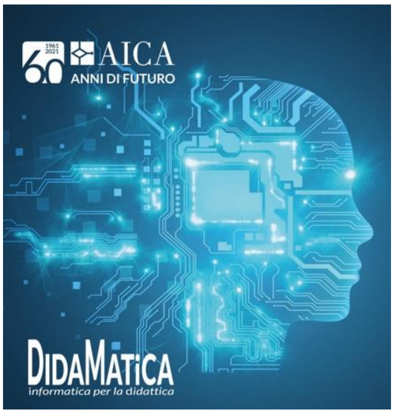
Giovanni Adorni, Mario Allegra, Salvatore Gaglio, Manuel Gentile e Nello Scarabottolo

AICA

Consiglio Nazionale
delle Ricerche

itd
ISTITUTO TECNOLOGIE DIDATTICHE

ATTI DEL CONVEGNO
DIDAMATICA 2021
7-8 OTTOBRE | PALERMO


1963
2023
60
AICA
ANNI DI FUTURO
DidaMatica
informatica per la didattica

Atti Convegno Nazionale

DidaMatica
informatica per la didattica
Artificial Intelligence
for Education
PALERMO | 7-8 OTTOBRE
CNR Istituto Tecnologie Didattiche
Consiglio Nazionale delle Ricerche
itd
ISTITUTO TECNOLOGIE DIDATTICHE
60 AICA
ANNI DI FUTURO

35ª edizione

Area della ricerca di Palermo
del Consiglio Nazionale delle Ricerche
Istituto per le Tecnologie Didattiche

Palermo, 7-8 ottobre 2021

A cura di

**Giovanni Adorni, Mario Allegra, Salvatore Gaglio,
Manuel Gentile e Nello Scarabottolo**

ISBN 978-88-98091-62-1

AICA

Consiglio Nazionale
delle Ricerche

itd
ISTITUTO TECNOLOGIE DIDATTICHE

Atti Convegno Nazionale DIDAMATiCA 2021

Area della ricerca di Palermo del Consiglio Nazionale delle Ricerche– Istituto per le Tecnologie Didattiche
Palermo, 7-8 ottobre 2021

A cura di: Giovanni Adorni, Mario Allegra, Salvatore Gaglio, Manuel Gentile e Nello Scarabottolo

ISBN**: 978-88-98091-62-1**

Risorse e aggiornamenti relativi a questi Atti sono disponibili all'indirizzo **www.aicanet.it/didamatica2021**

Piazzale Rodolfo Morandi, 2 - 20121 Milano
Tel. +39-02-7645501 - Fax +39-02-76015717 www.aicanet.it

Prima edizione: Ottobre 2021

Editing a cura di: Manuel Gentile e Mario Allegra

Prefazione

DIDAMATiCA - DIDAttica e infor**MATiCA -** (Informatica per la Didattica), organizzata annualmente da **AICA**, l'Associazione Italiana per l'Informatica e il Calcolo Automatico, è giunta quest'anno alla sua 35ª edizione. Negli anni **DIDAMATiCA** è divenuta un punto di riferimento per studenti, docenti, istituzioni scolastiche, professionisti ICT, aziende e Pubblica Amministrazione sui temi dell'innovazione digitale per la filiera della formazione. Ponte tra scuola, formazione, ricerca e impresa, tiene vivo il confronto su ricerche, sviluppi innovativi ed esperienze in atto nel settore dell'Informatica applicata alla Didattica, nei diversi domini e nei molteplici contesti di apprendimento.

DIDAMATiCA 2021 è una edizione particolarmente importante per **AICA**: l'evento prende il via nell'anno in cui l'*Associazione compie i suoi primi 60 anni* e, in continuità con le edizioni passate, vuole essere l'occasione per una riflessione concreta e strutturata sul tema dei nuovi scenari digitali imposti nel mondo della Scuola, del lavoro e della società dalle tecnologie digitali sempre più pervasive, immersive e sempre più "intelligenti".

In collaborazione con l'Istituto di Tecnologie Didattiche del CNR - Consiglio Nazionale delle Ricerche, **DIDAMATiCA 2021** viene organizzata in modalità mista, in presenza presso la sede dell'Istituto di Tecnologie Didattiche di Palermo e online per facilitare la partecipazione anche viste le criticità della situazione pandemica ancora in atto. Tale edizione si propone di aprire un confronto con i principali protagonisti del settore su un tema ormai non più di frontiera ma sempre più vicino alla realtà della didattica: come i rapidi avanzamenti nel campo dell'AI - *Artificial Intelligence / Intelligenza Artificiale* influiranno sui processi di insegnamento e apprendimento e della formazione in genere.

Il concetto di AI potrebbe sembrare qualcosa di molto distante dal mondo della scuola: un tema lontano dalla realtà odierna, soprattutto nel momento in cui il sistema scolastico e della formazione è impegnato a fronteggiare i problemi connessi all'emergenza pandemica. Tuttavia, l'impatto che questa tecnologia innovativa sta producendo in diversi ambiti della nostra vita (marketing, finanza, salute e sicurezza solo per citarne alcuni) impone una riflessione anche al mondo della scuola e della formazione.

Come indicato nel recente report dell'UNESCO "*Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*" (UNESCO, 2019), l'integrazione della AI nell'ambito *education* solleva diverse questioni. Ai temi di natura etica legati alla raccolta massiva di dati utili per la profilazione degli studenti e la personalizzazione dei percorsi di apprendimento, si aggiunge la necessità di approfondire le riflessioni, già avviate in altri settori, sulla trasparenza dei processi decisionali dei sistemi e/o algoritmi di Artificial Intelligence (settore di ricerca identificato con il termine *explainable AI*). L'integrazione delle tecniche di AI nei processi educativi richiede un ulteriore approfondimento sui temi del "digital divide" e dell'inclusione sociale, sui rischi connessi a tali innovazioni ma anche sulle opportunità che le tecnologie offrono per gestire questi temi con approcci nuovi. Infine, occorre riflettere anche e soprattutto sul ruolo dei docenti e su quali competenze debbano avere e su quali strumenti fornire loro per renderli attori consapevoli di questi processi di innovazione.

DIDAMATiCA 2021 "Artificial Intelligence for Education" indica una sfida e un'opportunità per rendere la Scuola e il mondo del lavoro produttivi e *smart*, rendere Studenti e Docenti attori consapevoli e capaci di mettere in atto comportamenti sicuri e pronti ad affrontare le sfide e minacce del futuro in un mondo *sempre più smart*. **DIDAMATiCA 2021** vuole essere un'opportunità non solo per fotografare lo stato attuale nel processo di integrazione fra il mondo della Scuola e del lavoro attraverso la raccolta di buone pratiche, ma vuole anche aprire il confronto teorico e metodologico su come pensare i processi di apprendimento e insegnamento per trarre beneficio da tali innovazioni.

Due giornate dedicate al confronto in presenza e a distanza e che proseguiranno per una ulteriore settimana con discussioni moderate in rete attraverso apposito portale. Obiettivo del confronto sarà la definizione di un'agenda di ricerca per tutti gli attori dell'innovazione didattica che sono interessati a esplorare i temi connessi all'Artificial Intelligence.

I contributi scientifici, selezionati dal Comitato Scientifico sulla base della doppia valutazione effettuata per ogni singolo lavoro sottomesso da parte del Comitato dei Revisori, sono suddivisi in sette sessioni:

- Algoritmi e modelli di Intelligenza Artificiale nel mondo della Scuola
- Intelligenza Artificiale in classe
- Le tecnologie educative e la loro evoluzione nell'era dell'Intelligenza Artificiale
- Le Tecnologie Educative nella Scuola
- Serious Games
- Innovazione delle pratiche educative
- Il docente, il formatore e le tecnologie didattiche

Come tradizione di **DIDAMATiCA**, gli abstract di tutti i lavori presentati al Convegno e contenuti in questo volume vengono pubblicati su un numero speciale della rivista **MONDO DIGITALE** che ospiterà, in forma estesa, anche i Best Paper che verranno selezionati con una ulteriore doppia revisione da parte del Comitato Scientifico.

Vogliamo ringraziare quanti hanno reso possibile **DIDAMATiCA**. In modo particolare lo staff di AICA, che ha avuto modo di esprimere ancora una volta alta professionalità e capacità di soddisfare tutte le necessità organizzative, adeguandosi anche ai cambiamenti legati alle nuove forme di comunicazione, i colleghi della Sezione AICA Sicilia, i colleghi dell'Università degli Studi di Palermo, e i colleghi dell'Istituto per le Tecnologie Didattiche del CNR di Palermo per il lavoro svolto e per l'ospitalità al Convegno.

Giovanni Adorni
Mario Allegra
Salvatore Gaglio
Manuel Gentile
Nello Scarabottolo

Sommario

Informing predictive models against Students Dropout

Andrea Zanellati[1], Stefano Pio Zingaro[1], Francesca Del Bonifro[1],
Maurizio Gabbrielli[1], Olivia Levrini[2], and Chiara Panciroli[3]
[1] DISI, University of Bologna, Via Mura Anteo Zamboni 7, 40126 Bologna, Italy
fandrea.zanellati2, stefanopio.zingaro,
maurizio.gabbriellig@unibo.it
[2] DIFA, University of Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
olivia.levrini2@unibo.it
[3] EDU, University of Bologna, Via Filippo Re 6, 40126 Bologna, Italy
chiara.panciroli@unibo.it

Abstract

Students' dropout is a complex widespread phenomenon which often lead to conditions of social, educational and professional exclusion. The design of Early Predictive Analytic Models can be a valid tool to counteract this phenomenon, which can be further enhanced by using Machine Learning. In this position paper we aim to contribute with two main points. First of all, we introduce the prominent position of the skills assessment, considered both as a target or as input data for the model, as essential integration to demographic and of economic, social and cultural status variables, often used as predictors for dropout risk. This leads us also to give a definition of implicit dropout, i.e. failure to achieve the expected skills, applicable in different educational contexts. Furthermore, we highlight the importance of integrate the predictive models in a broader framework described as a sequence of phases. The framework stresses the need to make the model "informed" at three levels: a reference pedagogical theory (a theory-laden dimension in a data-intensive approach); the persistence of the initial information and their integration together with the life cycle of the model (its creation, use and update); the guidelines to enable the explainability and transparency of the model outcomes, in accordance with the principles of Trustworthy AI. These contributions are presented both through an abstract description and an undergoing case study in Italian school system.

1 Introduction

Dropout in the educational and training field is a complex phenomenon characterised by different forms according to social environment, gender, age and geographical location. It manifests itself in different forms of students educational exclusion: get lost from one cycle to another, be not intercepted or dispersed in the first two years of the High School and at the beginning of the academic career, do not learn enough or acquire uncertain, fragmented and never consolidated knowledge, evade the obligation or attend occasionally and passively, etc. These aspects are often characterised by some co-morbidity and lead to an Achievement Gap, i.e. the unequal or inequitable distribution of educational results and benefits.

A further element of complexity is due to the involvement of different stakeholders, i.e. politicians, managers, teachers, parents and students, each of which is driven by different interests. Policy makers, for example, could be interested in using predictive models as descriptors of the most relevant factors with respect to the risk of dropout; they could have the main purpose of reforming curricula or initiating system actions to correct, stem or reduce the factors most causally involved in the dropout effect. As for school managers, they may be interested in improving the quality indices compared to other schools in the same area or of the same type; predictive information on dropout can therefore be useful to support initiatives for school guidance, but also for the drafting of guidelines in school planning documents. Another perspective guides teachers and educators, who may be interested in targeted recovery or enhancement actions on individual students. Finally, students and their families may have an interest in personalised indications, perhaps to be carried out even outside the school context.

The different interests in contrasting the students dropout listed above can justify the social and political relevance of the study of this phenomenon. This is further supported by the fact that students dropout is closely linked to the prospects of cultural and professional growth, to the employment opportunities and to the level of well-being and life satisfaction. For this reason, the report "EUROPE 2020 - A strategy for smart, sustainable and inclusive growth" [1] indicates reducing the rate of the Early Leavers from Education and Training (ELET) [2] and increasing the share of the population aged 30-34 having completed tertiary education among the targets to be achieved in educational attainment.

The ELET rate, just mentioned, refers to young people between 18 and 24 with a qualification lower than upper secondary. It is one of the most used indicators for estimating the students dropout, but it is important to underline that it is not the only indicator. In particular, it underestimates the complexity and quantitative consistency of this phenomenon: first of all, it considers only school education by excluding academic paths or other post-diploma qualifications; moreover, it does not consider those forms of the phenomenon that determine a Learning Gap, i.e. the disparity between what students have actually learned and what they were expected to learn at a particular age or grade. This may motivate the need to define different metrics suitable for estimating the different forms in which dropout occurs.

A reference to this "silent" form of students dropout is found in [4] which introduces the *Implicit Dropout* through an operational definition strongly based on the type of data available to the INVALSI[1]. However, this definition can be easily extended and generalised to other educational and training systems. In practice, students for whom a considerable Learning Gap occurs implicitly increase the dropout rate: even if they did not actually drop out of school or university, their stay did not produce the expected effects on their skills.

In recent years AI tools and algorithms have been applied to counteract and better understand this phenomenon. In particular, according to [3]:

AI systems hold promise to improve early warning systems, which are increasingly based on longitudinal datasets that are emerging in education. Even though identifying risks does not imply solving them, AI solutions help school principals to use existing data in new ways and design interventions to predict and prevent dropout more efficiently.

More generally, the use of Early Predictive Analytic Models can support decision-making processes in charge of all educational system stakeholders.

The concept of *Early Prediction* particularly characterises the models used in the context of students dropout in two main perspectives. On the one hand, a forecast on dropout at a certain school grade using data at a lower school level it is important to be able to activate the recovery, support and consolidation interventions at the right moment in the growth and training path. On the other hand, if it is impossible to use data of a previous level, it becomes necessary to make a forecast with those that can be collected in the first months of the beginning of a training course, in order to

[1] INVALSI stands for Istituto Nazionale per la VALutazione del Sistema d'Istruzione (National Institute for the Evaluation of Education Systems). The institute administers tests of Italian, Mathematics and English on a national scale every year in different school grades

be able to carry out the supplementary interventions in parallel with the normal course of the undertaken path.

This idea of using information and data from the previous school level to characterise an early prediction arise a question about what information and what data are available and can be exploited. Research so far has often referred to demographic or to economic, social and cultural status variables [5]. However, from the point of view of a predictive system that allows actions aimed at eliminating or reducing the aforementioned risk, this type of variables is rigid and difficult to be affected by the pedagogical and formative actions of the school.

Summarising, the complexity of the phenomenon, its political relevance, the deployment of AI techniques and the variety of stakeholders and interests involved raise several issues with an epistemological, methodological and ethical nature. In this panorama we want to place our contribution through this position paper, inspired by an undergoing case studies about the estimation of the risk of Implicit Dropout at grade 13th of the Italian school system through the INVALSI data.

As for the first contribution, we aim to highlight the prominent position of the skills assessment for the understanding and prediction of students dropout. More specifically, the skills (or rather the estimate of their level) can be used as a target with which to identify the occurrence of dropout, or they can be used as features for training the model. This leads us both to suggest a general definition of implicit dropout that can be declined in different educational and training contexts and to suggest, also through our case studies, possible representations, measurements and encoding of skills for their inclusion in the creation of the predictive models.

The opportunity to include skills assessment in the predictive models for dropout brings the need to think them as a part of a broader system. Therefore, as second contribution, we aim to introduce a reference framework that we name *Informed Model System*. The main section of this work is dedicated to describe the different phases of this framework and to motivate the choice to focus on what it means "to inform" the model in different steps.

In the remainder of the paper, we try to substantiate these claims by providing both a more detailed description in support of more formal definitions of the concepts just introduced, and their declination and exemplification through the undergoing case study.

2 Skills Assessment for Early Dropout Prediction

The most natural use that can be made of a predictive model is the prediction of the probability that a certain event will occur, i.e. the probability with which a student will run into one of the many forms of the dropout. However, this is not the only possible interesting outcome of the model; in fact, if it is sufficiently interpretable and transparent it can be used for a greater understanding of the phenomenon. More precisely, one can try to understand which factors are most significant for predicting dropout, identifying for example the features that were most discriminating to predict the output of the model.

This second perspective on the use of the model makes it significant as a decision-support system, especially if it allows us to identify concrete actions to modify those factors that contribute to the occurrence of an undesirable event, in our case the students dropout. The main task of schools and universities is the training of students, and this means that the interventions that are most spontaneously implemented in these institutions are of a formative, educational and cultural nature. These interventions are therefore designed and implemented in order to build knowledge, skills and competences.

In accordance with [8] position paper, increasingly complex needs require skills extended to the motivational, socio-emotional and meta-cognitive components. Skills therefore acquire a central role also in terms of building competences suited to the complexity of the XXI century society. Hence, the need to design educational actions that connect the acquisition of knowledge to the development of skills that can support the quality and effectiveness of training choices and apply knowledge in increasingly evolving contexts.

Considering both the high significance of skills and the type of interventions that can actually be implemented through teaching, we believe that the assessment of skills should be included in the predictive model for dropout risk, at least to supplement the variables used for information of a demographic nature or on the social and economic status. There are at least two levels at which this can occur.

On the one hand, the skill level can be used as a feature in the dataset, whose values can be used for a representation of the student. In other words, the skills level, appropriately detected and represented, can become a useful element for the students encoding and can be integrated in the inputs for the model. Moreover, a further way to exploit the skills for a representation of students and their learning is to monitor their levels progression over time through appropriate variables. For this integration of skills as input for the model, the problem of encoding becomes central; in the last section of this contribution we present how we tackle this issue in a case study.

On the other hand, the estimation of the skill levels can become the target for the model, if we consider the forms of dropout linked to the Learning Gap, which we have referred to with the term implicit dropout. The implicit nature is due not only to the fact that there is no actual leaving from school or training, but also to the fact that the measurement of skills cannot take place directly, but requires the definition of indicators that are considered significant [9,10]. In general, we can mean by implicit dropout the failure or partial achievement of the expected skill levels at the end of a training or educational path or when a summing-up evaluation occurs. To update this definition in contexts which can differ in many ways (for example by type of training, by order of study, by duration or by stakeholders), it is necessary to carry out a sequence of steps.

The steps can be summarised as follows: definition of the framework of skills, definition of their expected levels, definition of the indicators for their measurement, definition of the criteria by which it is established that implicit dropout has occurred. The case we present below regarding the predictive model with INVALSI data helps to exemplify these steps. Furthermore, the need to resort to these definition steps highlights, through a concrete case, the need for a reference theory when observing and interpreting an educational phenomenon. This aspect will be the starting point for the description of the Informed Model System as a framework in which to integrate the AI we use as a predictor.

3 Informed Model System

As a second contribution in this paper we want to introduce a framework in which to contextualize the design and the use of predictive models for the risk of dropout through ML methods. The need to insert the model into a broader system was suggested by a meta-analysis of the case studies that we present in the last section. We have identified three main reasons for this framework, each of which motivates one of its macro-phases.

The first reason has already been mentioned in the previous section: interesting and relevant factors in education are often not measurable or directly detectable. This requires identifying possible indicators with which to collect data attributable to the factors of interest. In the case of student dropout, for example, we have identified skills as a factor of interest. This intrinsic feature of educational research means that a purely data-driven approach is not applicable and a theory-laden component should be considered [7.

The explanation of this theoretical basis becomes decisive both on the way in which the data is collected or pre-processed during the features extraction or selection phase and for the determination of any encoding or embedding for the input to the model. In other words, it is a first level of information on which the actual predictive model is based. The steps that make up this first phase of the framework that we are going to define in the case of students dropout have already been described in the previous section; we could indicate this phase as the identification of a reference theory.

The second motivation for looking for a larger system in which to integrate the ML model lies in its positioning within a Decision Support System (DSS). There are two aspects that we can consider. Firstly, there are several possible users for the DSS. We have already highlighted that the stakeholders of the training and educational environments are many with different interests, professionalism and awareness. Each of them can be supported by a predictive model on the risk of dropout in an appropriate way to their expectations and this is reflected, for example, on the choice of the ML technique to use, the type of metric to consider to evaluate the predictive effectiveness, the target choice and representation or on data preprocessing. In addition to this, if the support translates into an intervention aimed at reducing this risk, it becomes useful to foresee that the model can be updated, considering new variables that can be included to represent the status of this intervention.

These considerations are also found in [6] which describe a three-steps ML model cycle (the creation, which includes training and testing, the use, i.e. the act of predicting, and the update) strictly concerning the life of the predictive model for the risk of dropout and which constitute the second macro-phase of our framework. We emphasise that even in this phase there are two steps in which the system "is informed": in the creation step design choices on the model are made by taking into account the knowledge we have about the stakeholders and their interests; moreover, the update step is based on a "fallout" of the knowledge generated by the model on itself.

The third reason is linked to the issue of interpretability, one of the key principles of Trustworthy AI. This issue includes the Transparency and Explainability of the model: the first refers to the possibility of understanding the logic and criteria learned by the AI model to solve the prediction task; the second can be seen as the possibility to gain insight from the machine learning model, eventually by using statistical methods, which helps to understand the outcomes.

On the one hand, the search for transparency for the model mainly influences the choice of the ML technique while, on the other hand, the problem of the interpretability of the model means safeguarding the undeniable interaction between human-users and the AI-model by providing information tools and supports that allow communication between these two parts. The third macro-phase of the framework that we propose is the effective integration of the ML model in the DSS which includes the information tools that allow its interpretation.

As a final observation, we want to highlight that in the different macro-phases (the identification of the reference theory, the three-steps ML model cycle and the integration that can be interpreted in the DSS) an information phase always intervenes. In all three phases there is an information component external to the system which is intentionally added as a methodological design tool. Moreover, in the second phase there is a further information component generated by the AI itself. This leads us to name the framework described as *Informed Model System*.

In the next section we will try to exemplify the macro-phases and the steps with a case study for which the research is under development and the final integration phase is still missing.

4 School Dropout with INVALSI data

The case study we present refers to a predictive model for implicit dropout in the Italian school, exploiting the INVALSI dataset. In particular, it was decided to examine the problem of the Learning Gap with respect to maths skills level, developing a predictive model that uses as input data the results of the INVALSI tests at grade 8th to predict the risk of implicit dropout at grade 13th. It is therefore a model for the dropout of skills at a disciplinary level, in this case mathematics, which can be easily transferred to Italian and English, which are the other two disciplines of which INVALSI aims to detect learning with specific tests.

As for the first phase of the Informed Model System, a solid basis that can be used for the explication of a reference theory is found in the INVALSI guidelines[2] which describe a skills framework. More specifically, four main areas are identified (numbers, space and figures, data and predictions, relations and functions), eight processes (for example "know and use algorithms and procedures") and three macro-processes (interpreting, formulating, using). The same documents also describe the expected levels, which are detected through items that are classified according to the area-process-macro-process framework just described. The items proposed by the INVALSI

[2] The reference framework are available at https://invalsi-areaprove.cineca.it/index.php

test are considered as skills indicators. This leads to a generalisation problem since the test administered in different school year has different indicators for detecting the skills levels. The last step of this first phase is the definition of the dropout occurrence criterion. For this case study we start from the assumption underlying the design of the INVALSI tests themselves, for which the expected skills level is considered achieved with a score equal to 3 out of 5; therefore it is considered that the implicit dropout occurred with a level less than or equal to 2.

As regard the ML model cycle, we begin by describing what are the inputs and outputs for the model and their representations, in a first case considered as a baseline. For the training and validation step, we have used a dataset obtained with an inner-join between INVALSI data at grade 8th in the school year 2013/14 and level of INVALSI at grade 13th in the year 2018/19. This excludes some students from the cohort of data, in particular those who suffer from other forms of dropout (compared to the implicit one) or not traceable for other reasons (e.g. the loss of a school year or the completion of studies in a foreign country). It has not yet been possible to test the models, as INVALSI data at grade 13th are not yet available for a cohort of students other than that used for training and validation.

In this step of model creation, and in particular for the training and validation phase, there were three main design choices. The first one is about encoding students to make them inputs for the model and it is "informed" by the reference theory described above. Each student is represented by some variables that contain demographic data and information on the economic, social and cultural status directly collected through a survey (data on the province and region of residence, profession and educational qualification of parents, gender, year of birth, school grade in mathematics and Italian). To these is added the score obtained in the INVALSI test at grade 8 and a variable for the estimation of cheating.

In addition to these, we compute some variables for the representation of the skills levels at grade 8. These variables are extracted from the original INVALSI dataset by computing a "correctness rate" for each skill by considering all the items attributable to that skill. The simplest of the possible encoding considers a variable for each area, process and macro-process. For example, item D1 of the reference test (maths test June 2014 – 8th grade) is classified as area "numbers", process "knowing and mastering algorithms and procedures" and macro-process "interpreting". The concatenation of these extracted features represents the encoding for students' learning. Finally, a Boolean label for the implicit dropout at grade 13 is computed in accordance with the criterion set out above, i.e. the conversion of the score in the maths grade 13 test of the student converted from INVALSI to a level lower than or equal to 2.

The second and the third design choices depends on the stakeholders who are imagined as users of the model and on the impact to be pursued by integrating this model into a DSS. In this first phase of the research, we imagined to address the model to policy makers or managers of educational institutions; therefore the transparency of the model is one of its main goal. For this reason the baseline has been developed with a Decision Tree on which an optimisation is applied through validation set and pruning.

Furthermore, the interest pursued by the stakeholders influences the choice of the most informative metrics. In binary classification False Negatives (FN) and False Positives (FP) are two kind of errors that can occur and, from the policy makers and managers points of view, FN are the most dangerous ones because a non-dropout is predicted while the student actually drops and we lose the opportunity to treat and prevent it. A good performance measure should take this difference among errors into account so we decided to monitor Recall as main metric. In fact, Recall is defined as the ratio between the True Positives (TP) and FN $\frac{TP}{TP+FN}$ and a high value indicates both a reduction in FN (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 possible intervention areas.

On the other hand, if we considered the interest of a family or a student the reliability of the model could be better represented by the Precision $\frac{TP}{TP+FP}$ as a high value is obtained for a reduction of FP and this confers a reliability to positive predictions.

As for the train/val split, it is performed splitting the dataset in a 3 ÷ 1ratio by preserving the original classes distribution. In fact, the dataset is composed by 34% of no-dropout students and, as we used supervised learning tools, there is the need to balance with respect to the two classes (dropout/no-dropout or True/False) the portion of the dataset used to train the algorithms. To do so we randomly sampled the majority class (False) in order to contain the same number of examples belonging to the minority class (True) in the trainset while validation set maintain the original distribution among the two classes. The under-sampling of the trainset is repeated several (10) times in order to obtain different *trials* and obtain the final results averaging on these different situations. This should decrease the dependency from a single random choice of the majority class sample. The preliminary results on the validation set, the baseline model has a recall of 0.77.

As regard the use and update steps of the ML Model Cycle and the third phase of the Informed Model System (easy-to-interpret integration in a DSS), they have not yet been tested on this case study but represent one of the main developments of the research we are conducting on this topic. To test and use the model we aim to use data on the tests administered in the school year 2020/21 as soon as the INVALSI made them available; alternatively, we could redefine the problem of implicit dropout using as input the data of grade 5 and as target the skills levels at grade 10, term of the compulsory education. With reference to the update of the model, it can be done by including the analysis of the progression on the skills levels of the students at different grades (in our case by entering the data on grade 10). In practice, the representation of the student with the data at grade 8 and the output of the model (dropout or no-dropout label) could become inputs for a second prediction, enriched by data at grade 10, with which the risk of implicit dropout at grade 13 is recomputed.

5 Conclusion

With this contribution, we wanted to emphasise the need to overcome the exclusive use of data on the social, economic and cultural or demographic context for the prediction of the risk of dropout through the inclusion of features for the estimation of skills. This seems necessary to us to make concrete interventions possible at the didactic level. Furthermore, we have described a possible reference framework for the integration of predictive models for dropout based on machine learning techniques in an “informed” system that considers the peculiarities of applied research in education and favours the impact within the DSS.

We have supported these two theses also through an undergoing case study on school dropout. The preliminary results seem to support the two contributions we have proposed with this position paper, even if the actual integration phase of the model in a DSS and its impact are only at a design state that has yet to be realised. In addition to this, we propose to take up a case of predictive model for the academic dropout already studied [6], trying to understand if and how we can estimate the skills levels to include them in an Early Predictive Analytic Model and place it in the Informed Model System we have described.

Furthermore, we believe that the Informed Model System can be experimented with other possible topics of interest for Educational Data Mining.

References

[1] European Commission. (2010). EUROPE 2020 - *A European strategy for smart, sustainable and inclusive growth*. https://tinyurl.com/59s8nwfp

[2] Flisi S., Goglio V., Meroni E., Vera Toscano M. (2015). *School-to-work transition of young individuals: what can the ELET and NEET indicators tell us*. Luxembourg (Luxembourg): Publications Office of the European Union; JRC95223.http://dx.doi.org/10.2788/161168

[3] Vincent-Lancrin S. and van der Vlies R. (2020). *Trustworthy artificial intelligence (AI) in education: Promises and challenges*, OECD Education Working Papers, No. 218, OECD Publishing, Paris, https://doi.org/10.1787/a6c90fa9-en

[4] Ricci R. (2019). *La Dispersione Scolastica Implicita*. INVALSIopen. https://www.invalsiopen.it/wp-content/uploads/2019/10/Editoriale1_ladispersionescolasticaimplicita.pdf

[5] 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

[6] 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. https://doi.org/10.13128/form-9767

[7] Pietsch W. (2014). *Aspects of Theory-Ladenness in Data-Intensive Science*. Philosophy of Science, 82, https://doi.org/10.1086/683328

[8] Organisation for Economic Cooperation and Development (OECD). (2018). *The future of education and skills: Education 2030*. OECD Education Working Papers.

[9] H jgaard T. (2009). *Competencies, skills and assessment*. In R. Hunter, B. Bicknell, T. Burgess (Eds.), Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia (Vol. 1). Palmerston, North, NZ: MERGA.

[10] Castoldi M. (2016). *Valutare e certificare le competenze*. Carrocci Editore. D. (2010, April). *graphicx: Enhanced support for graphics.* Retrieved from http://www.ctan.org/tex-archive/ help/Catalogue/entries/graphicx.html