



Early prediction of students at risk of failure: a Machine Learning Approach

Fiammetta Noccioli – Michele Marsili – Marcello Napoli

INVALSI

X SEMINAR

“DATA FROM AND FOR THE EDUCATIONAL SYSTEM: TOOLS FOR RESEARCH AND TEACHING”

ROME, 19 – 20 – 21 NOVEMBER 2025

Literature

There is an increasing use of predictive models aimed at detecting and preventing academic failure, which has contributed to a deeper understanding of students' careers and academic performance (e.g., Cannistrà et al., 2022; Priulla et al., 2024; Delogu et al., 2024):

- In the Italian context, studies are still based on single-university case studies or datasets limited to administrative records from individual institutions.
- Many Italian predictive studies lack detailed school-level information limiting insights on how pre-university schooling affects later university trajectories. Linking administrative university data with school assessment sources is emerging but not yet widespread.



Objective

The study aims to evaluate the performance of supervised machine learning models, highlighting the potential of these approaches, which have gained popularity in recent years.

01. Create machine learning models that represent a possible baseline for predicting student career, providing a "neutral" and replicable comparison point for all universities

02. Discover recurring factors that influence students' risk of dropout, beyond local differences, to guide public policies or targeted prevention programs

02.1. Explore the role of school-related variables in students' academic trajectories during their university careers



Dataset

Individual Student Information

- Repeating student at G13
- Foreign student
- Gender
- Mother education
- Father education

School Information

- Type (Lyceum, Vocational...)
- Private/public school

University Information

- Disciplinary Area
- Enrolments
- Career events
- Credits
- Geographical macro-area
- Geographical proximity

Student performance at school

- Diploma grade

Results in the National Surveys

- Italian Grade 13 INVALSI score
- Math Grade 13 INVALSI score
- English Grade 13 INVALSI score

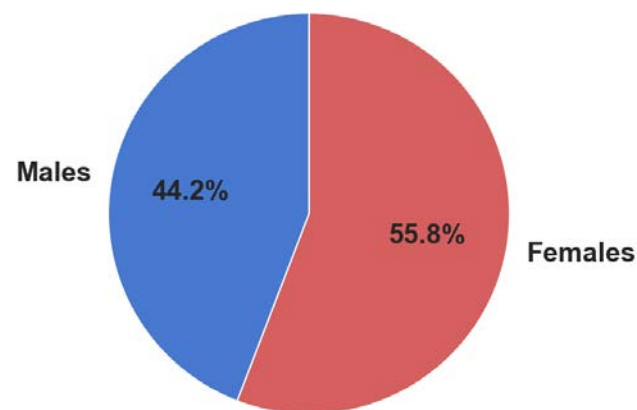
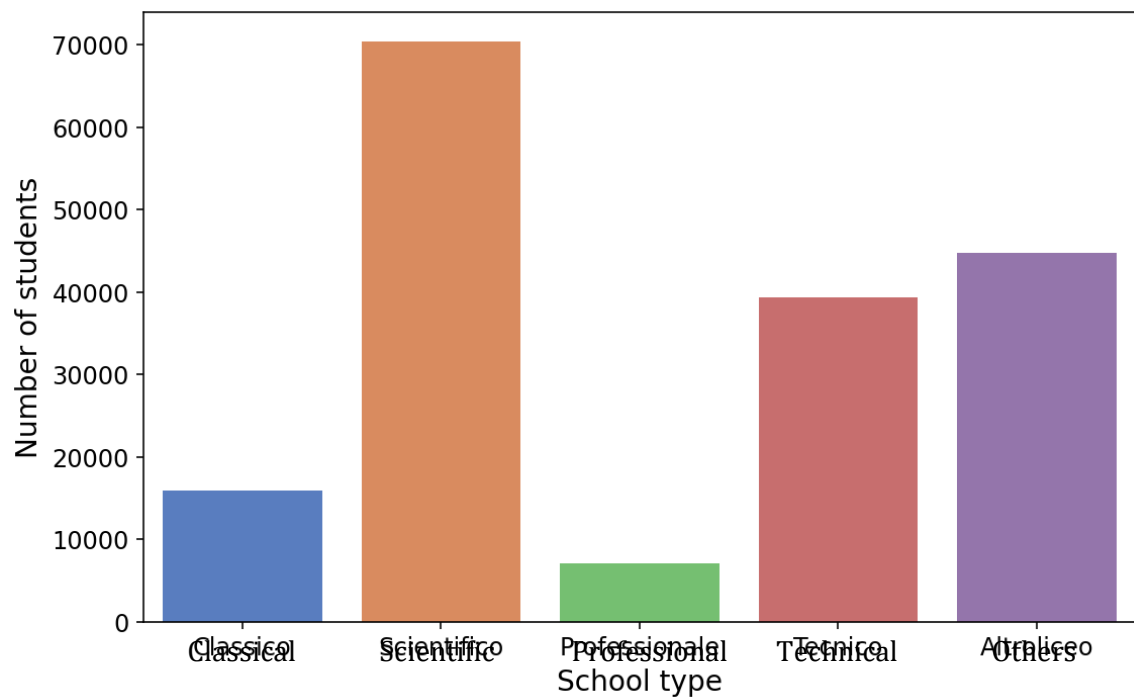
Target variable

- Risk of «failure»

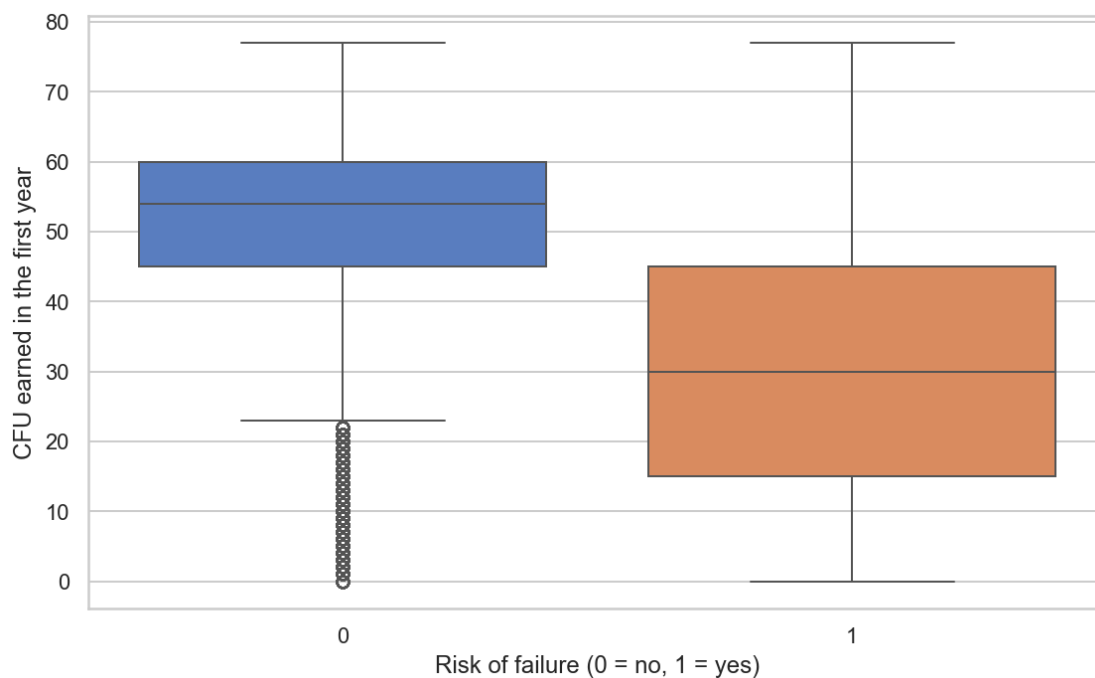
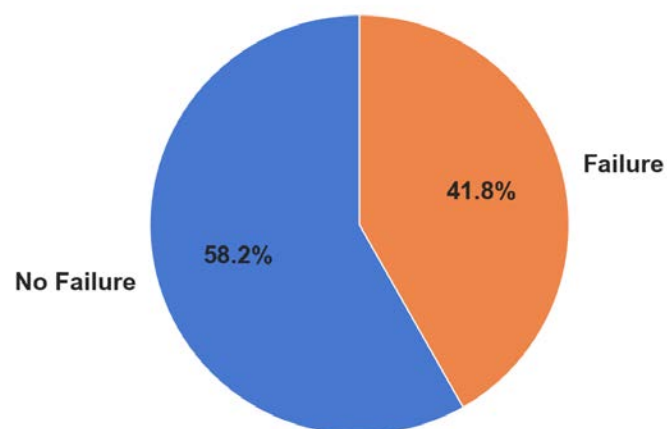
**1 = University
unsuccess**

**0 = University
success**

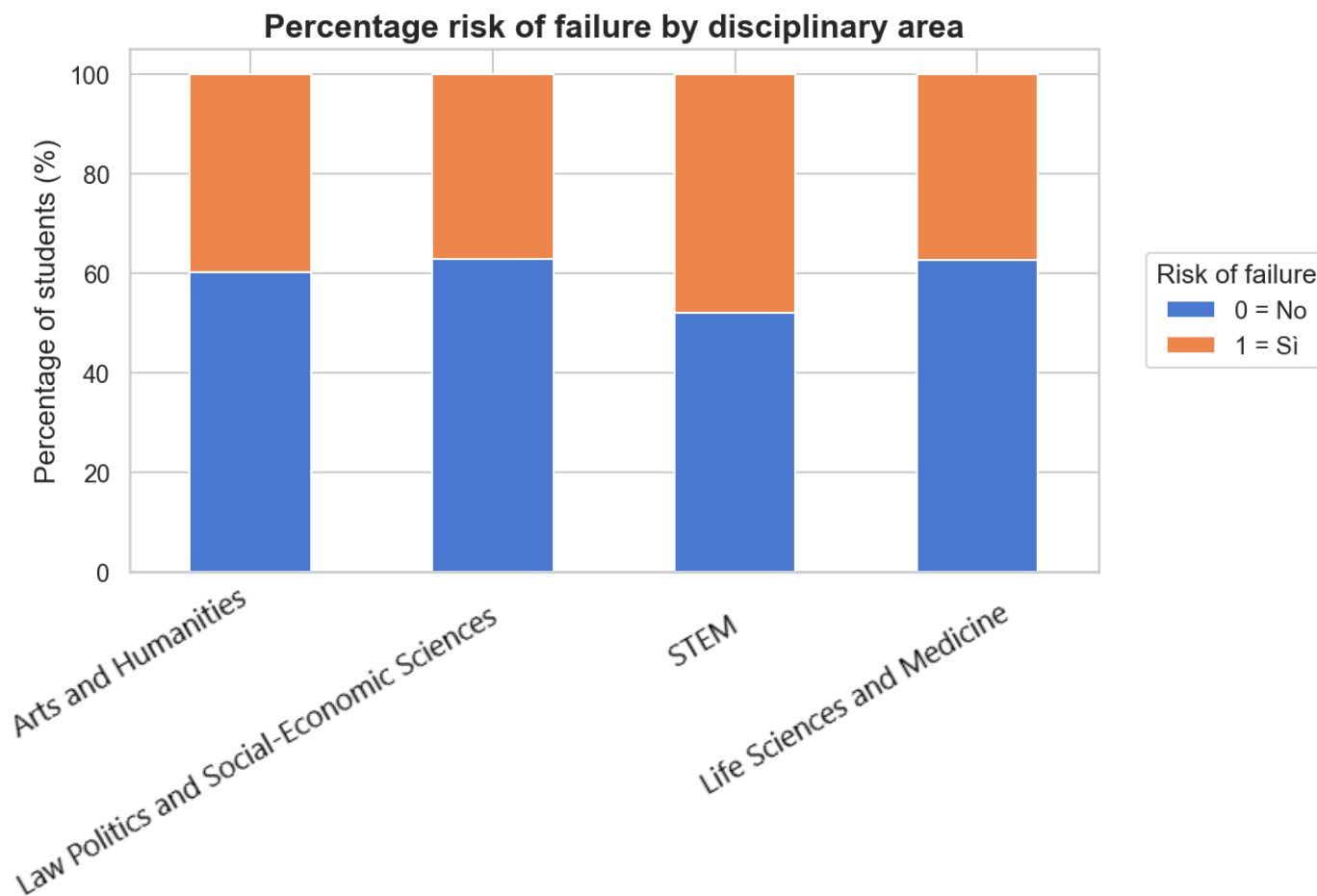
Data exploration



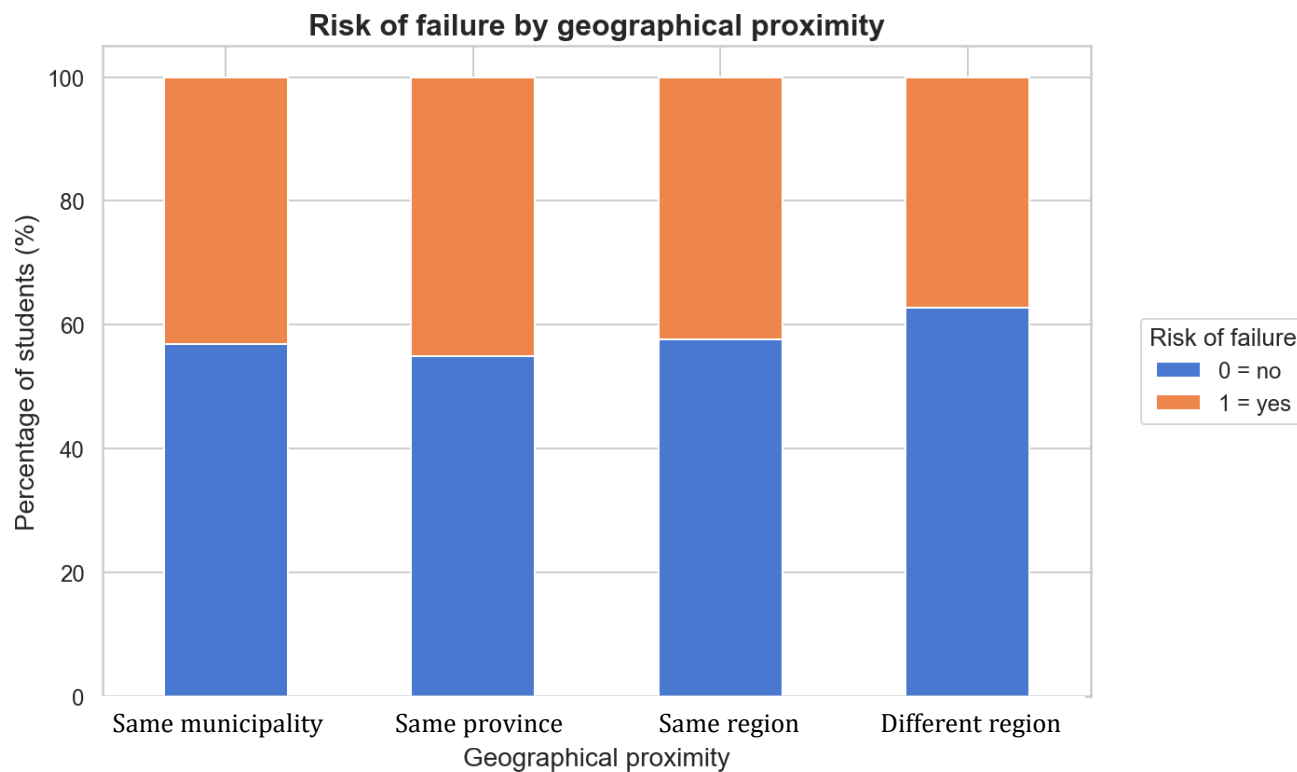
Data exploration



Data exploration



A categorical proxy for the distance between the student's residence and the university.



Methodological approach

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.

The main types of machine learning models are:

- 
- **Supervised:** the algorithm learns from labeled data, where inputs are associated with known outputs (e.g., classification, regression).
 - **Unsupervised:** the algorithm identifies patterns and structures in data without predefined outputs (e.g., clustering, dimensionality reduction).
 - **Reinforcement learning:** the model learns through interaction with an environment, receiving rewards or penalties based on the actions it takes (e.g., games, robotics, optimization).

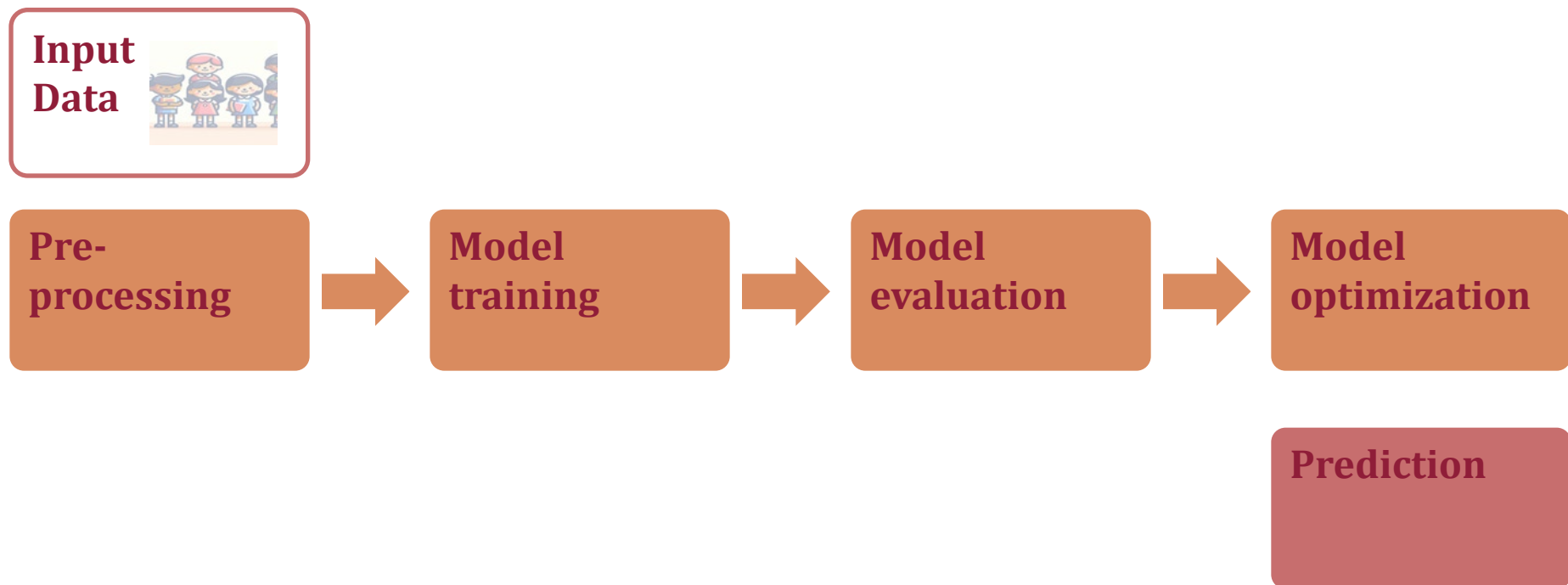
Classification problems are a type of supervised learning where the goal is to predict a discrete label (the output) based on patterns learned from the labeled training data.

The model assigns each unit to one of the predefined classes



Dataset is partitioned in Training set (80%) and Test set (20%)

Phases



Among the tested models at the end of first a.y., CatBoost achieved among the highest performance across evaluation metrics.

	Accuracy	Precision	Recall	F1	AUC
CatBoost	0,7693	0,7649	0,7564	0,7594	0,8312
LightGBM	0,7690	0,7645	0,7562	0,7591	0,8307
XGBoost	0,7682	0,7640	0,7548	0,7580	0,8298
GBM	0,7682	0,7635	0,7555	0,7584	0,8297

CatBoost Model

CatBoost is a gradient boosting algorithm that uses symmetric decision trees and ordered boosting to achieve fast, accurate, and robust machine learning models.

Pros:

- Handles categorical features automatically (no need for one-hot encoding)
- Robust to overfitting
- Fast and efficient
- Supports advanced hyperparameter tuning (implemented through Bayesian Optimization using the Optuna library in Python)

Cons:

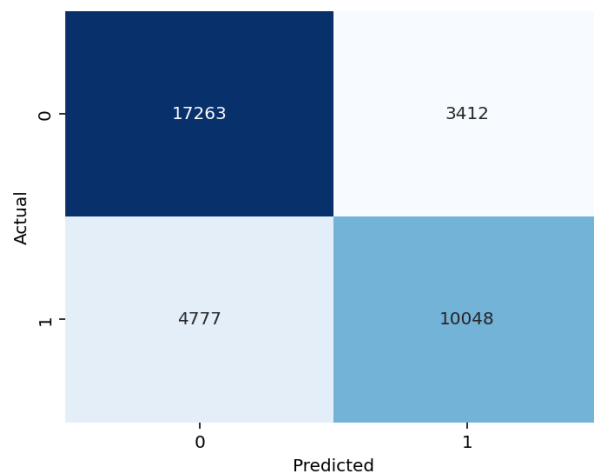
- Longer training time compared to simpler models (especially with large datasets).

After selecting CatBoost as the study model, we tested it on the cohort by examining three key stages — enrolment, first academic year, and second academic year — in order to estimate the evolution of performance levels over time.

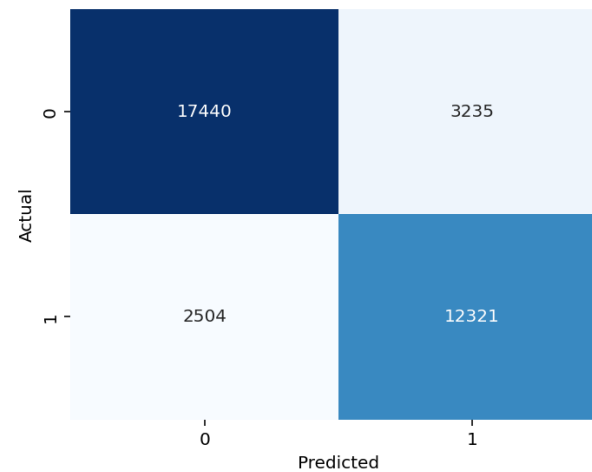
	Accuracy	Precision	Recall	F1	AUC
Catboost 2° a.y.	0.8383	0.8332	0.8373	0.8349	0.9129
Catboost 1° a.y.	0.7693	0.7649	0.7564	0.7594	0.8312
Catboost enrolment	0.6623	0.6526	0.6358	0.6367	0.7040

Confusion Matrices

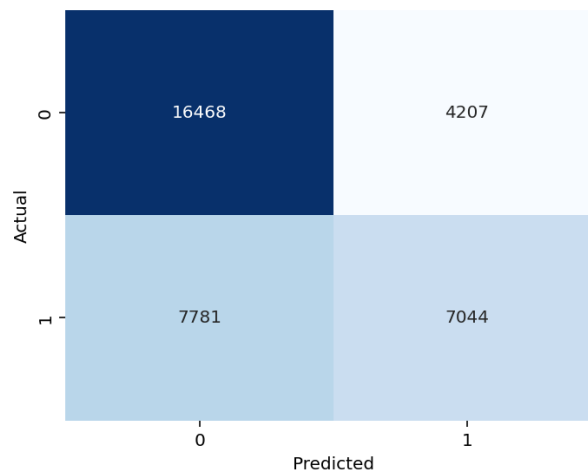
First Year



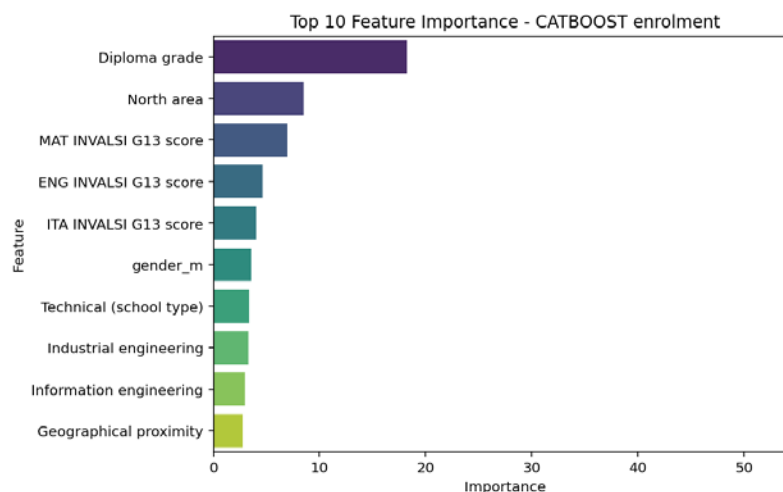
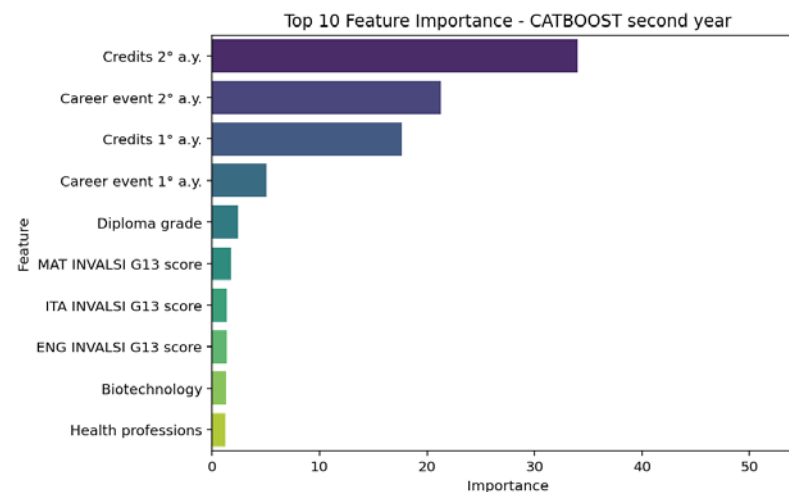
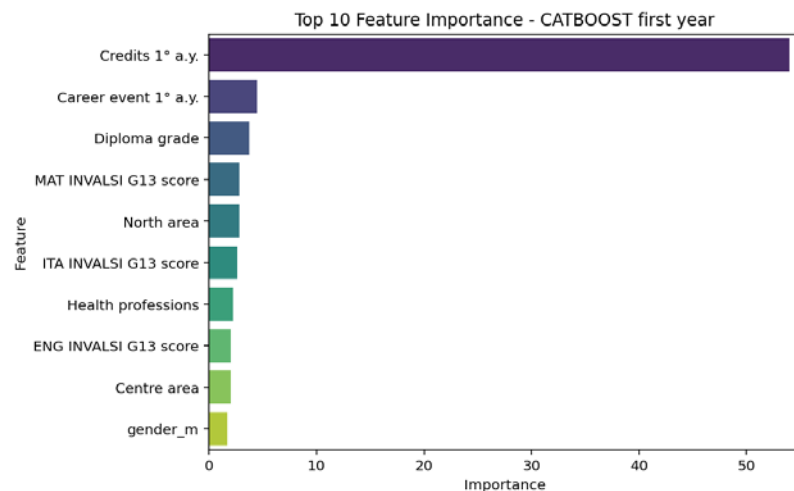
Second Year



Enrolment



Feature Importance



Results

- Model performance is stable across different parameter configurations, suggesting robustness of the CatBoost approach in this context
- CatBoost can serve as a reliable reference for future studies
- Feature importance analysis confirms the relevance of key variables (e.g. credits)
- The school-related variables, Diploma grade and INVALSI G13 scores, are among the top ten predictors in all CatBoost models.

Final thoughts

- A small set of variables (additional or more detailed variables might improve model accuracy such as taxes, family income, etc.)
- University-level differences cannot be captured



X SEMINAR
“DATA FROM AND FOR THE EDUCATIONAL SYSTEM: TOOLS FOR RESEARCH AND TEACHING”
ROME, 19 - 20 - 21 NOVEMBER 2025

Thank you for your attention!

michele.marsili@invalsi.it
marcello.napoli@invalsi.it
fiammetta.noccioli@invalsi.it