

# STUDENT LOW ACHIEVEMENT PREDICTION

ZANELLATI A., ZINGARO S.P., GABBRIELLI M.

Department of Computer Science and Engineering  
University of Bologna



In 2019, a study conducted by **INVALSI\*** found that 20% percent of Italian students had a lower-than-expected achievement and, eventually, dropped out of school.

**RQ1** Is it possible to quantitatively represent students' knowledge level and build a model of their skills attainments?

**RQ2** Is it possible to develop a suitable AI-tool to predict, at an early stage, the risk of low achievement at secondary school for primary school students?

\*Italian National Institute for assessment of the education system



# Objectives



## **CASE STUDY FOR EARLY UNDERACHIEVEMENT PREDICTION**

Italian case study with INVALSI dataset.  
Low achievement in math of high school students (K-10) with data collected at primary school (K-5)



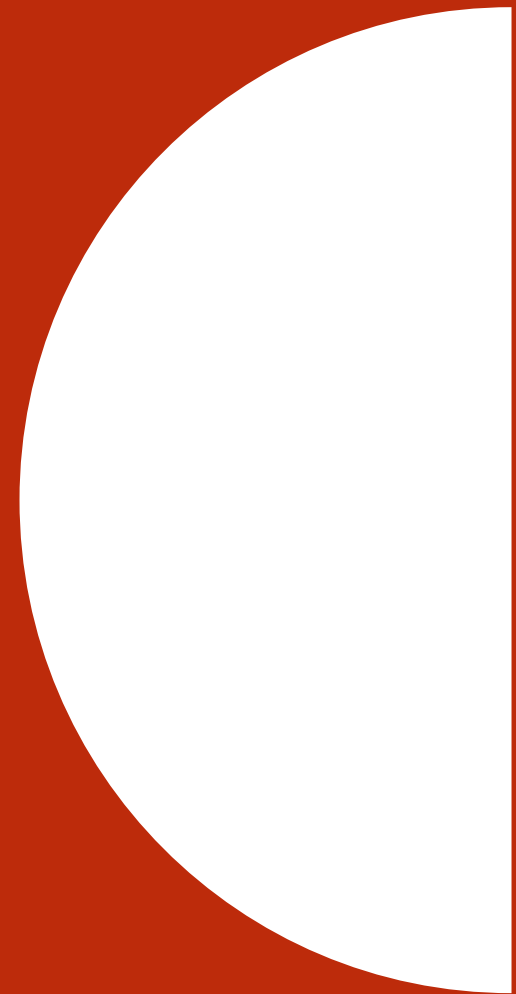
## **THREE AI TOOLS PROTOTYPES**

Comparison of three machine learning models (random forest and two neural networks)



## **ENCODING STUDENTS' LEARNING**

Extraction of features directly related to students' learning in terms of knowledge and skills



# Methodology

Dataset characteristics, Students' learning encoding,  
and Machine Learning Techniques

# The INVALSI dataset

## LONGITUDINALLY TRACK STUDENTS' LEARNING


Data on maths test from two cohorts of students K-5 of the s.y. 2012/13 and of the s.y 2013/14

## UNDERACHIEVEMENT TARGET

Students' data in grade K-10, for the definition of low achievement target: grade in the test  $\leq 2$  on a scale from 1 to 5

## RICHNESS OF DATA

706633 students  
Demographic data, socio-economic and cultural variables, boolean features for correctness of each test item



# Students' learning encoding

Handle the transferability between different students' cohorts

**Table 1.** Maths INVALSI framework for question encoding.

Areas
(NU) Numbers
(SF) Space and figures
(DF) Data and forecasts
(RF) Relations and functions
Process
(P1) Know and master the specific contents of mathematics
(P2) Know and use algorithms and procedures
(P3) Know different forms of representation and move from one to the other
(P4) Solve problems using strategies in different fields
(P5) Recognize the measurable nature of objects and phenomena in different contexts and measure quantities
(P6) Progressively acquire typical forms of mathematical thought
(P7) Use tools, models and representations in quantitative treatment information in the scientific, technological, economic and social fields
(P8) Recognize shapes in space and use them for problem solving
Macro-process
(MP1) Formulating
(MP2) Interpreting
(MP3) Employing

# Students' learning encoding

Handle the transferability between different students' cohorts

**D1. Look at the following numbers.**

3060. 315. 312. 96.

**They are**

- A. All even
- B. All multiples of 3
- C. All multiples of 5
- D. All less than 1000

**Area:** Numbers (NU)

**Process:** Know and mastering the specific contents of mathematics (P1)

**Macroprocess:** Employing (MP3)

**Table 2.** Example of the student's learning final encoding.

<b>Id</b>	<b>NU</b>	<b>SF</b>	<b>DF</b>	<b>RF</b>	<b>P1</b>	<b>P2</b>	<b>P3</b>	<b>P4</b>	<b>P5</b>	<b>P6</b>	<b>P7</b>	<b>P8</b>	<b>MP1</b>	<b>MP2</b>	<b>MP3</b>
<b>1</b>	0.86	0.75	0.90	0.80	0.71	0.80	1.00	0.89	1.00	0.67	0.91	0.75	0.81	0.73	0.94
<b>2</b>	0.50	0.25	0.50	0.53	0.29	0.60	0.50	0.22	1.00	0.33	0.73	0.25	0.50	0.47	0.44

# Machine Learning Techniques



## **RANDOM FOREST**

Training through bootstrap aggregating (bagging) to reduce overfitting and increase precision



## **CATEGORICAL EMBEDDINGS NEURAL NETWORK**

Input treated depending on its type: categorical inputs are passed through an embedding layer, numerical ones are fed to a dense layer



## **FEATURE TOKENIZER TRANSFORMER**

Identification of the group of inputs that most influence the output, thanks to attention maps





# Experimental setup and results



# Experimental Setup



## TRAIN, VAL, TEST

Two students' cohorts:  
- K-5 2012/13 - Train and Validation sets (351746 students)  
- K-5 2013/14 - Test set (354987 students)



## FEATURE SELECTION AND EXTRACTION

34 features



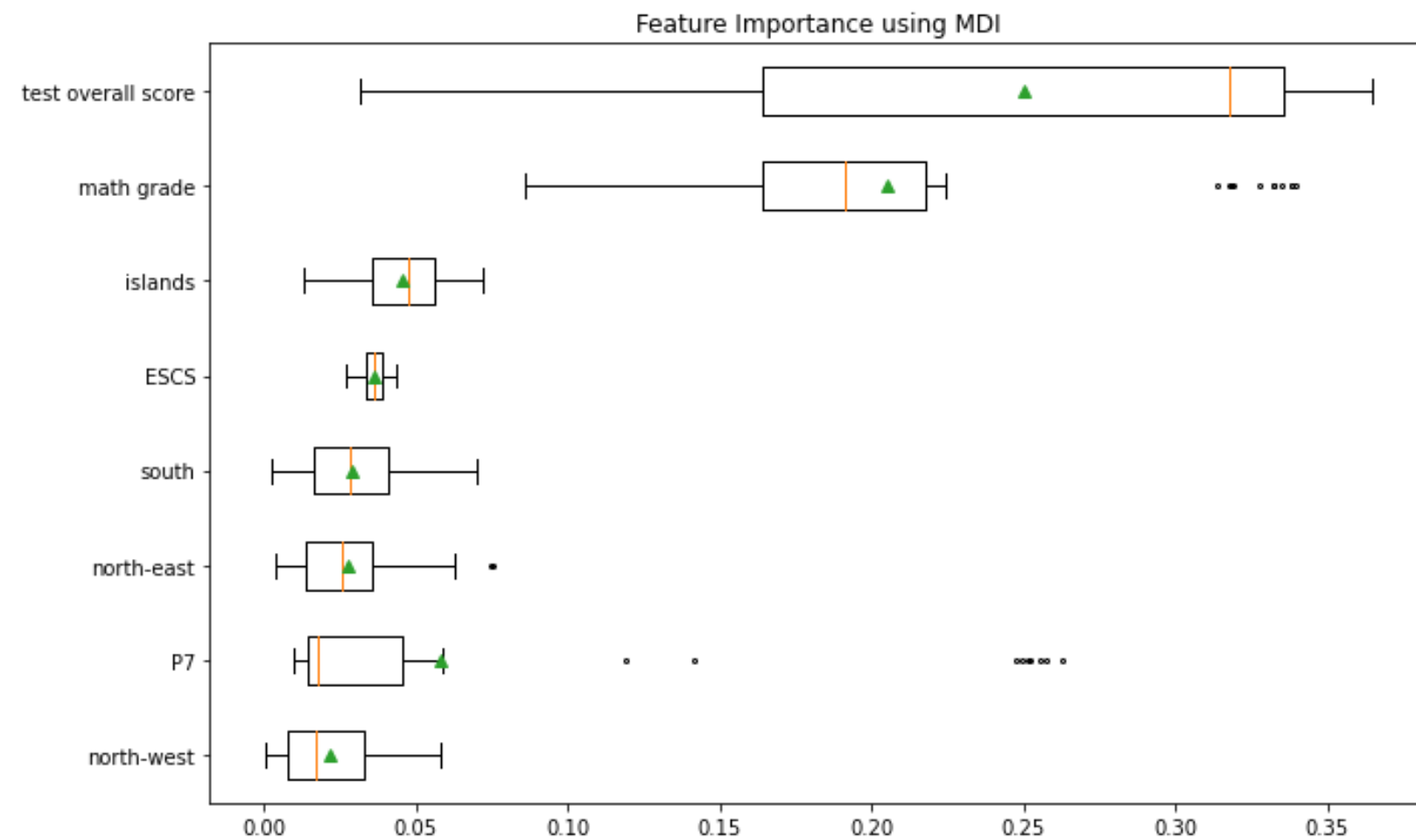
## UNBALANCED DATASET

Random Forest model ->  
random undersampling  
technique  
Neural networks -> weighted  
random sampler

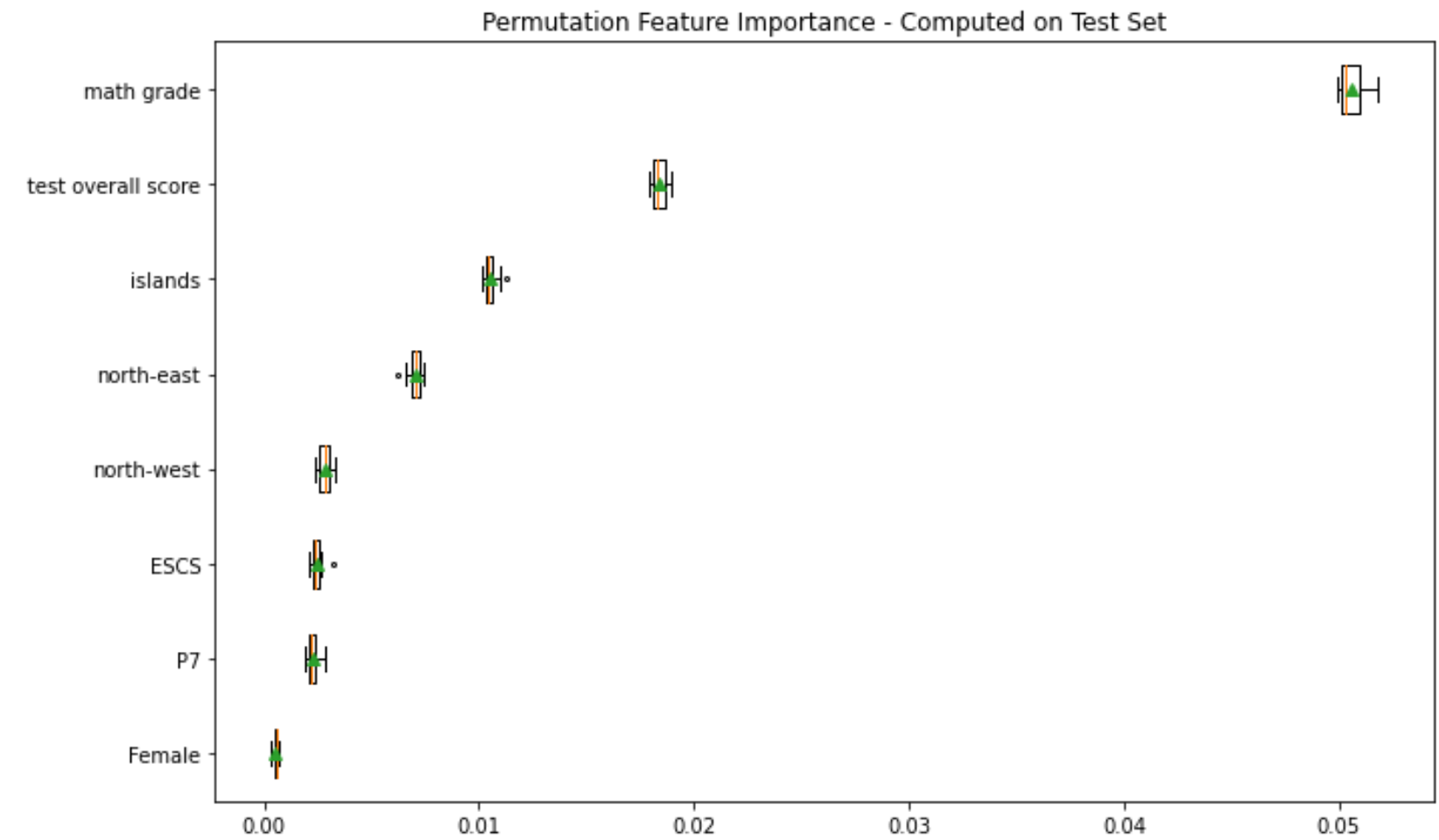
# Results

Models	Accuracy	Precision	Recall
Random Forest	0.77	0.62	0.67
CE neural network	0.76	0.76	0.76
FTT neural network	0.78	0.77	0.78

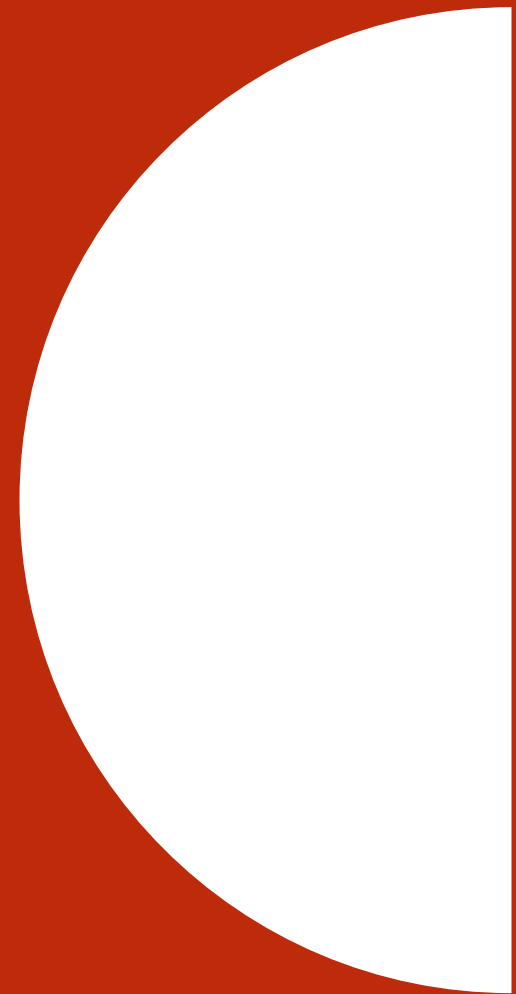
## PERFORMANCE ON TEST SET



FEATURE IMPORTANCE (MDI)



PERMUTATION FEATURE IMPORTANCE



# Conclusion and Future works

# Future works



## **TRANSFERABILITY TO OTHER DISCIPLINES**

Using a similar representation for students' learning in other disciplines.



## **IMPROVE DATA QUALITY**

Training and testing the model with students from different cohorts.  
Defining learning encodings not knowledge-based.



## **IMPROVE INTERPRETABILITY**

Analyzing the feature importance computed on the RF model in comparison with the weights in neural networks.



Thanks for your attention