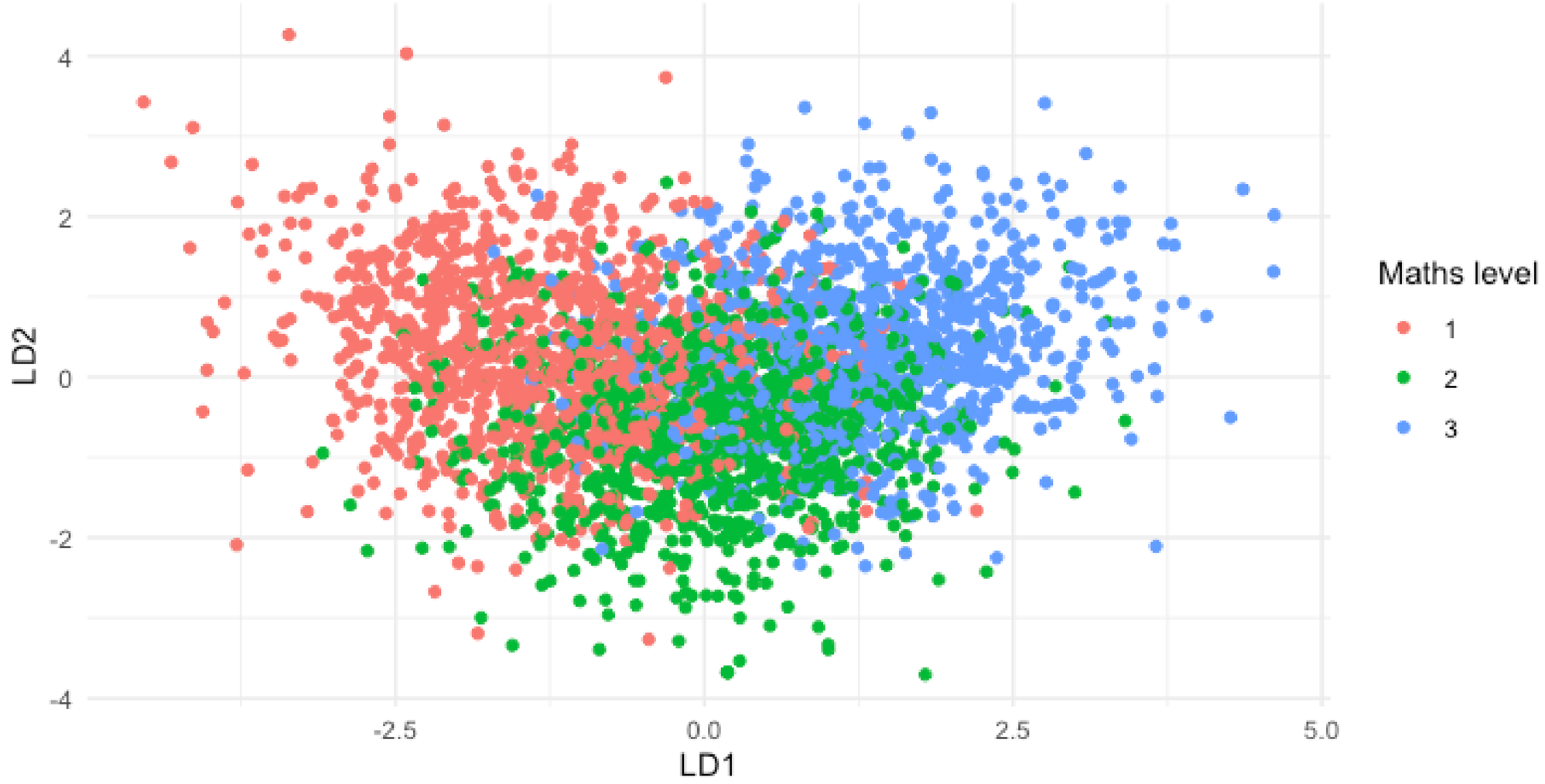
- Response variable: *Level_MAT.*
- 4.722 observations and 194 variables.
- Not affected by the presence of multicollinearity.

Random forest

- Response variable: *Level_MAT.*
- 4.722 observations and 194 variables.
- Not affected by the presence of multicollinearity.

LDA main results

1. We get **two** linear discriminants (LD) that explains 83.9% and 16.1% of the between-group variance in the dataset, respectively.
2. Having a **high Reading test** score and belonging to an **"Extended time" school** has a positive influence on the student's performance in Maths, while having repeated several times has a negative effect, as expected.
3. 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.
4. The model has a prediction accuracy of 56.2%

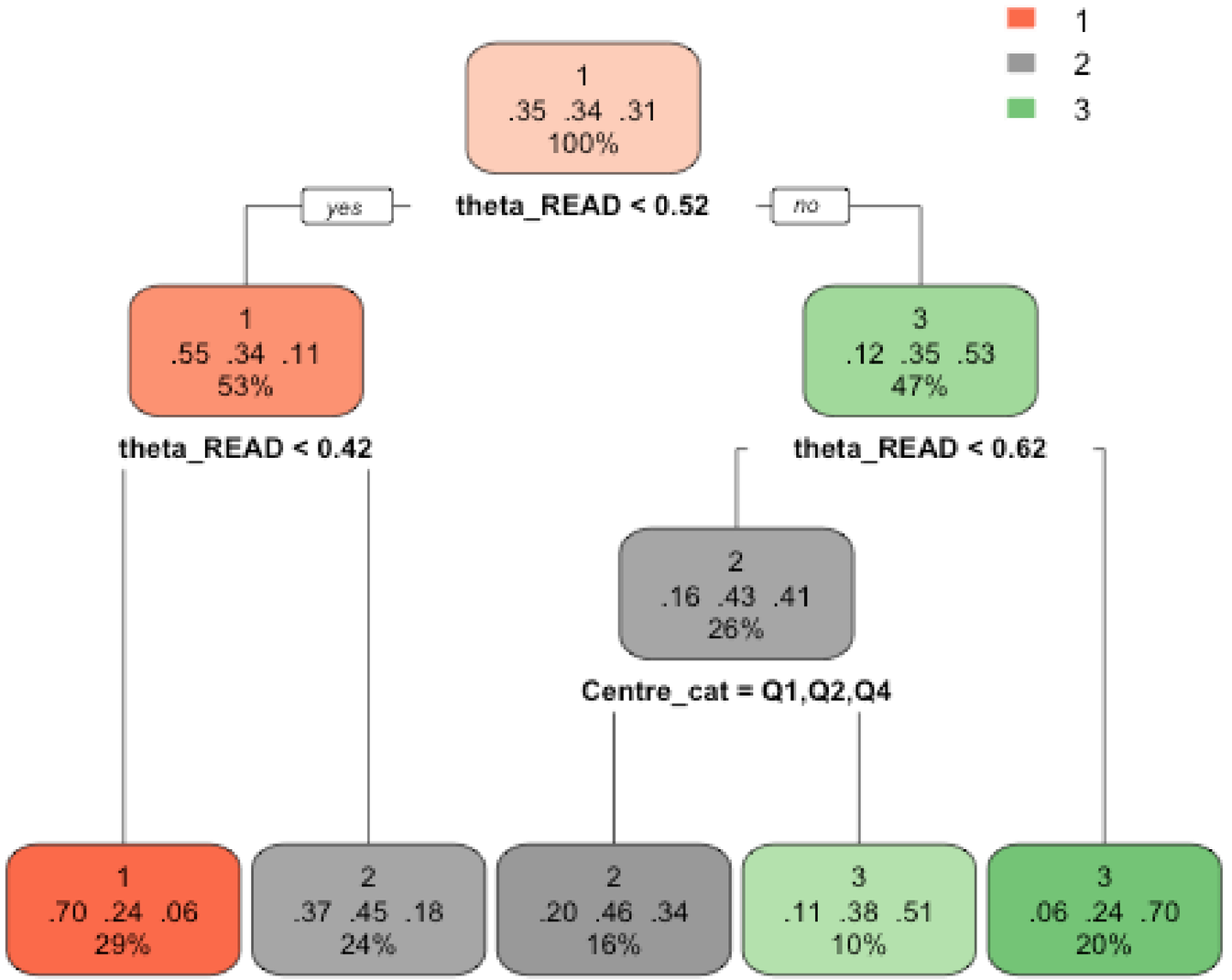


The first linear discriminant separates the three levels of Maths score while the second one distinguish mostly the extreme levels (1 and 3) from level 2.

Decision trees main results

1. The first decision tree (size 3) has the following characteristics:
    a. 4 decision splits along values of *theta_READ* and *Centre_cat*
    b. 5 prediction regions
    c. *Context* and some socio emotional variables were also considered to split the nodes.
    d. The model has a prediction accuracy of 56.2% and a validation error of 67%.

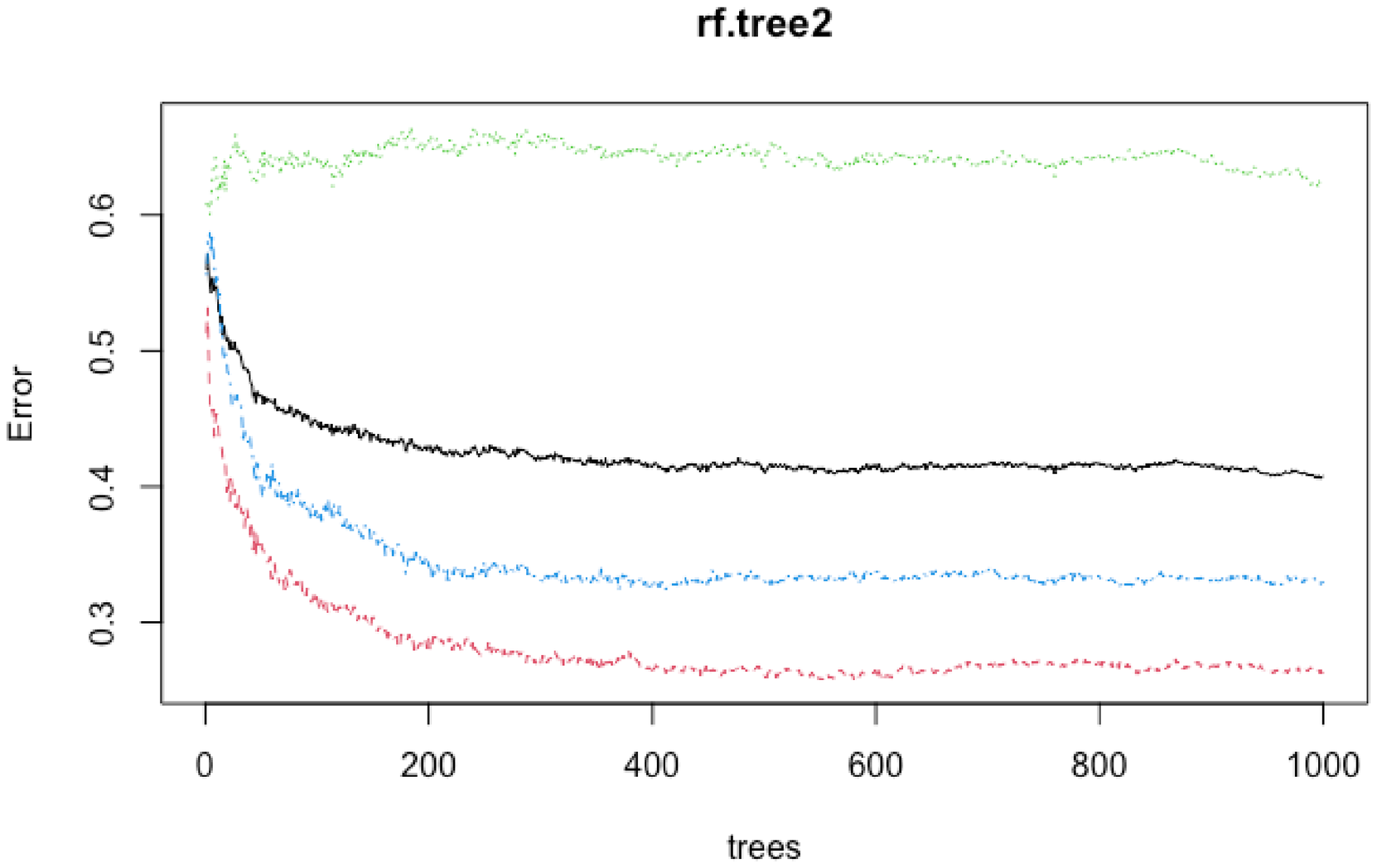
2. By incrementing the size of the tree to 7, the cross validation error lowered a bit (66.4%) and the accuracy was improved (62.1%).
    a. students who strongly agree that “A good student does not need to work hard to do well in school” are predicted to belong to the lower level in Maths.
    b. students that said that sometimes “We all made the class rules” are predicted to have the highest level score (level 3).



Decision tree with size = 3

Random forest main results

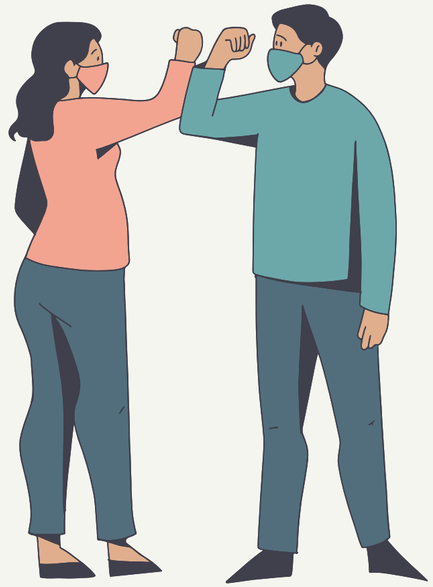
- The model fitted with 500 trees has an error rate of 41.9%.
- **Most important features**: theta_READ, Context, Centre_Cat, “A good student does not need to work hard to do well in school”, “Thoughts about repetition”, “Expectations about the school trajectory”, “Number of people living in the household”, and Centre_type.
- By making predictions over the test set, we get an accuracy of 60.5%.
- The results do not make a big improvement in terms of accuracy if we increase the number of trees (1.000) or change the number of variables randomly sample (from 13 to 20).
- A Random Forest including only the non rejected variables by the feature selection algorithm Boruta, found that theta_READ, Centre_Cat, Context, “Thoughts about repetition”, “A good student does not need to work hard to do well in school”, “Expectations about the school trajectory”, and “Reading on weekends” are among the most relevant features.



Error by number of trees

Conclusions

- Reading test score is the most important variable to explain Level_MAT according to all the models fitted.
- Variables related to the socio economic and cultural context of the school (Centre_cat, Context) are also relevant.
- Socio emotional variables that measure the student feelings are related to the level in Maths.
- As expected, repetition is another key factor that can explain the student's performance.
- Relevance of variables related to the COVID-19: capacity of the student to do homework, student-teacher relationship and school support activities participation.



Part B: Unsupervised analysis

PCA

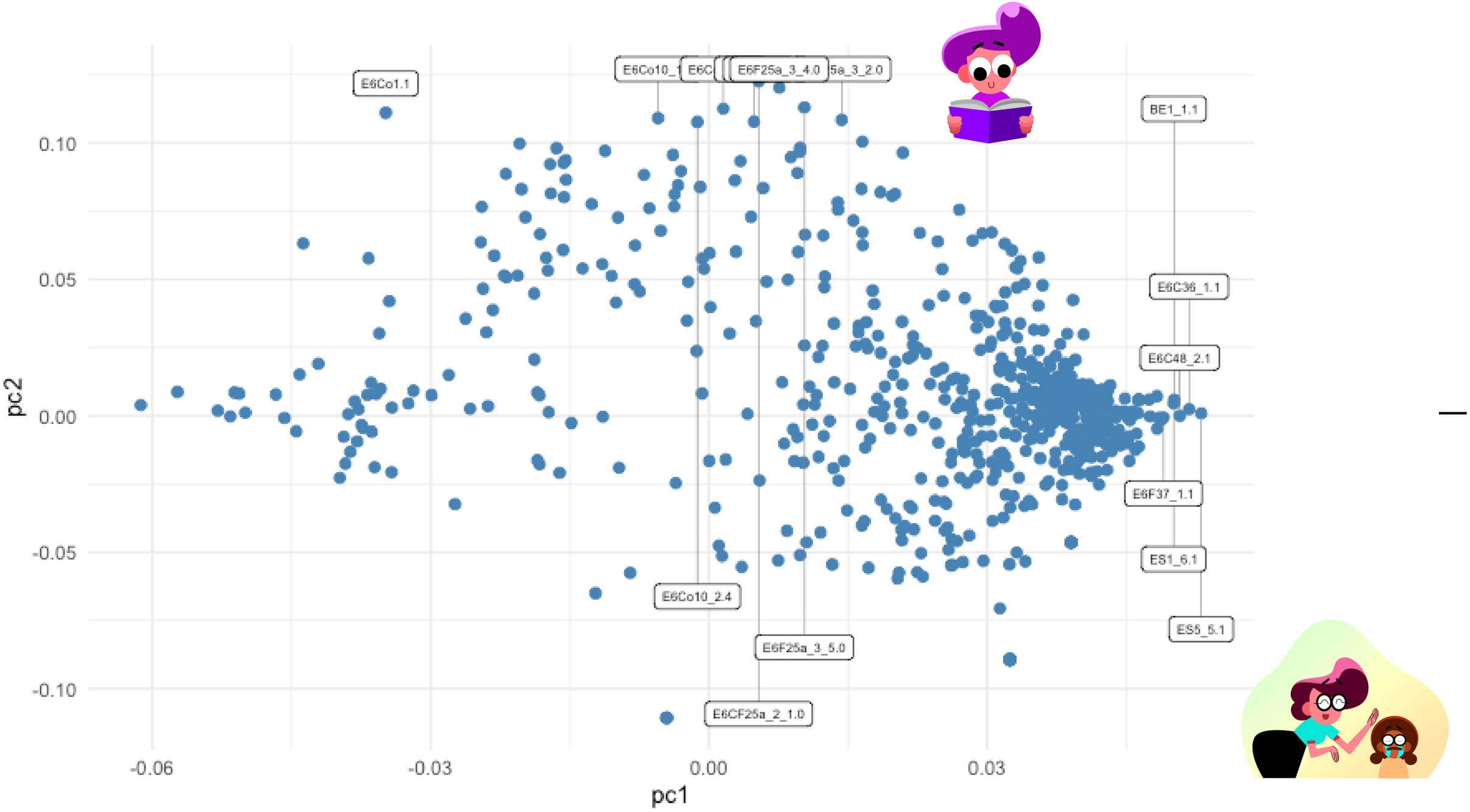
- Identify variables that better describe all the features related to the students.
- *Level_MAT* as another feature.
- *Generalized Low Rank Models* optimization.

k-means

- Classify students into mutually exclusive groups or clusters.
- Level_MAT excluded to compare rankings.
- Gower distance measure (K-medoids) was applied.

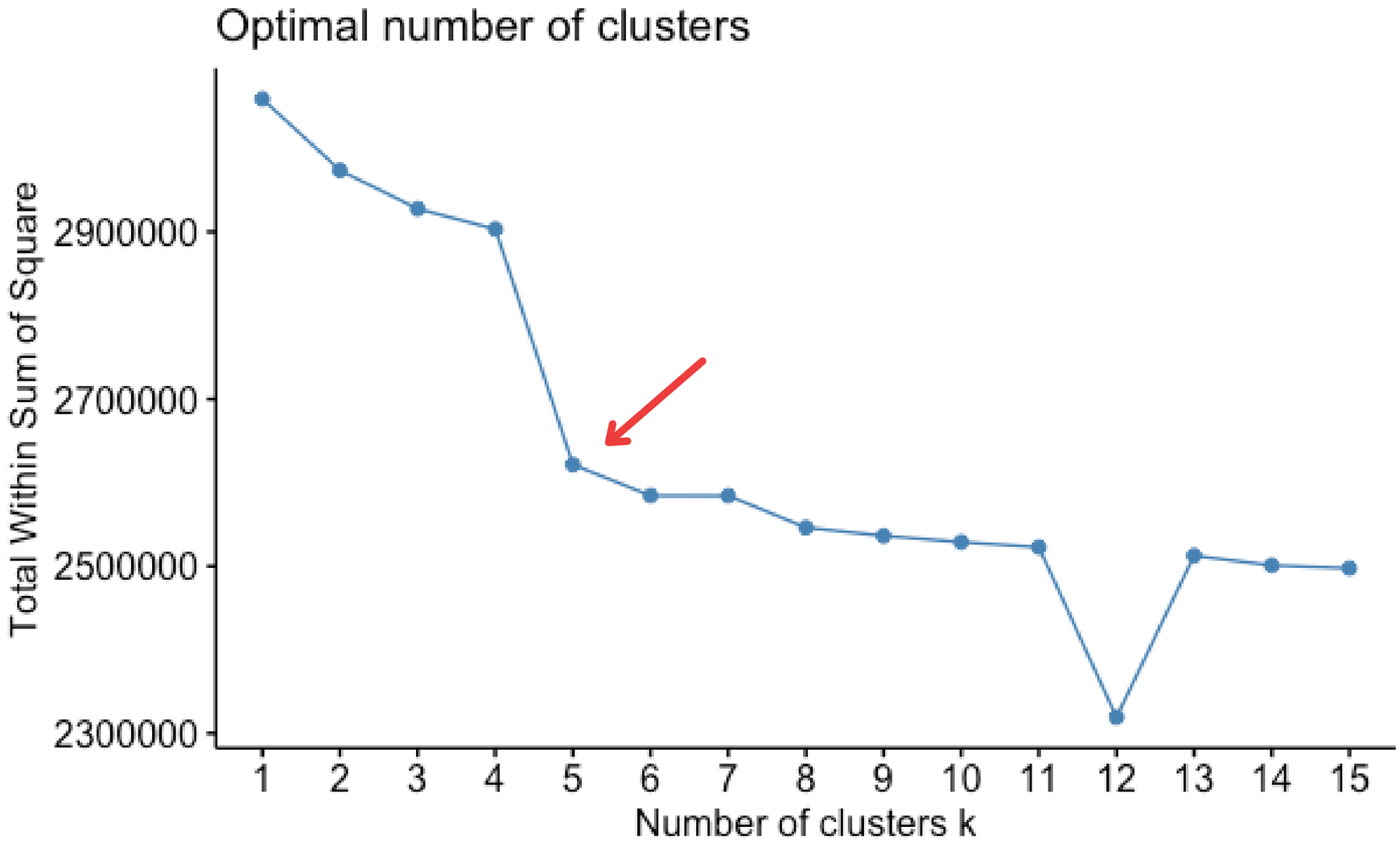
PCA main results

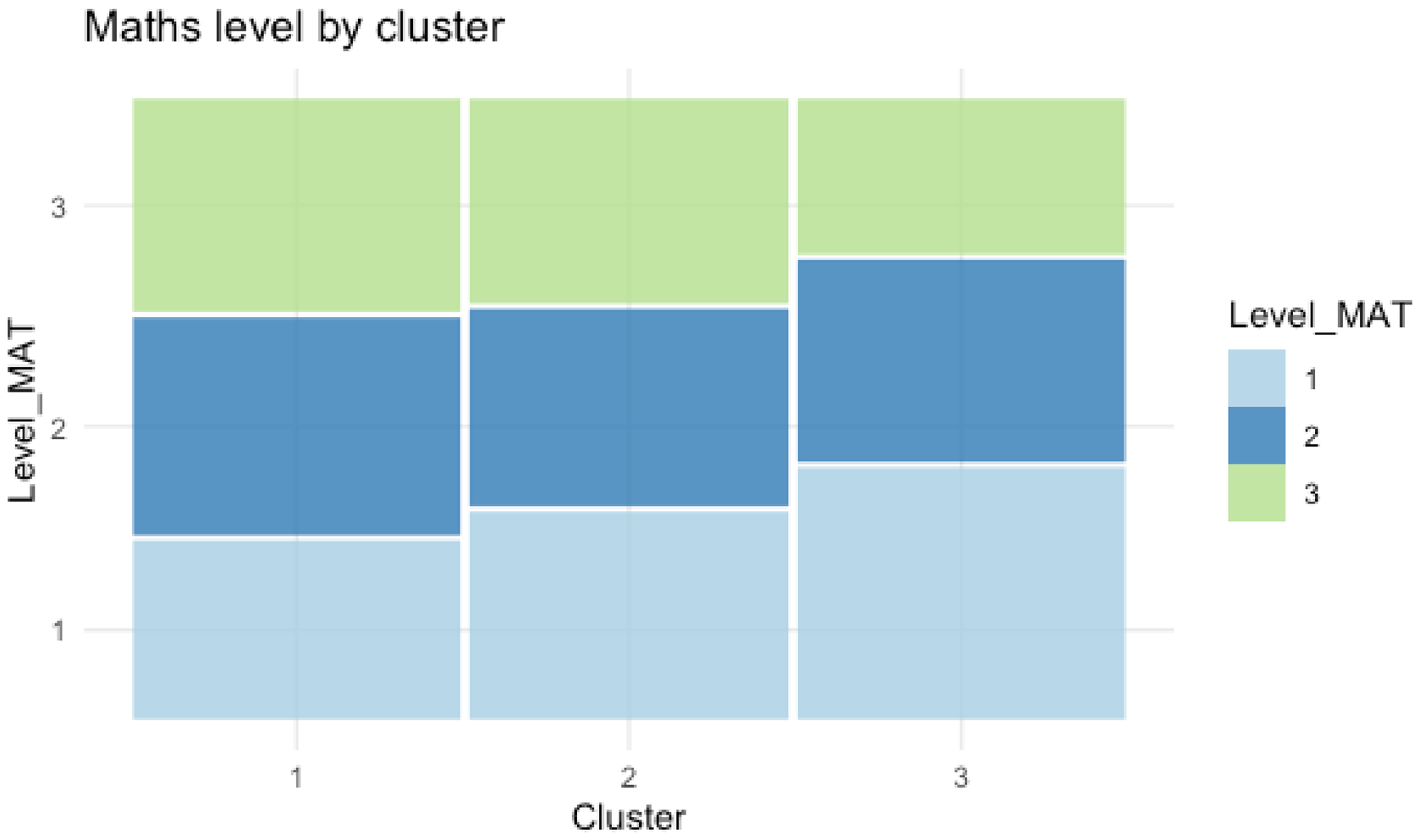
1. There is a first principal component that explains more than 75% of the variance, while the second component explains only 5.2% and the third 3.1%.
2. After using 3 criteria to choose the number of components (sum eigenvalues, PVE and "Elbow"), the first 2 PC were chosen (explain 80% of the variability in the original dataset).
3. PC1 most relevant features are related to negative aspects about the experience of the students at school.
4. PC2 most relevant features are related to having reading habits and back to school feelings after lockdown.


E6Co1.1
E6Co10_1
E6C
E6F25a_3_4.0
5a_3_2.0
BE1_1.1
E6C36_1.1
E6C48_2.1
E6F37_1.1
ES1_6.1
ES5_5.1
E6Co10_2.4
E6F25a_3_5.0
E6CF25a_2_1.0
0.10
0.05
0.00
-0.05
-0.10
pc2
-0.06
-0.03
0.00
0.03
pc1

K-means main results

1. The optimal number of clusters is 5, because the bend or "elbow" in the plot of total within-cluster sum of squares for each number of clusters (k) appears to happen when k = 5.
2. We get five balanced clusters, but the dataset is poorly structured, according to the overall average silhouette criterion.
3. Then, a 3-means clustering was fitted to be able to compare the similarity between the level in Maths test and the cluster that each student belongs to. We get three balanced clusters with an overall average silhouette also very low (0.027).
4. There is a significant relationship between the variables cluster and Level_MAT, but there is a weak association (Cramer's v).


Optimal number of clusters
Total Within Sum of Square
2900000
2700000
2500000
2300000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of clusters k



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 in the lower level.

Conclusions

- The first two principal components explains a great proportion of the variability in the data, more than 80%.
- 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.
- More than 40% of the students with the worst performance in Maths were grouped in the first cluster.

Thank you.

AUTHOR

MATHIAS CARDARELLO FIERRO

mathias.cardarellofierro@studenti.unimi.it

Student number: 963346

All the files can be easily accesed from this repository.