Research on Education and Media
Vol. 15, N. 1, Year 2023
ISSN: 2037-0830 – DOI: 10.2478/rem-2023-0014

Representation of learning in the post-digital: students' dropout predictive models with artificial intelligence algorithms

Andrea Zanellati[a], Anita Macauda[b], Chiara Panciroli[c], and Maurizio Gabbrielli[d]

[a] *University of Bologna, Italy, andrea.zanellati2@unibo.it, https://orcid.org/0000-0001-6171-0397*
[b] *University of Bologna, Italy, anita.macauda@unibo.it, https://orcid.org/0000-0003-1522-5440*
[c] *University of Bologna, Italy, chiara.panciroli@unibo.it, https://orcid.org/0000-0001-7173-670X*
[d] *University of Bologna, Italy, maurizio.gabbrielli@unibo.it, https://orcid.org/0000-0003-0609-8662*

Abstract

Within scientific debate on post-digital and education, we present a position paper to describe a research project aimed at the design of a predictive model for students' low achievements in mathematics in Italy. The model is based on the INVALSI data set, an Italian large-scale assessment test, and we use decision trees as the classification algorithm. In designing this tool, we aim to overcome the use of economic, social, and cultural context indices as main factors for the prediction of a learning gap occurrence. Indeed, we want to include a suitable representation of students' learning in the model, by exploiting the data collected through the INVALSI tests. We resort to a knowledge-based approach to address this issue and specifically, we try to understand what knowledge is introduced into the model through the representation of learning. In this sense, our proposal allows a students' learning encoding, which is transferable to different students' cohort. Furthermore, the encoding methods may be applied to other large-scale assessments test. Hence, we aim to contribute to a debate on knowledge representation in AI tool for education.

Keywords: *representation of learning; post-digital; low achievement; artificial intelligence algorithms; predictive model*

1. Artificial Intelligence for representation of learning

We live in a "datafied" society, based on the systematic extraction, aggregation, and manipulation of data about people. The "datafication" process strongly also affects education (Klašnja-Milićević et al., 2017) and has motivated a scientific debate on post-digital and education (Lacković 2020; Ryberg 2021; Panciroli, 2021). The educational environment involves a lot of learning-related processes which generate vast amounts of potential rich data. This datafication process in education has been accelerated in the last two years due to the emergency imposed by the COVID-19 pandemic. The need to make up for traditional teaching with distance learning has led to a more intensive use of Learning Management Systems (LMS), which allow the collection of numerous data for tracking learning processes. However, the datafication process had already started previously and it has not been bounded to the use of LMS: at the administrative and educational policy level, architectures for a more structured data collection, which are suitable for longitudinal studies, have spread. An example is the datasets built through national and international large-scale assessments tests.

Faced with the growing availability of data in the educational field, we have witnessed the development of research sectors such as learning analytics (LA) and educational data mining (EDM) (Romero & Ventura, 2020). Both disciplines make use of data-driven methods to tackle research problems in the educational field. This also implies the use of data mining techniques based on machine learning algorithms and, sometimes, can lead to the design of AI tools. According to Aljawarneh and Lara (2021), the deployment of these techniques "may result of great interests for involved stakeholders (students, instructors, institutions, …) since the extracted knowledge from educational data would be useful to deal with educational problems such as students' performance improvement, high churning rates in educational institutions, learning delays, and so on". In their review paper, they refer to several original contributions of studies where data science techniques have been applied to extract knowledge of interest for educational stakeholders. However, in all the studies, the knowledge extracted is only applicable to the problem addressed. A main challenge for the research in LA and EDM areas is generalizability (Baker, 2019), that is, obtain a general model that can be applied in other scenarios.

Data collected through national and international large-scale assessments lend by their nature to greater generalizability than traditional educational or psychological studies, which often rely on convenience samples (Ertl et al., 2020). However, these data are often used to support educational policy decisions (Fischman et al., 2019) or in studies aiming to determine the relationship between socioeconomic factors and school performances (OECD, 2018). Nevertheless, they are designed to measure students' knowledge and skills and often to track longitudinally the students' learning path, for example, INVALSI (Branchetti et al., 2015; Panero 2019) and National Educational Panel Study, NEPS (Blossfeld & Roßbach, 2019), two national large-scale assessments tests, respectively, in Italy and German. Therefore, these data may be promising also to extract knowledge of interest for students and teachers, concerning students' learning process such as low-achievement prediction or dropout risk, and not for policymakers alone.

With this position paper, we want to present an undergoing research project that exploits the INVALSI dataset to design a predictive model for students' low achievement in mathematics based on machine learning algorithms. The model has as predictive target the risk of underachieving the expected skill level in math at the end of high school (K–13) by using the students' performances in math INVALSI tests at the end of middle school (K–8). Specifically, we want to contribute to the debate on what it means to represent knowledge in an AI system for education. To address the low-achievement problem and therefore mitigate its socioeconomical impact, we aim to build a predictive model with high interpretability and explainability. The model is expected to inform students and teachers on learning factors that most influence a high risk of underachievement. In fact, learning factors can be more influenced by the school effect than students' family socioeconomic status. In other words, we want to address the two following research questions:

- (RQ1) Is it possible to use data collected through large-scale assessments for a representation of students' learning that can be integrated into a predictive model on their skills achievements?
- (RQ2) What knowledge is introduced into the model through this type of representation of learning and how does AI impact the production of new knowledge?

In Section 2, we present the theoretical framework in which we place our contribution with reference to three strands concerning the scientific debate on post-digital in education, machine learning to reference to learning processes, and representation of knowledge in relation to predictive models. Section 3 is dedicated to the methods: first, we describe the INVALSI dataset and which cohorts of students we consider in our study; second, we briefly present the machine learning techniques used for the design of the predictive models; finally, we focus on how we have integrated some extracted features from the data to represent students' learning. In Section 4, we briefly present the results obtained with our baseline model and discuss them according to the RQs. The last section is used for the conclusion of this position paper and highlights open challenges and future works.

2.Theoretical framework

2.1. Post-digital and education

The reference literature in post-Digital science and education (Lacković, 2020; Ryberg, 2021) focuses on a broad scientific debate on the relationship between technology and educational project (Rivoltella & Rossi, 2019; Panciroli, 2021). In the post-digital era, everything is connected and transferable to other fields, according to new relationships of exchange that lead to a process of reconstruction of new forms of fluidity, capable of overcoming great divisions and consolidated dichotomies—nature / culture, subject / object, individuality of the subject / social context conventions—which have characterized previous eras. Overcoming this logic requires new ideas, generated centered on a "human project," capable of reconstructing the meaning of our daily practices in relation to a "homeodynamic world" open to transformation (Eugeni, 2015). This also involves a relocation of the distinct functions of technology (research, storage, manipulation, transmission, communication, display, and reception of information), applied to different sociocultural areas in which, increasingly, the digital and people's lives are characterized by flows of intangibility within material forms. In this regard, Pisano considers post-digital as a condition characterized by nomadism, fluidity, and micro-materialism of listening in the contemporary world (Pisano, 2015). Specifically, Knox (2019) provides two general interpretations of the post-digital: the "post" in the sense of "post-digital," which suggests a different phase in the perception and use of technology (Cramer, 2015; Fuller & Jandric, 2019); the "post" as a signal of a critical evaluation of the assumptions embedded in the general understanding of digital. In this case, post-digital is a theoretical research approach specific to digital in the contemporary world. In this sense, the importance of reflecting on the impact of post-digital theories in the field of educational research emerges. The educational project is predominantly seen as a question of human development, in the form of individual behaviors, cognitive processes or social constructions, in which technology has a supporting role, an uncritical "empowerment" for learning (Bayne, 2015). Nevertheless, Knox (2019) suggests that the post-digital perspective in education may refer to the context of a larger society made up of pervasive digital technologies, which have not exempted the educational practice (Panciroli, Rivoltella, Gabbrielli, & Zawacki-Richter, 2020). In fact, the close intertwining between the datafication process leads to a massive collection and analysis of data and the pervasiveness of artificial intelligence (Janssen et al., 2020). In educational practice, this relationship has developed intending to enable metrics and insight processes on a national or international scale.

2.2. Machine Learning and predictive analysis

The datafication process in the educational field is linked to more intensive use of Learning Management Systems, which allow the collection of numerous data for the tracking of learning processes; accelerated use of distance learning developed during a health emergency. Specifically, the considerable amount of data produced with e-learning platforms (electronic register, Google for Education suite, teams, ...) requires artificial intelligence and machine learning (ML) solutions to be easily managed (Popenici & Kerr 2017; Zawacki-Richter et al. 2019). Artificial intelligence has moved toward statistical, probabilistic, data-based systems and, thanks to new machine learning algorithms, is able to deduce correlations and models in data sets, to carry out predictive analysis, speech recognition, and other advanced solutions of human–machine cooperation. In fact, the ML mathematical algorithms allow machines to learn so that they can carry out and complete a required task, without being previously programmed through a code that tells them exactly what to do. Learning thus becomes an iterative process that allows the machine to derive information (knowledge models) from the data that is provided (Panciroli, Rivoltella, Gabbrielli & Zawacki-Richter, 2020). In this sense, ML algorithms also make it possible to extract information from data by providing forecasts on the future trend of a given phenomenon. In fact, by combining "predictive algorithms" with historical data, it is possible to carry out a predictive analysis that calculates the probability with which an event can occur (Kumar, Garg, 2018). In this sense, predictive models make it possible to identify patterns and recurring trends in the data available and to use them to guide the decisions of the present (Shi-Nash, Hardoon, 2016). These models are becoming increasingly complex as the data that drives them is increasingly complex, both in terms of volume and in their nature and dimensionality, no longer just numbers or number vectors, but also images, paths, and texts. This first requires a description and modeling of the domain of the variables involved to then focus attention on the forecasting part. With reference to the subject of this article, predictive systems are functional to improve students' learning levels. They are based on a selection of metrics extracted from the INVALSI dataset, to explore correlations.

2.3. Representation of knowledge

In the context of AI studies, knowledge representation is a specific study and research sector that deals with organizing knowledge to make it usable by a machine or by an automatic intelligent system through a language and commands that are always more precise and detailed. In fact, the representation of knowledge does not only mean the basic knowledge that the machine acquires through information but also the broad knowledge that the machine acquires through the experience created by people and between people. However, the representation of knowledge on a large scale is overly complex because it is necessary to build a general ontology capable of understanding different topics and subjects (Inozemtsev, Ivleva & Ivlev, 2017). To overcome this difficulty, researchers prefer to work on single domains of knowledge or expert systems with the aim of identifying a language to be used to represent knowledge (Brachman & Levesque, 2004; Zaytsev, Khalabiya, Stepanova & Bunina, 2020). An expert system provides answers only on a particular field of knowledge and the representation of knowledge is much simpler because the subdivision and hierarchy of categories are limited to a single domain. In this sense, the representation of knowledge is an area of specific interest in the scientific debate in the education area. It "oscillates between two types of metaphors: those that favor synchrony, the ordered state of the complex that results from certain distinctions and oppositions represented spatially (…); those mainly diachronic which favor the connection between the elements "and the relationships that exist between the elements of a system" (Cammozzo, 2011). Each teacher or trainer or training designer must in fact reflect on his idea of knowledge and how certain knowledge is represented through predictive models to support the teaching/learning processes. Specifically, the reference here also is to the injection of knowledge (Borghesi, Baldo, Lombardi, Milano, 2021) into the machine learning model, where human knowledge is explicitly used to feed the data-driven model. In particular, a key point of our project is the inclusion of new variables directly linked to students' learning among the predictive features, to integrate those commonly used to represent the economic, social, and cultural status (ESCS) (Thomson, 2018). This increases the expressiveness of the model and enables its integration into a decision support system (Zingaro et al., 2020; Del Bonfiro et al., 2020). The problem of feature construction is not new in data-intensive research and is part of a feature engineering process aimed at optimizing the representation of knowledge to improve the tool performance (Sondhi 2009; Motoda & Liu, 2002).

3. Methods

3.1. The INVALSI Dataset

INVALSI is the national institute for the evaluation of the Italian school system. Since 2003, the institute organizes large-scale assessment tests in Italian, Maths, and English for grades K–2, K–5, K–8, K–10, and K–3 in a systematic manner. In addition to this, there is a student survey aimed at gathering information on the socioeconomic and cultural context. We can sum up the data collected through the test on individual students into five main categories of features: (i) Identification marks, such as school code, class code, and SIDI invalsi code; (ii) Boolean variables on the correctness of the answers given to each item of the test; (iii) features about students' demographic information, that is, gender, month, year and place of birth, country of origin, province and region of residence, type of school attended, Italian and Math grades; (iv) categorical data about parents' demographic information, that is, educational qualification, job, and birthplace; (v) synthesis indices used to express the degree or level of certain aspects of interest, such as WLE (Weighted Likelihood Estimation of ability according to Rasch Model) and ESCS (Economic, Social and Cultural Status).

An important strength of the data set is represented by the possibility to longitudinally link data collected in different school years. In fact, starting from 2011/12, students are identified through a unique code, called "SIDI invalsi," which allows tracking all the tests of the same student over the different school years. In our case study, the model was trained and validated using the INVALSI test data in the s.y. 2013/14 at K–8 (hereinafter reference test) and the prediction target is computed based on the skills levels achieved by the same students at the K–13 grade five years later. Specifically, on a 1–5 scale, low achievement occurs when the global skills level does not exceed 2. After preprocessing, the data set consists of 228,993 students and 40 features. More information about preprocessing and feature extraction is provided in the next subsections.

3.2. Decision Tree and Experimental Setup

As a baseline for our predictive model, we choose to use decision trees (Breiman. et al., 2017) as a classification algorithm due to their flexibility, their relatively low computational complexity, and highly interpretable outputs if compared to other classification algorithms. The output of a decision tree is a flowchart-like structure where each internal node represents a "test" on an attribute that splits the instance space into two sub-spaces. The leaves, that is, the terminal nodes of the tree, represent the target class (in the case of classification) for each specific path. "Tests" can be either Boolean conditions on numerical features (e.g., the feature is "less than" or 'greater than a certain threshold) or belonging to categorical feature (e.g., feature "belongs" or "does not belong" to a certain class). The "splits" happen by optimizing an arbitrary quantity, in our setting the entropy—that is, a measure accounting for information brought by that feature—a common choice for supervised problems (Priyam et al., 2013). To improve the model performance, we used to prune by looking for the best max depth for the tree that reduces the complexity of the final classifier and overcomes overfitting problems. The tuning was based on the grid search technique (Lerman, 1980).

The data set was preprocessed cleaning missing values and highly correlated features (computed by□2 measure above 0.5). We carried out the experiments using the Google Collaboratory Notebook environment using the Python programming language, and popular machine learning libraries, such as scikit-learn and pandas. A comparison between the students in the K–8 dataset of the 2013/14 school year and those in the K–13 dataset of 2018/19 shows that about 30% of the students are missing. We can assume that this is due to several causes. Some students may have repeated one or more classes and this did not get them to grade K–13 in the expected five years. Others may have chosen vocational training paths that end at grades K–11 or K–12. Someone else could be part of the phenomenon known as school dropout. Other students may not have taken the K–13 test because they were absent on the day it was administered. A minority may have escaped the INVALSI tracking because they completed their studies in another country or entered academic studies earlier. In our case study, we decided to focus on a predictive model for low achievements, so we removed students who were missing over the five years for whatever reason from the study. Even after this operation, the cohort is still made up of 228,993 students. However, the two classes used to define the target were unbalanced with 34% in the low achievement class against 66% for standard or high achievement. To tackle this problem, which can affect the model fairness and introduce a bias that favors the oversampled class, we used a random undersampling technique from the library Imblearn. The resulting sample was used to define both the training set (75%) and the validation set (25%). For the test set, we are going to use the cohort of K–13 students in the school year 2020/21 as soon as the data are made available.

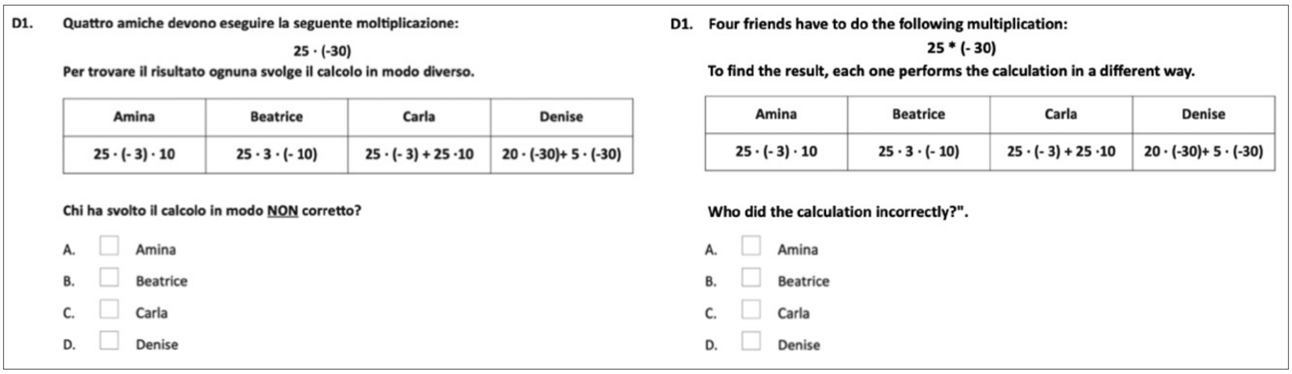
3.3. Students' Learning Representation

A key methodological point in the design of this predictive model is the way in which students' learning is encoded. There are two main reasons that make this point crucial. First, to make the model transferable to other cohorts of students or adaptable to data sets obtained through other large-scale assessments tests, it is necessary to release the data set from the individual items that constitute a certain test, in favor of a more abstract and general representation. Furthermore, being able to represent students' learning through features whose semantics are interpretable and explainable allows for better integration and usability of the system in the educational process and by the stakeholders (teachers, students, policymakers).

In this work, we used a knowledge-based approach for students' learning encoding, that is, we refer to a theoretical framework to obtain a suitable representation that can be exploited by the model. Specifically, we consider the classification of the items in terms of areas, processes, and macroprocesses according to the INVALSI framework for the design of maths tests. Each item is also associated with a level of difficulty on a scale from 1 to 3, provided with the correction guide for teachers. Below we show a table (Table 1) as a reference overview of the areas, processes, macroprocesses, and difficulty levels that have been used in the encoding of the questions. Moreover, in Figure 1 we present an item example, the first one in the reference test, with its classification in area, process, and macroprocess. This item is classified as area NU (numbers), process P1, macroprocess P1, and difficult level D1.

Table 1. Mathematical framework for items classification

Areas	Process	Macro-process	Difficulty
(NU) Numbers	(P1) Know and master the specific contents of mathematics	(MP1) Formulating	(D1) Easy
(SF) Space and figures	(P2) Know and use algorithms and procedures	(MP2) Interpreting	(D2) Medium
(DF) Data and forecasts	(P3) Know different forms of representation and move from one to the other	(MP3) Employing	(D3) Difficult
(RF) Relations and functions	(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 tyical forms of mathematical thought		
	(P7) Use tools, models and representations in quantitative treatment information in the scientific, technological, economie and social fields		
	(P8) Recognize shapes in space an,d use them for problem solving		


D1. Quattro amiche devono eseguire la seguente moltiplicazione:
25 · (-30)
Per trovare il risultato ognuna svolge il calcolo in modo diverso.

Amina	Beatrice	Carla	Denise
25 · (- 3) · 10	25 · 3 · (- 10)	25 · (- 3) + 25 ·10	20 · (-30)+ 5 · (-30)

Chi ha svolto il calcolo in modo <u>NON</u> corretto?

A. ☐ Amina
B. ☐ Beatrice
C. ☐ Carla
D. ☐ Denise

D1. Four friends have to do the following multiplication:
25 * (- 30)
To find the result, each one performs the calculation in a different way.

Amina	Beatrice	Carla	Denise
25 · (- 3) · 10	25 · 3 · (- 10)	25 · (- 3) + 25 ·10	20 · (-30)+ 5 · (-30)

Who did the calculation incorrectly?".

A. ☐ Amina
B. ☐ Beatrice
C. ☐ Carla
D. ☐ Denise

Figure 1. Item example from the INVALSI math test in the school year 2013/14. On left, there is the original item, on the right its translation.

For each area, process, macroprocess and level of difficulty, a new feature has been defined that expresses a correctness rate based on the answers provided by the students. For example, the new feature "Numbers" would assume the correctness rate of the items that belong to the area "Numbers". The correctness rate denotes the percentage of correct answers given to the items of that feature, so the value of "Numbers" is the ratio between the number of items belonging to "Numbers" for which the student's answer is correct and the total number of items for "Numbers". Last, we concatenate the computed values to obtain a new flattened representation of learning, where each item is a possible indicator and not its unique representative. Following our strategy, we represent each student's learning level in the space of 18 dimensions, one for each area, process, macroprocess, and difficulty level. Table 2 shows two examples of students' learning encoding.

Table 2. Students' learning encoding examples.

Id	NU	SF	DF	RF	P1	P2	P3	P4	P5	P6	P7	P8	MP1	MP2	MP3	D1	D2	D3
1	0.55	0.333	0.83	1.00	0.8	0.50	1.00	0.50	0.50	0.50	1.00	0.50	0.50	0.79	0.73	0.84	0.67	0.17
2	0.73	0.78	0.67	0.91	1.00	0.75	1.00	0.75	0.50	1.00	0.78	0.50	0.75	0.79	0.82	0.95	0.50	0.83

4. Results and discussion

The main contribution of this position paper lies in the methodological proposal on the representation of students' learning and its effective integration in an AI system. Therefore, in this section we present two main points: the evaluation of our baseline model on the validation set, which is promising in supporting our methodological strategy, and the suggestion of some open challenges arising from our experiment, which may contribute to a wider debate about human and machine knowledge interaction.

As for the results of our baseline model, we considered three metrics for its evaluation:

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{TP+FN}$$

where TP is the number of true positives, FN the number of false positives, TN the number of true negatives, and FN the number of false negatives. In our model, students are positive if their predicted class is "low achievement"; TP means that also the value of the actual class is "low achievement," FP means that the actual class is "no low achievement." In analogy, we can define TN and FN. In our experiment, we obtained 76% accuracy, 60% precision, and 78% recall. A particular interest is addressed to recall; in fact, a high recall score indicates both a reduction in false negatives (those who would need a support intervention and are not intercepted by the model) and validates the selection criteria learned from the model as effective indicators of intervention areas. The decision tree obtained after optimization through pruning has a depth of eight levels. In Figure 2, we present the first four levels of the tree. The economic, social, and cultural context variables play a significant role in the selection criteria. However, there are other key factors. The first split of the tree is based on the school's grade in mathematics. In the second level, and in many of the following, a split criterion on the D1 factor is used. Starting from the fifth level of the tree, many split criteria are defined with the other features added to our students' encoding process. All macroprocesses are involved, while the DF area does not appear, and among processes, only P2, P4, P5, and P7 are considered.

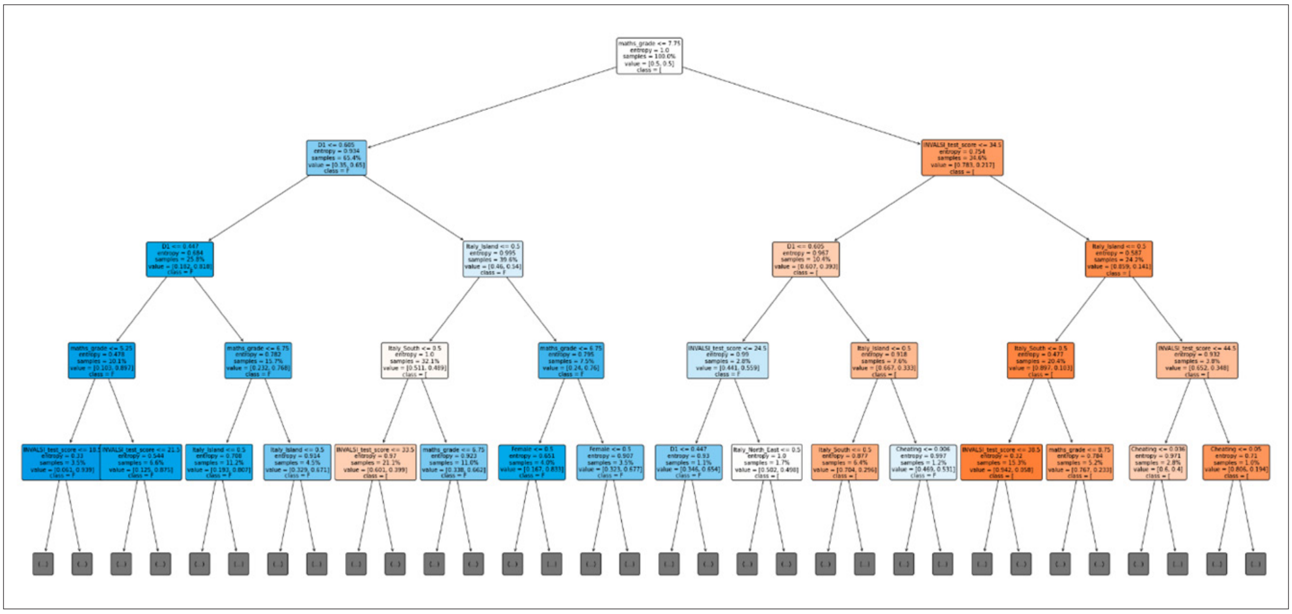


Figure 2. Decision Tree obtained for the model.

A reliable and truthful interpretation of these pieces of evidence raises critical issues (Zanellati et al., 2021). Of particular interest in this contribution is the theory-ladenness (Pietsch, 2015) due to how students' learning is represented and encoded. The definition of new features as described above refers to a theoretical-didactic framework, which is explicitly reflected on the data used as input for the model. This is a knowledge injection in the machine learning model, where human knowledge is explicitly used to feed the data-driven model. Furthermore, the way this representation occurs and is included in the model affects the interpretation of its outputs, in terms of transparency and explainability. For example, the use of the complexity levels of the items as independent features allows their explicit use to define selection criteria in the decision tree. An alternative representation may consider them as weights to use in the definition of the correctness rates of each area, process, or macroprocess; in this case, they could no longer be used transparently in the selection criteria in favor of a different, but plausible, representation of students' learning.

In other words, we are wondering what kind of knowledge is introduced in the model and how it affects the knowledge produced by the designed AI. This opens a further question on how to reduce the weight of this bias or to deal with it when interpreting the results.

Conclusions

The reflections on post-digital highlight the need to redefine our relations with the digital (Knox, 2019) and specifically, in the educational field, to recognize how this technology is already incorporated and intertwined with didactic practices. This case study highlighted the need to reflect on what it means to represent students' learning when it has to be included in an AI system for education. The problem of how to represent students' learning is certainly not new in education (Wang, Sy, Liu & Piech, 2017) and requires an assumption of responsibility, which is played out on several levels, including the design of learning indicators, the ability to analyze and interpret these indicators according to an experiential and cultural background, the communication of its evaluation, its inclusion in a dynamic and changing training path and many others. With respect to RQ1, we can affirm that data collected through large-scale assessment tests seems to be promising to represent students' learning in an AI tool to tackle the problem of students' underachievement. In our proposal, the representation is the result of a feature engineering process based on a theoretical framework of areas, processes, and macroprocesses for the classification of the items. This method could be transferred to other large-scale assessment tests or other disciplines, with the appropriate adaptations to the specific theoretical framework used in designing the test items, and it is one of the future developments of this research. Furthermore, we aim to exploit these students' learning encoding to develop predictive models based on neural networks.

As for RQ2, we can affirm that our model shows a flow of knowledge between that supplied with input data, after the preprocessing to obtain the students' learning representation, and that returned in output and that should be interpreted by a human expert. The students' learning representation becomes a fundamental key to making these points of dialogue possible. In particular, the features used for the representation define the selection criteria that determine the decision tree learned by the machine. On the one hand, they are the scheme with which to "reorganize" the raw data, on the other, they become a key to interpreting the predictive results of the model from an educational point of view. In this sense, post-digital can also be considered a return to fundamental educational concerns in a renewed didactic context.

Author Contributions

This contribution, developed and shared jointly by the four authors, was drawn up as follows: §1 by Chiara Panciroli; §2 by Anita Macauda; §3 by Andrea Zanellati; §4 by Maurizio Gabbrielli and Andrea Zanellati; the "Conclusions" have been written jointly.

References

Aljawarneh, S., Lara, J.A. (2021). Data science for analyzing and improving educational processes. *Journal of Computing in Higher Education*, 33, 545-550. https://doi.org/10.1007/s12528-021-09299-7

Baker, R. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining*, *11*(1), 1-17. https://doi.org/10.5281/zenodo.3554745

Bayne, S. (2015). What's the matter with 'technology-enhanced learning'? *Learning, Media and Technology*, *40*(1), 5-20. https://doi.org/10.1080/17439884.2014.915851

Blossfeld, H. P., & Roßbach, H. G. (Eds.). (2019*). Education as a Lifelong Process: The German National Educational Panel Study (NEPS*), Cham: Springer.

Borghesi, A., Baldo, F., Lombardi, M., Milano, M. (2021). Injective Domain Knowledge in Neural Networks for Transprecision Computing. *Machine Learning, Optimization, and Data Science: 6th International Conference* (pp. 587-600), LOD 2020, Cham, Springer. https://doi.org/10.1007/978-3-030-64583-0_52

Brachman, R. J., & Levesque, H. J. (2004). *Knowledge Representation and Reasoning*. Amsterdam: Elsevier.

Branchetti, L., Ferretti, F., Lemmo, A., Maffia, A., Martignone, F., Matteucci, M., & Mignani, S. (2015). A longitudinal analysis of the Italian national standardized mathematics tests. In *Proceeding of CERME9*, 1695-1701.

Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. (2017). *Classification and regression trees*. London, UK: Routledge.

Cammozzo, A. (2011). *Né rizoma, né albero: metafore autopoietiche nella rappresentazione della conoscenza*. ZeL Edizioni: Romeo Galassi.

Cramer, F. (2015). What is 'Postdigital'? In D. M. Berry & M. Dieter (Eds.), *Postdigital aesthetics: art, computation, and design*. New York: Palgrave Macmillan

Del Bonifro F., Gabbrielli M., Lisanti G., & Zingaro S.P. (2020). Student Dropout Prediction. *AIED* (1), 129-140.

Ertl, B., Hartmann, F. G., & Heine, J. H. (2020). Analyzing Large-Scale Studies: Benefits and Challenges. *Frontiers in Psycholog*y, 11. https://doi.org/10.3389/fpsyg.2020.577410

Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy 34(2)*, 470-499. doi: 10.1080/02680939.2018.1460493

Fuller, S., & Jandrić, P. (2019). The postdigital human: making the history of the future. *Postdigital Science and Education*, *1*(1), 190-217. https://doi.org/10.1007/s42438-018-0003-x.

Inozemtsev, V., Ivleva M. and Ivlev, V. (2017). Artificial Intelligence and the Problem of Computer Representation of Knowledge. *Advances in Social Science, Education and Humanities Research*, 124, 1151-1157.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3), 101493.

Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2017). Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education*, 25, 1066-1078. https://doi.org/10.1002/cae.21844

Knox, J. (2019). What Does the 'Postdigital' Mean for Education? Three Critical Perspectives on the Digital, with Implications for Educational Research and Practice. *Postdigital Science and Education* 1, 357-370. Doi:10.1007/s42438-019-00045-y

Kumar, V., Garg, M. L. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, 182, 31-37.

Lacković, N. (2020). Postdigital Living and Algorithms of Desire. *Postdigital Science and Education*. https://doi.org/10.1007/s42438-020-00141-4

Lerman, P. M. (1980). Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 29, 1, 77-78

Motoda, H., & Liu, H. (2002). Feature selection, extraction, and construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan), 5*, 67-72.

OECD (2018). Equity in Education: Breaking down barriers to Social Mobility, Paris: PISA, OECD Publishing.

Panciroli, C., Rivoltella, P.C., Gabbrielli, M. & Zawacki-Richter, O. (2020). Artificial Intelligence and education: new research perspectives. *Form@ re-Open Journal per la formazione in rete*, *20*, 1-12.

Panciroli, C. (2021). Il postdigitale. Società, cultura e didattica. *Scholé. Rivista di educazione e studi* culturali, LIX(2), 157-166.

Panero, M. (2019). Un'analisi longitudinale dei dati INVALSI di matematica di una stessa coorte alla scuola primaria. *Working Paper*, 41, 1-19.

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, *82*(5), 905-916.

Popenici, S., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*. https://doi.org/10.1186/s41039-017-0062-8.

Priyam A., Abhijeeta G., Rathee A., & Srivastava S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology,* 3(2), 334-337

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wires Data Mining and Knowledge Discovery*, 10, e1355. https://doi.org/10.1002/widm.1355

Ryberg, T. (2021). Postdigital Research, Networked Learning, and Covid-19. *Postdigital Science and Education* 3(4), 266-271. doi: 10.1007/s42438-021-00223-x

Rivoltella, P.C., & Rossi, P. G. (2019). *Il corpo e la macchina. Tecnologia, cultura, educazione*. Brescia: Morcelliana.

Shi-Nash, A., Hardoon, D.R. (2016). *Data analytics and predictive analytics in the era of big data*. Palo Alto, Hwaiyu Geng.

Sondhi, P. (2009). Feature construction methods: a survey. *sifaka. cs. uiuc. edu*, *69*, 70-71.

Thomson, S. (2018). Achievement at school and socioeconomic background - an educational perspective. *Science Learn* 3, 5. https://doi.org/10.1038/s41539-018-0022-0

Wang, L., Sy, A, Liu, L., & Piech, C. (2017). Learning to Represent Student Knowledge on Programming Exercises Using Deep Learning. *Proceedings of the 10th International Conference on Educational Data Mining,* 324-329.

Zanellati, A., Zingaro, S.P., Del Bonifro, F., Gabbrielli, M., Levrini, O. & Panciroli, C. (2021). Informing predictive models against Students Dropout. *Conference DIDAMATICA 2021*, Palermo.

Zaytsev E.I., Khalabiya R.F., Stepanova I.V., & Bunina L.V. (2020). Multi-Agent System of Knowledge Representation and Processing. In: Kovalev S., Tarassov V., Snasel V., Sukhanov A. (eds) Proceedings of the Fourth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'19). IITI 2019. Advances in Intelligent Systems and Computing, vol 1156. Cham: Springer. https://doi.org/10.1007/978-3-030-50097-9_14

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 1-27. doi:10.1186/s41239-019-0171-0

Zingaro, S., Del Zozzo, A., Del Bonifro, F., & Gabbrielli, M. (2020). Predictive models for effective policy making against university dropout. *Form@ re-Open Journal per la formazione in rete*, *20*(3), 165-175.