A predictive model of school failure

Patrizia Falzetti - Michele Marsili

VII Seminar
"INVALSI data: a tool for teaching and scientific research"
Rome, October 27th – 30th, 2022

Introduction

School failure is often understood only as early school leaving (ESL), in fact it means the student who leaves school during the year and then is outside the education system in the following years.

A further aspect of school failure, however, is that is related to low performances in some of the basic skills, Italian language (reading comprehension) and Mathematics mainly, but also in English.

We also have seen, over the years, emerge another phenomenon, outlined through INVALSI data, which is the implicit dispersion

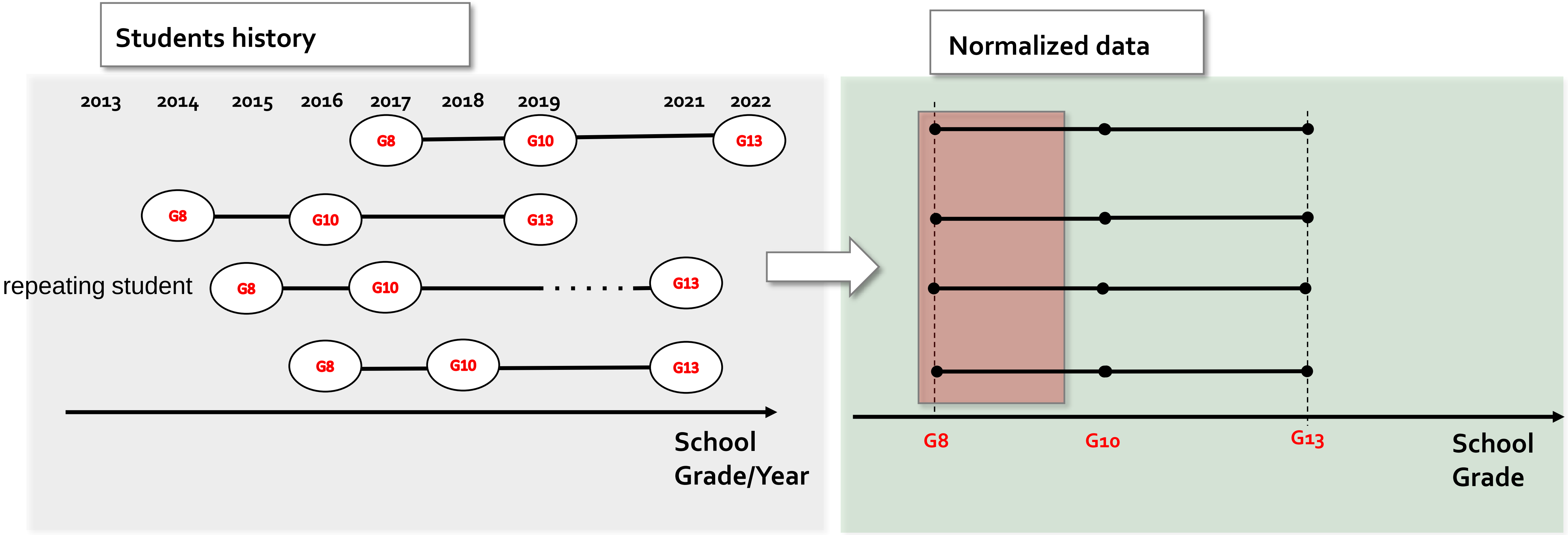
Dataset Description

The data used in this work are INVALSI data of 3 cohorts, the one outgoing in **2019**, **2021** and **2022**; since these are outgoing students from grade 13 and the students' entire career is considered backwards.

For each student, the previous scores and all the information of family background, geographical and school context available over time were retrieved in order to have a dataset as complete as possible.

We remove an observation if there is a missing value anywhere in that row.

Data preparation



Methodological approach

In this work we propose an approach based on a supervised machine learning algorithm to identify students at risk of school failure.

Supervised learning is a subcategory of machine learning where Input variables (X) and an output variable (Y) are known and an algorithm is used to learn the mapping function from the input to the output. This mapping function is used to predict the output variables (Y) given new input data (X)

What is machine learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

There is a lot of data: pictures, music, words, videos etc.

.

The volume of data is so high that we will increasingly turn to automated systems that can learn from the data and make better decisions in the future, based on the examples that we provide.

How Machine Learning works?

Input data is processed in order to obtain structured data, on which a machine learning algorithm is trained.

Input data → Processed data → Machine Learning Algorithms → Trained model

The trained model is applied on new data to make predictions.

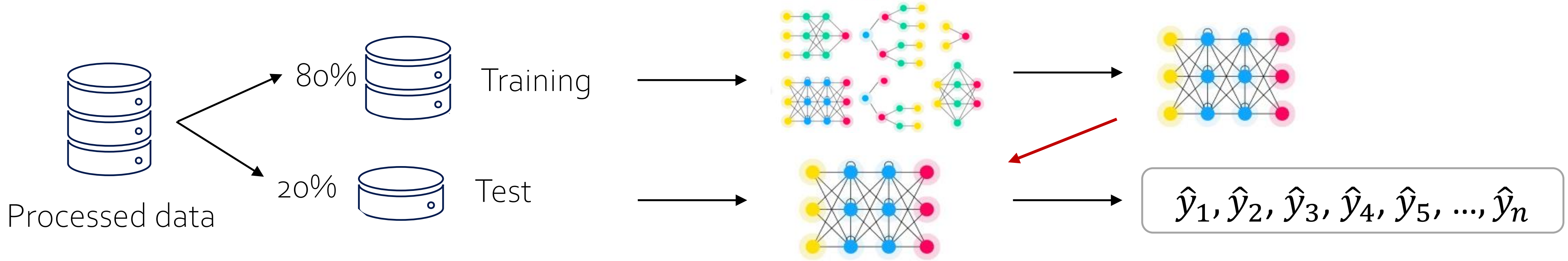
New data → New Processed data → Trained model → Prediction

Supervised machine learning algorithms

(X and Y are given)

1. The processed data is divided into training and test.
2. Top performing parameters are determined on the training set, in order to build the best model for that data.
3. The trained model is used to perform predictions on the test set.

Multiple statistical metrics can be used to assess the performance of Machine Learning algorithms.
The final validated model is saved and used to perform predictions on new input data (until a new model is trained).



Decision tree

The goal of using a Decision Tree is to create a training model that can predict the target variable by learning simple decision rules inferred from prior data.

Decision tree algorithms belongs to the family of **Supervised Learning algorithms.**

$$\overbrace{\begin{matrix} X_1 & X_2 & X_3 & \dots & X_n \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{nn} \end{matrix}}^{X} \qquad \overbrace{\begin{matrix} Y \\ y_1 \\ \dots \\ y_n \end{matrix}}^{Y}$$

Problems that Decision Tree can solve:

- **Classification**: a classification tree will determine a set of logical if-then conditions to classify the target variable that is categorical.

- **Regression**: a regression tree is used when the target variable is numerical or continuous. A set of conditions based on the sum of squared errors are used to make the prediction.

How Decision tree works

Suppose we want to predict if the following student school success for next School Years.

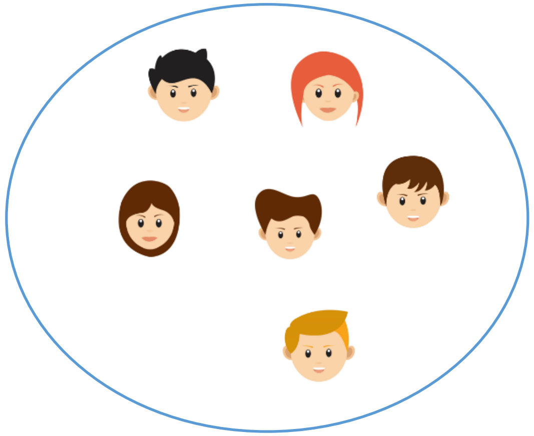
ID: 1234

SY : Grade 10 in SY 2022-23

Sex : Male
ESCS (2023): 1,20
ZWLE Maths G10 (2023): 1,13

We have an initial dataset that we used to build the model. For which we know all the information:

Sex	ESCS	ZWLE MAT G10
M	0,22	-1,52
F	1,41	1,02
M	1,53	1,21
M	-1,22	-0,56
F	2,13	1,82
M	-1,56	-1,3



How Decision tree works

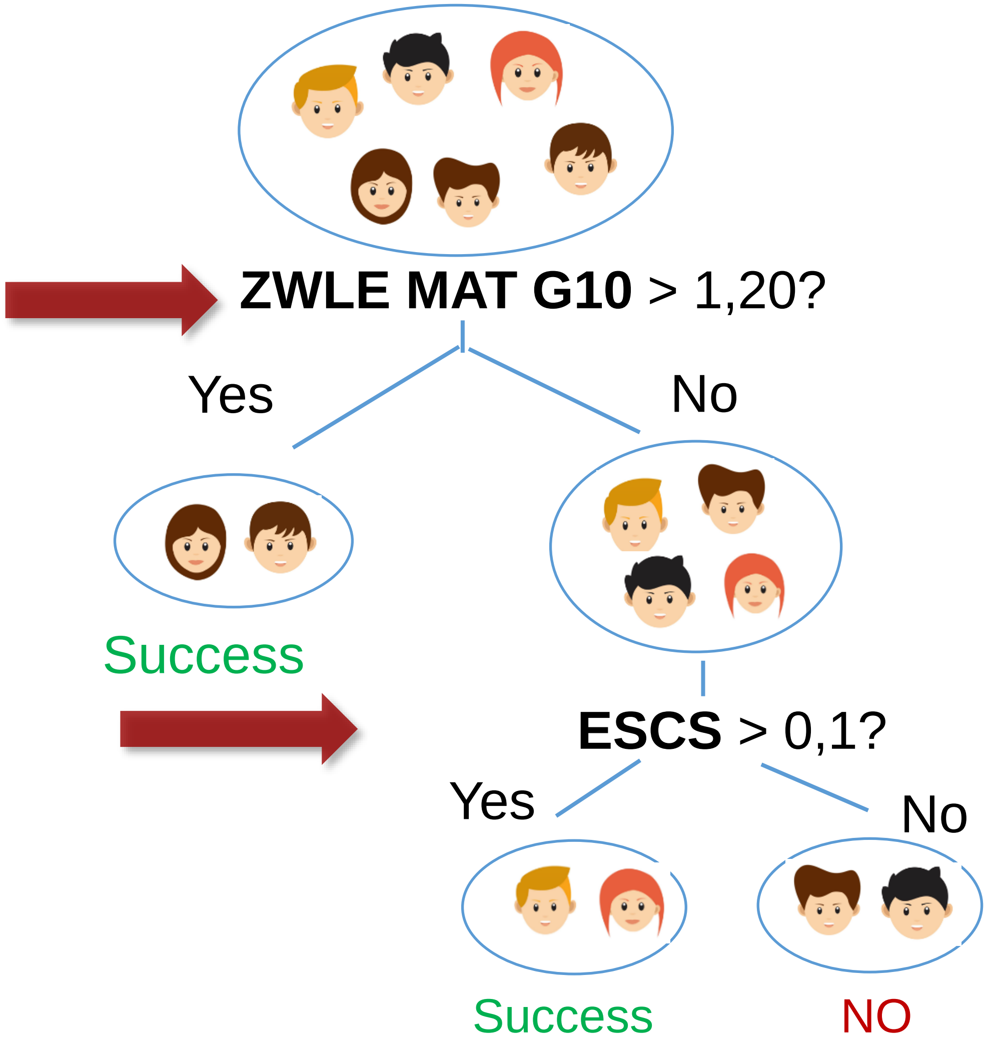
We have an initial dataset:

1. To give more information about the prediction of the target variable a decision rule is used to split the data into two subgroups:

The resultant sub-nodes are more homogeneous

2. We can also use the **ESCS** variable to make another split:

Thanks to this grouping in predicting we will make a smaller error.



How Decision tree works

In this way we can use this decision tree to predict our interest unit.

ID: 1234

SY : Grade 10 in SY 2022-23

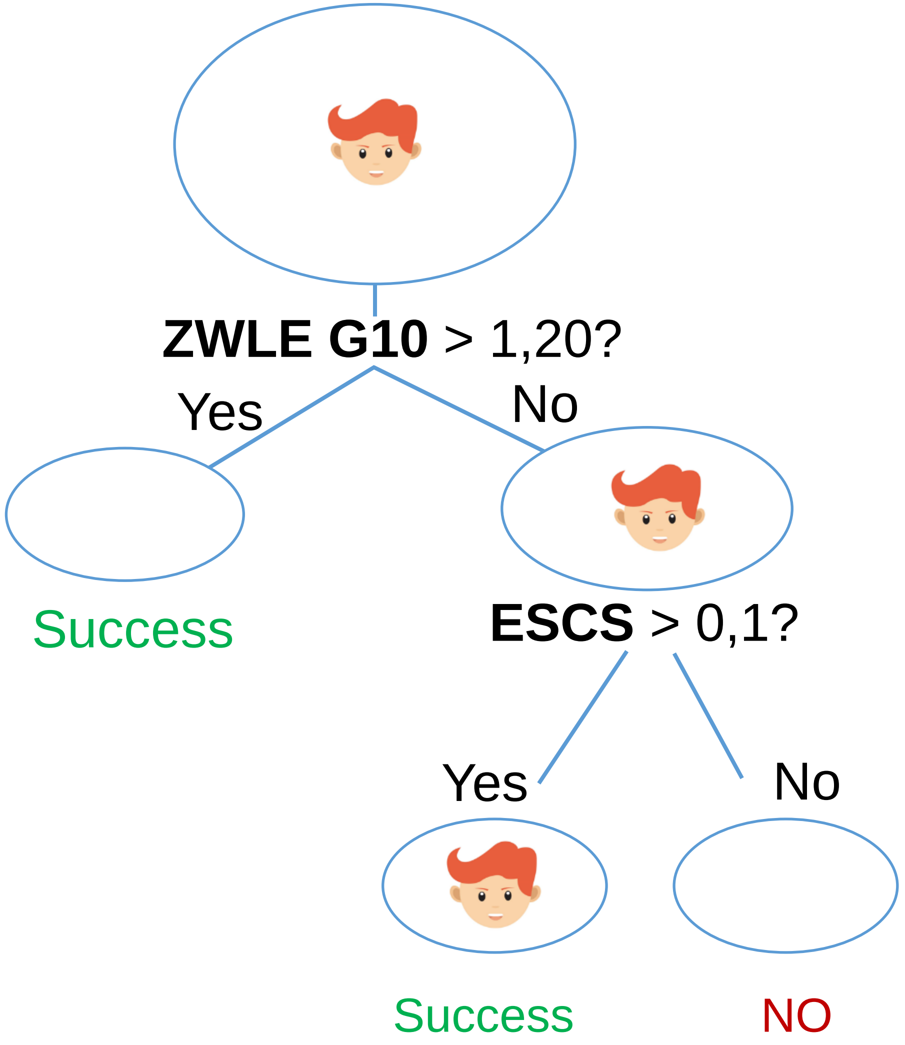
Sex : Male

ESCS (2023): 1,20

ZWLE G10 (2023): 1,13

Will the student ID 1234 have school success?

YES



Overfitting problem

The parameters are optimized to obtain the model that best fits the data structure.
Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points.
Overfitting the model generally takes the form of making an overly complex model to explain the structure of the data.

Training → Model has good capability to predict the target variable in the Training data → Low error

Test set → Model has bad capability to predict the target variable in the Validation data → High error

Beyond a single tree

- Tree-based methods are simple and useful for interpretation.

- They are not competitive with the best supervised learning approaches in terms of prediction accuracy.

Random Forest

This method grow multiple trees which are then combined to yield a single prediction, reducing the variance and increasing the prediction accuracy.

Bootstrap Sample

The bootstrap method involves iteratively, resampling a dataset with replacement. Obtaining **n-sub-samples** with equal sample size of the initial dataset.

Stud ID	ESCS	zwle	sex
1	0,22	-1,52	1
2	1,41	1,02	0
3	1,53	1,21	1

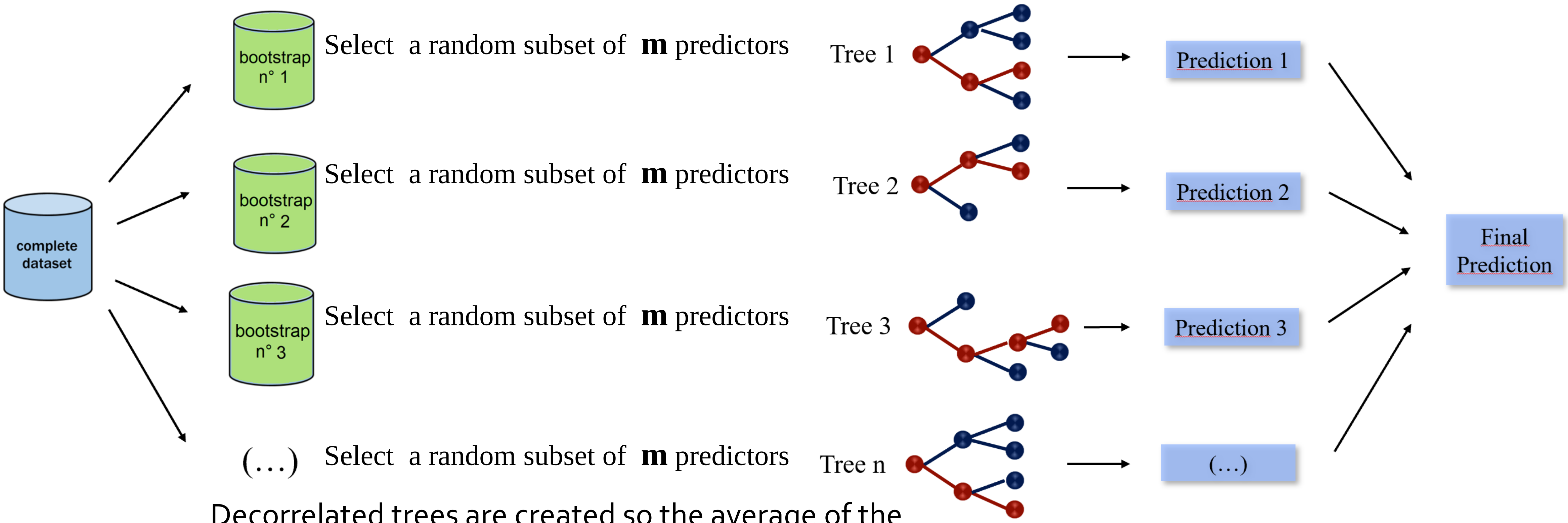
Stud ID	ESCS	zwle	sex
1	0,22	-1,52	1
2	1,41	1,02	0
2	1,41	1,02	0

Stud ID	ESCS	zwle	sex
1	0,22	-1,52	1
1	0,22	-1,52	1
3	1,53	1,21	1

Stud ID	ESCS	zwle	sex
2	1,41	1,02	0
2	1,41	1,02	0
3	1,53	1,21	1

(...)

Random Forest



Decorrelated trees are created so the average of the resulting trees is less variable and hence more reliable.

Model features

Individual Student Information

- Reapiting student at G10
- Foreign student G10
- Sex G10
- ESCS G10

School Information

- Type (Lyceum, Vocational…)
- Geographical macro-area

Target variable

- Low performances in some of the basic skills (in Italian language and Mathematics)
- Repeating student
- Dropout

1 = school failure

- Else

0 = NO school failure

Student performance at school

- School marks in Italian
- School marks in Mathematics

Results in the National Surveys in current year and previous years

- Italian Grade 10 INVALSI score
- Math Grade 10 INVALSI score
- Italian Grade 8 INVALSI score
- Math Grade 8 INVALSI score
- Italian Grade 5 INVALSI score
- Math Grade 5 INVALSI score

Results - Model performance

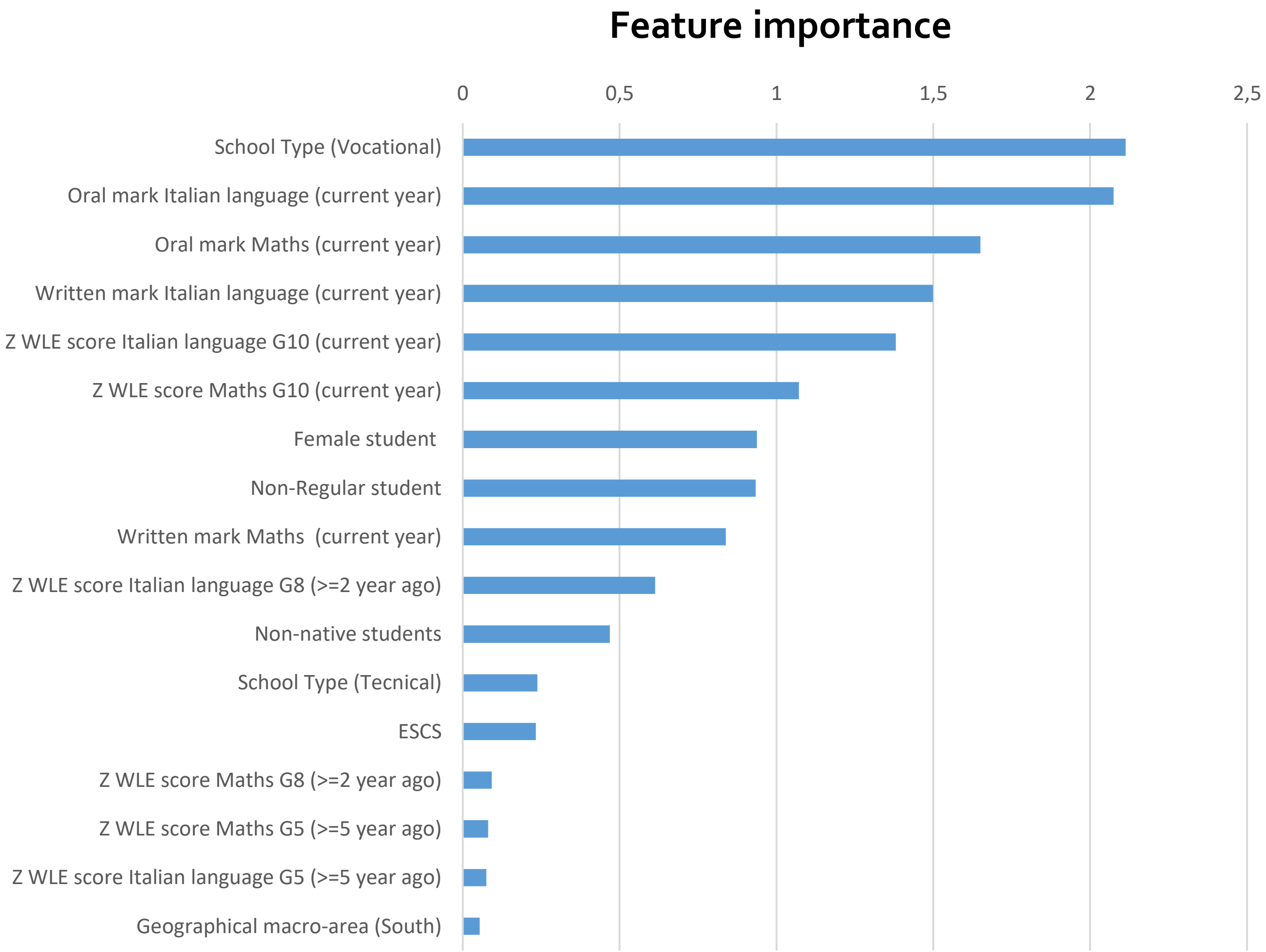
	Model performance
Accurancy	0,863
(TP+TN)/N. of obs. (%)	85,96

Results - Feature importance

After the model training it's possible to analyze the importance that each feature (variable) has in the model.

In the Random Forest algorithm the features importance measures the (normalized) **total reduction of the criterion brought by that feature** (es. Gini or Entropy).

Feature importance is a useful tool to **understand the inner workings** of a ML model and **helps in the interpretation** of the phenomenon under analysis.



Conclusion

The results show that the algorithm is able to predict with a good level of accuracy students at risk of school failure.

The analysis of classification performance metrics should be considered thoroughly before predicting potential cases of abandonment and a possible design of mechanisms for improvement interventions.

This kind of approach could be great interest as it allows for predicting the possible dropout or low performance of a student and being able to take corrective actions both at a global and individual level.

Thanks for your attention

patrizia.falzetti@invalsi.it
michele.marsili@invalsi.it

https://invalsi-serviziostatistico.cineca.it/

https://www.facebook.com/Servizio-Statistico-INVALSI-354920338928978