Learning analytics: an investigation on the influence of school quality in overcoming social inequalities

FINAL REPORT

Nonparametric Statistics course – A.Y. 2020-21

Federico Capello (10564653) – Federico Fatone (10530963) – Filippo Fedeli (10534669) – Gianmarco Genalti (10560999)
MSc Mathematical Engineering – Statistical Learning Major – Politecnico di Milano

POLITECNICO
MILANO 1863

INVALSI

1. INTRODUCTION, DATASET AND EXPLORATORY ANALYSIS

1.1 Preliminaries

The purpose of this work is to investigate **at school level** the main factors affecting the performance in the INVALSI tests at 8th grade, providing actionable insights and instruments to the School Inspectorate (*provveditorato*).
In Italy, the Italian Institute for the Evaluation of Educational System (INVALSI), assesses students' abilities in **reading** and **mathematics** at different stages. This report focuses on data collected in 2013 on the standardized tests at the end of the first and third year of lower secondary school (**6th grade** and **8th grade** in international notation), where students are requested to answer questions with both multiple choices and open-ended questions, that test their ability in reading and mathematics. In addition, they are requested to compile a questionnaire about their **parents' educational level** and their **socio-economic** situation with the aim of building an indicator about their background (namely ESCS [1]; Economic, Social and Cultural Status).
Our data is completed by the results of a questionnaire filled out by **principals** about themselves, their management style, infrastructures and school environment.

The work is organized as follows: Section 1 presents the dataset and the preprocessing necessary to effectively work at school level; in Section 2 the main differences between Northern and Southern Italy are explored; in Section 3 we provide some insights to school management through nonparametric hypothesis testing; in Section 4 an insightful prediction model for the INVALSI absolute score and improvement in scores is developed through nonparametric regression; Section 5 contains discussion and conclusions.

1.2 School Level Aggregation Process and Feature Engineering

As previously stated, our main focus was on school level analysis. This required a method to group the data starting from a student level, while losing the least amount of information as possible. After a first data cleaning, where we fixed some corrupted text data and converted each feature to its correct data type, we proceeded by subdividing the features in two types: the features that are yet **school-wise**, like the management's answers to the questionnaire and some school-related information, and the features that are **student-related**. The first type features have been simply grouped taking the first apparition in the student-level dataset for each school, the latter type instead has been grouped using **multiple statistics** describing as much as possible their behavior for each school: we decided to bring on a school level the mean, the standard deviation, the skewness, the minimum, the maximum, and the main quantiles of each numerical feature. But we were still not finished, as this approach still lacked a school-level transposition of student-level categorical features; moreover, many categorical school-level features are very granular and difficult to be used in models. For these reasons, we decided to provide a series of **indices** grasping, in a quantitative way, information coming from such features.

1.2.1 Indices from Students-Level Data

prop_maschi Proportion of male students
*prop_laurea_padre/*madre Proportion of male/female parent with university degree.
prop_diploma_padre/madre Proportion of male/female parents with high school diploma.
prop_padre/madre_ita Proportion of male/female parents being Italian citizens.
prop_padre/madre_disocc Proportion of male/female parents being unemployed.
prop_bocciati Proportion of students being 1+ years late in their school path.

1.2.2 Indices from Management Questionnaire Answers

For each question, we aggregated the answers in a custom fashion, depending on the presence of sub-questions to be answered. For "ordinal" and "binary" answers we encoded them with integers, taking the mean whenever more sub-questions are present. Few questions have been condensed in indices in a custom way, which will be specified.
opinione_invalsi (D1) How good is the opinion of the principal on INVALSI.
program.*utilizzo_invalsi* (D2) How much the principal declares to use INVALSI outcomes in school management.
discussi_insegnanti (D4_a) If the INVALSI outcomes are discussed with teachers.
discussi_genitori (D4_f) If the INVALSI outcomes are discussed with parents.
sodd_pon (D6) For Pon schools only, if the principal is satisfied with the program.
attivita_preside (D11) A score quantifying how much the principal is involved in school dynamics.
pressioni_genitori (D15) If most of the parents puts pressure on the school about performances.
coinvolgimento_genitori_prop (D16) How much the principal tries to involve parents in school dynamics.
opinione_associazioni_genitori (D17) What is the opinion of principal about parents' associations.
coinvolgimento_genitori_eff (D18) How much the parents are effectively involved in school dynamics.
condotta_studenti (D19) A score quantifying the number of misbehaviors coming from students.
registro_elettronico (D20) If a digital record of students' activities is present.
numero_plessi (D21) The number of buildings forming the school.
infrastrutture (D22) A score quantifying the state of school's infrastructures.

strumenti (D23) The number and the quality of instruments available for students at school.
preside_maschio (D24) If the principal is male.
eta_preside (D25) The age of principal computed as 2014.

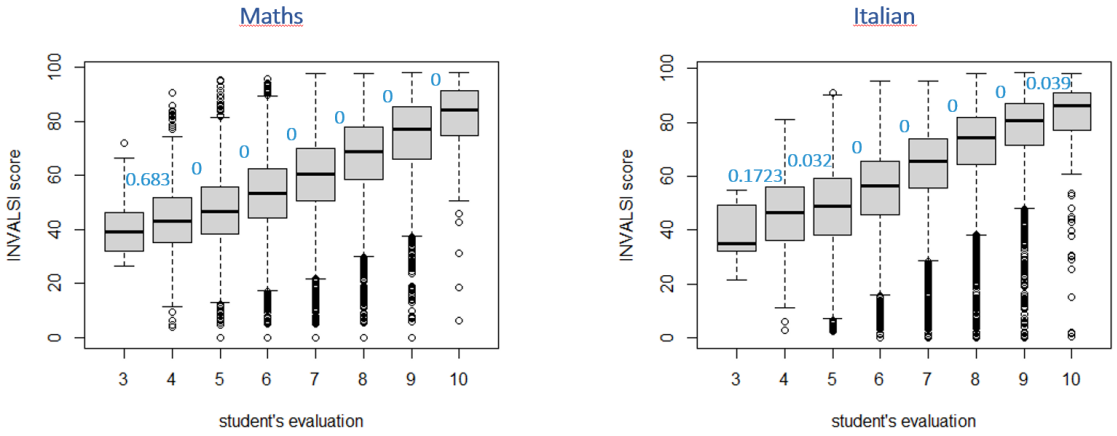
1.3 Student Level Analysis

1.3.1 Introduction

As part of the exploratory analysis and in order to confirm some well-known results in literature (Masci, De Witte, Agasisti – 2016), we performed some **Student Level analysis**. In particular we applied **nonparametric regression** and **testing** to spot differences among different categories of students. We report here some of the most interesting insights, for more results and a deeper analysis, please refer to our GitHub folder ('Student_Level_Exploratory.R').

1.3.2 INVALSI Scores and Students' evaluations

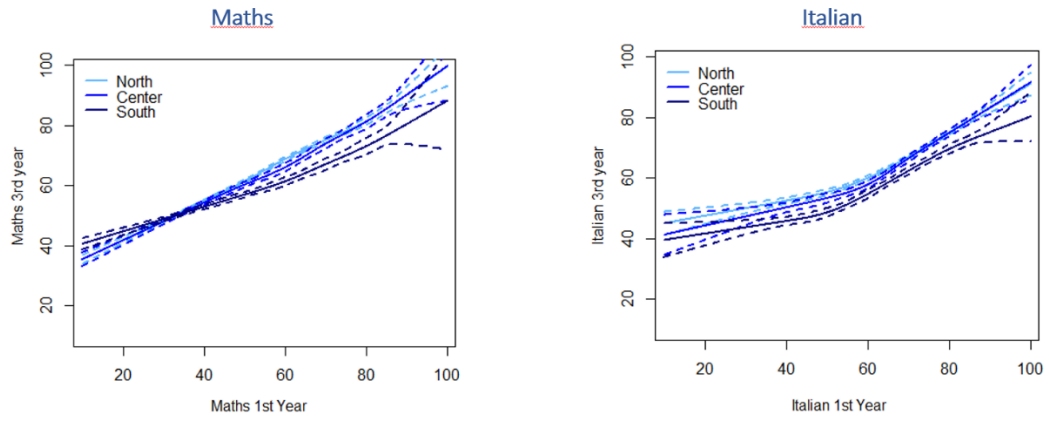
First of all we considered the outcomes of INVALSI tests at a student level together with the students' evaluations given by their teachers. We found both for Maths and Italian a **significant difference** (confirmed by the permutational coupled ANOVA p-values, reported in the image) in INVALSI outcome between students having different school evaluations.



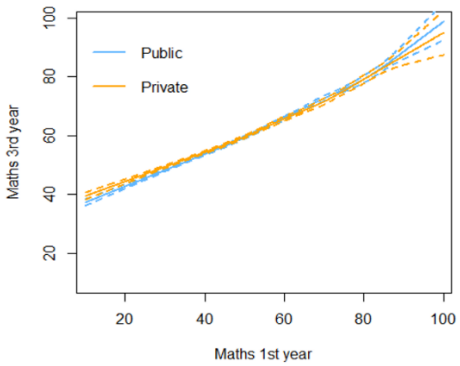
This confirms the goodness of the INVALSI tests scores in getting the performance at student level.

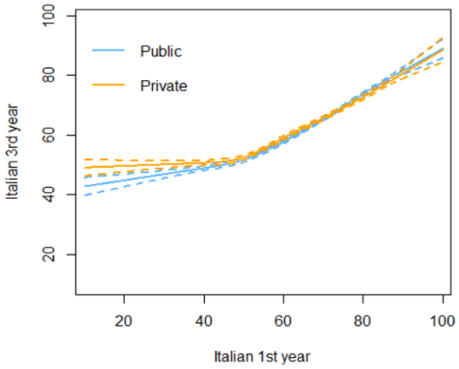
1.3.3 A Study on the impact of different factors on INVALSI Score improvement

We then focused on the **identification of the factors driving the improvement** of students starting from their first-year result in Maths and Italian, using **natural splines regression** (2 standard deviations confidence lines can be found in the plots).
We started considering a possible **geographical difference** (well documented in literature), finding disconcerting results: there seems to be no difference in the improvement trend regarding its slope, but there's a clear geographical shift affecting it, thus meaning that for the only fact that a student is born in the South, after three years he/she would lose in mean 4.43% of INVALSI score in Maths and 6.73% of INVALSI score in Italian w.r.t. a student starting from the same first year result, but living in Northern Italy. This **affects mainly students getting intermediate results**, representing an even more serious problem, as they are the majority and those who could actually drive a general improving trend in basic skills, as it is desired looking at the Italian position in international rankings.



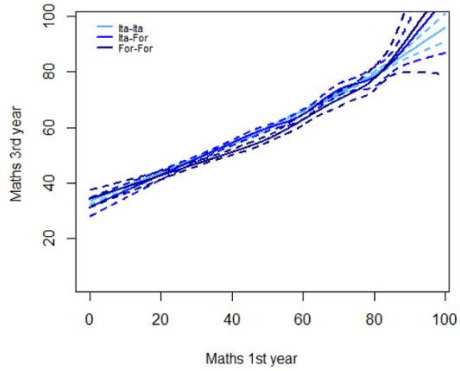
Other meaningful differences in improvements patterns can be found between **public** and **private** schools:

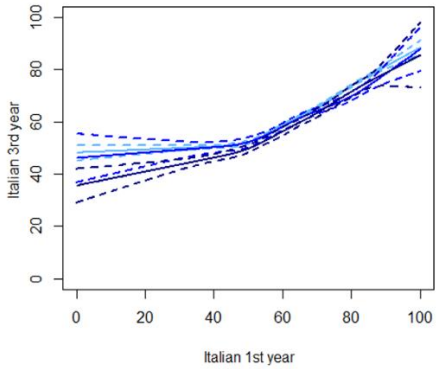




Where we find a significance difference for improvement only for the worst-performing students. This can be justified with the larger amounts of resources private schools can dedicate to students needing special attention.

Also, unfortunately, there's a difference in the improvement of students with **both Italian parents** with respect to those having **both immigrant parents**.



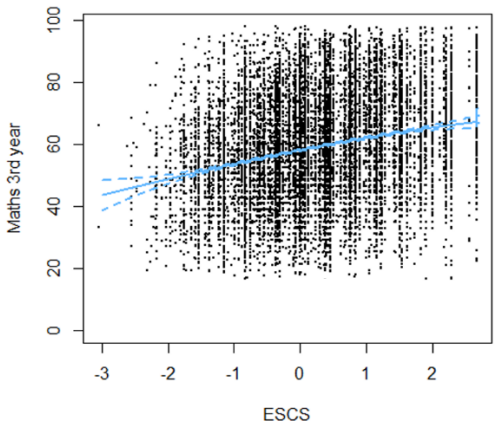


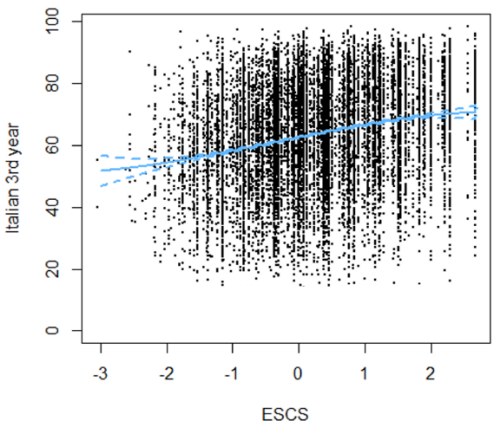
What is most worrying about it is that students with both foreigner parents (in particular the worst-performing ones, likely to be the first-generation new Italians) improve less in Italian w.r.t. Italian students by attending Italian middle schools, thus showing their **poor inclusivity performance**.

We performed a certain amount of other tests, some of which significant, please refer to the attached codes to see them in detail.

1.3.4 Socio-Economic Status of students and INVALSI performance

We focused then on **social differences** between students, finding the following trend explaining Italian and Maths performances for students according to their families' **ESCS score**.





There's a clear positive correlation between higher ESCS score (taking into account the family's economic status, the culture of parents, their working status and other socio-economic indicators) and the INVALSI score. We performed a bootstrap test for the

significance of the ESCS factor for linearly predicting the outcome of the INVALSI score finding a 0 p-value and [3.749, 4.406] RPI for Maths and [3.584, 4.202] for Italian, thus meaning that the improvement of 1 point in ESCS score for the family corresponds to a ~4 points improvement in INVALSI score.

1.3.4 Other findings

Using the same nonparametric regression approach, we found that in **bigger schools**, students with the same ESCS perform better (confirming the results in Masci, De Witte, Agasisti – 2016 from another perspective). This may be due to socio-economic factors not taken into account by the ESCS score, for example the fact that bigger schools are usually located in bigger cities, and living in a bigger city is often associated with a higher socioeconomic status, but the dimension of the city of residence is not considered in the ESCS score computation [1].

Furthermore, we found that the students having **ESCS score under the median** perform worse in **private schools** w.r.t. **public schools**, thus debunking (in the middle school context) the myth of private schools acting as social climbing factors.

2. DIFFERENCES BETWEEN NORTHERN AND SOUTHERN ITALY

We now outline **key differences** between the **North and the Center-South of Italy**. These structural divergences might be the underlying explanation of the link between factors and outcomes (as we will describe in section 3). It is thus of great importance considering these when performing further analysis.

2.1 Data and Methods

We use the school-level dataset described in 1.2; we consider the following as possible targets of school outcomes:

- *mate8_mean* (Grade in Mathematics - INVALSI - in the 8th grade)
- *ita8_mean* (Grade in Italian - INVALSI - in the 8th grade)
- *mate8_std* (Standard deviation of the grade in Mathematics - INVALSI - in the 8th grade)
- *ita8_std* (Standard deviation of the grade in Italian - INVALSI - in the 8th grade)
- *improv_mate_rapp_mean* (Mean average improvement, as ratio of grades in Mathematics - INVALSI in 8th grade and 6th grade)
- *improv_ita_rapp_mean* (Mean average improvement, as ratio of grades in Italian - INVALSI in 8th grade and 6th grade)
- *improv_mate_rapp_std* (Standard deviation of average improvement, as ratio, of grades in Mathematics - INVALSI in 8th grade and 6th grade)
- *improv_ita_rapp_std* (Standard deviation of average improvement, as ratio, of grades in Italian - INVALSI in 8th grade and 6th grade)

As possible factors we first consider the indices presented in section 1.2.2.

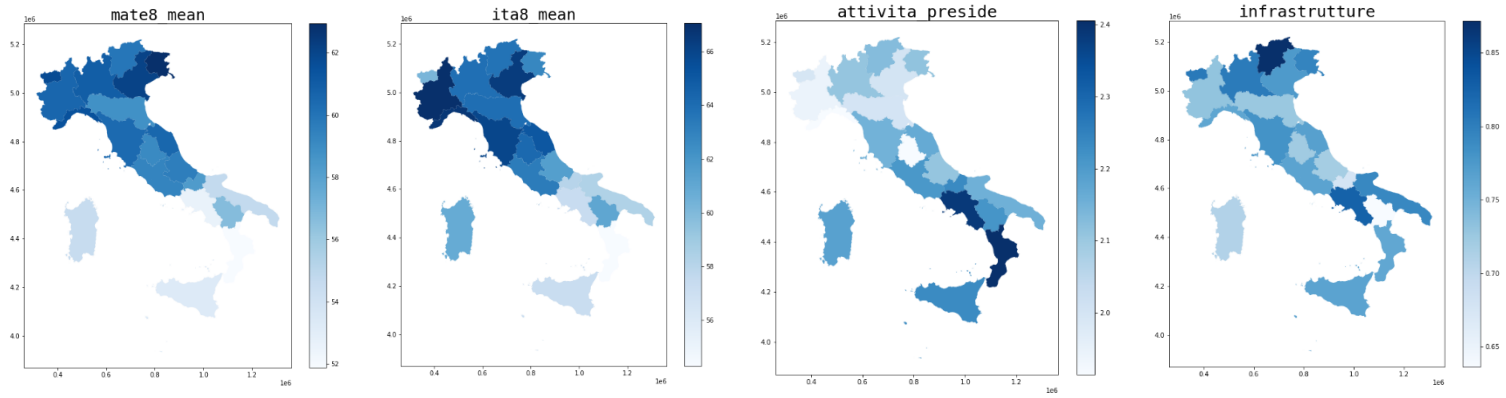
We employ the Mann-Whitney U-Test (MWUT) with 10^4 simulations to find possible significant differences between the North and the Center-South of Italy.

2.2 Results

The following tables and maps summarize the geographical differences of outcomes, factors and indices:

GEOGRAPHICAL DIFFERENCES OF OUTCOMES		
Outcome	**Where is greater?**	**p-value**
mate8_mean	North	0
ita8_mean	North	0
mate8_std	Center-South	5e-4
ita8_std	Center-South	0.004
improv_mate_rapp_mean	Center-South	0.0428
improv_ita_rapp_mean	North	0
improv_mate_rapp_std	North	0.0332
improv_mate_rapp_std	/	0.3057

GEOGRAPHICAL DIFFERENCES OF FACTORS AND INDICES		
Indices / Factors	**Where is greater?**	**p-value**
condotta_studenti	/	0.3572
attivita_preside	South	0
pressioni_genitori	South	0.0134
coinvolgimento_genitori_eff	/	0.248
opinione_associazioni_genitori_binary	North	0
infrastrutture	North	0.0047
eta_preside	/	0.1135
preside_maschio	/	0.1511
utilizzo_invalsi	North	0.0865
coinvolgimento_genitori_prop	South	0
strumenti	North	0.0029
strumenti_binary	North	0.0347



2.3 Differences in INVALSI scores

After having found such a significant difference between differently located schools, we were interested in characterizing with a deeper detail level the **difference in INVALSI outcome among Northern, Central and Southern schools**. In particular, we considered depth measures for (mate8_mean, ita8_mean) joint distribution for differently located schools, finding some interesting results.

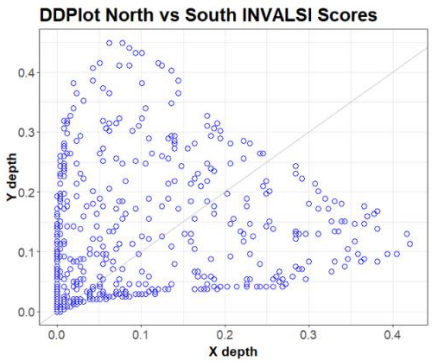
We first report the location (**Tukey Median**) and the **scale estimate** ($\frac{1}{n}(x-\nu)(x-\nu)^T$) for North, Center and South Italy:

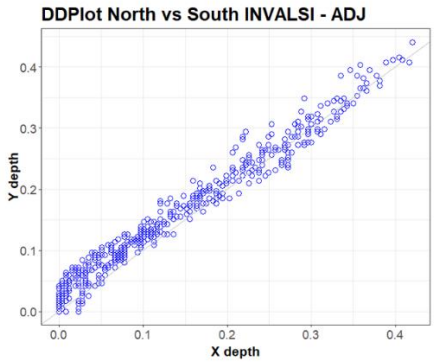
LOCATION ESTIMATES		
North (Maths, Italian)	**Center (Maths, Italian)**	**South (Maths, Italian)**
(61.44593, 67.25585)	(59.78921, 64.86440)	(53.55620, 59.88419)

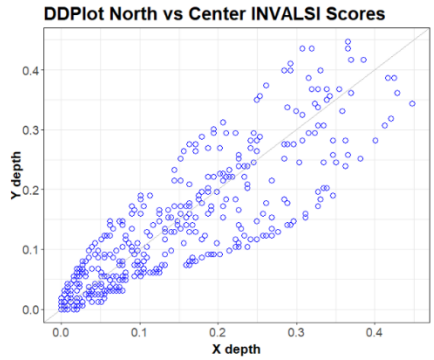
SCALE ESTIMATES		
North (Maths, Italian)	**Center (Maths, Italian)**	**South (Maths, Italian)**
$\begin{bmatrix} 47.250753 & -6.522819 \\ -6.522819 & 97.632972 \end{bmatrix}$	$\begin{bmatrix} 77.0618633 & 0.9436745 \\ 0.9436745 & 43.4245738 \end{bmatrix}$	$\begin{bmatrix} 179.28904 & 68.89979 \\ 68.89979 & 174.30175 \end{bmatrix}$

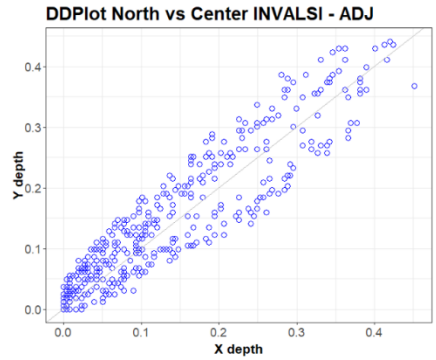
It's clearly possible to see the major difference in terms both of location and scale for different regions. With a general negative trend as the latitude decreases.

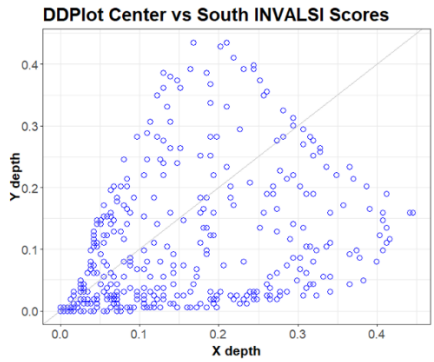
We then considered the joint distributions with DDPlots to better characterize the differences, both in the standard sense and considering location and scale adjustments:

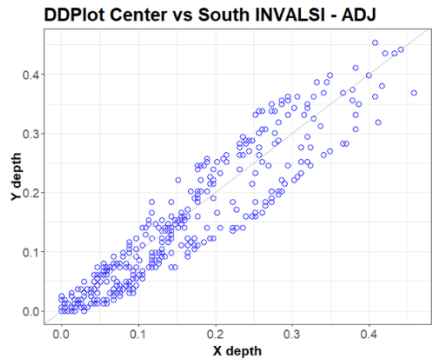












We see that after the adjustments (Location and Scale for North-South and Center-South and only Location for North-Center), the distributions are comparable, thus suggesting a **difference only affecting the parameters of the underlying distribution and not the distribution itself**.

2.4 Summary of Major Differences

Generally, mean grades for both Maths and Italian are higher in the North. Also, improvement in Italian is higher in the North, whereas for Maths it is higher in the South. Generally, **absolute outcomes are better in the North**. Regarding indices and factors, we highlight **more self-reported activity from principals in the South**, while for more **tangible indices** (instruments and infrastructure), we found statistical significance **in favor of** the richer **North**.

3. ACTIONABLE INSIGHTS FOR SCHOOL MANAGEMENT

Continuing with our goal of offering **actionable insights** to school principals and supervisors, we now focus on finding **possible connections between managerial practices in schools and school outcomes**.

3.1 Data and Methods

We employ the school-level dataset described in 1.2; we consider as targets the school outcomes and as factors the indices described in section 2.1.

As methods to connect factors to outcomes we use:

- For continuous-continuous data: residual bootstrap estimation of the slope coefficient of linear regression, by means of **reverse percentile intervals** with 10^4 simulations (BRPI).
- For univariate continuous-binary data: **univariate permutational tests** with absolute difference of the means of the groups as test statistics (UPT) then refined by a **Mann-Whitney U-Test** (MWUT). Both with 10^4 simulations.
- For multivariate continuous-binary data: **permutational MANOVA** with the Wilk's lambda as test statistics with 10^4 simulations (PMAN).

We investigate factor-outcomes first on a national level, then, in light of the findings at 2.2 on the inhomogeneity between regions, we focus only on the Northern schools.

Afterwards, we try to see if we can pick specific factors among the specific answers to the principals' questionnaire. Again, keeping in mind the results of section 2.3, uncovering the different user biases of principals in Northern and Southern schools, we debias and **investigate principals' answers to question D11**.
After removing user bias from each of the 14 sub-questions, we try to correlate these new debiased items to *improv_mate_rapp_mean* and mate8_mean for the whole Italy and for Northern Italy only.

3.2 Principal Debias Procedure

We describe the debiasing procedure of question D11: "***In riferimento a questa scuola, qui di seguito sono enunciate alcune attività che possono riferirsi al Suo modo di dirigerla. Indichi, per favore, la frequenza con cui si sono verificate nel corrente anno scolastico***".
We start by numerically encoding answers for each of the 14 sub-questions with 1 for "*Mai o quasi mai*", 2 for "*Qualche volta*", 3 for "*Spesso*", 4 for "*Sempre o quasi sempre*". We compute the principal's bias at the average of the scores across the 14 items. Then, for each principal and item **we subtract their user bias from the actual score**.
We can interpret these de-biased scores as numbers that indicate the **relative strengths/weaknesses of the management practices of each principal** (e.g.: a positive score on a certain item means that the principal is particularly keen on that practice, while a negative one means that, relatively to other practices, it's not at the top of their to-do list).

3.3 Experiments and Results Selection Method

We ran 260 experiments, according to the specifics described above. Of these, 29 showed statistical significance (p-value ≤ 0.05) and were selected for further investigation. We then **stress-tested each of these 29 results** in the following way:

- We investigated the relationship between the alleged significant factor and North / South to see if we had a significant geographical difference in the factor.
- We investigated if we had a significant North South difference in the outcome.
- We investigated if the factor - outcome relationship also holds in the regional settings (if we found something for the whole Italy) or for the whole Italy if we found something at the regional setting.

From these 29 preliminary results, we removed those that, for example, showed a significant difference of the factor for North / South but didn't within the 3 regions. In the end, after a further grouping, we are presenting **7 results** in the following sections.

3.4 Results

3.4.1 Does the principals' activity score index (*attivita_preside*) impact the variance of Maths scores?

We found statistical significance that **more activity** actually **lowers** the **variance** for Math scores, nationally (p-value = 0.01 with BRPI). This could be hindered by an inter-regional effect, because in a North vs Center-South experiment we found more self-reported activity in the south than in the north and more variance for Math scores in the North than Center-South. However, **this also holds in the Center region** (p-value = 0.02 with BRPI).

3.4.2 Does a male principal (*preside_maschio*) impact on the variance of Maths and Italian scores?

We found statistical significance that a **male principal increases** the **variance** of the scores in both Italian and mathematics, at a national level (p-value = 0.0028 for Italian, p-value = 0.006 for Maths with MWU). There was no significant difference between North and Center-South for principal male proportions. We also found that the results also hold in the South for Italian (p-value = 0.0063 with MWU) and in the North for Maths (p-value = 0.003 with MWU). This was also confirmed by a PMAN for both standard deviations of Maths and Italian (p-value = 0.0075).

3.4.3 Does the parents' involvement (*coinvolgimento_genitori_prop*) impact on the variance of Maths scores?

It does, negatively, on a national level (p-value = 0.01 with BRPI). No statistical significance for the variance of Italian scores. However, there is a significant geographical difference of both variance and parents' involvement. Still, we found a significant **negative effect in the Center** (p-value = 0.05 for Maths with BRPI).

3.4.4 Does pressure from parents (*pressioni_genitori*), as seen from principals, impact on Maths scores?

Yes, for Maths, **positively in the North** (p-value = 0.0201 with MWU), and not at a national level. Additionally, a MPAN test on average Maths and Italian scores confirm significant differences both in the North (p-value = 0.0025) and in the South (p-value = 0.1038).

3.4.5 Does the principals' usage of INVALSI scores (*utilizzo_invalsi*) impact on Maths scores?

Yes, for Maths, positively, but not on a national level, **only in the North** (p-value = 0.01 with BRPI) and **in the Center** (p-value = 0.05 with BRPI). It's also interesting to notice that there is no geographical significant difference regarding *utilizzo_invalsi* between North and Center-South.

3.4.6 Does a tendency to responsibilize teachers on school objectives impact on average Maths scores?

We investigated the debiased version of question "*Parte importante del mio lavoro è quello di assicurarmi che i docenti si considerino responsabili dell'attuazione degli obiettivi della scuola*" and found an interesting **negative** correlation with average Maths scores **in the North** (p-value = 0.05 with BRPI), but not in the Center or South.

3.4.7 Does a tendency to outline the objectives of the school personnel impact on average Maths scores?

We investigated the debiased version of question "*Definisco gli obiettivi che il personale della scuola deve realizzare*" and found a **positive** statistical significance (p-value = 0.02 with BRPI) **in the North** and **negative** (p-value = 0.05 with BRPI) **in the South**. No difference in the whole Italy and no geographical difference North vs Center-South for this debiased answer.

3.5 Practical Advice to School Management

From the statistical results at point 3.4 we can further group advice in 4 categories:

1. Direct impact of principals with strong judgement:

With evidence of the direct impact of the principals' activity on Maths scores at a regional level (see 3.4.1), a negative link between the tendency of responsibilization of teachers and Maths scores (see 3.4.6) and a positive link between the tendency of setting objectives for everyone and math scores (see 3.4.7), at least in the North, we can advise the School Inspector (*provveditorato*) to promote as principals **objectives-driven people with strong judgement**, that might even lean on **autocratic tendencies**, but who seem to be helpful in a school environment.

2. Relationship with parents:

From 3.4.3 we see that parents' involvement lowers the variability of the scores, which we consider a good thing and in 3.4.4 we see a possible positive direct link between parents' pressure and absolute scores. We can therefore advise school principals to invest time in **making sure that parents are involved**, as this helps foster homogeneous classes and higher scores, at least in Maths.

3. Usage of INVALSI:

In 3.4.5 we find a positive association, twice at a regional level, between Maths scores and INVALSI usage/consideration. We advise principals to **make use of the INVALSI results**, as this might have a positive effect on school outcomes.

4. Gender differences:

Finally, in 3.4.2 we gather consistent evidence, for both Italian and Maths, that a male principal is connected with a higher variance in the test results. As a practical advice, we might suggest the School Inspectorate (*provveditorato*) to **promote** as principals **more females in schools with a greater disparity of results.**

4.1 School performance driving factors

A useful insight for the school inspectorate would be to find out which are the **schools that perform better than others** and what are the **differences** among the best performing and the worst performing ones.

In order to tackle such a general problem we first fitted a **linear mixed effect regression model**, trying to predict the result at third year by considering the one at first year and the effects given by schools, as follows:

$$score_{third\ year} \sim score_{first\ year} + school.effect\ + school.effect * score_{first\ year}$$

In this way, we achieved an adjusted R^2 of 0.5655 in Maths and 0.6015 in Italian, w.r.t. 0.4505 of the model without school effects in Maths and 0.3045 in Italian.

In order to avoid the simple identification of schools blamed of cheating thus artificially increasing the R^2 of the linear model, we decided to exclude from the pool the schools having a *fattore_correzione* smaller than 0.85 for Maths and 0.95 for Italian. In this way we were confident **not to misclassify** schools with a high **cheating penalization** with **poor performing** ones. Unfortunately, considering

Italian results, we couldn't get rid of the significance of the *fattore_correzione* factor and therefore we decided **not to include** the results for it.

We then considered **poor performing schools** the ones having a significant random effect on the model (p-value < 0.01) and in the worst 20% of the random effects values and **high performing** ones those having a significant random effect on the model (p-value < 0.01) and in the top 20% of the random effects values.
We also considered poor **performing schools w.r.t. students effects** the ones having a significant mixed effect with the students' first year scores on the model (p-value < 0.01) and in the worst 20% of the mixed effect with the students' first year score values and the **high performing** ones the same way we did for high performing schools.

After this division, we performed **permutational tests** for the difference in mean and standard deviation of the distribution of some school factors that could be related with such differences.
We found the results summarized in the following tables:

MATHS – SCHOOL RANDOM EFFECT		
Significant factor	**P-value**	**What is 'better'?**
INVALSI results discussed with teachers - mean	0.0274	Discuss less
Proportion of fathers with degree or high school diploma - mean	0.0278	Higher education
Minimum ESCS in school - mean	0.0376	Higher ESCS minimum
INVALSI results discussed with teachers - std deviation	0.0162	Less variance
Minimum ESCS in school - std deviation	0.0420	Higher variance

Besides the expected effects of socio-cultural indicators on the school, we found a significant effect on the school performance driven by the discussion of INVALSI results with the teachers, penalizing the schools that discuss more. As this feature was referred to past school president's experience, it might indicate a poor value of outcomes' discussions with teachers, suggesting to School Inspectorate that they may be **not sufficient** as a **contrast measure** for poor performances, at least for Maths results.

MATHS – SCHOOL MIXED EFFECT WITH STUDENT PERFORMANCE		
Significant factor	**P-value**	**What is 'better'?**
INVALSI results discussed with teachers	0.0264	Discuss less
Proportion of fathers with degree or high school diploma	0.0334	Higher education
Minimum ESCS in school	0.0376	Higher ESCS minimum
INVALSI results discussed with teachers - std deviation	0.0158	Less variance
Minimum ESCS in school - std deviation	0.0412	Higher variance

Mixed effects strictly resemble the random effects, showing no significant difference of school performance between the **global level** (intended as a mean outcome of all of its students) and the **student level** (intended as how it magnifies the performances of single students).

4.2 School Results Prediction

4.2.1 Purpose and description of the work

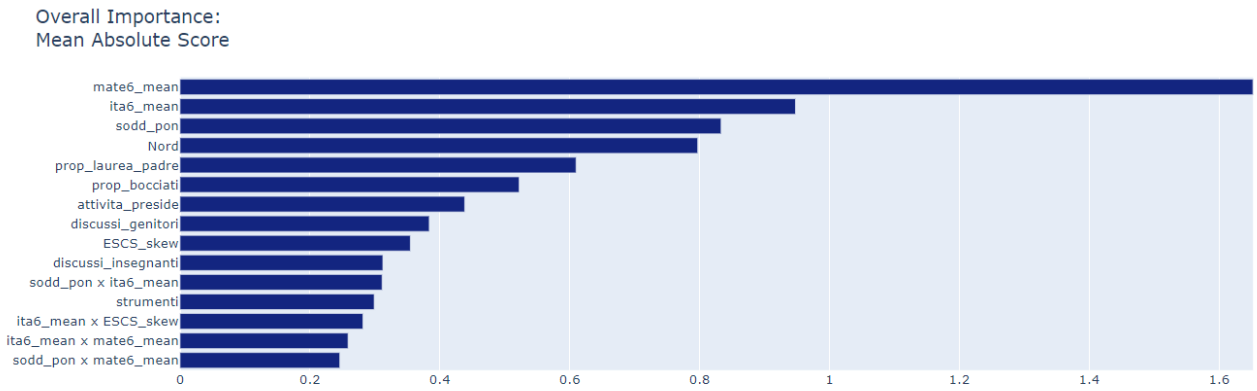
For the School Inspectorate, it would be of great interest to be able to **anticipate** the behavior of a school's cohort, in order to monitor constantly the level of the education provided in the school. It would, therefore, make sense to use the scores of the **6th grade** INVALSI tests of the specific cohort, alongside **socio-economic** and **geographic** information about the school and **principals**' characteristics to predict the expected results of this cohort at the **8th grade** INVALSI tests with two years in advance, in order to assess the critical situations and try to intervene dynamically.

In order to achieve this task, we first setup a **regression model** to try and predict *mate8_mean*.
First, we drop all the schools with *fattore_correzione* > 0.85. From now on, this will be the standard procedure when dealing with prediction for mate8 or ita8, as the impact of *fattore_correzione*, which is defined as 1 - probability of a school having cheated and computed by class (considering the school) looking at the distribution of the grades, is **huge**, as our final target scores are the raw scores multiplied by *fattore_correzione*.
Then, the observations with **missing data** for the variables we want to use in the model (usually in *mate6_mean*, *ita6_mean* and *ESCS_mean/std/skew*) are dropped, passing from 620 schools to 350-450 schools, depending on the covariates. We have chosen not to impute the missing data, even though they were missing only in few of the relevant variables, because the missing at random **(MAR) hypothesis did not seem to be verified**, as there was high correlation between data missing in different columns, which might be an indicator of a systematic procedure.

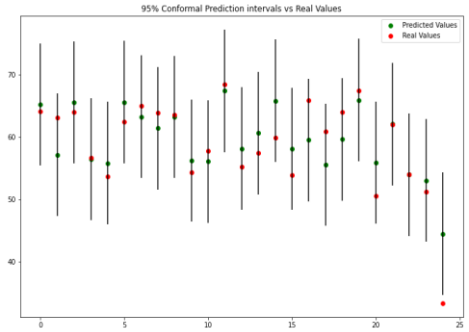
Finally, in order to achieve the task of predicting *mate8_mean*, we want to use a model which, at the same time, is powerful enough to spot some **nonlinear relationships and interactions** between variables, but **doesn't overfit** too easily and keeps a **high level of interpretability**.

After trying Generalised Additive Models (GAM), we moved to **GA2M** [2], GAM with automatic recognition of significant interactions between variables. In fact, we do see a spike in performance by inserting some **interactions** which are meaningful, as also seen in [3]. Moreover, while staying in a non-parametric framework, we move from using splines to model the nonlinear terms to using **boosted trees**, with each tree operating on a **single variable**, for the function approximation, with trees being estimated in a **round-robin** manner which forces the model to sequentially consider each variable as an explanation of the current residual rather than greedily selecting the best feature.
These kind of shape functions and the intuitive addition of interactions seem to be more suited for a problem of this kind, where variables have more of a 'stepwise' effect, which would result in oscillating terms by approximating with splines with many degrees of freedom. The resulting technique is named **EBM (explainable boosting machine)**.
[4][5]

The resulting model stays interpretable, through **exact partial dependency plots** (PDP), while maintaining an accuracy on par with **state-of-the-art** boosting techniques (and, on this dataset, with few data available, it manages to outperform consistently both GAM, because of the sheer inductive power, and boosting, as EBM tends not to overfit). After building a baseline model with 30/40 pre-selected variables, we perform variable selection through a **hill climbing wrapper** approach, while trying to lower the number of used interactions without hindering the performance. We will adopt this procedure for every model in order to have powerful and compact explainable models.
The final regression model works really well, reaching ~**0.6 R^2** in 4-folds **cross-validation**, working consistently better than the first, oversized one (~0.5), which still suffered from overfitting, and the baseline GAMs (~0.4). It does, therefore, do its job in providing an accurate pointwise estimate, while being a **glass-box**, which we can inspect by looking at the PDP both for single variables and found interactions.
From the PDP and the overall importance plot, we see the main variables affecting the results in mathematics at the 8th grade are the results the **6th grade** in Mathematics and Italian, with a similar, increasing trend, the **geographical** variables (North, South, Pon), some of the **socio-economic** indexes (proportion of graduated parents, mean and, interestingly, **skew of the ESCS**), some of the **principal** behavior indexes, alongside a limited number of interactions (10 in the final model).



Therefore, we then built prediction intervals for the model using **inductive conformal prediction** with mean absolute error as nonconformity score. The resultant intervals, while **valid** (and indeed, covering more points than the nominal cover), are quite wide and do not differentiate much between schools at risk and schools not at risk.
Moreover, even though the model is quite accurate, it tends to always make predictions **too close to the mean** of the results, as we can see from the local explainer for the singular predictions.



Still, it learns a meaningful ranking between the observations, so, while we can use this model to have some valid prediction intervals for the mean result for all schools, to tackle our initial goal, we moved to a classification framework, where we tried to **predict directly the schools at risk (SaR, lower 20th percentile).** Our interest, in this case, is to learn a meaningful ranking for the schools, with a particular attention in spotting the lowest 20th percentile, so we are mainly interested in three evaluation metrics.

First, the Area Under the Curve (**AUC**), as it insensible to imbalance and, as our problem can also be seen as a ranking problem, the AUC can be interpreted as the probability it would rank a random positive instance over a random negative instance; then **Sensitivity** (Recall), as we are interested in seeing if the model is able to spot the schools at risk and a metric linked both to sensitivity and ranking problems, **Recall@20th**, defined as the Sensitivity (Recall) the model would reach by classifying as 1 only the top 20th percentile of the output probabilities, as we are mainly interested in seeing how many schools in the 20th percentile are actually ranked there by the model.

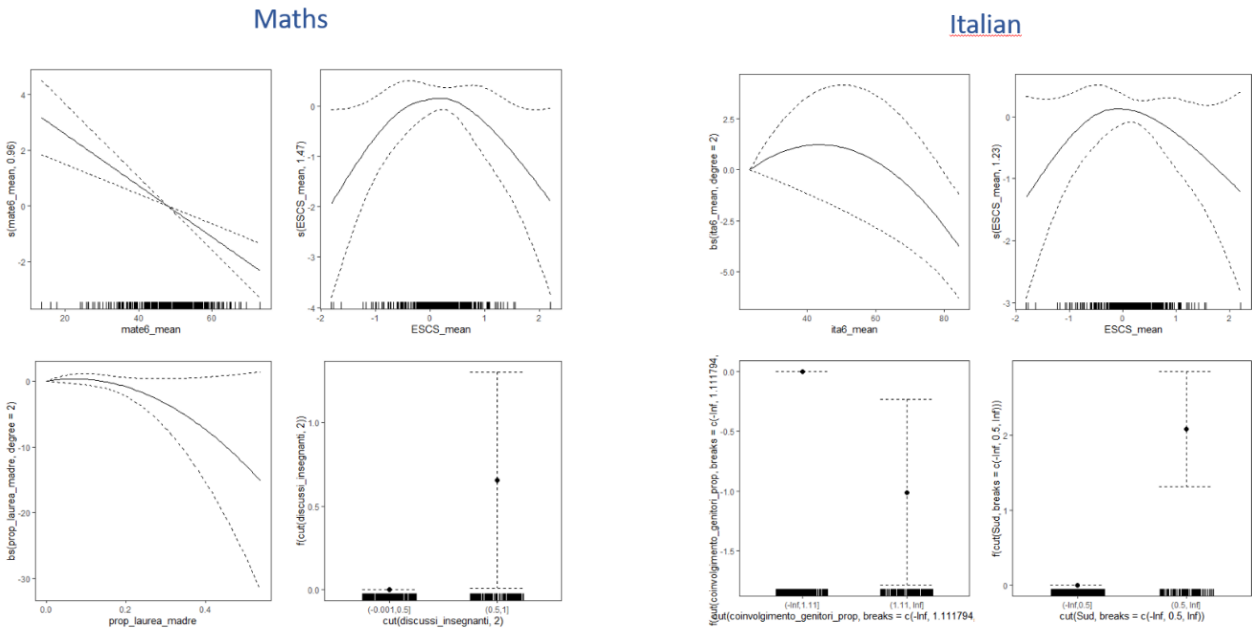
In this case, we first tried to use some classical logistic GAM, which we use as a benchmark to assess our final model.

4.2.2 Generalized Additive Models

We selected 2 different **GAM logistic models** to **predict SaR**. In particular, the selected significant variables were chosen assessing the performance of the model on a **4-folds stratified cross-validation**.

For the model used to predict SaR for Maths we used **Maths scores at first year**, **school location** (South=1, North/Center = 0), the **proportion of parents** (mothers and fathers, distinctly) with **higher education** in the school, the **mean ESCS** score of the school, the **usage of INVALSI** indicator, the **INVALSI tests discussion** with teachers and the level of **school president activity**.
For the model used to predict SaR for Italian we used **Italian scores at first year**, **school location** (South=1, North/Center = 0), the proportion of **mothers** who were **born in Italy** in the school, the proportion of **parents** (mothers and fathers, distinctly) **with higher education** in the school, the **mean ESCS** score of the school, the **pressures exerted by parents on school's president**, the **proposed** and **effective parents' involvement** and the **opinion on parents' associations**. For further details on the significance of the regressors, please have look at our GitHub folder in the 'GAM_rischio.R' file.

We report here the **more significant** partial dependency plots for Maths and Italian SaR GAM predictors:



We see that students' **socio-economic status** plays an important role. But, quite surprisingly, for both maths and Italian, the risk is classified as higher for schools having a middle-low socio-economic status w.r.t. those having the worst ESCS score, even if we think that this is partially due to the lower number of data points for very low ESCS schools. Also, the proportion of **higher educated mothers** plays a much more important role w.r.t. **higher educated fathers** (not reported). This might be explained with the fact that in Italy mothers are usually more involved in children education than fathers.
For both Maths and Italian, schools located in the **South** are significantly classified with a higher risk w.r.t. those in **Northern Italy**. Regarding Italian performance, we found that the proposed involvement of parents is associated with a lower SaR classification risk, thus suggesting that a strategy based on increasing parents' involvement in school activity could be beneficial for the results in Italian. Investigating Maths performance, finally, we retrieved the results showed in 4.1: **discussing** the INVALSI results with teachers is **associated with a higher SaR classification probability**.

These models show very promising performances in stratified 4-folds cross-validation. We achieved the following results (in the form MEAN (STD DEV) over the 4 folds; accuracy and sensitivity are reported with a cross-validation chosen threshold):

Maths

	$predF$	$predT$
$trueF$	257	64
$trueT$	19	52

Accuracy: 0.7883 (0.0227)
AUC: 0.7670 (0.0054)
Sensitivity: 0.7328 (0.0212)
Recall@20th: 0.5240 (0.0195)

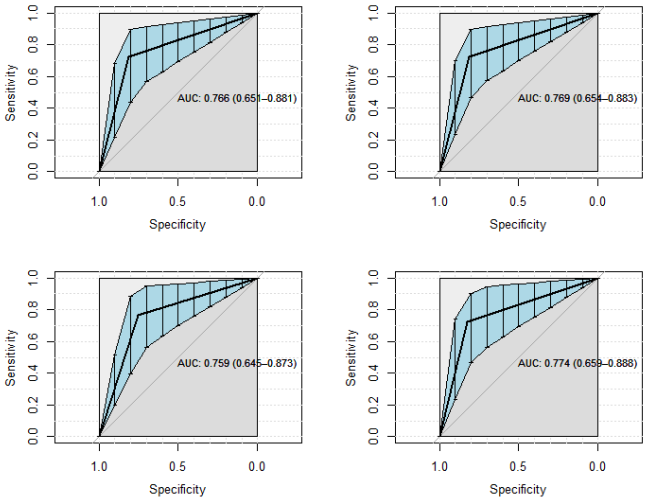
Italian

	$predF$	$predT$
$trueF$	255	71
$trueT$	20	50

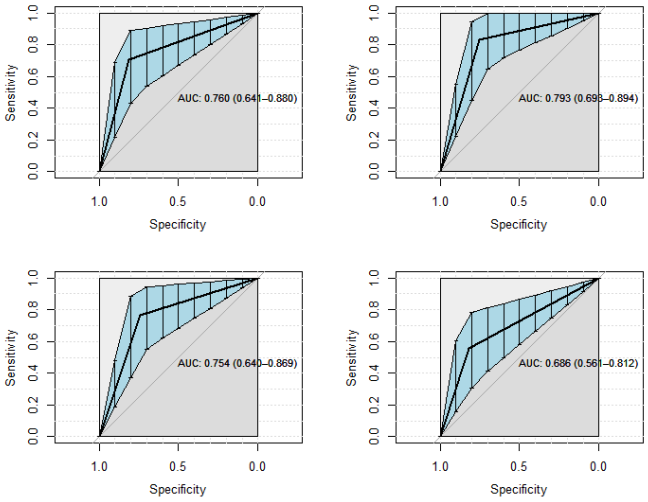
Accuracy: 0.7703 (0.0199)
AUC: 0.7483 (0.0389)
Sensitivity: 0.71493 (0.1183)
Recall@20th: 0.5742 (0.0325)

Furthermore, we report the **ROC curves** for both the Maths and the Italian classification models, together with bootstrapped confidence intervals for each fold:

Maths ROCs for 4 folds



Italian ROCs for 4 folds



4.2.3 Explainable Boosting Machines for SaR

For the final prediction model for SaR, we adopted the same approach we used for regression, fitting an EBM model after careful feature selection through **hill-climbing** in stratified **4-folds cross-validation**, optimizing for the Recall@20th.
The resulting models use a similar set of covariates with respect to the regression model and the GAM model, with some interesting additions which are directly related to the school's buildings with an intuitive increase in risk with a lower reported level of infrastructures.
The results, reported below, are really good, showing decent robustness between the different folds. In particular, in both cases, the **AUC is much higher**, so this means the model learns a really good ranking between the schools. In general, the results are really good, showing high predictive performance:

Maths[1]

	$predF$	$predT$
$trueF$	248	52
$trueT$	15	51

Accuracy: 0.8166 (0.0732)
AUC: 0.8742 (0.0188)
Sensitivity: 0.773 (0.0236)

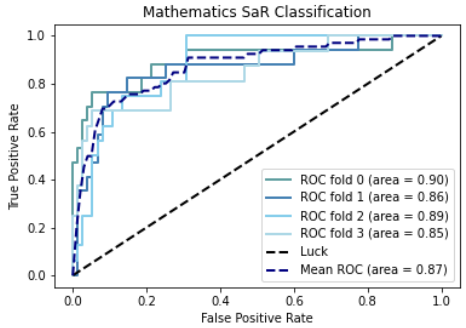
Italian[1]

	$predF$	$predT$
$trueF$	263	43
$trueT$	16	44

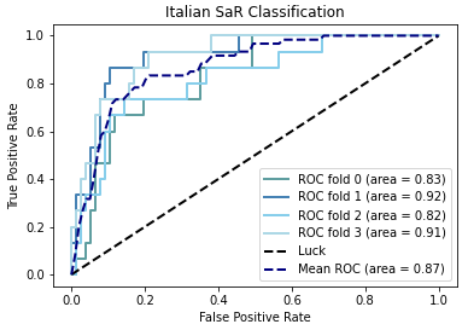
Accuracy: 0.8553 (0.0189)
AUC: 0.8701 (0.0441)
Sensitivity: 0.7028 (0.1977)

[1] The number of instances in the two confusion matrices are, indeed, different. This is due to the drop of the nas, which is performed voluntarily after the computation of the 20th percentile, as our aim is classifying the overall worst schools . This also explains the lower performances in sensitivity and recall@20 in italian, as there is more class imbalance

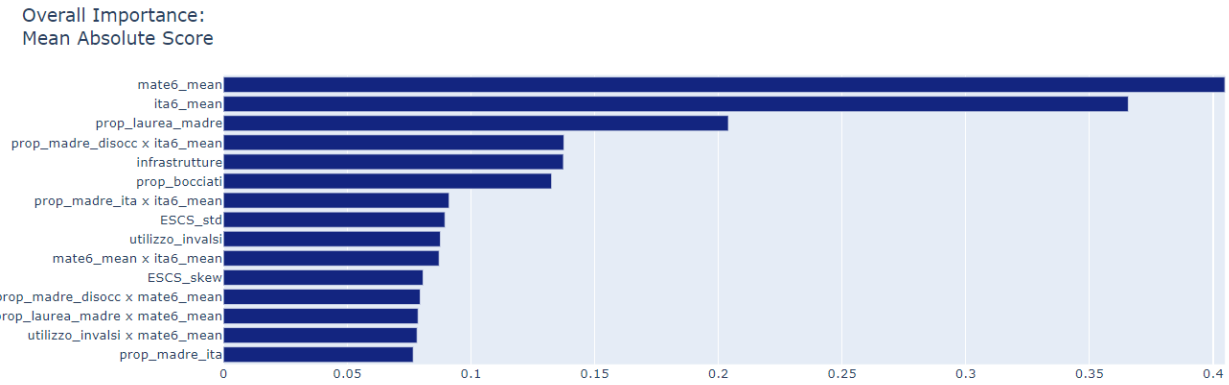
Recall@20th: 0.6389 (0.0621)



Recall@20th: 0.5694 (0.0461)



Further details on the procedure and the PDP can be found in the notebook "Classification_schools-ita.ipynb" and "Classification_schools-mate.ipynb".



4.2.4 Explainable Boosting Machines for INVALSI Scores Improvement

Another task of interest for the School Inspectorate would be trying to predict not only the absolute performance of the schools and the relative ranking of the schools, but also whether a school shows **students' improvement** with respect to their previous results or not (accounting for the different mean score for INVALSI tests at 6th and 8th grade by performing a 1/0 split with respect to the national median increase). After with GAM we could not find a statistically significant model for this task, we proceeded by using Explainable Boosting Classifiers to **classify schools having the better improvements** in performances. Such a task is aimed not only at detecting the most promising schools, but also at gaining **valuable insight about what characterizes the improvements in such subjects**. We used the variables *improv_mate_diff_mean* and *improv_ita_diff_mean*, nominally the school averages of differences between the performances of third and first year. After feature selection using Hill Climbing trying to optimize AUC, we obtained two models which are effectively decent at discriminating between schools, with the following performances:

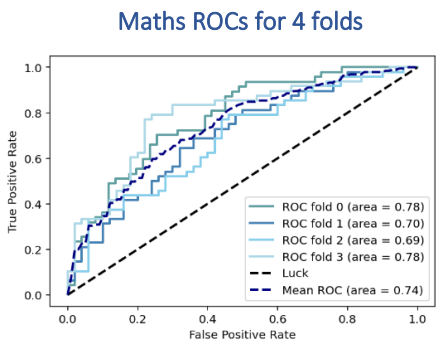
Maths

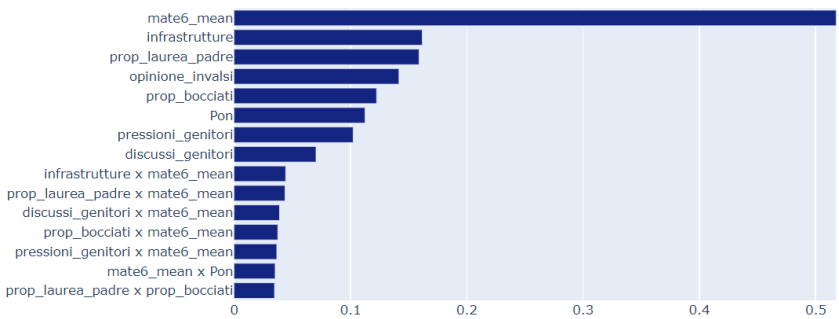
AUC: 0.74 (0.0454)
Recall@20th: 0.8026 (0.0436)

Italian

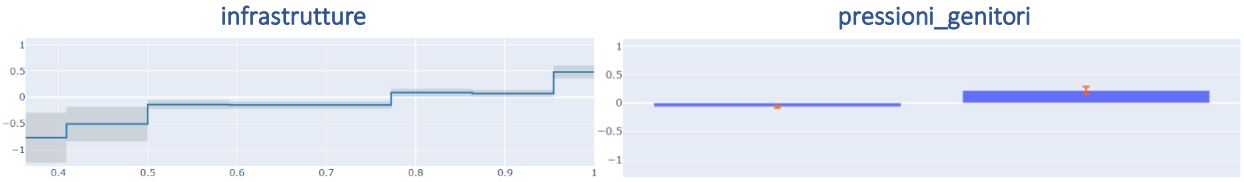
AUC: 0.63 (0.0408)
Recall@20th: 0.7222 (0.0393)

Exploiting the information that Explainable Boosting Machines are able to provide we decided to report, together with the ROC curves, the Mean Absolute Score of features' importance and selected partial dependency plots in order to gain insights over our target.

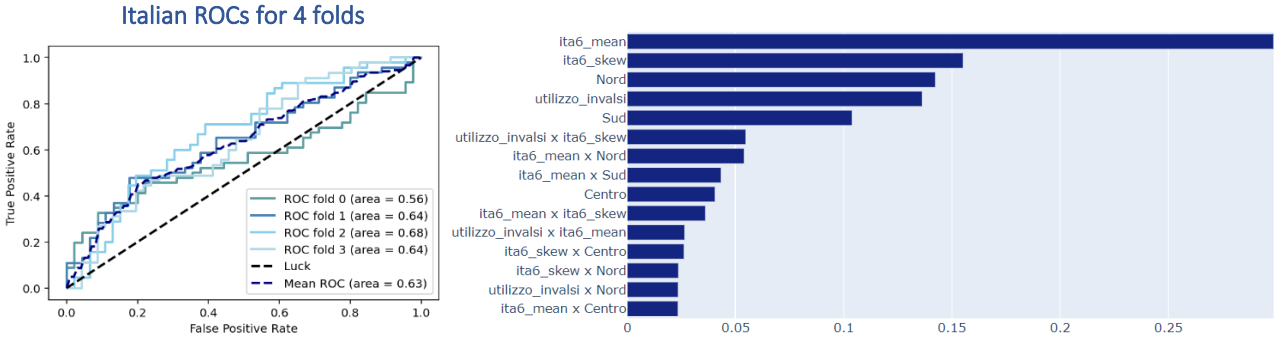




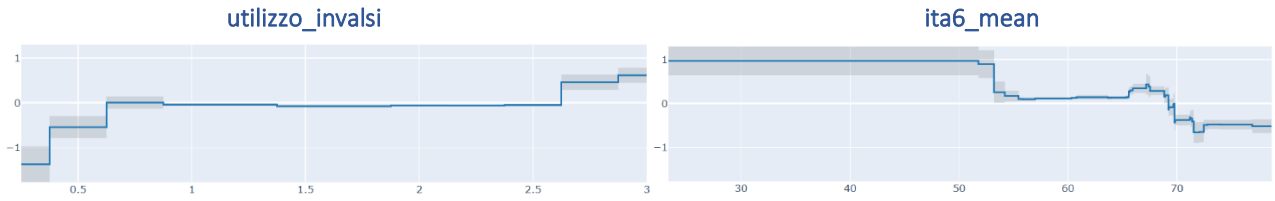
The improvement in Math is strongly dependent on the starting level; indeed, a lower starting level will most likely result in an improvement, probably due to the phenomenon of regression towards the mean. It is also interesting to look at the partial dependency plot of the second most important variable, the one regarding school's infrastructures:



This plot shows that a **better level in infrastructures** will usually help students improve in Maths, showing a possible effect of the whole school's environment. Also **the family background of students**, as for the absolute value, captured by the proportion of graduate fathers, parents' pressures and by having the INVALSI outcomes discussed with parents (the second biggest factor of total explanation, summing up this three variables) plays an important role in this: all these three features have an impact on the Maths score improvement in students.



In Italian, it is **really hard to build a functioning model**. Like in Maths, the improvement in Italian is strongly driven by the starting point, the lower the starting point the easier will be to improve, this time also the skewness of the first year's score is influent, in a coherent way with the mean. Apart from that, the main factors driving the improvement in Italian are **the geographic area** and **how much the INVALSI outcome are deployed by the school's management**.



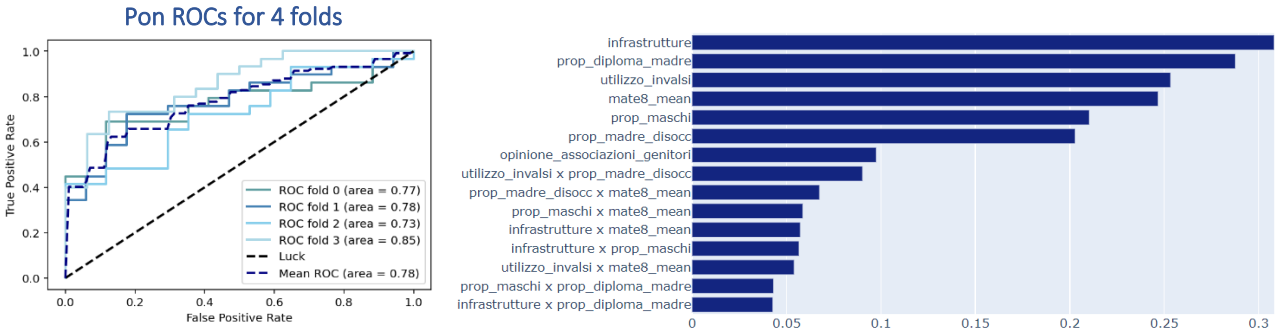
The first partial dependency plot shows that using the INVALSI outcomes in school's management plays a role in improving the Italian performances of students, while not using them properly is associated with lower levels of improvement.
Concerning geographical area, **the main distinction is between North and South of Italy**: while North is related to better improvements in Italian the South goes in the opposite way. These results are in line with the ones already found in Section 2.2.

4.2.4 Explainable Boosting Machines for Pon Classification

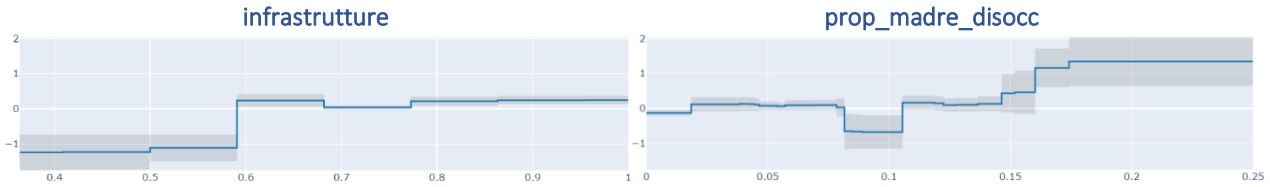
We decided to deploy Explainable Boosting Machines to gain insights over the Pon [6] project 2007-2013, a **help plan** for some Southern Italy schools in difficulty, funded by the European Union. Restricted to schools in the South, we are interested in **classifying whether a school was involved in the Pon project**, in order to see which factors provide explanation for this and if these schools are in any way different from the other schools. With the usual hill climbing procedure for feature selection, even though our focus is not really the predictive performance, we manage to outperform the dummy predictor in stratified cross-validation, with the following results:

Pon

Accuracy: 0.6848 (0.0565)
AUC: 0.78 (0.0461)



The most important feature for the model is, interestingly, **infrastructures**, which is something that we can expect since the Pon project is in part aimed at renewing the public school's infrastructures and this seems to confirm the effectiveness of the project. Also, the *average background of the students* seems to play a role to distinguish the Pon project's schools.



We can see how the principals of the schools that received the Pon declare better infrastructures, probably thanks to targeted investments. The schools receiving Pon are, in average, in a socio-economic condition lower than the national mean, so the fact that the feature corresponding to unemployment of the mother is one of the most relevant for our model may be in accordance with this fact and may attest that the funds might be received by the right schools.

5. CONCLUSIONS

5.1 Summing up

In this report we first pointed out **structural regional differences** between North, Center and South of Italy; not only from an **outcome perspective**, where the South performs worse than the North, but also from a **factor / indices point of view**. Keeping in mind our goal of assisting the School Inspectorate in improving school outcomes with managerial practices and prognostic models, these dissimilarities were instrumental in our subsequent task of highlighting possible **influences of managerial practices on school performance**, ending with a selection of 4 strategic suggestions. Additionally, we successfully implemented good GAM/GA2M prediction models which the School Inspectorate and interested parties can use to obtain an **accurate estimate of the future performance of schools**. These models could be used, for example, to foresee potentially subpar schools and quickly step in with tailored interventions to support them.

5.2 Further Advancements

- While we have tried to find causal links between variables by inspecting their relative impact manually, it would be worth trying to use more suitable models for nonparametric causal inference.
- We could reanalyse indices and more management-related questions in light of the user debiasing procedure.
- It would be interesting to replicate the analysis contained in [7] using DEA (Data Envelopment Analysis), in order to build some more compact indexes and investigate their effect in prediction of overall level/improvement in INVALSI tests at a school level.
- It would be interesting to test the quality of the analysis using data from INVALSI tests of the following years and, if possible, to build a model which takes into account the time-dependency of the phenomenon, hopefully helping us isolate even better the effect of the school on the different cohorts.

6. REFERENCES

[1] L'indice di background socio economico culturale (ESCS): Che cos'è ESCS?, Ministero dell'Istruzione, dell'Università e della Ricerca (https://www.istruzione.it/snv/allegati/01_A_INVALSI_escs_slide.pdf)

[2] Fritz Schiltz, Chiara Masci, Tommaso Agasisti, Daniel Horn. Using regression tree ensembles to model interaction effects: a graphical approach, Applied Economics, 50:58, 6341-6354, DOI: 10.1080/00036846.2018.1489520, 2018

[3] Yin Lou, Rich Caruana, Giles Hooker, Johannes Gehrke, Accurate Intelligible Models with Pairwise Interactions, KDD'13, August 11-14, 2013, Chicago, Illinois, USA, 2013

[4] Harsha Nori, Samuel Jenkins, Paul Koch, Rich Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223, 2019

[5] Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, Rich Caruana, How Interpretable and Trustworthy are GAMs? arXiv preprint arXiv:2006.06466, 2020

[6] Il PON, Ministero dell'Istruzione, dell'Università e della Ricerca (https://www.istruzione.it/pon/ilpon.html)

[7] Chiara Masci, Kristof De Witte, Tommaso Agasisti, The influence of school size, principal characteristics and school management practices on educational performance: An efficiency analysis of Italian students attending middle schools, Socio-Economic Planning Sciences, Volume 61, Pages 52-69, 2018

Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni. Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. MOX Report, Politecnico di Milano, 2015.

Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni. Does class matter more than school? Evidence from a multilevel statistical analysis on Italian junior secondary school students, MOX Report, 2015

Chiara Masci, Anna Maria Paganoni, Francesca Ieva, Semiparametric mixed effects models for unsupervised classification of Italian schools, Journal of the Royal Statistical Society: Series A, 2019