A predictive model of school failure in Italy

Michele Marsili (INVALSI), Patrizia Falzetti (INVALSI), Emanuele Marsili (Civico Polo Scolastico Manzoni)

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. By definition, the students part of this phenomenon are those who, even if they obtain a high school diploma, do not have the appropriate skills to deal easily with adult life, in short, those who leave high secondary school with the basic skills provided at the end of low secondary school. Unfortunately, data, although slightly improving, tell us that at the end of high secondary school this phenomenon stands at around just under 10%, a considerable number of students, therefore, obtain the diploma without reaching the basic levels provided by the National Indications in all the subjects investigated, Italian language (reading comprehension), Mathematics and English (Listening and Reading). Research object and hypothesis: The present work aims to create a model that allows to well identify in advance the so-called "at risk" students, i.e. those students on whom the school, through the work of teachers and school leaders, can intervene in order to reverse the forecast of school failure. A statistical model of this type allows to identify, with a reasonable margin of error, the students who may fall into one of those categories at risk, namely abandonment, implicit dispersion or low performer. We intend to analyze the phenomenon from different aspects, also from a geographical point of view, to understand if there are areas more at risk, but the final goal is certainly to attribute a probability of risk for each student and provide a disaggregated and aggregated indicator to schools to allow them to intervene. Preventing school failure at this point would be "possible", or at least there would be the premises to start doing so. It could be done, for example, for students who enter a school cycle, the lower secondary school (the first class, grade 6) or the secondary school (the first class, grade 9), and make available to teachers a measure that identifies the students most at risk and the subjects in which they are most in difficulty so that they can intervene at the beginning of the school year on the student without wasting additional time; thus giving the student himself, for example, the opportunity to catch up, to recover the gaps accumulated in previous years. Data used: The data used in this work are the 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, the data on absences from the Ministry of Education has also been added.

All the datasets have been harmonized and queued in order to create a single database useful for preparing the model. 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. The various cohorts are distinguished through the year variable. Method: In this work we propose an approach based on a supervised machine learning algorithm to identify students at risk of school failure. In particular, a Random Forest model was used, one of the most widely used algorithms for classification tasks. Using the data of the three cohorts to train the model it is possible, given a new dataset, to make predictions and thus be able to identify students at risk. The variables used concern both the context data of the students and the results in the National Surveys in previous years. The assessment of the importance of these variables in the classification provides further indications on what are the potential causes of school failure. Results: 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. The analysis of the importance of the most influential variables for the classification shows that the school performance in the previous surveys makes the greatest contribution to the forecast.