ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

DATA SCIENCE AND COMPUTATION

Ciclo 36

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

UNDERSTANDING PREDICTIVE MODELS IN EDUCATION: ENHANCING TRANSFERABILITY, EXPLAINABILITY, AND GENERALIZABILITY

Presentata da: Andrea Zanellati

Coordinatore Dottorato

Daniele Bonacorsi

Supervisore

Maurizio Gabbrielli

Co-supervisore

Olivia Levrini

Esame finale anno 2025

Abstract

This thesis explores the application of Data Science and Artificial Intelligence (AI) within the field of educational sciences, focusing specifically on leveraging machine learning (ML) techniques to enhance educational outcomes. The primary case study revolves around predicting the risk of low achievement among students using data from the national INVALSI test, which serves as a practical application to address common challenges in AI integration in education: transferability, explainability, and generalizability.

In addition to low achievement, the thesis considers other educational issues, such as academic dropout and knowledge tracing, reflecting a broader perspective on student performance. A key aspect of this work is the incorporation of Informed Machine Learning (IML) principles, which facilitate the infusion of domain expertise and other knowledge sources into predictive modeling. This methodological and epistemological reflection underscores the importance of understanding how predictive models can be effectively employed in educational settings to inform policy and practice.

Throughout the thesis, various strategies are proposed to tackle the identified challenges. For transferability, the potential for adapting models to different educational contexts is examined. The explainability of ML models is emphasized as essential for fostering trust among stakeholders and supporting informed decision-making processes. Additionally, generalizability is addressed through innovative approaches to student representation across diverse cohorts.

By interconnecting these themes, this research aims to contribute to the understanding of predictive analytics in education and provides a framework intended to support the thoughtful implementation of AI solutions in educational settings, with the aspiration that it may lead to improved outcomes for students.

Contents

Chapter 1

Introduction

This thesis marks the culmination of four years of dedicated doctoral research, during which I delved into the applications of Data Science (DS) within the realm of educational sciences. During the last decades, we have witnessed a process of digitization of society and the widespread adoption of data-driven practices in research across most sectors. This renewed focus has led researchers and educators to explore innovative solutions, including the integration of Artificial Intelligence (AI) techniques [1] into educational decision-making. AI, particularly through machine learning (ML), has shown exceptional capability in predictive and diagnostic tasks across numerous fields. The application of AI in education yields substantial benefits for the system, supporting the interests of various stakeholders. In presenting the results of this thesis, we will highlight these benefits in the context of the case studies considered.

This thesis fits into this research landscape, waving on various contributions —some already published, others in the process of publication in international journals or conferences. Most of the chapters are closely tied to a key publication in my research path. The presented work is entirely my own, with specific mentions in the text and Acknowledgements section regarding collaborative efforts with other researchers. Collaborative contributions are transparently indicated, and comprehensive references are provided for all supporting literature and resources. At times, I integrated sections of these papers into the thesis in their original form, particularly when detailing the methodology applied to a specific case study or presenting results and their discussion. In other instances, I found it necessary to rephrase, add content, or provide a different perspective on these prior contributions. This adjustment is often made in the introduction and conclusion of the papers, reflecting the evolution of these contributions during my Ph.D. journey. It outlines how they influenced subsequent phases of my research among various possibilities and contributed to the primary goals of this thesis.

Furthermore, I have also incorporated unpublished sections, serving as extensions of previous works identified in the papers as future endeavors.

The starting point of this research was strongly motivated by my professional interests. For close to a decade, I dedicated myself to teaching mathematics in Italian high schools —a subject often marked by students facing challenges in meeting expected learning outcomes. This struggle is more pronounced among Italian students than their peers in other OECD countries, as highlighted by the students achievement in international Large Scale Assessment Tests [2]. Thus, my research embarked on an exploratory journey, aiming to assess the potential informativeness of data collected through such assessments in predicting the risk of *low achievement* at an early stage.

We used data collected through the INVALSI test, a national Large Scale Assessment Test administered annually by the Italian Ministry of Education [3]. While this exploration did yield satisfactory results in terms of predictive performance metrics, it also brought to light certain limitations, which will be discussed in greater detail in Chapter 2. The chapter is dedicated to presenting the main case study of this thesis on machine learning (ML) methods for predicting at grade 5 the risk of underachievement in mathematics five years later, at the end of the Italian compulsory education.

Here I sum up which are the three emerging issues in our main case study: the transferability of the proposed predictive models, their explainability, and their generalizability. These terms are widely used in the literature, and there are varied interpretations regarding their usage [4, 5]. To clarify them in the context of this thesis, in the following I briefly introduce specific meanings for each term. Moreover, I utilize this introduction to define three goals that I aim to address in this thesis related to these issues.

1.1 Challenges and goals

When using machine learning and data-intensive methods for developing predictive tools, several common challenges can undermine users' trust in the model and its large-scale implementation. As mentioned earlier, in the development of this thesis research, I encountered three main issues that I will now introduce, characterizing them with a sense resonating in the context of AI applications in education.

A first challenge concerns *Transferability*. It is related to the applicability and effectiveness of AI tools or models when they are applied in a different educational setting or context. It estimates how well the modeling process gained from one environment can be transferred and utilized in another educational context. We

have interpreted this as having an educational data science pipeline that is applicable to other relevant problems. The transferability issue led to the formulation of the first goal for this thesis as follow:

G1 Evaluate the possibility of transferring the machine learning pipeline used in the initial case study to other educational contexts and identify necessary precautions.

In this thesis we addressed the transferability challenge, moving from students' low achievement to academic dropout [6]. While this problem differs from the main case study, involving a distinct educational segment (higher education) and a different educational outcome (dropout instead of underachievement), it serves as a suitable transferability context due to similarities in the processed data. Both the educational outcomes can be detected analyzing tabular data; additionally, they both show intrinsic differences in the collected data among cohorts of students from different years, as I will elaborate on later.

As second issue, let us consider *Explainability*. It refers to the ability of a model or system to provide understandable and transparent explanations for its predictions, decisions, or outcomes. It involves making the inner workings of the AI model accessible and comprehensible to users, especially stakeholders in the educational domain, such as teachers, students, or administrators. It is crucial in gaining trust, facilitating user comprehension, and supporting informed decision-making based on the insights generated by the AI tools.

This issue is crucial in the development of predictive models for both low achievement and dropout. In fact, the significance of these models often extends beyond simply predicting a success or failure label for an individual student. Rather, the value lies in comprehending the underlying reasons that influenced the model's prediction of a specific outcome. This understanding may support various stakeholders in implementing actions of different scales to counteract undesirable outcomes. I formulate here a second goal in this thesis:

G2 Explore explainability strategies for predictive models in the educational domain that provide value for the various stakeholders involved in designing policy and educational actions to counteract undesirable outcomes

In this thesis, I addressed this challenge by conducting an analysis using various techniques to determine feature importance, i.e. selecting the features that most significantly impact the model's predictions. Our main contribution for this objective concerns the application of three post-hoc explainability techniques, namely permutation fetures importance, attention map-beased explanations, and SHAP both to our main case study on student low achievement prediction and academic dropout.

The third emerging challenge concerns *Generalizability*. It pertains to the degree to which findings, patterns, or models derived from a specific dataset —essentially a subsample of the entire population— can be extrapolated to other similar datasets or the entire population. This assesses the ability of insights to remain valid beyond the specific conditions of the original study, including the capacity to apply models to diverse student cohorts or other learning domains. This challenge is particularly evident in the INVALSI dataset selected as main case study of this thesis. The INVALSI tests, along with Large-Scale Assessment Tests in general, assess the same skills and/or knowledge of students but through different questions each year of administration. Consequently, the data collected on a cohort's responses is specific to that particular group of students, making direct comparisons with any other previous or subsequent cohort challenging. This prompted the need for a feature engineering process or other embedding solutions that facilitate the comparison of students across different cohorts, allowing the models to be applied to cohorts not encountered during the training phase. In light of this, we can establish the third goal of this thesis:

G3 Enable the application of developed predictive models on diverse student cohorts by evaluating suitable forms of student representations derived from the available tabular data.

We tackled this objective with two main attempts. Firstly, we drew up a feature engineering process aimed at computing new and more general features, which can be calculated for all student cohorts, exploiting their set of responses to items in the administered tests. The process was guided by the reference to a specific domain knowledge, i.e. a taxonomy for the classification of the items through preestablished categories by experts in mathematics education. Furthermore, this feature engineering process, guided by theory, has outlined the possibility of investigating the presence and effectiveness of additional domain knowledge sources, as well as exploring alternative methods for their integration into the standard ML pipeline. The injection of domain knowledge or, more in general, of other prior knowledge sources, has been explored in the literature under various names, including Informed Machine Learning (IML) [7]. This methodological approach seems promising for overcoming the issue of generalizability in student representation, i.e. creating an abstract and robust student representation applicable to diverse student cohorts. Thus, as second attempt, we dive into this direction through a systematic literature review [8] and the proposal of a graph-based data engineering process [9] to apply to the main case study considered in this thesis.

As a final remark for this introduction, I want to point out the strong intertwining between transferability, explainability and generalizability, in the sense that addressing one meant also impacting the other two. Simultaneously, to pursue any of the three paths, it was needed to consider the others. In the development

of this thesis I try to assume a reductionist approach, considering the three challenges separately along the various experiments, analyses and discussions. The connections between these three aspects are reconstructed at a later stage, as final outcome of this research journey.

1.2 Thesis structure

The thesis structure unfolds as follows. In Chapter 2, the primary focus is on the early prediction of students at risk of low achievement. This involves the application of two well-established ML techniques: Random Forest and Neural Networks. Addressing the challenge of generalizability, the chapter introduces a feature engineering process, which is detailed along with information about the available data and the framework guiding their collection. The chapter discusses the model results using various performance metrics. Furthermore, a feature importance analysis was performed for the Random Forest model, establishing a foundation for the subsequent work on explainability. The conclusion of this chapter provides insights into the motivation behind the three challenges and the research directions previously introduced.

Chapter 3 significantly contributes to the issue of transferability. It presents and discusses the results of the same architectures used in the main case study, this time applied to academic dropout for one of the major Italian universities. The conclusions drawn in this chapter are leveraged to highlight the precautions necessary for the transferability of the ML pipeline described and implemented in the previous chapter.

Moving on to Chapter 4, the focus shifts to the explainability problem, examining its application to both case studies presented in the preceding chapters. This approach not only contributes to Goal G2 but also monitors the transferability of the explainability techniques considered, which nevertheless fall within the proposed Educational Data Science pipeline for tabular data in this thesis.

Chapter 5 provides an overview of the Systematic Literature Review conducted on Informed Machine Learning (IML) approaches for student modeling [8]. The primary focus is on understanding the consideration and integration of prior knowledge sources into the standard machine learning pipeline when dealing with student modeling in artificial intelligent systems. The challenge of generalizability in student modeling here is explored in connection with the knowledge tracing (KT) problem [10]. KT has different characteristics from those of predicting underachievement through Large-Scale Assessment Tests. However, there are relevant points of similarity that are discussed in this chapter. In the conclusion of this chapter, I point out the state-of-the-art, shedding light on strengths and

potential research directions in the application of IML to student modeling and, consequently, to our main case study.

In Chapter 6, I introduce a potential alternative to feature engineering for achieving generalizability. The main concept is to harness the similarity connections among test items, either based on their content or the types of skills they engage in students. This approach aims to derive a representation of the student's cognitive state, which can be captured through their responses in the test. The chapter begins by presenting the pedagogical and didactic assumptions that form the foundation of these student models. Subsequently, I delve into the implementation of these assumptions and explore how they influenced the performance metrics and explainability of the models proposed in Chapter 2. In the conclusion of Chapter 6 I propose a two-step framework for AI systems like those considered in this thesis, aimed at improving both the explainability and performance of such systems. This framework is the result of interdisciplinary collaboration with many fellow travelers during this journey. Our proposal has a dual perspective: first, we decompose the scope of our predictive system into an epistemic aspect (hypothesis testing for knowledge gathering), and then we consider a pragmatic one (knowledge deployment in real-case scenarios). The framework is employed to discuss the data science pipeline used in the previous chapter and the results obtained.

The Conclusion of this thesis is used to summarize and highlight the contributions presented in the preceding chapters concerning the three goals stated in this introduction. Within the same chapter, in addition to discussing individual strengths and weaknesses regarding the three themes of transferability, explainability, and generalizability, attention is given to how they intertwine. The focus is on how advancements in one aspect may have positive repercussions on the others. In addition, I provide some future works that may serve as possible extensions for this thesis.

Chapter 2

Students' Low Achievement

In this chapter, I present the foundation of my research throughout my doctoral studies. As mentioned in the Introduction, my initial goal was to explore how Data Science and Machine Learning techniques could effectively analyze and address the issue of students' low achievement. This chapter's content is primarily derived from my paper presented at the AIED 2022 conference, co-authored with Dr. S.P. Zingaro under the guidance of Prof. M. Gabbrielli [3]. There are three significant enhancements from the original conference paper.

Firstly, I've expanded the introduction, now presented as the background and motivation for my research activity (section 2.1). I aim to provide a more thorough rationale for investigating low student achievement and applying predictive models to this issue.

Secondly, I have included a more comprehensive description of the dataset compared to the conference paper. The dataset used in this case study is extensive and includes several features, based on a didactic framework that underpins the test design. I thoroughly introduce the dataset in section 2.3, to support the numerous references throughout the thesis. Moreover, the description of the data collection choices that initiate the data-driven process enhances transparency and fairness in interpreting the results [11, 12].

Thirdly, I have reorganized the presentation of the results. More space is dedicated to the feature importance analysis (section 2.5.2), which had been compressed due to page limits in the conference paper. Additionally, section 2.6 is structured to highlight the gaps in the responses to the RQs that motivated the subsequent steps in my research activities selected for this thesis.

2.1 Background and motivation

Student low achievement refers to the scenario where a student fails to meet established learning objectives. This widespread issue has long-term repercussions for both individuals and society. In 2016, more than 28% of students in OECD (Organization for Economic Co-operation and Development) countries scored below the baseline proficiency level in at least one of the three core subjects—reading, mathematics, and science—assessed by PISA (Programme for International Student Assessment) [13].

Low achievement is closely linked to school dropout, which is the premature abandonment of the educational path by students. This connection is twofold. Firstly, there is a causal link between underachievement and school dropout. Poorly performing students often become trapped in a vicious cycle of demotivation and low grades, leading to further disengagement from school. This increases their risk of dropping out [14] undermining both their prospects of cultural and professional growth, and their future as engaged citizens [15]. Studies indicate that the impact of low achievement on dropout rates can start as early as first grade, where poor performance already serves as a significant predictor of future dropout risk [16, 17].

Secondly, low achievement is not only a primary factor contributing to dropout but also a form of dropout itself. INVALSI (the National Institute for the Evaluation of the Italian Education System) supports this thesis, presenting data in their 2019 report [18]. They observed that 7.1% of students, by the end of their school careers, exhibited underachievement in all subjects covered by national assessments. These students remain in school but fail to meet the expected learning standards, resulting in fragmented and uncertain knowledge despite sometimes receiving passing grades. Thus, low achievement can be considered an "implicit" form of school dropout. While these students may not leave school explicitly, they do not acquire the necessary skills according to national standards. When considering both explicit and implicit dropouts, 2019 Italian data indicates that about 20% of students are affected by this phenomenon.

The strong connection of low achievement with dropout is one of the main reasons why predicting it in advance is important. Implementing preventive measures for underachievement could, in turn, help mitigate the subsequent problem of school dropout, which is much more challenging to address. Indeed, it is evident that once students leave the educational system, it becomes more difficult to intervene and ensure their right to education and to support them in building the skills necessary for active citizenship. Therefore, having predictive tools for the risk of low achievement can be useful as an alert system.

This consideration led us to formulate our first research question:

> **RQ1.1** Is it possible to develop a suitable tool to predict, at an early stage, the risk of low achievement at secondary school for students of the primary school?

By using "early", we mean as soon as possible, that is several years in advance so that suitable countermeasure can be taken. This includes designing interventions aimed at reducing the risk of underachievement and, consequently, the risk of dropout [19].

The previously mentioned statistics on the incidence of low achievement among students in OECD countries, and particularly in Italy, highlight the social impact of this phenomenon. This raises the question of what causes underlie this problem and how to effectively address and mitigate them. Answering this question is complex and multifaceted. Within the field of mathematical education, for instance, Abd Algani and Eshal propose five categories of factors that can contribute to underachievement: student-related factors, teacher-related factors, curriculum-related factors, school-related factors, and family-related factors [20].

In this study, we focus on the Italian context using data from the INVALSI national large-scale assessment test. This test is administered annually in three core subjects common to all school curricula in Italy: Italian, mathematics, and English. In this case study, we focus on assessing the risk of low achievement specifically in mathematics, which is the subject most prominently affected by this issue. According to the 2023 INVALSI report [21], by the end of compulsory schooling (K-10), 44.1% of students do not achieve the expected level of competence in mathematics, and 38.5% do not reach the expected level in Italian. The English test is administered only at the end of grade K-13.

The INVALSI dataset includes information related to most of the previously mentioned categories, specifically demographic data on students, their responses to the administered test questions, information about their family and socio-economic context, and data about their schools. The dataset is described in greater detail in the dedicated section 2.4. Here, we highlight two key strengths: firstly, its extensive scope, with data on approximately 1000000 students over two years of test administrations; and secondly, the detailed information on student responses, showing how they performed on each question of the administered test. This allows us to investigate whether it is possible to obtain a representation of students' skills achievement at the time of the test administration, and to determine if there are any key educational factors in predicting the risk of future low achievement. Thus, we can formulate these two additional research questions:

> **RQ1.2** Is it possible to quantitatively represent the level of knowledge of students and build a model of their skills achievement?

> **RQ1.3** Is it possible to derive which are the factors that are most related to students' low achievement?

In diving into these research questions, we must consider the issue of generalizability. Specifically, we need to find a representation that is valid for students from different cohorts who have taken different tests, thereby overcoming the specificity of the administered test items. This approach also enables us to identify factors relevant in determining the model predictions that are consistent across different student cohorts. Such factors would have general value and potentially greater significance. Indeed, delving into the three research question we aim to align with the interests of various stakeholders, including policy-makers, principals and coordinators, teachers, students and their families [22].

On one hand, the early predictor can serve as a warning system, notifying teachers or families of the necessity to implement recovery and precautionary measures. On the other hand, the underlying model derived from statistical learning provides valuable knowledge that applies not only at the individual student level but also at the systemic level. Identifying the key features influencing the model's predictions is particularly valuable, as it can highlight potential areas for intervention in social and educational policies aimed at curriculum reform and improving school quality indices.

Regarding methodological decisions, we opted for a data-driven approach, confident that emerging digital technologies, especially artificial intelligence, would offer valuable operational support. To strike a balance between explainability and performance, aligned with the principles of Trustworthy AI [23, 24], we evaluated two state-of-the-art machine learning techniques: random forests and neural networks. We intend to leverage random forests [25] to extract rules that facilitate interpreting model outcomes and to develop protocols for mitigating student underachievement. Additionally, we explore two different neural networks for their flexibility and potential performance enhancements.

The remaining of the chapter is organized as follows. Section 2.2 provides an overview of the literature concerning the development of automatic tools—based on machine learning techniques—that aim to exploit educational data to prevent low achievement. Sections 2.3 and 2.4 describes the dataset and the techniques used to build our predictive models, while Section 2.5 presents the results of the experiments we carried out to validate our approach. Finally, in Section 2.6, we discuss the main findings and conclude with some possible future directions for this work, some of which are further developed in the subsequent chapters.

2.2 Related Works

The topic of students' low achievement is a widely studied phenomenon in the social sciences and education [26, 27]. The problem was also addressed in terms of predictive models for student performance or dropout risk both at school,

high school and academic levels. These models exploit different machine learning techniques, including supervised learning, e.g., random forests, support vector machine and Bayesian network, unsupervised learning, e.g., k-means and hierarchical clustering, and recommender systems, e.g., collaborative filtering [28, 29]. Moreover, several kinds of data have been used to tackle the problem. In [30] the training set is built with demographic data of the students and their grades in some tasks. Other studies are based on students performance during first semester courses [31, 32]. Some datasets include behavioural data supplemented with other features related to learning results [33, 34, 35], in a mix of cognitive and non-cognitive characteristics.

It is worth noting that in all the researches considered so far, the performance prediction is made within a relatively short period, i.e., one school or academic year or cycle of studies. Furthermore, even when the datasets include cognitive features, these are expressed in terms of previous or present education marks, total university score and final or admittance exams. We aim to enrich the cognitive features expressiveness by including variables for the representation of areas of knowledge and skills, privileged indicators for the study of learning [36]. To overcome the limitations mentioned above, we decided to consider data collected through the administration of large-assessment tests. This kind of data lend by their nature to greater generalizability than traditional educational or psychological studies, that often rely on convenience samples [37]. These datasets are often used to support educational policy decisions [38] or in studies aiming to determine the relationship between socio-economic factors and school performances [39]. Nevertheless, they are designed to measure students' knowledge and skills and often to track longitudinally the students' learning path [40, 41], as in our interest.

In some studies data collected through large-scale assessment tests were used to design predictive models of student performance through several machine learning techniques. In [42, 43], for example, the authors refer to data collected through the PISA international large-scale assessment tests. In both cases, attributes directly concerned with the students' attitudes, intentions, behaviour as well as data about out of school lessons, concepts familiarity and overall experience, have been used as classification features. Moreover, the data refers to a single cohorts of students, both for training and testing the models.

To the best of our knowledge, our proposal has three main novelty points. It is the first in the literature that aims to develop a machine learning-based predictive model targeting low achievement at secondary school when students are still in their primary cycle of studies. Secondly, we aim to use the data collected through a national large-scale assessment test to extract features directly related to student learning in terms of knowledge and skills. In this way the explainability of the model may enable useful information for teachers and didactic coordinators.

Finally, it uses different students cohorts to train and test the models, striving for a transferability of our tools.

2.3 The INVALSI dataset

In this section we introduce the INVALSI dataset, used as data source for our tools. It collects data from disciplinary tests and student surveys administered across various school grades —at the levels K-2, K-5, K-8, K-10, and K-13— in Italy in a census manner. These tests aim to evaluate students' skills levels in specific subjects, according to specific frameworks for each discipline. On the other hand, the surveys collect socio-economic and cultural information that significantly impacts educational outcomes, as supported by numerous international studies, including those by the OECD [44]. Since the initial tests in the 2002/03 school year, the scope and format of these assessments have evolved significantly.

Key milestones in the development of the INVALSI tests include:

- 2013/14: Introduction of separate tests for Italian and mathematics at grades 2, 5, 8, and 10.
- 2017/18: Transition to computer-based testing and the inclusion of an English test at grades 5 and 8
- 2018/19: Expansion to include Italian, mathematics, and English tests at grade 13, marking the end of secondary education.

While the primary objective of these tests is to measure students' skills development at various educational stages, their broader goal is to assess and improve the quality of the national education system. This goal is emphasized by recent legislative changes, such as Legislative Decree no. 62/2017, which removed the requirement for these tests to be part of the final exams for grades 8 and 13.

A significant advantage of the INVALSI dataset is its ability to track students longitudinally across different school years [40]. Since the 2011/12 school year, each student has been assigned a unique identifier, the SIDI[1] INVALSI code, allowing their test results to be traced over time. This longitudinal tracking enabled, for instance, the 2018/19 administration of grade 13 tests to evaluate the school effect on skill improvement for students who completed the first cycle of secondary education in 2013/14. The school effect refers to the school's contribution to changes

[1]SIDI is the acronym of Sistema Informatico Dell'Istruzione (Information Technology System of Education)

in students' skill levels. This comprehensive dataset thus provides a robust foundation for evaluating educational outcomes and informing policy decisions aimed at enhancing the overall quality of education in Italy.

Data from the Math, Italian, and English tests provide extensive information on individual students. The features in the dataset can be categorized into five main groups:

1. **Identifying Marks**: this includes school code, class code, and SIDI INVALSI code.

2. **Items Boolean Variables**: these indicate the correctness of the answers given to each test item by the student.

3. **Student Demographic Information**: this includes gender, month, year and place of birth, country of origin, province and region of residence, type of school attended, and grades in Italian and Math.

4. **Parent Demographic Information**: This includes educational qualification, job, and birthplace.

5. **Synthesis Indices**: These express the degree or level of certain aspects of interest, such as WLE (Weighted Likelihood Estimation of ability according to the Rasch Model) and ESCS (Economic, Social and Cultural Status).

The socio-economic-cultural survey data aim to gain information on the parents studies and work or to explore the context and methods of learning, such as the study environment at home, personal or external motivations for studying, and meta-reflection on study methods and school lessons. These questions serve various purposes and have evolved over time.

In our case study, we considered data on maths test from two cohorts of students: K-5 of the 2012/13 school year (485225 students) and K-5 of the 2013/14 school year (477236 students). For each cohort we also had data from five years later at grade K-10, to be used for the definition of the underachievement target. We define the occurrence of low education achievement when the students level in the test is less than or equal to 2 on a scale from 1 to 5, according to the INVALSI interpretation of the large-scale assessment tests outcomes. We applied a feature selection process to determine a subset of relevant features. In Section 2.4.4 we will provide further information on the preprocessing steps which lead to the selection of a subset of features.

As already mentioned, the datasets also contain a boolean feature for each test item, where the students' answers correctness are recorded. In order to enable the use of our predictive models on different cohorts of students and to provide a coherent representation of their learning in terms of areas of knowledge and skills,

it is necessary to release the dataset from the individual items that constitute a certain test. Therefore, we introduced a set of new features for students' learning encoding, which trace back to the same learning representation space students belonging to different cohorts.

2.4 Methods

In the following, we motivate our methodological choices. First, we introduce our knowledge-based approach to ensure generalizability across different student cohorts for our models. Then, we briefly describe the selected ML algorithm, and the chosen performance metrics. Finally, we present the experimental setup, with particular emphasis on the necessary dataset preprocessing steps.

2.4.1 Students' Learning Encoding.

Aiming to encode students' learning, we used a knowledge-based approach considering the items classification in terms of areas, processes and macro-processes according to the INVALSI framework for the design of math tests[2]. In Figure 2.1, we present a translated (Italian to English) item of the maths INVALSI test for the year 2012/13, with its classification. In Table 2.1, we give for reference an

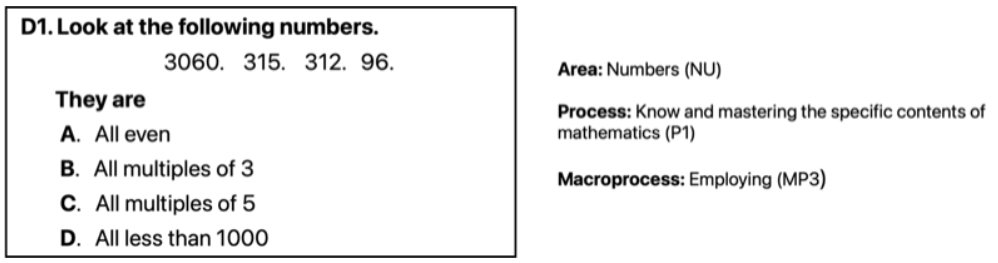

Figure 2.1: Example of item in INVALSI maths test with its classification on the right

overview of the areas, processes, and macro-processes that have been used in the encoding of the questions.

We propose a novel learning encoding by defining one new variable for each area, process, and macro-process. Each of these new features takes the value corresponding to the percentage of correct answers provided by the student for that

[2]The framework is available on the INVALSI website at https://invalsi-areaprove.cineca.it while the items' classification for the reference test is available at https://www.gestinv.it/Matematica.aspx.

Table 2.1: Maths INVALSI framework for question encoding.

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

Table 2.2: Example of the student's learning final encoding.

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

specific group of items, namely, correctness rate. For example, the new feature "Numbers" would assume the percentage of correct answers given to the items that belong to the area "Numbers"; 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 fifteen (15) dimensions, one for each possible area, process, and macro-process. This multi-dimensional representation offers a more structured approach to analyzing students' learning, as it moves beyond raw scores to capture performance across specific cognitive and mathematical dimensions. The rationale for choosing fifteen dimensions stems from empirical research in educational assessment [45], where categorizing student performance into distinct conceptual areas has been shown to improve both predictive modeling and pedagogical insights. By structuring learning performance along these dimensions, we align our representation with INVALSI theoretical frameworks that emphasize domain-specific competencies and cognitive processing.

In table 2.2, we show an example of applying our learning encoding strategy for two students (identified by Id=1 and Id=2). Indeed, in the general case, once items are classifiable using the same scheme we selected (area, process, and macro-process), one can use the same encoding and obtain a representation aligned for students belonging to different cohorts.

2.4.2 Machine Learning Techniques

We decided to exploit two state-of-the-art machine learning techniques to develop an AI predictor for the risk of low achievement. As the first technique, we used Random Forest (RF) It is a kind of ensemble learning classification algorithms, which integrate the classification effect of multiple decision trees [25]. We trained our models through bootstrap aggregating (bagging), i.e., a random sub-sample with fixed size and a limited number of features are used to fit each tree. This reduces the overfitting of datasets and increases precision. To tune the model, we performed a grid search [46] for finding the best hyper-parameters set-

ting (number of estimators, max depth, percentage of features and percentage of samples to use for each estimator). Moreover, we exploited k-fold cross-validation to assess the quality of the model [47].

RF algorithm is widely used in Educational Data Mining and Learning Analytics for the high degree of explainability and effortless interpretation of the results. For this purpose, we compare the features selected by two well-known technique, namely, feature importance based on Mean Decrease of Impurity (MDI) [48] and permutation feature importance (PFI) [49]. MDI is defined as the total decrease in node impurity—weighted by the probability of reaching that node, which is approximated by the proportion of samples reaching that node—averaged over all trees of the ensemble. This is highly informative about how the ensemble model provides its predictions. PFI directly measures feature importance by observing how random re-shuffling of each predictor, thus preserving its distribution, influences model performance. This allows to fairy treat our dataset, whose features are both categorical and numerical and have very different cardinality. Moreover, PFI can be computed on a left-out test set, removing bias due to the training set. By a comparison of the selected features, the involved stakeholders, such as teachers, school principals and coordinators, can easily identify the most relevant risk factors for the prediction of low achievement, as we will discuss in Section 2.6.

Then we relied on neural networks. The use of neural networks has recently become widespread also in the field of Educational Data Mining and has also been applied in predictive models for student performance [50]. We implemented two neural networks based on different data transformation approaches.

Firstly, we used Categorical Embeddings (CE). It is a neural network that treats the input depending on its type: if the input is categorical, we pass it through an embedding layer; if the input is numerical, we feed it to a dense layer. Secondly, we relied on a variant of the Feature Tokenizer Transformer (FTT). It is based on attention mechanism [51], able to identify the input or the group of inputs that most influence the output, thanks to attention maps. Moreover, in its architecture it exploits a feature tokenizer function to extract tokens from the input and then fed these tokens to a Transformer architecture [52] for classification.

Neural networks are known to be less interpretable than statistical learning models. However, recent advancements in explainable AI now enable post-hoc analysis of feature importance for these models, though at a higher computational cost. This approach is the focus of the study presented in Chapter 4.

2.4.3 Performance metrics

To evaluate our prediction models, we employ several common performance metrics for binary classifiers: Accuracy, Sensitivity (Recall), and Specificity. These

metrics provide a comprehensive assessment, especially crucial in our imbalanced dataset.

Accuracy quantifies the proportion of correct predictions across both classes relative to the entire dataset. While useful, its reliability can be compromised in the presence of class imbalance, where the model may favor the majority class. Therefore, we use additional metrics to provide a more nuanced evaluation.

Sensitivity (or Recall) measures the model's ability to accurately identify students at high risk, specifically those prone to low achievement in this case study. This metric is crucial because, despite being a minority, the high-risk group holds significant importance for educational stakeholders. High sensitivity ensures the effective identification of students who need targeted interventions.

Specificity assesses the model's performance in correctly identifying students at low risk of dropping out. It complements sensitivity by indicating how well the model avoids false positives.

2.4.4 Experimental setup and preprocessing

We carried out all the experiments using the Google Colaboratory Notebook environment, with the Python programming language and popular machine learning libraries, such as scikit-learn and pandas.

The dataset for all the experiments was preprocessed in a sequence of steps. Firstly, we cleaned features with many missing values, highly correlation (computed by R^2 measure above 0.5) or specifically referred to a cohort of students, preventing the model to be transferred to new cohorts (e.g. identification code for a class). This features selection process, together with the elimination of the attributes related to the items in the tests in favor of the coding of students' learning results in a set of 34 features, plus the target feature. These features refers both to socio-economic and cultural context—as this significantly affects student formative career [44]— demographic data and learning dimension —the core domain for school actions. Table 2.3 lists all the 19 selected features, along with the 15 features we defined previously for students' learning encoding (Section 2.4.1) .

To deal with our hybrid dataset, we include a further preprocessing step, aimed at encoding the values of categorical variables into numerical values. We selected the "one-hot" encoding algorithm that encodes each variable of n categories into n binary variables, whose value is 1 only for the variable corresponding to the transformed category while it is 0 for the remaining ones.

Moreover, a comparison between the students in the K-5 datasets and those in the correspondent K-10 datasets shows that about 27% of the students are missing. We can assume that this is due to several causes—e.g. remedial students, dropout, transition to vocational training, anonymization errors, and many more.

Table 2.3: List of features to add to those for student's learning encoding. From top to bottom, they include: demographic information about the student, information about their academic performance, information about their previous academic career, and information about the parents.

Feature Name	Type	Range/Categories
Student's Gender	Categorical	Female or male
Student's month of birth	Categorical	One of the twelve months
Student's place of birth	Categorical	4 categories (Italy, EU country, european no-EU country, other)
Student's Citizenship	Categorical	4 categories (italian, EU, european no-EU, other)
ISTAT Province Code	Categorical	ID code for the residence province of the student (108)
Region Code	Categorical	ID code for the residence region of the student (19)
Student Geographical Area	Categorical	5 categories (north-west, north-east, central, south, island)
ESCS	Numeric	Index for Economic, Social, and Cultural Status
Italian grade	Number	Student's grade in Italian at the end of the first semester (range 1-10)
Math grade	Number	Student's grade in Math at the end of the first semester (range 1-10)
Overall math score	Number	Score obtained in the mathematics test normalized (range 0-1).
Attendance of nursery school	Boolean	True if the student attended nursery school
Regularity	Boolean	True if the student has not repeated any school years
Father's Place of Birth	Categorical	4 areas (italian, EU, european no-EU, others)
Father's Educational Qualification	Categorical	6 levels of education from primary school to bachelor's degree or higher
Father's Occupation	Categorical	Type of profession (9 available categories)
Mother's Place of Birth	Categorical	4 areas (italian, EU, european no-EU, others)
Mother's Educational Qualification	Categorical	6 levels of education from primary school to bachelor's degree or higher
Mother's Occupation	Categorical	Type of profession (9 available categories)

Considering our focus on a predictive model for low achievements, we removed from the dataset all the students who were missing over the five years for whatever reason. Even after this preprocessing operations, the K-5 2012/13 cohort is still made up of 351746 students, the K-5 2013/14 cohort of 354987.

For the definition of the training set we used the data from 2012/13 K-5 cohort. For the models based on neural networks we split this cohort to generate both training and validation sets (split in 80% and 20% respectively). The K-5 2013/14 cohort was used as test set, to evaluate the transferability of the models on different cohorts, i.e., obtained with a different INVALSI test. This allowed us to evaluate the validity of the proposed learning encoding that is, the effectiveness of abstraction from specific items to learning in terms of areas, processes and macro-processes.

The dataset is unbalanced between underachievement/non-underachievement classes; therefore balancing techniques were applied. In the development of the RF models, a random undersampling technique was used, implemented in the imblearn library. We trained neural networks using a weighted random sampler, that samples the data to balance classes ratio in the training batches.

2.5 Results

In this section we sum up the results of our experiments. We present them in two subsections. The first one is dedicated to the performance metrics. We selected three well-known metrics for classification: accuracy, precision and recall. In the second subsection we show the results of the analysis we conducted on features importance in the RF model.

2.5.1 Predictive models performance.

In Table 2.4, we present the overall results on the test dataset, i.e., the data from cohort 2013/14, on three different models: Random Forest (RF), Categorical Embedding neural network (CE) and Feature Tokenizer Transformer neural network (FFT).

The models are compared accordingly to three standard metrics. Accuracy expresses the percentage of good predictions for the models, but can be misleading when the dataset is imbalanced. Precision is a measure of quality for the predictor, and high value indicates that the model is not overfitting on the target class. Recall is highly informative for our purpose, because it is related to the quantity of students in underachievement condition who are retrieved correctly by the model.

Table 2.4: Performance on test set

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

For RF, we considered the best hyper-parameters setting determined with the grid search technique: 50 estimators in the forest, trained with 30% of random samples, 60% of random features and max depth set to 11.

The FFT outperforms the other predictive models. The accuracy values are similar for the three models, but FTT increase of the 1% with respect to the best RF. Moreover, this architecture maintains higher performance on the test set in terms of precision and recall, with increases of 15% and 11% respectively compared to the RF model. The CE model performs slightly lower, trailing the FTT model by 1-2 percentage points across all metrics.

2.5.2 Features importance

RF offer great interpretability, especially through the analysis of rule-based features importance. In Figure 2.2 we show the 8 top relevant features selected with a MDI higher than 2%.

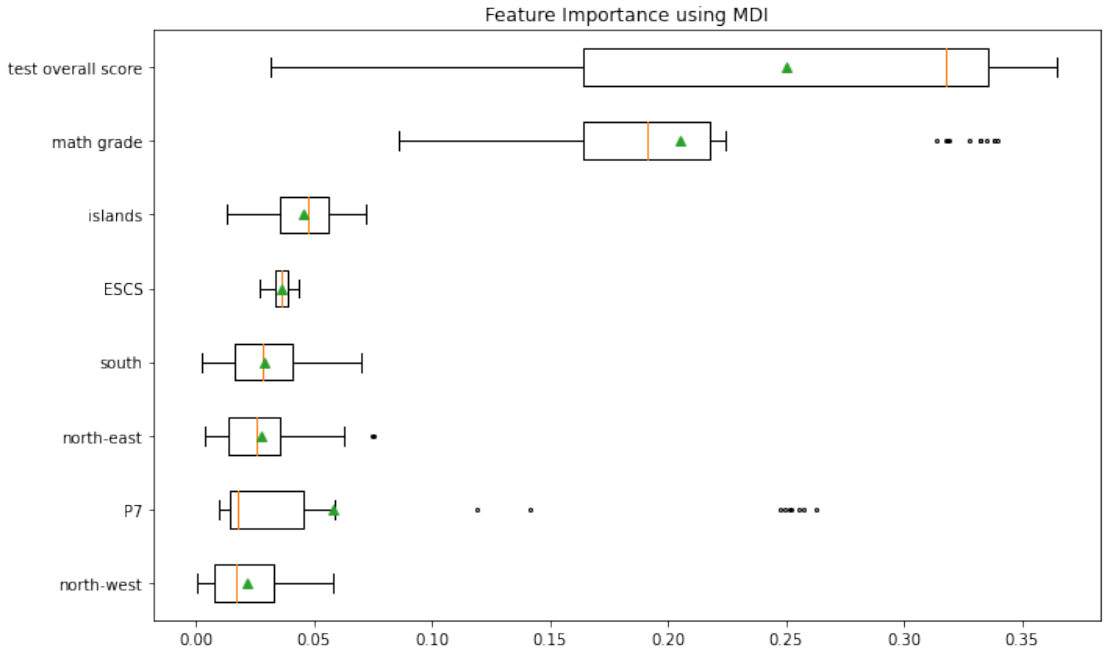

Figure 2.2: Feature Importance model with MDI. The legend for the box-plot is the following: the central rectangle in each line shows the second and the third quartiles together, where the orange line represents the 50% threshold; the green triangle is the average value for the importance of the selected feature on all the decision trees in the RF model; dots represent outliers.

The first two selected features are the overall test score and the math grade. Their average MDI is higher then 20%, which is significantly relevant with respect to the other features, whose MDI values is always under the 5%. Four out of the eight features selected with an MDI threshold greater than 2% pertain to the student's geographical area of residence. These are boolean features resulting from the preprocessing of the categorical variable geographical area, conducted on the dataset for the RF model. The only student learning encoding feature selected is the one corresponding to process P7 — use tools, models and representations in quantitative treatment information in the scientific, technological, economic and social fields. However, we note an anomaly: the average MDI value for this feature is near the end of the fourth percentile due to the presence of some out-

liers—specifically, certain decision trees in the ensemble for which the MDI of P7 exceeds 25%, making it one of the most important features.

For the features importance analysis, we also adopted PFI. We aimed to compare the top 8 features extracted using this technique with those selected by MDI. For the PFI computation, for each feature we compute the mean difference between the default estimator score (accuracy) and the scores obtained by replacing 10 times the values of the feature with their permutation in a random way. In Figure 2.3 we show the results.

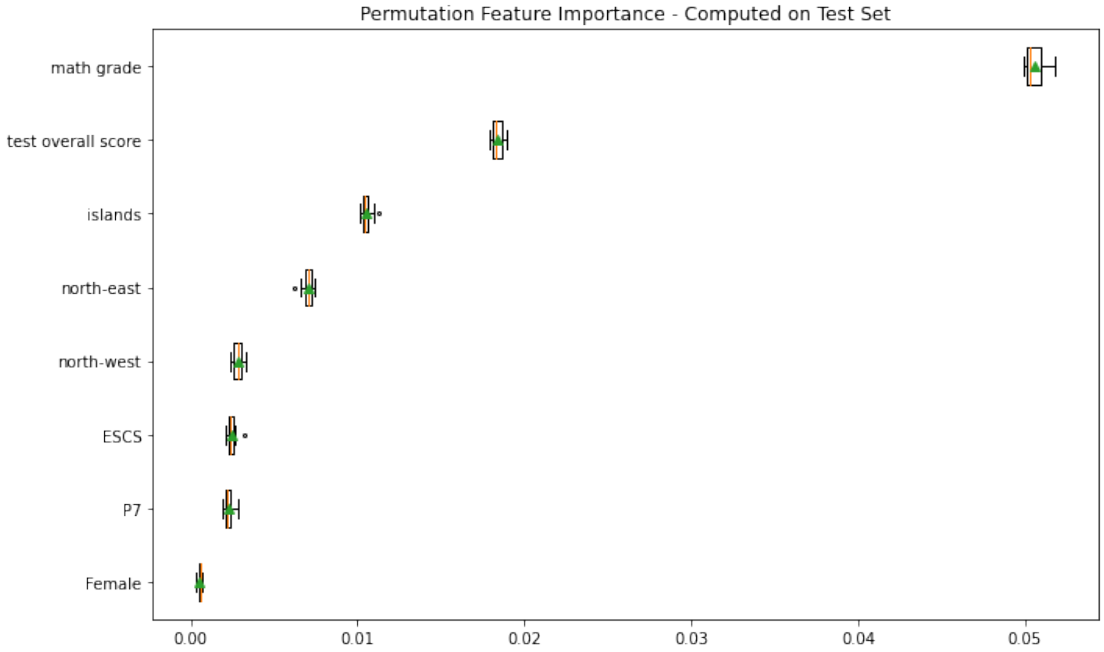

Figure 2.3: Feature Importance model with PFI. The legend for the box-plot is the same of figure 2.2.

In this case as well, the two most significant features are the math grade, with random reshuffling causing an accuracy drop of over 5 percentage points, and the overall test score, which decreases accuracy by about 2%. Many of the selected features are again related to the student's geographical area of residence; P7 is the only student learning variable that appears, and student gender is identified as the eighth most important feature by PFI. It is worth noting that, after the first two selected features, only the "Island" feature impacts accuracy by at least 1%; all other features fall below this threshold.

2.6 Discussion

Our research aims to understand how systems based on artificial intelligence can be used to build early automatic prediction tools and derive new and useful information on the phenomenon of low educational achievement. In this casestudy, we selected information from the Italian Ministry of Education reporting both demographic and socio-economic information and INVALSI test scores for a large number of students. Concretely, we exploited this information to produce a dataset which was then used to train three predictive models. The first one is based on random forest algorithm, from which we derive useful rules for the understanding of the underachievement phenomenon. The others are based on neural networks, which improve the performance of the first one in terms of accuracy, precision and recall, and is a good candidate for integration in automatic systems for teaching support.

Our results demonstrate that the challenge of predicting low achievement risk for primary and secondary school students can be effectively addressed through the use of well-curated datasets and the choice of reliable predictive models: using data from level K-5 we were able to predict the low achievement risk at K-10 with accuracy and recall of 0.78 with our best model. Therefore, we can affirmatively answer our **RQ1.1** regarding the use of ML techniques for early prediction of the risk of low achievement. A particular interest is addressed to recall; in fact, a high recall score —also named sensitivity— 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 possible intervention areas. This ability to predict an underachievement five years in advance with a reasonable precision offer, we believe, a practical tool to policy makers, managers and educators who want to tackle this problem.

A key point in our approach is the abstract representation of (INVALSI) tests and the related encoding of students learning that we defined: this allowed us to generalize our models on different cohorts and therefore to obtain a meaningful prediction, addressing **RQ1.2**. The chosen representation is knowledge-based, meaning it is guided by the theory behind the design of test items according to the relevant educational framework. It embodies the expertise of scholars and mathematics education professionals who have refined a description of student learning outcomes over the years. Although this approach relies on the implicit assumption of isolating individual aspects of learning (areas, processes and macroprocesses), without considering their interactions and overlaps, it has proven to be an efficient baseline for addressing the problem of generalization. This has enabled all proposed models to achieve good predictive performance.

As regards **RQ1.3**, the analysis of the feature importance highlights some significant elements which enhance the RF model explainability. As expected,

the most significant features for predicting achievement in mathematics are the maths grade provided by the school and the overall score in the maths INVALSI test. These are summative evaluation indices, designed to summarize learning outcomes in a single value. While their relevance strengthens confidence in national evaluation processes on the school system by satisfying an interest of school policymakers, on the other hand they do not highlight specific critical issues in learning. With our students' learning encoding, we sought greater granularity to verify if some areas or skills were more crucial than others for low achievement prediction. Through a statistical factor analysis the didactic coordinators and teachers could plan targeted enhancement and recovery actions to prevent underachievement. As already noted in other studies, an impact on the phenomenon is due to the socioeconomic and cultural context. The relevance of the features on the geographical area suggests, for example, that a study on a regional basis, rather than a national one, could highlight further determining factors on low achievement in that territory.

2.7 Chapter Conclusion

The results discussed in the previous section contribute to the goals of this thesis and lay the groundwork for the studies presented in subsequent chapters.

This case study successfully addressed the issue of generalizing the representation of student learning across different cohorts, particularly in terms of the model's predictive performance, thus contributing to goal G3. However, the chosen representation did not highlight didactic-disciplinary features among the most relevant ones in the explainability study. This limitation affects the model's utility for stakeholders such as teachers and students, who may be interested in early interventions of didactic nature to counter the risk of underachievement. Thus, we tried to increase the quality of the information provided as input to the predictive models, improving the students' learning encoding. To achieve this, among the many possible paths, we decided to explore strategies that mitigate a purely data-driven approach by incorporating additional knowledge sources during the model training phase. This direction will be presented in Chapter 5 and Chapter 6, following a methodological approach known as Informed Machine Learning [7].

Another important point is that both the RF and neural network-based models demonstrated satisfactory predictive performance, with the FTT model emerging as the most effective. The tabular data examined in this study are characteristic of many other educational contexts. The pipeline used here —starting from the generalization process across different student cohorts, the training and testing of models with the same ML techniques used in this case study, and their interpretation and discussion — could therefore be extended and validated in other contexts

that need to solve predictive problems in education using tabular data on students and their learning outcomes. This approach enables us to investigate the issue of transferability (goal G1), which we address in the next chapter through a case study on academic dropout.

As for the contribution of this work to goal G2 concerning the explainability challenge, here we limited to the feature importance analysis for the RF model. However, advancements in Explainable AI now allow for post-hoc analysis of feature importance even for neural network models [53, 54]. As a development of this work, it is worthwhile to compare post-hoc explainability techniques for feature importance applied to both models, i.e. RF model and neural network-based ones. This allows us to compare them not only in terms of predictive power but also in terms of their transparency for various stakeholders. Additionally, comparing the features identified as significant by models based on different techniques can provide insights into the reliability of the explainability analysis. In this regard, further research will be introduced in Chapter 4.

Chapter 3

Academic Dropout: testing transferability

The primary aim of this chapter is to test whether machine learning techniques used to address the prediction of low achievement risk, as described in the previous chapter, can be transferred to other similar contexts. The similarity is based on two main factors: firstly, the task remains a risk prediction within an educational context; secondly, the available data are tabular in nature, comprising heterogeneous and mostly categorical information about students.

In this chapter, we examine the phenomenon of academic dropout, specifically the prediction of a student's risk of prematurely leaving their educational path. This issue meets the similarity criteria mentioned above, but differs in two essential aspects from the problem studied in Chapter 1. Firstly, it pertains to a different educational segment, namely higher education instead of high school; secondly, the target to predict is different, as it involves the actual early leaving of the educational path rather than the quality of student learning outcomes.

The interest in this new case study is driven by two distinct reasons. The first motivation is the significance of dropout at the societal level and its impact on individual well-being. The second reason is more practical, stemming from the interest of a major Italian university in developing tools, policies, and educational actions to counteract student attrition. In the next section, we will delve deeper into the motivations and background for this new case study.

A critical aspect of this investigation is understanding the conditions under which predictive models trained on K-12 educational data can be effectively transferred to higher education. The feasibility of this transfer depends on several key assumptions. Firstly, while both contexts involve students' academic trajectories, the underlying factors influencing performance and dropout may differ signifi-

cantly. The primary predictors of low achievement in high school, such as attendance, parental involvement, and early academic struggles, may not hold the same predictive power in a university setting, where factors like financial constraints, self-regulation, and institutional support play a greater role. Secondly, the distribution and nature of available data may vary: while K-12 datasets often come from standardized testing and school records, higher education data may include more self-reported and administrative variables. Lastly, the interventions available in each context differ; while early interventions in high school often involve direct teacher or parental support, universities rely more on academic advising, financial aid, and institutional policies. These considerations frame the transferability of models and highlight the necessity of adapting feature selection, model tuning, and evaluation metrics to align with the specific characteristics of higher education dropout prediction. However, despite the necessary considerations just mentioned, the transferability we test is methodological in nature, meaning, in a nutshell, the application of the same type of machine learning algorithms to obtain predictive models based on educational tabular data.

This chapter presents the research work detailed in [6]. The work was developed in collaboration with Dr. Zingaro and Prof. Gabbrielli, expanding on their previous research [55]. The paper addresses two complementary aspects: the transferability of techniques that have proven effectiveness in tackling the issue of student low achievement, and the explainability of such models. In this chapter, we focus on the first aspect, deferring the second part to Chapter 4. The ML techniques considered are random forest (RF) and a Feature Tokenizer Transformer (FTT), compared against a naive benchmark model. In the previous case study, we also considered Categorical Embedding as an embedding technique for categorical features in neural networks, but it was discarded here due to its inferior performance across all considered metrics. The analysis of the model's predictive performance has been supplemented with a fairness analysis, which examines the model's equity concerning students belonging to specific demographic groups.

All sections have been reorganized to emphasize the issue of transferability, which is the main focus of this chapter. In the sections on related work, materials and methods, and results, the content primarily involved selecting from the contributions already presented. The results section, along with its discussion, has been expanded with a more detailed analysis of fairness. In the journal version, this analysis had been presented in a condensed form due to editorial constraints. The introduction and conclusion of this chapter underwent the most significant revisions to better align with the overarching thesis and its objectives. This restructuring creates a more cohesive narrative, providing clearer insight into how the methods can be applied across different educational contexts.

3.1 Background and motivation

Dropout is a critical issue in the field of education, with significant consequences for individuals and society as a whole. The complexity and importance of this phenomenon have prompted research efforts since the 70s [56]. A key indicator for dropout is the ELET rate, which measures the percentage of Early Leavers from Education and Training [15], referring to young people aged 18 to 24 who have not attained an upper secondary qualification. Early leavers are more likely to be unemployed or employed in low-paid jobs with few or no prospects for training and further career progression; they are more prone to social exclusion and to experience lower levels of health, wellbeing and life satisfaction; they are also more likely to experience limited civic participation. In [57] the European Commission indicated as one of the main targets to be achieved in education the reduction of the ELET rate from 15% to 10% in the decade 2010-2020. While this target was reached in several European countries, Italy in 2019 still had an ELET rate of 13.5%[58].

These findings motivated us to choose academic dropout as a new case study. We focus specifically on the outcomes of first-year students, using machine learning techniques to analyze real data from a prestigious Italian university in a in-person learning setting. We define dropout as a situation where a student does not re-enroll in the same study program for the following academic year. Therefore, the dropout target is always assessed after 12 months of enrollment.

To test the transferability of the techniques used for addressing student under-achievement, as presented in Chapter 2, we establish our baseline as RF, and we compared the results with the results gained by the FTT approach. RF is known for its predictive performance and explanatory power in previous case studies of academic dropout prediction [59]. Using the transformer technique, we explore flexible strategies for handling the categorical data prevalent in our dataset to adequately represent the findings. To the best of our knowledge, there are no other studies in the literature applying FTT to address the prediction of academic dropout risk. As a result, we address the following research question:

RQ2.1 To what extent does the use of FTT improve predictive models of student dropout compared to state-of-the-art techniques?

We considered data on approximately 40,000 students. They cover multiple information, including demographics, prior schooling, enrollment, and first-year academic performance, to identify patterns in students' academic trajectories and predict dropout risk at an early stage. To facilitate early intervention, we include data on the academic performance of the same cohorts of students at different time intervals, i.e., at enrollment, after three, six, nine, and twelve months. By

measuring the performance of the model at each time interval, we aim to answer the second research question:

RQ2.2 To what extent does post-enrollment academic career information improve model performance?

Given the current model of university assessment and education, we hypothesize that students' academic career characteristics can provide valuable insights into predicting dropout risk. The validity of this hypothesis is tested using our set of predictive models, trained at different stages, as a simulation tool. We use appropriate performance metrics, including precision, recall, and F1, to test and analyze the hypothesis. Our experiments were conducted using all available data, including all students enrolled in any undergraduate course at the university over three academic years, with no sub-sampling or data exclusion.

The chapter is structured as follows. Section 3.2 provides an overview of related approaches. In Section 3.3, we describe the dataset, and introduce the preprocessing pipeline. The ML techniques used for the AI predictor implementation are not presented here because they overlap with those used for the previous work described in Chapter 2. Section 3.4 presents the results, comparing predictive performance. Section 3.5 discusses the results in relation to the research questions. Finally, Section 3.6 provides concluding remarks and highlights future research directions.

3.2 Related Works

The study of understanding and decreasing dropout rates within higher education has advanced significantly, with numerous investigations utilizing diverse analytical methodologies and data sources [1, 60]. This review focuses on research in conventional *in-person* classrooms, categorized by machine learning algorithms (RQ2.1) and the role of academic career information (RQ2.2).

3.2.1 Machine Learning in School Dropout Prevention

The landscape of machine learning algorithms for academic dropout prediction has evolved significantly [55, 61, 62], with a growing emphasis on the adaptability and performance of deep architectures [63]. Early work by Anand et al. [64] used recursive clustering to evaluate student performance in programming courses, identifying underperforming students early. Alban and Mauricio introduced neural networks for university dropout prediction, using multilayer perceptrons and radial basis function networks to achieve high accuracy rates [65]. Nabil et al. [63]

compared various machine learning algorithms, finding that DNNs outperformed traditional methods due to their ability to capture non-linear correlations between student characteristics.

Baranyi et al. [66] extended the utility of deep learning by focusing on interpretability. They used deep neural networks and gradient-boosted trees, achieving high prediction accuracy and providing feature ranking through permutation importance and SHAP values. Tang et al. [67] introduced KIDNet, a knowledge-aware neural network model that combines factorization machine and deep neural network algorithms to capture both lower-order and higher-order feature interactions, demonstrating its effectiveness on a real-world dataset.

In summary, the empirical validation of deep architectures for predicting academic dropout has enriched the state of the art and opened avenues for future research. Our work aligns with this trend by adopting the FTT model [68], exploring the potential of attention-based neural networks for tabular data in the context of academic dropout prediction. This research aims to leverage these architectures to develop more effective and nuanced models to mitigate dropout rates.

3.2.2 Data Sources and Features for Predicting Academic Risk

We review the types of data sources and features used in existing literature to predict academic risk, focusing on academic history information. Dekker et al. [69] used structured data, such as student grades and attendance records, to predict dropout in electrical engineering programs, achieving 75% to 80% accuracy with decision trees. Kiss et al. [70] incorporated both structured and unstructured data, including pre-enrollment achievement measures and first-semester performance indicators, using artificial neural networks and boosting algorithms to highlight the incremental predictive validity of early university performance indicators.

Jayaraman [71] used unstructured data from counselor notes, employing natural language processing techniques to extract sentiments and using them as features in a random forest model, achieving 73% accuracy in predicting student dropout. Del Bonifro et al. [55] presented a prediction tool that uses machine learning techniques to assess the risk of first-year undergraduate students, incorporating a range of variables from personal data to proficiency credits. This study serves as a foundational reference, particularly in its methodological approach to using pre-enrollment and first-year academic data for predictive modeling.

Alwarthan et al. [72] conducted a systematic review of data mining techniques used to predict student academic performance, identifying random forest and ensemble models as the most accurate but noting a lack of consensus on the im-

pact of admissions requirements on student performance. Alam [73] introduced a multimodal neural fusion network combining structured and unstructured data to predict various student retention risks, reporting promising performance and investigating the fairness of the model.

Our research aligns with the existing literature on using structured data for predictive modeling, capturing the temporal aspects of academic performance through a time series approach. Our dataset comprises over 40,000 student careers, spanning three academic cohorts and including 110 different degree programs, enhancing the predictive power and generalizability of our models across different academic contexts.

3.3 Materials and Methods

In this section, we present the novel aspects in terms of materials and methods compared to the case study described in the previous chapter. One of the most significant contributions of this paper is the richness of the dataset considered, which we detail in subsection 3.3.1. The preprocessing operations performed on the dataset are described in subsection 3.3.2. Unlike the previous case study, there are no major updates on the ML techniques used, as we still employ RF and FTT, for which readers can refer to section 2.4.2. To put the performance of our chosen models into perspective, in this new case study, we also implemented a basic model that simply predicts the most frequent class for all instances. This naive baseline serves as a point of reference to evaluate the added value of the more sophisticated RF and FTT models, especially in the context of our dataset's class imbalance. The final subsection (3.3.4) introduces the methods used to conduct the fairness analysis of the proposed models.

3.3.1 Dataset description

The dataset used for this work was extracted from a collection of real data from one of the largest Italian universities. Specifically, we have considered pseudonymous data describing 44, 875 students enrolled in 110 courses in the academic years 2018/19, 2019/20, and 2020/21. The dataset is collected by the university, thanks to the informed consent provided by students at the time of enrollment. This allows the data to be used in pseudonymized form for research activities aimed at improving the teaching offer and academic services. However, the pseudonymization of the dataset ensures that students cannot be identified, thereby meeting the ethical requirements of the research.

Our analysis focuses on the first year. Statistical evidence from the source

data suggests a concentration of dropouts in the first year of the course, with the phenomenon gradually decreasing in subsequent years. For the 2018 cohort of students, the only one for which we have data three years after enrollment, the dropout rate after one year is 14.8% of the total number of enrolled students, while those who leave by the third year is 23.4%. This means that 63.2% of the registered dropouts occurred in the first year, confirming the importance of acting within the first year to prevent dropouts.

Table 3.1: Available features for each student in the original dataset, along with the possible values range. The first column uniquely identifies the corresponding feature.

UId	Features	Type	Range
AE	Age of Enrollment	Numeric	≥ 0
SG	Student Gender	Nominal	1, 2
GOma	Geographical Origin (macro)	Nominal	1–6
GOmi	Geographical Origin (micro)	Nominal	1–76
EFSI	EFSI	Nominal	1–8
HST	High School Type	Nominal	1–10
HSM	High School (final) Mark	Numeric	60–100
CD	First/Single Cycle Degree	Nominal	1, 2
AS	Academic School ID	Nominal	1–11
DN	Degree Name	Nominal	1–97
PT	Place of Teaching	Nominal	1–9
ALR	Additional Learning Reqs.	Nominal	1, 2, 3
WMA	Weighted Marks Average	Numeric	0 or 18–30
NH	Number of Honors	Numeric	≥ 0
ECTS	Number of Credits	Numeric	0–60
DO	Dropout	Nominal	True or False

Table 3.1 provides a comprehensive overview of the features of the dataset. The table is divided into four columns: the first column serves as a unique identifier for each feature, which will be referenced later in Section 4.4; the second column names the feature; the third column specifies its type (either nominal or numeric); and the fourth column outlines the possible values or ranges.

The features are categorized into four distinct groups.

Personal data includes characteristics such as gender, age, and geographical origin, as well as the Equivalent Financial Situation Indicator (EFSI), which measures the economic status of the family at the time of enrollment.

- *Age of Enrollment.* This numeric feature indicates the age of students at the time of their enrollment. It can offer insights into the relationship between age and academic performance or dropout rates.

- *Student Gender.* This feature captures the gender classes given as binary (male or female) encoding. This is used as a basis for stratified analyses to assess model fairness across gender categories.

- *Geographical Origin.* This feature is further divided into macro- and micro-categorizations. The macro-categorization identifies six modalities, distinguishing between four macro-areas in Italy, foreign students, and instances where the information is not available. The micro-categorization offers 76 possible values, corresponding to either the Italian region or the country of origin for foreign students.

- *EFSI.* The Economic Financial Situation Indicator (EFSI) feature is optional upon enrollment and is segmented into eight distinct financial bands. These bands are designed to encapsulate the economic status of both the students and their families. The bands are ordinal in nature, ranging from the lowest, which signifies the most financially disadvantaged situations, to the highest, indicative of more financially favorable conditions.

Educational background relates to the educational background attained at the upper secondary level. Specifically, this group includes two key characteristics:

- *High School Type.* This nominal characteristic delineates ten different types of high schools from which students graduated. It is used to capture the diversity of educational backgrounds and to potentially elucidate any correlations between the type of high school attended and academic performance or dropout rates in higher education.

- *High School Final Mark.* This numerical characteristic represents the final mark obtained by students at the end of their high school education. It is intended to provide an initial quantitative measure of academic competence that may be indicative of subsequent performance in higher education.

Academic program set relates to the characteristics of the program in which the student is enrolled. This group comprises several attributes:

- *First/Single Cycle Degree.* This ordinal characteristic categorizes the length of the program. A value of '1' represents first cycle degrees, which typically last three years, while '2' corresponds to single cycle degrees, which last five or six years.

- *Academic School ID.* This nominal feature identifies the academic school selected by the student. The dataset currently includes eleven different academic schools, each potentially offering a unique set of degree programs.

- *Degree Name.* This characteristic serves as a unique identifier for the specific program chosen by the student, allowing for granular analysis of academic pathways.

- *Place of Teaching.* This nominal characteristic indicates the geographical location of the program's headquarters, with nine different cities represented in the dataset.

- *Additional Learning Requ(irement)s (ALR).* This ordinal feature accounts for the possibility of mandatory additional coursework during the first academic year. Certain programs require an admission test, and failure to pass this test necessitates additional coursework and subsequent examinations. The ALR characteristic is coded as follows: '1' indicates programs without ALR; '2' indicates that the ALR exam was passed; '3' indicates that the required ALR exam was not passed.

Academic performance set relates to measures that capture students' academic progress after enrollment. This group is informed by three main variables, each available at different time intervals:

- *Weighted Marks Average.* This numerical characteristic represents the average examination mark, weighted by the corresponding ECTS credits for each examination. In the context of the Italian academic evaluation system, exam marks range between 18 and 30. Consequently, the weighted average also falls within this interval. If a student has not passed any exams, this average is set to 0.

- *Number of Honors.* This numerical characteristic quantifies the cases where an exam was passed with honors. Note that although honors are recorded, they do not affect the weighted average of exam grades.

- *Number of Credits.* This numerical characteristic indicates the total number of European Credit Transfer and Accumulation System (ECTS) credits earned by the student. The maximum number of ECTS credits that can be accumulated in a single academic year is 60.

The **target variable** for our predictive models is the 'dropout' characteristic, represented as a Boolean variable with values of 0 and 1, encoding `False` and `True`

respectively. Specifically, a value of 0 is assigned to students who exhibit canonical academic outcomes, characterized by the continuation of their studies and the successful acquisition of course credits. Conversely, a value of 1 encapsulates three distinct non-canonical outcomes, each of which indicates a form of academic withdrawal. The first category includes students who formally abandoned their studies without transferring to other Italian programs. The second category includes students who have transferred to other programs within the same academic institution. The third category consists of students who left their current program to enroll in another university.

It is appropriate to categorize these three non-canonical outcomes as forms of dropout, as they all represent a deviation from the student's original academic trajectory. The differences between them lie solely in the subsequent choices that students make after dropping out. Moreover, these non-canonical outcomes collectively constitute a minority in the dataset, accounting for 23.4%. Treating them as separate classes would exacerbate the existing problem of class imbalance in the dataset. Therefore, we opted for a binary classification framework where the target variable is set to false for canonical outcomes and true for all non-canonical outcomes, thereby simplifying the problem while retaining its essence.

In order to facilitate a nuanced analysis of students' academic progress, we have divided the data into five different time intervals, each of which captures a different phase of the first academic year. These intervals are defined at 0, 3, 6, 9, and 12 months after enrollment. Importantly, each student is represented in each of these intervals; no data points were excluded at any time. This approach resulted in five different versions of the dataset for each cohort of students. The versions are distinguished by the values of the fourth set of characteristics, which are updated to reflect the academic metrics at each time interval. This methodology allows us to strike a balance between making early predictions and capturing the evolving academic trajectories of students. In the subsequent sections, we will use the notation T0, T1, T2, T3, and T4 to distinguish between datasets collected at different time intervals after enrollment, respectively 0, 3, 6, 9, 12 months since the beginning of the academic year.

3.3.2 Dataset preprocessing

The dataset contains both numerical and categorical characteristics. Numerical features, such as age at enrollment and final school grade, are processed as floating point numbers. The target variable for classification, called the 'dropout' feature, is Boolean, as explained in the previous subsection.

Categorical features require distinct preprocessing techniques to suit the specific requirements of RF and FTT algorithms. For RF, one-hot encoding is em-

ployed to fit the training set, similar to the approach used in Chapter 2. In this new scenario, unknown categories may also be present due to changes in study programs offered across different cohorts of students, such as new institutions or curriculum reforms. Therefore, all categorical variables are managed to account for unknown categories by using a vector of zeros. In contrast, the FTT models use label encoding, which assigns unique numerical labels to each category within a feature. This method is advantageous for algorithms that benefit from ordinal relationships between categories. Similar to RF, FTT models are trained on the training set, and the same encoding scheme is used for validation and testing. Unknown categories are coded as zero.

The dataset exhibits class imbalance, with dropout instances representing only 15.4% of the total. Such imbalance can negatively affect the performance of binary classification models [74]. For RF models, the imbalance is mitigated by class-weighted options. The weights are calculated based on the bootstrap sample for each decision tree and are inversely proportional to the class frequencies. These weights influence both the entropy criterion for splits and the "weighted majority vote" of the terminal nodes [75]. In the case of FTT, random weighted batch sampling is used to counteract the imbalance. This technique adjusts the selection probabilities based on class frequencies, thereby improving the representation of the minority class during training. This eliminates the need for data replication and mitigates the effects of class imbalance.

3.3.3 Experimental setup

The dataset was split into training, validation, and test sets. We used data from the academic years 2018/19 and 2019/20 for training and validation, while data from 2020/21 was reserved for testing. The split ratio for the training and validation sets was 70:30.

For training the RF models, we performed grid search cross-validation to identify the optimal hyperparameters, including the number of trees, maximum depth, and minimum samples per leaf. The models were trained using 5-fold cross-validation to ensure robustness and generalizability, i.e., the dataset was divided into five subsets, the model was trained on four subsets and validated on the remaining one.

The FTT models were trained using the Adam optimizer with a learning rate schedule that decayed the learning rate based on the validation loss. We used early stopping to prevent overfitting, monitoring the validation loss and halting training when no improvement was observed for a set number of epochs. Random weighted batch sampling was employed to handle class imbalance during training.

Evaluation metrics included accuracy, sensitivity, specificity, as described in

Section 2.4.3, and weighted F1 score. These metrics were computed for both the validation and test sets to assess model performance. We considered the *Weighted F1 Score* important because it balances sensitivity (recall) and precision (positive predictive value), providing a harmonic mean of these two metrics. This measure is particularly useful in imbalanced datasets as it accounts for both false positives and false negatives, giving a comprehensive view of the model's performance. A high F1 score indicates that the model is both accurate and sensitive, effectively identifying students across risk categories.

To ensure the reproducibility of our results, we set random seeds for all random processes involved in data splitting, model training, and evaluation. All experiments were conducted using Python with libraries such as scikit-learn for the RF models and PyTorch for the FTT models.

3.3.4 Fairness analysis

Fairness analysis was conducted to identify and mitigate potential biases that may differentially impact specific demographic groups. We focused on several protected attributes, including gender, geographical origin (categorized into macro regions), and economic status as indicated by the Economic and Financial Situation Indicator (EFSI). These attributes were chosen due to their relevance in reflecting the diverse backgrounds of the student population and their potential influence on academic outcomes.

For each protected attribute, we performed stratified analyses to evaluate the performance metrics across different subgroups. The metrics analyzed included accuracy, precision, sensitivity, specificity, false positive rates, and false negative rates. This analysis was aimed at detecting any disparities in model performance that could indicate bias.

To quantify fairness, we utilized several fairness metrics following recent pivotal research [76, 77]: (i) *demographic parity*: Ensures that the prediction rate is similar across different demographic groups; (i) *equalized odds*: Requires that the true positive rate and false positive rate are similar across different demographic groups; (iii) *predictive parity*: Ensures that the precision is similar across different demographic groups.

Confidence intervals at the 95% level were determined using bootstrap resampling techniques. These intervals provide a measure of the variability in our fairness metrics and help in assessing the statistical significance of any observed disparities. In the stratified analysis, we used the following steps: (i) define groups based on the protected attributes (e.g., male vs. female for gender); (ii) calculates the performance and fairness metrics for each group; (iii) compare the metrics across groups to identify disparities and analyze the potential causes of any detected bias.

The fairness analysis ensures that our predictive models not only achieve high accuracy but also uphold the ethical standards essential in educational settings. By systematically addressing fairness, our study contributes to the broader discourse on ethical AI in education, ensuring that our models are both effective and equitable.

3.4 Predictive Performance Results

Table 3.2: Summary of Accuracy and Sensitivity on Test Set

	Accuracy		**Sensitivity**	
Time Step	**RF**	**FTT**	**RF**	**FTT**
October enrolment (`T0`)	0.72	0.78	0.48	0.44
End of January (`T1`)	0.75	0.78	0.65	0.51
End of April (`T2`)	0.84	0.83	0.59	0.65
End of July (`T3`)	0.85	0.86	0.75	0.74
End of October (`T4`)	0.85	0.87	0.80	0.81

In Table 3.2 we present the performance metrics, including accuracy and sensitivity, for both the RF and FTT models at different time intervals after enrollment, as described in subsection 3.3.1. These metrics are evaluated on the test set. The FTT models generally demonstrate superior accuracy compared to their RF counterparts, except when assessed six months post-enrollment. Conversely, the RF models show improved sensitivity capabilities under certain conditions.

The best-performing model overall is the FTT variant trained on data available twelve months after enrollment, achieving an accuracy of 0.87 and a sensitivity of 0.81. For comparative analysis, we also introduce a naive baseline classifier that predicts the majority class label from the training set across all test instances. This classifier achieves an accuracy of 0.84 and a sensitivity of 0.25 over all time intervals considered.

Figures 3.1 and 3.2 illustrate the time trends in the performance metrics for the RF and FTT models. The weighted F1 score, identified as the most equitable metric in Section 3.3.3, also shows a general improvement over time. This suggests that the quarterly updates of student career information contribute significantly to the predictive power of the models.

For RF models, the most notable improvement in performance occurs between the zero and six-month intervals, with a slight decrease in sensitivity thereafter. For the FTT models, the metrics show a more consistent upward trend, reaching satisfactory levels even at the time of enrollment. The performance of the naive

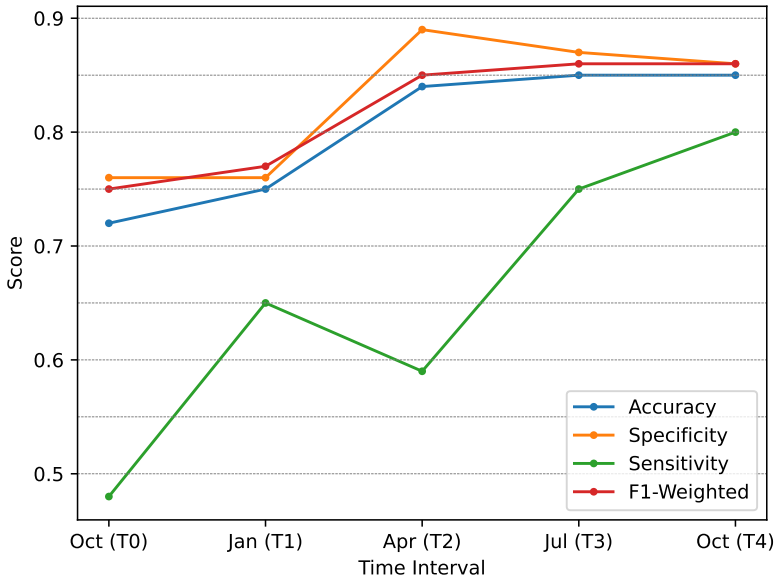

Figure 3.1: **RF Model Performance Over Time**. Variation in accuracy, sensitivity, specificity, and weighted F1-score at different intervals from October enrollment. Actual weighted F1 score values are provided.

classifier serves as a baseline to help interpret the effectiveness of the RF and FTT models, particularly in scenarios with unbalanced datasets.

3.4.1 Fairness analysis results

In Figure 3.3, we present a comprehensive comparison of the fairness analysis of the best model in terms of predictive performance. The figure consists of 15 box plots, systematically arranged in a grid. Each row in this grid is dedicated to the analysis of a particular feature, while each column corresponds to one of the selected evaluation metrics—namely accuracy, recall, precision, false positive rate, and false negative rate—all evaluated at a decision threshold of 0.5. This visual representation serves as a robust tool for examining the performance of the model across different subgroups, thereby facilitating a nuanced understanding of its fairness attributes.

We conducted the fairness analysis by segmenting the dataset based on key demographic and academic factors. Specifically, we focused on three variables to define the subgroups: gender, geographical origin, and the Economic and Financial Situation Index (EFSI). These variables were chosen as they are the most relevant

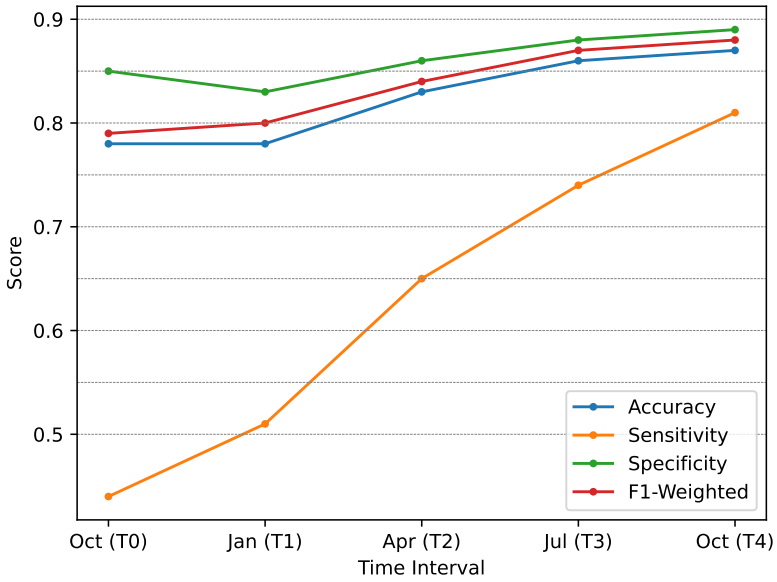

Figure 3.2: **FTT Model Performance Over Time**. Fluctuations in accuracy, sensitivity, specificity, and weighted F1-score at specific intervals from October enrollment. Actual weighted F1 score values are included.

for identifying subgroups within populations at risk of discrimination. Unfairness often arises along income lines and demographic groups [78, 79], and this segmentation enabled us to assess the model's performance across different subgroups, ensuring that no particular group is disproportionately advantaged or disadvantaged. This fairness analysis ensures that our predictive models not only achieve high accuracy but also adhere to the ethical standards critical in educational contexts.

3.5 Discussion

This section synthesizes the predictive and explanatory evidence to address the research questions (RQs) outlined in the chapter introduction, incorporating findings from similar studies and reflecting on the practical implications of our results.

In response to our first research question (**RQ2.1**), our empirical study supports the effectiveness of FTT models in predicting academic dropout risk. As detailed in Section 3.4, the FTT models, particularly the T4-FTT, consistently

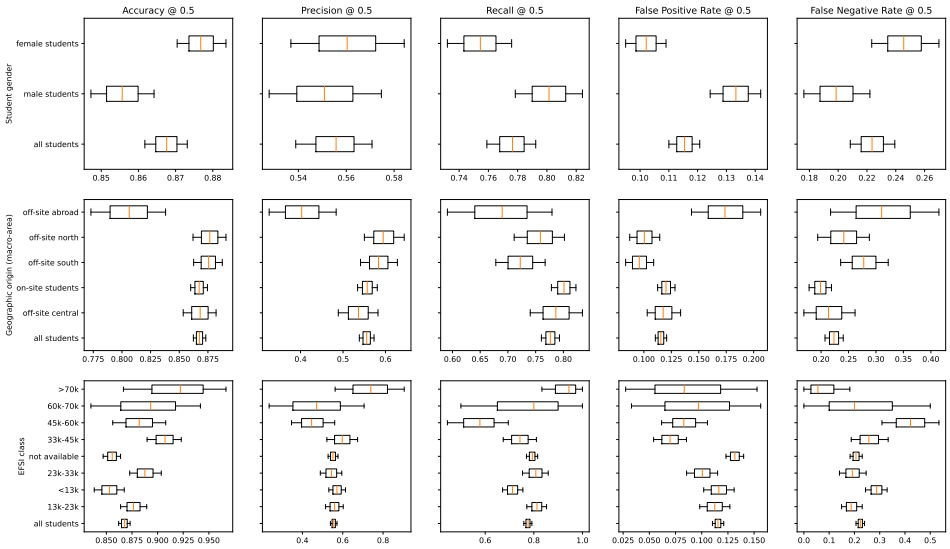

Figure 3.3: **Quantitative analysis of fairness across multiple features and metrics**. The boxplot matrix is organized into three rows (student gender, geographical origin, economic class) and five columns (accuracy, recall, precision, false positive rate, false negative rate), all evaluated at a decision threshold of 0.5. This layout provides a comprehensive view of model fairness across different subgroups and evaluation criteria.

outperform Random Forest (RF) models across various evaluation metrics, showing at least a one percentage point improvement. These results align with the trend of using deep learning algorithms for dropout prediction, recognized for their adaptability and sophistication [55, 61, 62, 63]. The balanced performance of FTT models in sensitivity and other metrics makes them suitable for dropout intervention strategies, consistent with findings in [64, 65]. Furthermore, our fairness analysis, which shows consistent results across different characteristics within a 95% confidence interval, complements the focus on interpretability seen in [66].

Our study expands the use of deep learning in educational data mining, achieving high accuracy while improving model transparency and fairness. The flexible architecture of FTT models, including an embedding component and attention mechanism, allows customization to different feature sets and data distributions, enhancing predictive fairness. This novel application of attention-based neural networks for tabular data in dropout prediction adds to the existing literature [68], confirming the efficacy of FTT models for this task.

Regarding our second research question (**RQ2.2**), our findings highlight the crucial role of European Credit Transfer and Accumulation System (ECTS) credits in predicting dropout risk, supporting prior research on the importance of academic history information [69, 70, 71]. Using structured data, including ECTS credits, our approach accurately predicts dropout risk, validated by our extensive dataset capturing diverse academic paths. The temporal sensitivity of our models to academic career characteristics underscores the dynamic nature of academic risk factors, with incremental improvements in predictive performance using data from progressively distant time points from enrollment. This aligns with Kiss et al.'s emphasis on early university performance indicators [70].

However, relying solely on academic characteristics like ECTS credits, which can change over time, has limitations. Building on the work presented in [71, 73], we propose incorporating immediate behavioral features for a more comprehensive assessment of dropout risk. Our study verifies ECTS credits as a reliable dropout risk indicator and underscores the significance of analyzing academic career characteristics over time. Combining structured data in a time series format, our methodology contributes to the current academic discourse, proposing ways to incorporate other information types for more comprehensive predictive models.

3.6 Chapter Conclusion

This chapter significantly contributes to the discussion on predicting academic dropout by showcasing the effectiveness of innovative machine learning techniques. Our research, aiming to evaluate the effectiveness of the Feature Tokenizer Transformer (FTT) compared to traditional models like Random Forest (RF) and to assess the impact of academic career data, illustrates the powerful potential of machine learning in identifying students at risk of dropout.

Our results indicate that FTT models exceed the predictive accuracy of RF models for academic dropout prediction, albeit with higher computational expense. The integration of academic career data markedly improves model performance, especially by enhancing the sensitivity and enabling a more detailed profiling of students likely to drop out.

For educational stakeholders, the implications are substantial. By harnessing data-driven insights, institutions can better tailor their student retention strategies. Utilizing comprehensive datasets and sophisticated models like FTT allows for a nuanced understanding of student behavior and risk factors, aligning with prior research suggesting that these methods refine retention strategies when paired with simulation-based analyses.

Despite these advances, further exploration is necessary. While ECTS credits provide valuable data, they may not fully explain dropout causes. Future studies

should incorporate qualitative aspects, such as student motivations and study habits, to offer a holistic view of dropout causation. Additionally, developing customized predictive models to cater to the unique dropout dynamics of different academic programs is essential.

Ethical considerations are crucial in the application of predictive models in education. Future research should address program-wide evaluations rather than focusing solely on individual students, incorporating broader contextual factors like environmental and support systems, as emphasized by previous studies. Combining insights from educational, cognitive, and psychological research with data-driven techniques enhances the ethical and effective use of machine learning in educational contexts, ultimately leading to a comprehensive understanding of academic dropout patterns.

In summary, our study underscores the importance of integrating varied data sources and advanced machine learning models to enhance the prediction of dropout rates. This approach not only improves predictive accuracy but ensures interventions are informed, ethical, and effective, contributing to a deeper understanding of academic attrition and strategies to mitigate it.

Within the broader objectives of this thesis, this chapter emphasizes the transferability of machine learning techniques across educational settings. By predicting low achievement risk in primary and secondary education and academic dropout in higher education, we demonstrate the adaptability and robustness of models like RF and FTT across varied environments and tasks.

This ability to transfer is critical, as it demonstrates that well-crafted machine learning models are not limited to singular datasets or problems, but can be adapted to tackle a range of educational challenges. The success of FTT in higher education—an area with limited existing literature—underscores the potential of emerging techniques that extend beyond traditional methods in capturing intricate patterns related to student achievement and dropout risks.

Achieving genuine transferability involves more than just applying machine learning techniques to new datasets. It requires a thorough comprehension of the entire modeling pipeline, including data preprocessing, feature selection, and model tuning, ensuring models not only perform well across different contexts but also provide interpretable and actionable insights. This process involves evaluating how model behaviors align or diverge across studies, informing necessary adaptations for each unique scenario.

In conclusion, our work highlights the value and challenges inherent in transferring machine learning methods across educational domains. By continuing to refine and optimize these approaches, we can leverage AI more effectively to identify students in need across various educational stages, ultimately facilitating more targeted and successful interventions.

Chapter 4

Academic dropout: explainability

In this chapter, we shift our focus from the predictive accuracy of machine learning models to their explainability, a crucial aspect that ensures transparency and trust in AI applications within educational settings.

Our examination builds upon the work already presented in Chapter 3, where we assessed the predictive performance of Random Forest (RF) and Feature Tokenizer Transformer (FTT) models to tackle academic dropout. Here, we delve into the explainability of these models, making advanced tools more interpretable to educators and stakeholders, who require clear insights to effectively apply AI to improve educational outcomes.

The research presented here introduces the explainability aspect through the study conducted in [6], thereby integrating the contribution from the same paper discussed in Chapter 3. Most of the content is directly drawn from the reference paper, with relevant sections selected to align with the chapter's objectives. The concluding section has undergone the major revisions to align with the approach proposed in this thesis. All pertinent information about the study's context, dataset characteristics, and developed models has already been covered in the previous chapter and is omitted here, as this chapter is intended to be an integral part of the thesis rather than a self-contained unit.

4.1 Background and motivation

In the previous chapter, we examined the predictive performance of the models. We now shift our focus to explainability, which is crucial for building trust in the proposed models and enhancing their applicability for stakeholders dealing with the dropout issue [80, 81]. Explainability involves the model's capacity to provide

transparent justifications for its predictions, allowing stakeholders to understand the underlying factors [4].

By evaluating both predictive accuracy and explainability, we ensure that the prediction models offer meaningful insights that can inform decisions and interventions. To achieve reliable measures of importance, we compare different post-hoc explainability techniques. We adopt both a local and global perspective in discussing explainability, taking into account the characteristics of the learning algorithm to determine which techniques are most effective in extracting relevant information. In section 4.3 we provide an overview on the selected methods.

Our aim is to assist educational stakeholders, including program coordinators and higher education policymakers, in their decision-making processes by quantifying the impact of each feature on predictions. This leads us to the primary research question addressed in this chapter:

RQ3.1 To what extent do the explanations provided by various post-hoc explanatory techniques contribute to the reliability of the hypotheses underlying our models and their results?

The chapter is organized as follows: the section 4.1 sets the stage by contextualizing the challenge of explainability within our academic dropout case study, and it formulates the primary research question addressed in this chapter. Section 4.2 provides a concise literature review associated with this topic. In Section 4.3, we detail the explainability methods used, noting the specific differences between those applied to the RF model and the FTT model. We explored three state-of-the-art explainability techniques: grouped permutation feature importance, attention map, and SHAP. Section 4.4 presents the findings, and the following section 4.5 provieds some insights on how these findings address the research question. The chapter conclusion repositions the contribution of this work in relation to Goal 2 of this thesis, concerning the explainability challenge for predictive models in education.

4.2 Related Works

Interest in explainability in AI (often referred to as XAI) has grown significantly in recent years. This is partly because AI technologies have become more widespread and partly because neural models, which often outperform standard machine learning techniques, tend to be perceived as "black boxes." Educational applications have followed this trend, often driven by the interests of various stakeholders involved in educational processes [82]. The importance of interpretability

and explainability in machine learning models is particularly pronounced in educational data mining, where the implications extend to human futures and career trajectories. Cohausz [83] emphasized the need for a nuanced, multi-stage approach to interpretability, advocating a fusion of artificial intelligence and social science methodologies, and extending LIME [84] for deeper interpretation.

Cannistrà et al. [85] highlighted the pivotal role of feature relevance in early dropout prediction, using an information-driven modeling strategy and considering the specific programs in which students were enrolled. Nagy and Molontay [86] used a range of explainable artificial intelligence (XAI) tools, such as permutation importance, partial dependence plots, LIME, and SHAP scores, demonstrating their utility in elucidating both global and local aspects of dropout prediction models. Delen et al. [87] presented a hybrid machine learning framework designed to provide actionable insights for individualized interventions, cautioning against the indiscriminate application of group-level insights for individual decision-making.

In line with these contributions, our research highlights the criticality of model interpretability and explainability. As detailed in Section 4.3, our methodology incorporates both global and local perspectives on explainability, emphasizing reliability and validity, underpinned by our comprehensive dataset and rigorous evaluation metrics.

4.3 Methods

One of the main contributions of this work is the implementation of explainability techniques to understand the predictions made by RF and FTT models. We focus on computing feature importance, determining how each feature contributes to the predictions. We applied our explainability strategies to the top-performing model in each family: Grouped Permutation Importance (GPI) for RF [88], Attention Map (AM) for FTT [89, 68], and SHAP [90] for both.

Random Forest (RF) models are valued for their interpretability. We used GPI [91], an adaptation of Permutation Feature Importance (PFI) [92], which addresses the consistency issue of one-hot encoded features by treating them as a single block during shuffling. GPI also incorporates feature weighting within each group to account for varying feature importance. This model-agnostic, post-hoc technique can be applied universally to explain trained black-box models [4].

For FTT models, we used Attention Map (AM) to compute feature importance. This model-specific technique relies on the attention mechanism in Transformers, averaging the attention weights for each token in the sample to determine feature importance.

In addition to GPI and AM, we employed SHapley Additive exPlanations (SHAP) [93], inspired by Shapley values from cooperative game theory, to pro-

vide local explanations for individual predictions. SHAP quantifies the influence of each feature on a model's prediction, offering nuanced insights into feature contributions for each instance.

GPI and AM offer global explanations, identifying which features drive overall model performance. GPI measures feature importance by the decrease in the model's sensitivity, while AM uses attention weights to estimate feature usage by the model. SHAP provides local explanations, detailing the impact of features on individual predictions.

We used GPI and AM to derive global explanations from the test set for RF and FTT models, respectively, and compared the features identified by both techniques. SHAP was used for local explanations on three selected students (early dropout, transfer, non-dropout) and for a global perspective using a beeswarm plot, summarizing how top features impact the model's output.

Due to the high computational cost of SHAP, we applied approximation strategies. Kernel SHAP [93] was used for general models, while Tree SHAP [90], optimized for decision trees, was applied to RF models. This allowed us to include all test instances in the RF beeswarm plot and limit FTT samples to 200 randomly selected instances.

4.4 Results

In this section, we outline the results of our explainability analysis, first adopting a global XAI framework similar to the methodology presented in [92], and then extending our investigation through the application of localized techniques. Among the RF-based models, we choose two versions: the model trained with data six months after enrollment (referred to as T2-RF model hereinafter) and the one trained with data twelve months after enrollment (T4-RF model). The first analysis aims to get insights into why the model registered a pitfall in the sensitivity performance to the advantage of specificity (see Figure 3.1). The second model has been chosen because it has the highest results according to all the performance metrics. For FTT, we considered the twelve months model (T4-FTT model), which is our best model according to the results presented in Section 3.4.

Our explainability results are organized as follows. In Section 4.4.1, we present the global explainability perspective with the techniques chosen for each model, i.e., GPI and AM for RF and FTT respectively. In Section 4.4.2, we present the results obtained with SHAP when used in its local explainability mode. We present its application to students in different conditions of continuation of studies as an example of the kind of insights that can be derived locally with SHAP. Finally, in Section 4.4.3, we present beeswarm plots for a global perspective through SHAP values, both for RF and FTT.

Table 4.1: Grouped Permutation Importance Results for Random Forest computed as mean decrease of sensitivity

Feature	**T2-RF**	**T4-RF**
Weighted Mark Average	0.248 (0.007)	0.001 (< 0.001)
Number of ECTS	0.036 (0.003)	0.493 (0.005)
Additional Learning Reqs.	0.025 (0.003)	0.005 (0.002)
Academic School	0.019 (0.002)	0.005 (0.001)
Age of Enrollment	0.001 (< 0.001)	0.005 (0.003)

4.4.1 Global Feature Importance

The use of Grouped Permutation Importance (GPI) in the Random Forest (RF) model facilitates the identification of salient features that contribute to the generation of predictions for each trained instantiation of the model. In this study, a feature is considered significant if its importance measure is greater than or equal to 0.01. This corresponds to a minimum 1% decrease in sensitivity due to random shuffling of that particular feature. Conversely, a feature is considered negligible if its importance measure falls below the 0.01 threshold. We performed 100 random shuffles for each feature to calculate the permutation feature importance. The mean and standard deviation of the most salient features for the T2-RF and T4-RF models are shown in Table 4.1.

In the T2-RF model, the most salient feature is the weighted mean grade, denoted by $\text{mean}_{\text{WMA}} = 0.248$, followed by the number of ECTS credits earned within 6 months of enrollment, denoted by $\text{mean}_{\text{ECTS}} = 0.036$; i.e. they contribute to a decrease in sensitivity of 25% and 4% respectively. The threshold of 1% is also exceeded by the allocation of Additional Learning Requirements, which are determined on the basis of admission tests, and by the categorization of the Academic School.

For the T4-RF model, we visualized an equivalent number of features as identified for the T2-RF model. However, only the number of ECTS credits earned twelve months after enrollment exceeds the 1% threshold criterion. Specifically, this feature shows an average decrease of 50% in the sensitivity metric, and thus emerges as the most important variable for identifying dropout risk. For the remaining features, the perturbation in sensitivity due to randomized reshuffling is insignificant, falling below the 1% threshold criterion.

The key difference between the two models is the impact of WMA and ECTS on sensitivity. As ECTS range increases along the academic year and WMA range remains stable, ECTS's importance grows relative to WMA@. The results (see Figure 3.1) indicate that reliable ECTS information, available by the end of the first

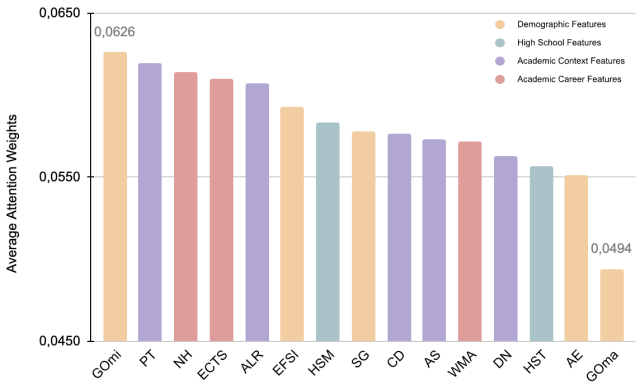

Figure 4.1: **Attention Map based Feature Importance for T4-FTT model**. Each bar shows the average attention weight for a feature in the training dataset, colored by information type as introduced in Section 3.3.

enrollment year, is crucial for model robustness compared to WMA. Indeed, where WMA matters the most, according to GPI analysis, there is a pitfall in sensitivity.

For the FTT model, we applied an attention map-based feature importance analysis to the T4-FTT model. In Figure 4.1, the average weight assigned to each feature in the attention map is shown. It is noteworthy that no single feature has a significantly higher average weight than the others. The average weights for all features are in the range $[0.0494, 0.0626]$. Nevertheless, it is worth noting that the top five features, in descending order of importance, are the geographical region of origin, the place of teaching, the number of awards obtained, the number of ECTS credits obtained, and information on additional learning requirements.

The two global explainability techniques applied to their respective models (T4-RF and T4-FTT) provide different insights. ECTS is by far the preeminent feature for T4-RF with GPI; on the other hand, it ranks fourth for T4-FTT with AM feature importance in a context where no feature stands out more than the others. However, the relevance of features related to the student's current academic career is consistent across both models. While the Global Feature Importance procedure offers a preliminary understanding of feature importance, it is worth noting that the results may not fully capture the explainability power of the models. The limitations of this kind of analysis suggest that more sophisticated techniques, such as SHAP, could be employed to provide a more comprehensive and interpretable understanding of the models' decision-making processes.

4.4.2 SHAP for Local Explanations

The SHAP explainability technique has been applied to both RF and FTT models. We chose SHAP because it is an agnostic state-of-the-art explainability technique. Thus, we considered the best models both for RF and FTT, i.e., T4-RF and T4-FTT models, and we present and compare their explainability outcomes. Firstly, we aim to introduce the results of local explainability gained from the models on some selected students, taken as examples. Figure 4.2 and Figure 4.3 display how the selected models came to the prediction correctly for three selected students, i.e., the predicted risk for the presented cases agrees with their actual value. Figure 4.2 refers to the RF model; Figure 4.3 refers to the FTT one. In each figure, we selected a student who early interrupts the academic career, a student who transfers to another degree program, and a student for whom dropout does not occur.

As for local explainability with the RF model, the number of ECTS is the one with the highest SHAP value (longest bar in the plot) both for the student who early interrupts the academic career (case *a* in Figure 4.2) and for the one for whom dropout did not occur (case *c*). The bar color and its orientation tell how this feature contributes to the predicted risk: for student *a*, pink and left-right oriented ECTS bar, not having accrued credits in twelve months raises the risk of dropout; for student *c*, blue and right-left oriented ECTS bar, having acquired 42 ECTS (out of 60 total) contributes to the prediction of a low risk of dropout. The same feature acts misleadingly for student *b*. In this case, ECTS is the second main feature according to its SHAP value. The attainment of 40 out of 60 ECTS credits is considered to be satisfactory as per the model. Typically, 60 ECTS credits represent the maximum amount of credits that a student can earn during the first year of enrollment. Thus it is used to downgrade the dropout risk prediction, although the actual target class for the student is positive to dropout. The most relevant feature for high-risk dropout for student *b* is the academic school, whose actual value is pharmacy and biotechnology. Statistics confirm that this academic school is affected by the highest number of transfers compared to other schools of the same university (36.9% in the three-year enrollment period 2018–2021 against a university average of 8.4%). This is because many first-year students choose pharmacy and biotechnology courses as a second study choice after being excluded from other degrees with restricted admission procedures, e.g., medicine and surgery, or veterinary medicine. As a final remark for the RF model, also WMA (weighted marks average) appears as a relevant feature for the dropout predictions for all the students (among the first five SHAP values).

As regards local explainability with the FTT model, we refer to the examples in Figure 4.3, and introduce the enabled explanations also in comparison with our observations for the RF model. Also for the FTT model, ECTS is the prominent

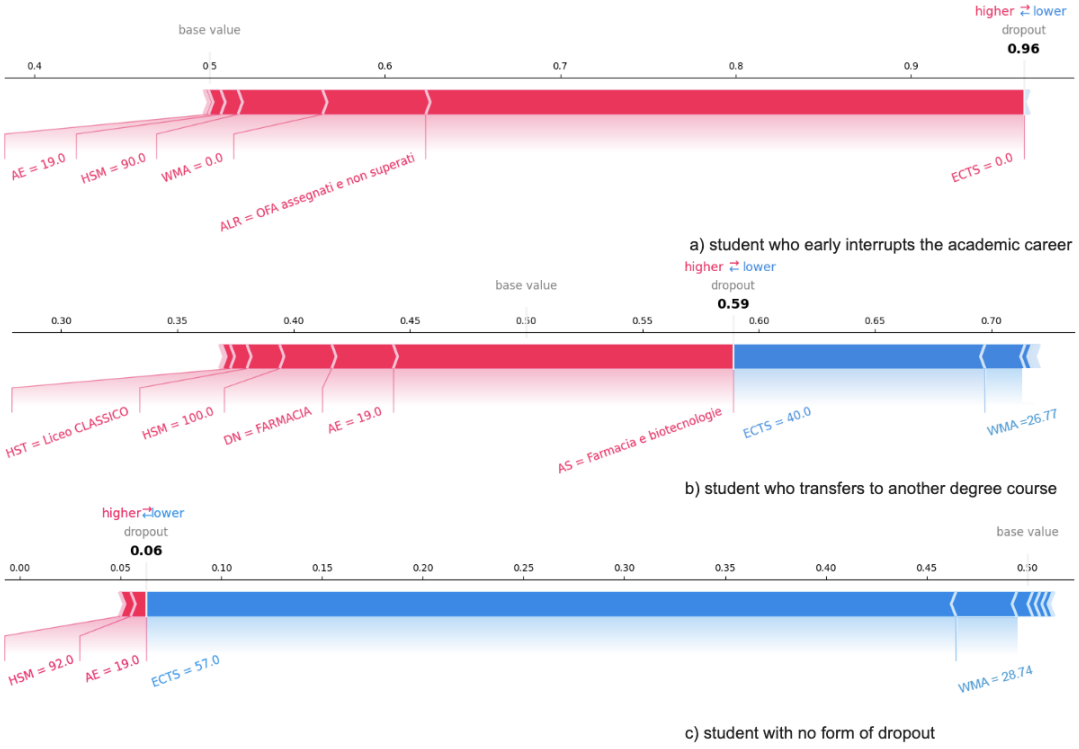

Figure 4.2: **SHAP Local explanations for RF model trained with data twelve months after enrollment**. Each line shows the main features impacting the predicted dropout risk for a student, with bar lengths proportional to their SHAP values. Pink bars indicate features that increase dropout risk, while blue bars indicate features that decrease it. The combined contributions determine the predicted value.

feature in the risk prediction for students *a* and *c*. The same feature is less relevant for student *b* (it appears as the seventh positive SHAP value). We have previously noted, based on the global analysis of feature importance using the AM method, that ECTS stands out as one of the most significant features, despite not being favored by the RF models. The different SHAP value of ECTS on different samples fits with this result. Furthermore, we want to underline that, unlike what was observed for the corresponding case for the RF model, the number of ECTS acquired by student *b* here contributes to raising the dropout risk, despite being an acceptable asset (40 out of 60). For student *a*, together with ECTS, the features associated with higher SHAP values are the assignment of ALR that have not been passed and the weighted marks average (equal to zero as no ECTS has been acquired). All these factors contribute, as expected, to raise the risk of dropout. The prominent feature for student *b* is DN, which identifies the degree program.

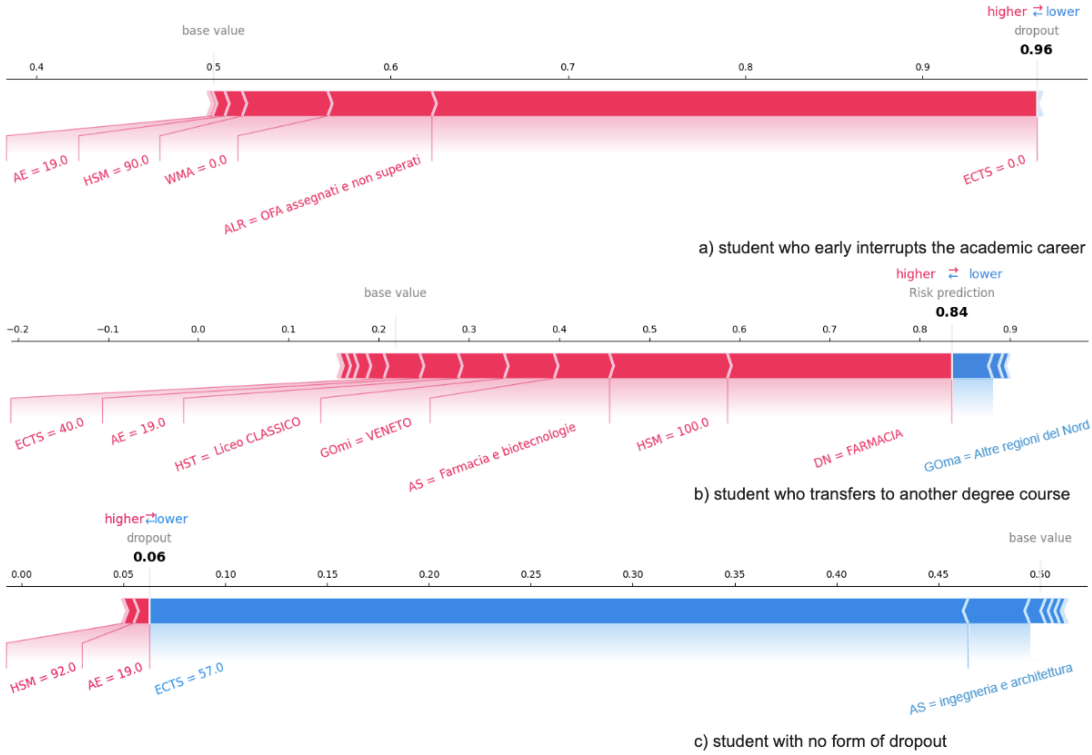

Figure 4.3: **SHAP Local explanations for FTT model trained with data twelve months after enrollment**. Each line shows the main features impacting the predicted dropout risk for a student. Refer to Figure 4.2 for instructions on reading the graph, which is similar for the RF model.

We found matching information for the RF model (related to the academic school), and we have already motivated how to interpret these results with some descriptive statistics. For student *c*, ECTS is definitely the prominent feature, followed by data on the academic school, which is engineering and architecture.

To sum up, we find two main similarities between the local explanations gained by the two models. Firstly, the relevance of ECTS for the dropout risk prediction of students *a* and *c*. Secondly, the relevance of information on the context of enrollment for student *b*, i.e., the enrollment in pharmacy. On the other hand, we have a main difference in how the number of ECTS (40 out of 60) is used for the dropout risk prediction of student *b*. One might wonder what interpretation to give to this difference. We hypothesize that the RF model struggles more in learning correlations between different features; the feature tokenizer module for input features embedding and the attention mechanisms of the FTT architecture provide greater flexibility, which allows, in the case of student *b*, to consider in a

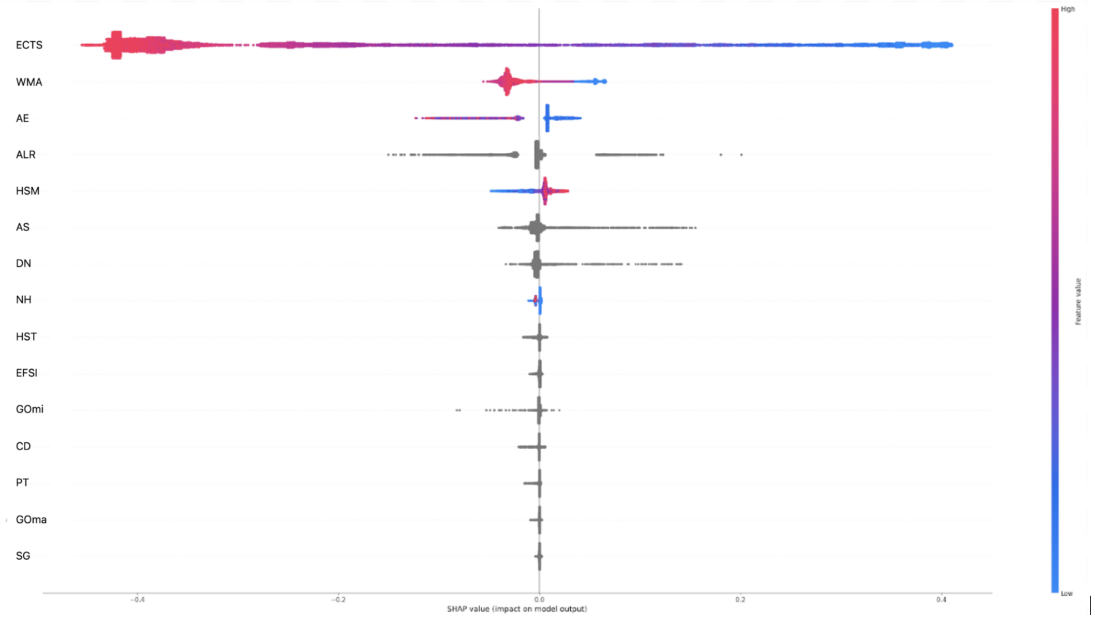

Figure 4.4: **SHAP global explanations for RF model trained with data twelve months after enrollment**. The beeswarm plot for the T4-RF model shows features ordered by average SHAP value. Each dot represents an instance, positioned by SHAP value; colors indicate numeric feature values.

"contextualized" way the weight and the orientation effect of the number of ECTS. We deepen this discussion in Section 4.5.

4.4.3 SHAP for Global Explanations

Let us move back again to a global explainability perspective, aggregating local explanations computed with SHAP in a summary plot, namely beeswarm. Figure 4.4 and Figure 4.5 refer respectively to RF and FTT models, trained with data on students' academic careers twelve months after enrollment. In a beeswarm plot, for each instance, i.e., a student in our case study, the provided explanation is visualized by a single dot on each feature row. The SHAP value of the row feature for each instance determines the horizontal position of the dots, whose distribution along each row shows a density graph. This information may be exploited to provide a global overview of the feature's importance. Features are in descending order according to their mean SHAP value. Moreover, each dot for numerical features is colored according to a chromatic scale to display the original value of a feature for each instance. Thus, a blue point on the right side of the ECTS row means that there is a student with a low number of acquired ECTS

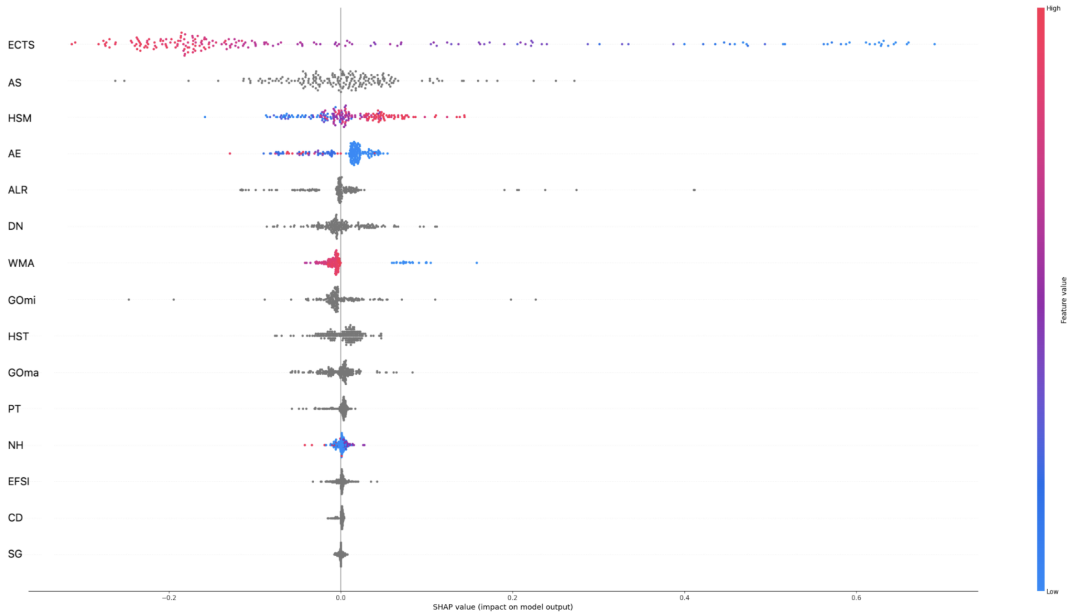

Figure 4.5: **SHAP global explanations for FTT model trained with data twelve months after enrollment**. The beeswarm plot for the T4-FTT model shows features ordered by average SHAP value for 200 randomly chosen samples. Refer to Figure 4.4 for guidance on interpreting the chart.

and this has a great influence in boosting her/his risk of dropout.

As for categorical features, both binary and non-binary, we consider each of them as a single factor of analysis, encompassing all its modes. Specifically, SHAP values are computed for each instance by summing the SHAP values of all binary features associated with that categorical feature. For this set of features, no color mapping has been set to avoid implying an order among categorical features.

The beeswarm plot for T4-RF model points out ECTS as the main feature; this is in line with the result retrieved with GPI. In particular, there is a high density of pink dots on the left, thus we can infer that the model often uses the acquisition of a high number of ECTS as an impact criterion to place the student in the low-risk class. The weighted average mark is another determining factor for the low-risk class, considering the high-density area of pink dots i.e., high weighted average mark, on the left side. For the applicability of the model, it would be of interest to determine the features that are most decisive for the high-risk class, i.e., we are looking for high-density areas in the right part of the plot. However, no such situation is evident in any row. We may observe a slight correlation between low age of enrollment or low weighted average mark with high dropout risk. It is worth noting a counterintuitive result among the explanations used by the model:

the final high school grade (HSM) when high, is often considered by the RF model as a rationale for high dropout risk. Nevertheless, its impact on the prediction is small according to SHAP values. Finally, we point out the relevance that numeric features play in the RF model with respect to the categorical ones. The top three consists of all numeric features, and this is even more impressive if we consider that the dataset has only five numeric features against ten categorical ones. Among the categorical features, the most effective (fourth position according to mean SHAP value ranking) is the one on the attribution of ALR.

We conduct a similar analysis also for T4-FTT model. We limited the beeswarm plot generation to 200 randomly chosen samples, due to time computational cost. This represents a limit in the SHAP global perspective on FTT but still allows us to obtain some insights. Similarly to the case of T4-RF, ECTS is the preeminent feature with a cluster of pink dots on the left side of the plot. Thus, in many cases, there is a correlation between a high achievement of ECTS and a lower dropout risk prediction. Other relevant data are those on the academic school of the students, their high school marks, age of enrollment, and information on additional learning requirements. We observe a blue cluster for the age of enrollment on the right side of the plot, revealing that young students are more likely to drop out. Also for the FTT model, the numerical features are relevant in determining the risk prediction. However, their distribution is less asymmetric than we observed for the RF model. In Figure 4.4 for the T4-RF model, we have four out of the first five features which are numerical. Moreover, the numerical features are all in the first half of the ordered features. In the T4-FTT model, we have a more homogeneous, albeit still not symmetrical, distribution.

4.5 Discussion

In examining the research question (**RQ3.1**), we explore the explainability of predictive models within educational data mining. Our findings align with and expand upon the current discourse on model interpretability. The significance of ECTS credits across different explainability techniques, such as Global Post-hoc Interpretability (GPI) and Shapley Additive Explanations (SHAP), supports previous studies [86, 85]. The detailed explanations provided by SHAP for the FTT model contribute to discussions about the adaptability of models to individual student cases, a topic also explored by Cohausz [83] and Delen et al. [87].

While both RF and FTT models focus heavily on ECTS credits, their ability to explain outcomes differs, highlighting the complexity of model interpretability. Our analysis showcases the FTT model's local explanatory power, especially regarding ECTS features, and emphasizes its versatility in adjusting to various feature sets and samples, allowing for a detailed understanding of factors contributing

to dropout rates.

Our study also focuses on improving the sensitivity of the T2-RF model by evaluating the relationship between contextual information and academic career data. Incorporating ECTS credits as a feature enhances the understandability and reliability of predictive models, as highlighted by Delen et al. [87]. By transparently utilizing ECTS credits in our predictions, we improve the credibility and transparency of these models, emphasizing the importance of fairness and adaptability in their implementation.

4.6 Chapter Conclusion

In this chapter, we explored a case study of explainability analysis for predictive models assessing the risk of academic dropout. Together with Chapter 2, this work extends the initial case study of the thesis, which focused on the risk of low academic achievement among high school students. We not only tested the transferability of the original methodology to a new context but also addressed the emerging challenges of transparency in ML models to achieve reliability and trust among stakeholders.

The explainability analysis and its results provide valuable insights that aid decision-making processes to counteract undesirable phenomena like low achievement or dropout. By evaluating the importance of different factors in determining risk, we offer quantitative evidence that supports the investigation of specific situations or the prioritization of certain actions.

For instance, the consistent significance of ECTS across multiple predictive models, as highlighted by various explainability techniques, suggests the need for a more detailed examination of the factors influencing their acquisition by students. Since the ECTS index is an indicator of academic success or failure, additional data could be collected on various aspects of study habits, such as study locations, timing, duration, resources used, and social learning networks. Furthermore, insights into the formal learning environment might include factors like study program organization, teaching methodologies, and the accessibility of lectures and materials. These represent just a few examples of areas where further investigation could yield beneficial information.

This work advances the exploration of explainability strategies for predictive models in education (Goal 2). First, the case study itself demonstrates the effectiveness of various explainability techniques applied in other contexts. Second, it equips the university with useful information for designing actions to combat dropout. Building on this, we propose that explainability analysis should not be viewed as the final step in the pipeline that merely provides model reliability. Rather, it should be part of an iterative refinement process where findings, such

as the significant role of ECTS, highlight areas for deeper investigation to enhance both the model's predictive performance and its explainability. This approach encourages consideration of whether additional data or knowledge sources related to ECTS could be integrated into the model during both the training and application phases. By doing so, we aim to achieve a more detailed and comprehensive understanding of dropout causes, potentially enhancing both the model's predictions and its transparency. In forthcoming chapters, we will further develop this iterative perspective and explore how it connects to the third goal of this thesis, which is the theme of adequate student representation.

Chapter 5

Informed Machine Learning for Knowledge Tracing

In the previous chapter, we introduced a central issue for this thesis. To outline the roadmap of the journey thus far, three key points stand out. First, we introduced predictive models for mitigating the risks of student low achievement and academic dropout, aiming to identify the most effective tools by comparing standard and innovative ML techniques. Next, we conducted an explainability analysis of these models, primarily to provide reliability and transparency for the benefit of various stakeholders. This analysis, however, raised epistemological questions regarding the type of knowledge introduced, processed, and generated by these models and their explainability analyses, and how this knowledge should be utilized.

Specifically, when employing a predictive AI model, one might ask: what knowledge was inputted for the model to learn its task? How is this knowledge processed and given meaning in the context of prediction? How does this knowledge interact with the educational setting or align with the stakeholders' existing knowledge?

These questions led us to conceptualize the development of predictive models not as a linear sequence of steps, but as an iterative process where the knowledge generated by the model can suggest improvements for subsequent training phases [9].

As noted in the conclusion of Chapter 2, a pivotal issue regarding this knowledge theme is linked to student encoding. Representing a student involves complex layers, such as their current learning status, socio-economic background, previous academic history, and the dynamic educational context influenced by peers and teachers. While on one hand, it is beneficial to collect extensive data across various factors to map learning influences, pedagogical-related theories may also provide

guidance for this data collection, and help in interpreting the results.

For these reasons, I turned my interest in exploring the knowledge tracing through the lens of Informed Machine Learning. Knowledge tracing, as we will detail later, is a well-established challenge in the AI literature within education, differing from the prediction of low academic achievement or dropout, yet sharing its reliance on student encoding. Its extensive literature history makes it a fertile ground for exploration. Informed Machine Learning offers a hybrid approach—not solely data-driven—focused on integrating diverse knowledge sources in AI tool development. This approach aligns well with the aim of addressing epistemological questions related to the development of predictive models.

In this chapter, I will present a comprehensive literature review on this topic conducted during my time abroad at the DIPF - Leibniz Institute for Research and Information in Education in Frankfurt [94]. This chapter lays the foundation for the research developments to be discussed in Chapter 6, framing the context and introducing key terminology required for understanding its impact. The only modification from the original paper is the addition of a paragraph in the chapter's conclusion to underscore this chapter's contribution to the overall objectives of the thesis.

5.1 Motivation for a Systematic Literature Review

Learner modeling –also called student modeling– is a widely studied problem due to its relevance in various technologies to enhance learning, including intelligent tutoring systems (ITS) [95] and adaptive educational hypermedia systems (AEHS) [96]. The problem's relevance has its roots in the theories for individualized learning, studied since the 1980s by Cohen et al. [97] and Bloom [98], which prove its effectiveness compared to traditional classroom learning. This motivated the search for technological support and strategies for learner modeling, which could promote individualized learning.

A Learner Model (LM) is an abstract representation of the learner that considers cognitive and non-cognitive characteristics. According to Vagale and Niedrite [99], the LM "contains all information that the system has on the user and maintains live user accounts in the system", i.e. it keeps both static and dynamic information about the learner. Specifically, static information is data that is not changed during the student and system interaction, e.g. personal data or pedagogical and preference data collected once and which stay unchanged during the system utilization. On the other hand, dynamic information is data on the student's learning progress and interaction with the system. It can refer to the student's performance, i.e. student achievements during the course session, and the actual state of the

student's knowledge concepts and skills. The dynamic data component in the LM determines a continuous flow of collecting and updating data about the learner.

As for the methodological approaches to tackle the learner modeling problem, there are two main families [100]. The first set of approaches relies on psychometric methods, e.g. Item Response Theory [101] and Cognitive Diagnostic Models [102], which are mostly based on static data. However, in the last decades, technological advances opened the possibility of collecting dynamic data while the student interacts with a learning system. This attracted the interest of computer scientists in Knowledge Tracing (KT) [10], which can be described as monitoring students' changing knowledge states during the learning process and accurately predicting their performance in future exercises.

However, there are two main challenges connected to the KT problem. The first one concerns its complexity due to its interdisciplinarity nature. According to Abyaa et al. [103], learner modeling is challenging since it is based on intertwining education science, psychology, and information technology. This led them to suggest that "to construct an ideal learner model, one should identify and select the learner's characteristics that influence their learning, then take into consideration the learner's psychological states during their learning and choose the most adapted technologies to model each characteristic with the best precision". This challenge is inherited by KT as a subclass of learner modeling.

The second challenge is spread out by the limitations that emerge from the implementation of ML techniques in KT. Although purely data-driven techniques for KT achieve satisfactory performance, they have some pitfalls regarding their applicability, reliability, or interpretability, that we do not find in psychometric models [104]. They mainly differ from purely data-driven approaches because they are grounded in a theoretical framework. In Item Response Theory, for instance, each item is associated with an a priori difficulty coefficient. Moreover, there is the assumption that learning doesn't occur during testing. Both these assumptions are used in designing the model. This theoretical advantage in psychometric models led us to suggest a possible methodological framework for addressing the pitfalls emerging by purely data-driven techniques, referring to Informed Machine Learning (IML), introduced by von Rueden et al. [7]. In a nutshell, IML aims to overcome a purely data-driven approach favoring hybrid ML models which integrate alternative knowledge sources to data in the ML pipeline.

In this chapter, we want to contribute to the two challenges just presented, by conducting a systematic literature review. Specifically, we aim to highlight if and how the prior knowledge due to the interdisciplinary and complex nature of KT can be used to overcome the limitations which emerge by using standard ML methods. For this purpose, we want to verify if the paradigm proposed with the IML is effective and productive for KT, i.e. whether it can be applied fruitfully to

develop KT models. Therefore, we want to find references in the literature that explicitly consider forms of prior knowledge injection to address the KT problem. On the one hand, we can outline the state of the art in deploying hybrid ML approaches for KT. On the other hand, we aim to point out the current gaps in the literature and suggest new avenues for research

The rest of the chapter is organized as follows. In section 5.2 we introduce the background. We describe the main ML techniques used for KT with their critical issues. Then, we outline the main features of the IML paradigm. Section 5.3 describes the methodology. We display and motivate the RQs tackled through our systematic literature review. We describe both the literature survey procedure and the classification process. In section 5.4, we present the results of our analysis. Firstly, we describe the taxonomy we distilled from the surveying of the papers. It is an adaptation of the one proposed by von Rueden [7] due to our focus on the educational field. Secondly, we show the classification results of the selected papers gained by applying our IML adapted taxonomy. Section 5.5 discusses key insights into the results concerning the RQs. In section 5.6, we summarise this chapter's main contributions and provide some concluding remarks.

5.2 Background

5.2.1 Knowledge tracing

Formally, the KT problem can be described as follows. Let us consider a learner's history exercise sequence $X = \{(q_1, r_1); (q_2, r_2); ...; (q_{t-1}, r_{t-1})\}$, where $\{q_i\}$ is the id for the question answered by the learner at the i^{th}-time step, and $r_i = 1$ if the student provided the correct response to the question q_i, 0 otherwise. The goal of KT is to predict the probability of correctly answering the question q_t at time step t, i.e. computing $P_t(r_t = 1|q_t, X)$. Hence, given the unknown function $f : \mathcal{X} \to \{0, 1\}$ which associate each learner's history exercise sequence (X, q_t) to 1 (q_t correctly answered) or 0 (otherwise), the KT goal is to determine a function $g : \mathcal{X} \to [0, 1]$ which is a good approximation of f. The prediction is based on hidden variables, whose values are updated at each time step and which model the student's knowledge state.

There are three main classes of ML techniques widely exploited in the literature and well described by Minn [104]: Hidden Markov models, Factor Analysis models, and Deep Learning-based models. The main exponent of the first class is Bayesian Knowledge Tracing (BKT) [10]. In this model, the learner's knowledge state is represented through a set of binary variables for each skill or knowledge component (KC), which assumes the true value if the student is in the learned state. The

observed data is the student performance, the latent variables are the student knowledge state for each skill. The truth value of the latent variable corresponding to the skill or KC k depends on four factors: (i) the initial learning factor $p(L_0^k)$, which is the prior probability that a student already masters k; (ii) the acquisition factor $P(T^k)$, which is the probability for the student to pass from the unlearned state to the learned state after the next practice opportunity; (iii) the guess factor $P(G^k)$, i.e. the probability the student guessed the correct answer despite being in the unlearned state; (iv) the slip factor $P(S^k)$, which models the probability that the student makes a mistake despite being in the learned state. The estimate of student mastery of k, i.e. the student knowledge state for k, is continually updated every time a student responds to an item [10]. In a nutshell, the student knowledge state for k after the n-th action of the student, indicated by $P(L_n^k)$, is computed considering both the posterior probability that the student was already in the learning state given the evidence (whether or not the n-th action is correct), and the probability that he will make the transition to the learned state if it is not already there. Then, the current student's knowledge state for k is exploited to compute the probability to perform a correct action taking into account the mitigation effect of the slip factor $P(S^k)$ and the positive effect of the guess factor $P(G^k)$.

Although BKT has been used successfully in many systems, it has some limitations, well summarized by Tato and Nkambou [105]. Specifically, as a starting point of BKT, there is a Bayesian network (BN) [106], which "sometimes implies to manually define apriori probabilities and manually label student interactions with relevant concepts". Also, "the binary response data used to model knowledge, observations and transitions impose a limit on the kinds of exercises that can be modeled". Furthermore, BKT is designed to model one skill or KC at a time, ignoring the interactions between skills and KCs and affecting a single performance.

As for the second class, it worths mentioning Performance Factor Analysis (PFA) [107], which is is a logistic regression model to predict accuracy considering the student's number of prior failures and successes on that skill. It is an extension of Learning Factor Analysis [108], designed to model multiple skills simultaneously, i.e. the prediction of the student performance relies on the conjunction or compensation of the skills needed in the performance by summing their contributions. PFA is competitive and outperforms BKT models[104]. However, PFA does not consider important behavioral factors such as the order of answers and the probability of students guessing or slipping. This may affect the reliability of the models prediction.

The third class of techniques is deep learning-based models, which have recently been widely used for KT, as in many other domains. There are two main

approaches in this class: Deep Knowledge Tracing models (DKT) [109], which is based on recurrent neural networks; and Dynamic Key-Value Memory Networks (DKVMN) [110], which is a memory-augmented neural network based on two memory matrices to exploit the relationships between underlying concepts and directly output a student's mastery level of each concept. As for the disadvantages of DKT, Yeung and Yeung [111] highlight two main points. Firstly, the model fails to reconstruct the observed input, i.e. the model predicts a failure for a student in a certain skill, despite the observation that a student on the same skill in the input data is a success, and vice versa. Secondly, the predicted performance for skills across time steps is inconsistent, i.e. there are sudden spikes and falls across time steps. This is intuitively undesirable and unreasonable as students' knowledge state is expected to transit gradually over time but not to alternate between mastered and not-yet-mastered. Moreover, neural networks have a high computational cost and are prone to overfitting. Sun et al. [112] point out some limitations also for DKVMN models. They ignore both the students' behavior features collected by their interaction with the learning system and the student's learning abilities, which affect the students' knowledge state.

In his review, Minn [104] compares IRT, BKT, PFA and DKT on three dynamic public datasets (ASSSITments 2009-2010 and ASSISTments 2014-2015, derived from the homonymous learning system, and Algebra 2005-2006, released in KDD Cup 2010 competition). He obtained the best performance with the IRT psychometric model, followed by the DKT model. The fact that the IRT model performs better than ML models is surprising in the first instance. Minn argues that "this result can be explained by the item difficulty factor, which is explicitly taken into the IRT models and not by DKT nor other student models without consideration of item information."

5.2.2 Informed machine learning

In section 5.1, we introduce two challenges. Firstly, the problem of KT lies at the intersection of several disciplines, including pedagogy, psychology, cognitive science, and information technology. Secondly, as supported by the previous subsection, standard ML models for KT show performance pitfalls that we do not find in psychometric models, which integrate a theory-ladenness.

We can expand the first issue by affirming that learning cannot be described only with information gathered directly from the learner or about the learner. Learning is influenced by the context in which it takes place, understood as a physical, relational, emotional, and disciplinary space [113]. The relevance of the context on learning has been considered since the first research on ITSs, which are based on domain models, pedagogical models and tutor-learner interface models,

Table 5.1: Informed Machine Learning taxonomy introduced in [7]

Source	Representation	Integration
Scientific knowledge	Algebraic equations	Training data
World knowledge	Differential equations	Hypothesis set
Expert knowledge	Simulation results	Learning algorithms
	Spatial invariances	Final hypothesis
	Logic rules	
	Knowledge graphs	
	Probabilistic relations	
	Human feedback	

together with the learner model [114]. However, these components are usually modeled independently, i.e. as 4 separate parts in the system. Little attention is paid to modeling how one can influence the others. Simplifying with an example, on one hand, the domain model can be seen as an organizational model of the repository for the system's educational resources. On the other hand, it can be understood as an epistemological model of an area of knowledge that may affect how the student learns, hence affecting also the learner model [115].

As for the second point, several contributions in the literature affirm the need to overcome purely data-driven approaches in machine learning [11, 116], especially in those contexts where the phenomenon is very complex, it is difficult to obtain sufficiently large and representative datasets. A priori or a posteriori forms of knowledge, acquired over years of research, are available [117]. All these factors exist for KT: its complexity has been motivated in the previous point; the challenge of quantity and quality of data [118] is quite common in the form of class imbalance [119] (e.g. correct answers on skills difficult to master), and a priori and a posteriori knowledge are usually available and are already used in ITSs and AEHSs. In these cases, it may be worthy to test hybrid learning techniques [120], which can be recognized as a strategy of *Informed Machine Learning* (IML).

von Rueden et al. [7] define IML as "learning from a hybrid information source that consists of data and prior knowledge. The prior knowledge comes from an independent source, is given by formal representations, and is explicitly integrated into the machine learning pipeline". With the term knowledge, they assumed a computer science perspective, defining it as "validated information about relations between entities in certain contexts". Moreover, they introduced a taxonomy for IML, outlining a scheme consisting of three types of knowledge sources, eight possible knowledge representations, and four forms of integration, as shown in Table 5.1. However, their paper did not refer to educational case studies. Hence, whether their taxonomy fits with the specificity of KT remains to be explored.

We suggest referring directly to their paper for a full description of the terms they introduced in the taxonomy. In section 5.4, we display the terms that we have distilled for our taxonomy, which is an adaptation of their proposal as a result of our analysis.

5.3 Methodology

5.3.1 Research Questions

To sum up, we are assuming that the complex nature of KT can be addressed by explicitly taking into account the information sources due to the different disciplines that deal with learning and the situation in which it takes place. This means finding a way to integrate these forms of prior knowledge in data-driven machine learning models. Therefore, we took as a reference the taxonomy proposed by von Rueden et al. [7] for IML, trying to apply it to the specificity of our topic. As already mentioned, in their framework the authors introduce three dimensions: knowledge source, knowledge representation, and knowledge integration (see Table 5.1). They also associate each dimension with an analysis question. Here we assume them as our research questions, focusing the field of study on the KT problem.

RQ4.1 Which source of knowledge can be integrated into machine learning models for knowledge tracing?

RQ4.2 How is the knowledge represented in those models?

RQ4.3 Where is the knowledge integrated into the machine learning pipeline?

We opted for a systematic literature review to highlight which avenues have already been explored, which trends are more common to design hybrid models for KT, and to identify new research methodological trajectories.

5.3.2 Literature surveying procedure

To perform our systematic literature review we followed the PRISMA statement. [121]. We included four main databases which contain relevant literature in the field: ACM Digital Library, IEEE Xplore, Scopus, and Web of Science. They are authoritative databases for the research sector in learning analytics and artificial intelligence in education, for which it is possible to carry out searches with articulated queries and by restricting the search field to some parts of the paper (e.g. abstracts and titles).

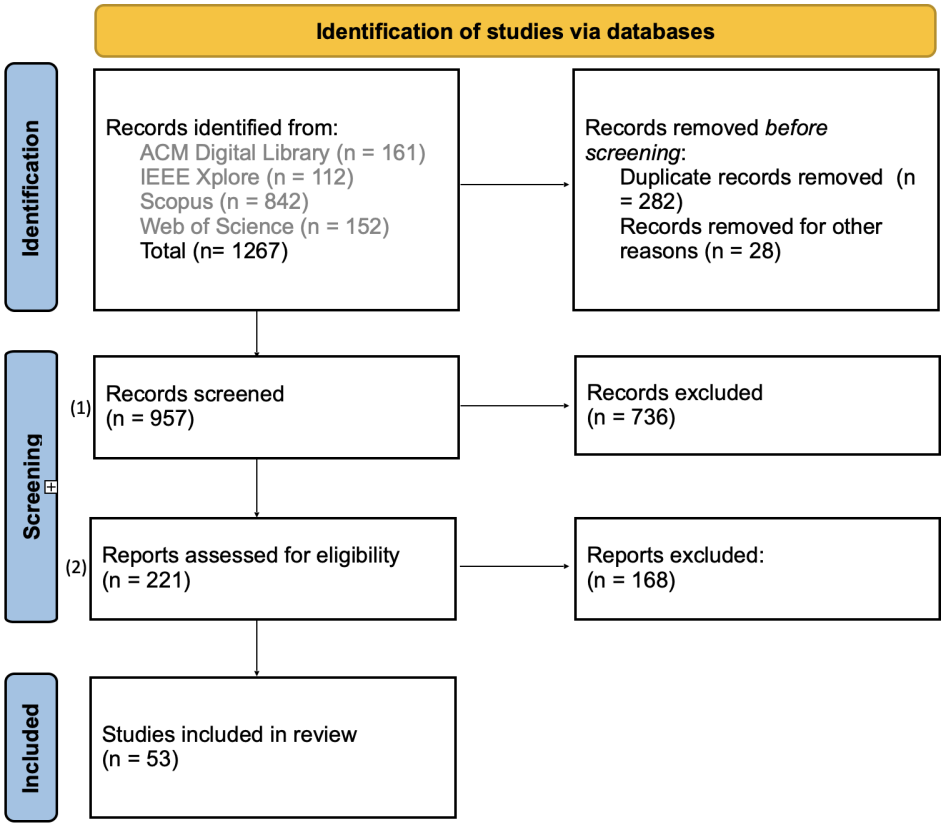

Figure 5.1: **PRISMA 2020 flow diagram for the screening process**. After the identification of the potential candidates, there are two screening steps. The first one (1) consists of applying 4 inclusion and 4 exclusion criteria considering only papers' titles and abstracts. The second (2) has been conducted with two inclusion criteria to focus on the prior knowledge injection problem, considering the papers' full texts. The selection criteria are described in section 5.3.

The query used to retrieve results in these databases is the following: *("skill development" OR "skill acquisition" OR "skill assessment" OR "knowledge tracking" OR "knowledge tracing" OR "knowledge assessment") AND ("machine learning" OR "artificial intelligence" OR computing OR "deep learning" OR "learning analytics" OR "data mining") AND (education OR educational)*. The search was limited to the titles, abstracts, and keywords of the documents in the databases to select only papers with the main focus on the topic of our interest. The query consists of the conjunction of three parts: the first is for keywords about the learning object under study; the second aims to bind the research methodology reference to machine learning and other related fields; the last one is used to disambiguate

the terms knowledge and learning, collocating them into the educational sciences. There is a fourth aspect characterizing our research questions regarding the use and integration of sources of prior knowledge. However, it is not easy to identify related keywords which are sufficiently general for an automatic filtering process taking this aspect into account. In a sense, one of this research's objectives is identifying which sources are most used as prior knowledge and which lexicon is used to refer to them. Therefore, the focus on prior knowledge was not considered in the first phase of the PRISMA checklist and was integrated later, as we will describe.

The query was run on August 5th, 2022, collecting 1267 documents. Figure 5.1 shows the main steps of the systematic review process according to the PRISMA flow. In the top-left box of the diagram, we summarized the numbers of retrieved documents with the query search, divided among the selected databases. After removing duplicates and documents in the form of full books or conference proceedings, the list of potential candidate papers was reduced to 957.

On this set of papers, we carried out a manual screening of titles and abstracts to assess the relevance of our study. Specifically, we considered the following inclusion criteria:

- the paper has a methodological focus on KT, i.e. aims to describe a technique, an algorithm, or a method to deal with KT problems rather than presenting a digital tech application (serious games, virtual reality systems, web platforms, etc.);
- the main methodological approach refers to the field of machine learning, computational intelligence, or data science;
- the data used to build the learner model are collected from the student's interaction with Learning Management System (LMS);
- the paper refers to human learning.

The first two criteria were chosen to pursue the methodological focus of the RQs. The third and fourth criteria were chosen to explicit a sufficiently broad but defined application target for knowledge tracing. Specifically, the third criterion narrows the interest of this study to school, academic and training contexts that use LMS as a teaching support tool. The fourth criterion disambiguates the word "learning", which can be used in AI concerning machine or robot learning. In addition, we considered four exclusion criteria:

- the full text of the paper is not available in English;
- the paper presents preliminary results, i.e. it is a position papers or the authors declare that they are describing an exploratory study;

- the paper has as its main objective a literature review;
- there are later more updated or complete versions of the paper by the same authors and on the same research project.

The application of these criteria led to the selection of 221 papers considered eligible. This set of papers has gone through a new screening phase on the full text, aimed at selecting only documents with a focus on using and integrating prior knowledge sources in ML tools. Specifically, it was decided to consider the following inclusion criteria:

- the authors explicitly consider the need to integrate apriori forms of knowledge with methods traditionally used to deal with KT;
- the paper includes a clear description of the methods, i.e. which prior knowledge is taken into consideration and how this is integrated into the ML pipeline.

To check the first criterion, we seek evidence in the text, particularly in the introduction and conclusion sections, where authors usually state their main contributions. For the second criterion, the methodology section of each paper was examined.

As a final result, we identified the 53 papers included in this review. All selected papers present case studies in which a close match occurs between the “informed" models and authoritative datasets exploited for KT (e.g. ASSISTment, KDD cup 2010, or data collected with commonly used LMS). The experiments described in the selected papers perform comparably to or better than the purely data-driven models, taken as a reference benchmark. This motivates our interest in exploring potentials and gaps of prior knowledge source integration into the ML pipeline.

5.3.3 The classification process

To classify the 53 papers considered eligible for the review, we tested and refined the IML taxonomy by von Ruedend et al. [7], which we have already introduced in section 5.2.

We felt comfortable using their notation about knowledge integration, which refers to the main steps in any ML pipeline [122]. As for the knowledge representation forms, we relied on the existing taxonomy, although it remains to distill which forms are actually used in the ML models for KT, and the possibility of expanding the initial list if other forms emerge. On the other hand, we immediately perceived the adaptation of the taxonomy for knowledge sources to our context as more delicate. von Rueden et al. proposed three main sources [7]: scientific knowledge,

world knowledge, and expert knowledge. According to their definitions, scientific knowledge mainly refers to science, technology, engineering, and mathematics, and it is validated through formal reasoning or scientific experiments. World knowledge alludes to facts from everyday life, which can be validated implicitly by human reasoning based on intuition; they also subsume linguistics as world knowledge, e.g. syntax and semantics of a language. Expert knowledge is common knowledge within a specific experts' community and is mainly validated through a group of experienced specialists.

Following these definitions, scientific knowledge does not apply to KT, while the other two forms fit with the educational context. However, it is sometimes difficult in our selected papers to distinguish whether a knowledge source is the result of general knowledge or is based on an expert-domain learning theory. Furthermore, the classification in the world and expert knowledge is extensive and does not capture some specificity of the sector, which a finer granularity of the taxonomy might capture. This refinement process was inspired by another taxonomy source borrowed from ITS and AEHS. These systems have four major components [123]: domain model, pedagogical model, learner model and tutor-learner interface model. The latter is mainly a model on a technical level: it determines the admissible inputs (e.g. click, typing, speech) and produces output in different formats (e.g. text, diagrams, animations, agents); it shapes the architecture through which data are collected; it mediates the interaction between the learner and the contents. On the other hand, the other three components are models for integrating information on different aspects that influence learning into ITS and AEHS. Hence, they are eligible as possible sources of knowledge specific to our topic.

Operationally, a first screening of the paper was made to build a set of labels suitable for classifying each of the three IML dimensions, taking into account the previous considerations. The most appropriate categories were gradually detected as the papers were analyzed. Once we derived the labels, we homogenized and reorganized them to arrive at a stable classification taxonomy. We conducted a second screening phase with this new label set by classifying the 53 papers. During this classification process, we identified the knowledge source enclosed in the model and how it was represented and integrated. Even more labels for each dimension of the taxonomy can be applied to a single paper if there are more types of knowledge sources or if the authors exploit different representation or integration strategies.

In the next session, we introduce the classification taxonomy obtained from the first qualitative analysis of the papers and the classification results in quantitative terms. Furthermore, we chose one of the papers included in our systematic literature review to show how our taxonomy can be applied to describe the prior knowledge injection flow in a real case study.

Table 5.2: References classified by knowledge representation and knowledge source

Source		Representation				
		Algebraic equations	Simulation results	Knowledge graphs	Probabilistic relations	Other data
Domain knowledge	Items difficulty	[119, 117, 124, 125, 126, 127, 128, 129, 130]		[131]	[132]	[133]
	Semantic similarity		[118, 134, 135, 136, 137, 138]	[139]		
	Knowledge structure	[140, 141, 142]	[143, 135, 144, 145, 128, 137, 142]	[117, 139, 124, 146, 147, 148, 149, 150, 151, 152, 153, 141, 140]	[105, 154, 155, 156, 157]	
	Class context			[139]	[139, 158]	
Learning knowledge	Pedagogical assumptions		[159]	[160]	[161]	
	Cognitive theories	[162, 112, 129, 130, 138, 111]	[163]		[150, 132, 164]	
Behavioural knowledge	Time	[165, 166, 167]			[125]	[133, 168]
	Scaffolding interactions	[169]			[157]	[168]
	Attempts			[162]		[112, 133, 168, 170]

5.4 Results

5.4.1 Taxonomy of Informed Machine Learning for Knowledge Tracing

Here we present the result of the qualitative analysis that led to the determination of the reference taxonomy for the selected papers' classification. In illustrating our results, we cite only the most recent paper among those included in the systematic review. The full classification, according to the introduced taxonomy, is offered in Table 5.2 and Table 5.3. The two tables classify the papers by knowledge representation and (path from) knowledge source and by knowledge representation and (path to) knowledge integration.

Table 5.3: References classified by knowledge representation and knowledge integration

Integration	Representation				
	Algebraic equations	Simulation results	Knowledge graphs	Probabilistic relations	Other data
Training data	[112, 165, 125, 127, 166, 128, 129, 130, 167]	[118, 134, 135, 143, 136, 159, 145, 128, 138, 137]	[160, 146, 147, 148, 149, 150, 151, 152, 153, 141]	[125, 154, 150]	[112, 133, 168, 170]
Hypothesis set	[162, 124, 129, 169, 138, 140]	[144, 163, 142]	[162, 117, 139, 124, 140, 131]	[139, 105, 161, 132, 155, 156, 158, 157, 164]	
Learning Algorithm	[119, 117, 126, 141, 111]				

Knowledge Source

The first focus in our taxonomy concerns pointing up the knowledge sources which can be considered when dealing with KT, i.e. other information retained valuable to integrate those generally used in the standard KT models, that are, the sequence of students' performances. We developed a two-level classification, which expresses different degrees of granularity, summarized in Figure 5.2. At the first level, we have three nodes inspired by the ITS components: domain knowledge (domain models), learning knowledge (pedagogical models), and behavioral knowledge (student knowledge).

With the term *domain knowledge*, we indicate both the disciplinary space, i.e. information related in some way to the content object of the learning, and the context where learning occurs. There are four kinds of information included under this umbrella term: items' difficulty, items' semantic similarity, knowledge structure, and class context. The first one refers to information about the difficulty level that characterizes each item used to track the students' knowledge development. It can be assumed either as an intrinsic property of the item, i.e. the level of difficulty is the same for all the students (e.g. [133]), or as a feature to be modelled properly for each student (e.g. [124]). Semantic similarity indicates the benefit of the items' texts as a source of knowledge. The general objective is to exploit semantic similarities between the exercises to highlight valuable relationships among them, as in [139]. In most cases, the integrated knowledge source concerns the knowledge structure, i.e. making explicit the relationships between knowledge concepts, skills, and exercises. This includes both links between concepts or skills considered to be communed by experts (e.g. [147]), and concept(s)-item or skill(s)-item links (e.g. [135]). In this way, we are considering the epistemological structure of a discipline to handle a typical issue in learning assessment: students' mastery level of a set of attributes cannot be measured directly but must be inferred from their pattern of responses to the items. The last option in this family, namely class context, indicates the use of information about the other students in the class to infer characteristics of the context in which the learning took place, assuming that this can influence each student's learning. For example, in [139], Tong et al. consider which exercises are often solved in sequence to infer hierarchical relations between the items. In [158], Wang and Beck try to create a model of the class because it can be representative of important information that affects a student's prior knowledge. For example, students in the same class have the same teacher and curriculum and have been assigned the same homework.

The second family of knowledge sources is named *learning knowledge*. It refers to expert knowledge about how learning occurs. We differentiate two types: pedagogical assumptions and cognition theories. In the first group, we enclose theories or hypotheses on learning from an external point of view. For example, in [160], Lee et al. cited knowledge space theory as a reference to capture the knowledge

structure. They assumed that if students correctly solve a tough exercise on a specific topic, they could even solve correctly other easier exercises on the same topic. Among the cognition theories, we include references to the individual learning process, e.g. the Ebbinghaus forgetting curves proposed in cognitive science studies [162].

As for *behavioural knowledge*, we refer to information concerning how students behave during the learning process in terms of interactions with the learning materials (mainly the items in a learning system). We see a connection with the learner model component of the ITS because this information is related to the student. Still, it enriches the exercise-performance sequence traditionally considered in KT. More granularly, we have identified three sub-labels: time, scaffolding interactions, and attempts. In [125], they used the average time of answer to estimate items' difficulty. Moreover, information about time is used to estimate the learner's skill mastery, as in [133]. As for scaffolding, in education, it refers to breaking up new concepts so they can be learned more easily. Hence, taking into account scaffolding interactions indicates the willingness to integrate the learners' data with information about how or when they use scaffolding materials during their learning process (e.g. [169]). Finally, considering the learners' attempts means monitoring their actions between two consecutive time steps, i.e. determining the knowledge state in a time step also through the attempts and wins/fails ratios that have occurred (e.g. [112]).

Knowledge Representation

As categories to define the forms for the representation of knowledge, we referred to the taxonomy introduced by von Rueden et al. [7] (see Table 5.1). Here we describe only the forms of knowledge representation found in the papers examined in the review.

In most cases, *algebraic equations* are functions to express a mathematical relationship between the variables and constants used to model the problem. Sometimes, the term algebraic constraints are more appropriate because the knowledge is represented through inequalities to determine a feasible set of values (e.g. [162]).

Simulation results is used to describe the numerical outcome of a computer simulation, intended as an approximate imitation of the behavior of a real-world process. There are two recurring forms in the analyzed papers. Firstly, embedding techniques (often pre-training) to obtain more informative representations from the data. Secondly, the use of attention mechanisms in neural networks. For example, in [135], the domain information on the exercises is integrated through two simulation processes: a pre-training embedding of their texts to gain semantic knowledge and an attention mechanism to model the relations between the items.

Knowledge graph is a common form of knowledge representation. A graph is a pair (V, E), where V is the set of its vertices (or nodes), which usually describe

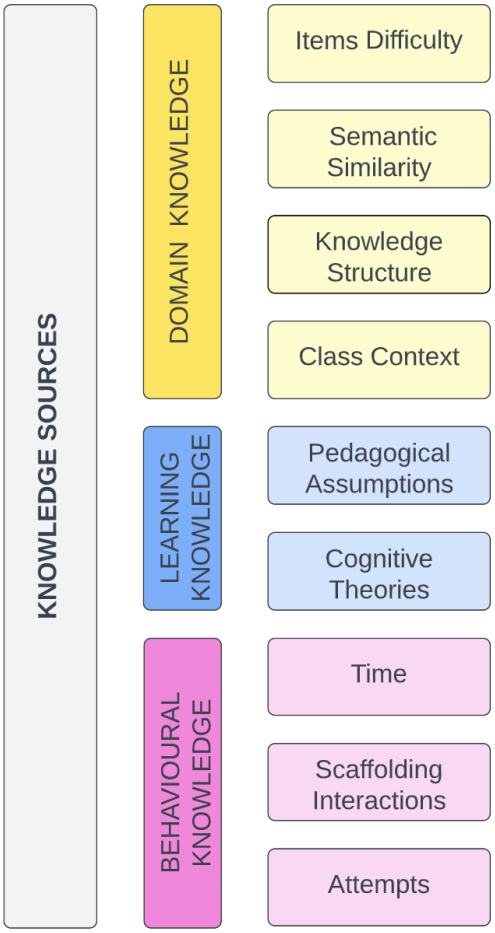

Figure 5.2: **Taxonomy of knowledge sources for the Knowledge Tracing Problem**. There are three main classes of knowledge sources for which we have identified some subclasses.

concepts, and E is the set of its edges, i.e. the abstract relations among them. A common knowledge graph within the KT is the Q-matrix, a binary matrix showing the relationship between test items and latent or underlying attributes or concepts [171]. It can be provided by an educational expert (e.g. [150]) or estimated directly in the embedding layer by exploiting Graph Neural Network (GNN) that extends existing neural network methods for processing the data represented in graph domains (e.g. [146]).

Another knowledge representation type is *probabilistic relation*. It determines the relations between two or more random variables according to their joint distribution. According to von Ruedend et al. [7], "prior knowledge could be assumptions on the conditional independence or the correlation structure of random vari-

ables or even a full description of the joint probability distributions". This form of knowledge representation is the milestone of Bayesian network models [106], very popular as KT techniques.

We add a new class of knowledge representation, named *other data*. There are some cases in which the integrative knowledge source is expressed directly by the collection of additional data to those usually considered in the KT problem (e.g. [112]. As can be seen in Table 5.2, this is quite common when we aim to integrate information on the learner's behavior during learning.

As a final remark to the results concerning knowledge representation, we highlight that there are four classes in the IML taxonomy (see Table 5.1) never used in our qualitative analysis: differential equations, spatial invariances, logic rules and human feedback.

Knowledge Integration

As for knowledge integration, we found three of the four steps of the ML pipeline [122] in the qualitative analysis of the papers.

Integrating prior knowledge sources in *training data* is intended as acting on the information provided as input to the model. There are several ways this can happen. Firstly, we mention data augmentation. In [160], for example, Lee et al. define synthetic data based on pedagogical rationales to deal with the complexity of knowledge acquisition. Another common integration practice is embedding the data with a feature engineering process. This process can be either expert-driven, e.g. in [112] the authors define correct, and error rates as new features to model the students' learning ability, or data-driven, e.g. in [143] a pre-training embedding architecture is designed to model the knowledge structure in the domain. Lastly, some papers expand the training dataset with new kinds of data. In [118], the author leverage knowledge in other domains, which can be transferred to the KT's domain (discipline) object. In other papers, the training dataset includes behavioural data obtained while tracking the learners' interactions with the learning system (e.g. [162].

The second step of the ML pipeline where prior knowledge can be integrated is the *hypothesis set*. It can be defined as the set of functions to choose to solve the initial problem. Relying on the notation introduced in section 5.2, the initial problem is estimating $f : \mathcal{X} \to \{0, 1\}$, and the hypothesis set $\mathcal{H}$ is the set of candidate functions among which to choose $g : \mathcal{X} \to [0, 1]$, as a result of the learning algorithm. It may be, for instance, the set of linear functions, the set of neural networks, or the set of logistic functions. Integrating the prior knowledge in the hypothesis set can be intended as bounding the form of the functions included in $\mathcal{H}$. For example, Liu et al. in [162] manage two explicit choices in this direction: they exploit recursive functions in their architecture to handle the knowledge master degree estimation according to constructive learning theories; they define a

graph convolutional network to include latent learning ability estimation influence on the learner's knowledge concepts states.

The last knowledge integration type found in our literature review is in the *learning algorithm*, i.e. how the model updates the parameters which define the functions in $\mathcal{H}$ during the training. In a neural networks-based model, this integration consists of modifying the loss function to force the model to consider a prior knowledge source. In [119], for example, the authors introduced a penalization term in the loss function to handle item difficulty.

It is worth noting that we do not have any models where knowledge integration occurs in the final hypothesis step. This kind of knowledge integration would occur when the output of the machine learning pipeline is validated against existing knowledge.

The distilled taxonomy of knowledge representation and integration for KT is summarized in figure 5.3.

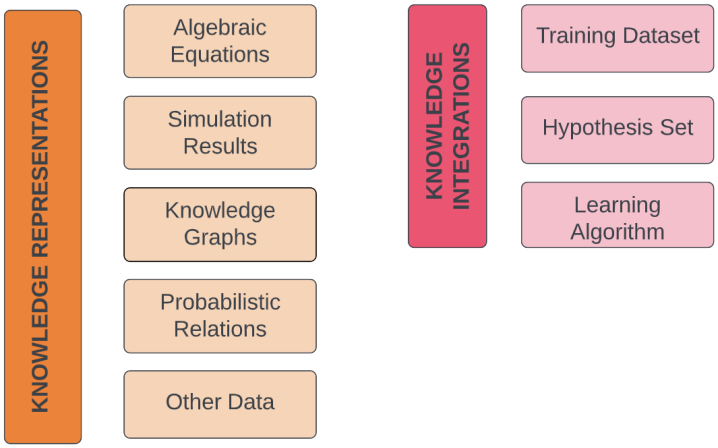

Figure 5.3: **Taxonomy of knowledge representation and integration for the Knowledge Tracing Problem**. Five main forms of knowledge representation and three integration steps in the machine learning pipeline are distilled from our systematic literature survey.

5.4.2 Quantitative analysis

We applied the taxonomy described in the previous subsection to the 53 eligible papers selected for our systematic literature review. One paper can have more than one path for prior knowledge integration. A path is defined by a triad "knowledge source-knowledge representation-knowledge integration". We counted each path separately for the quantitative analysis, identifying 77 paths. For instance, Tong

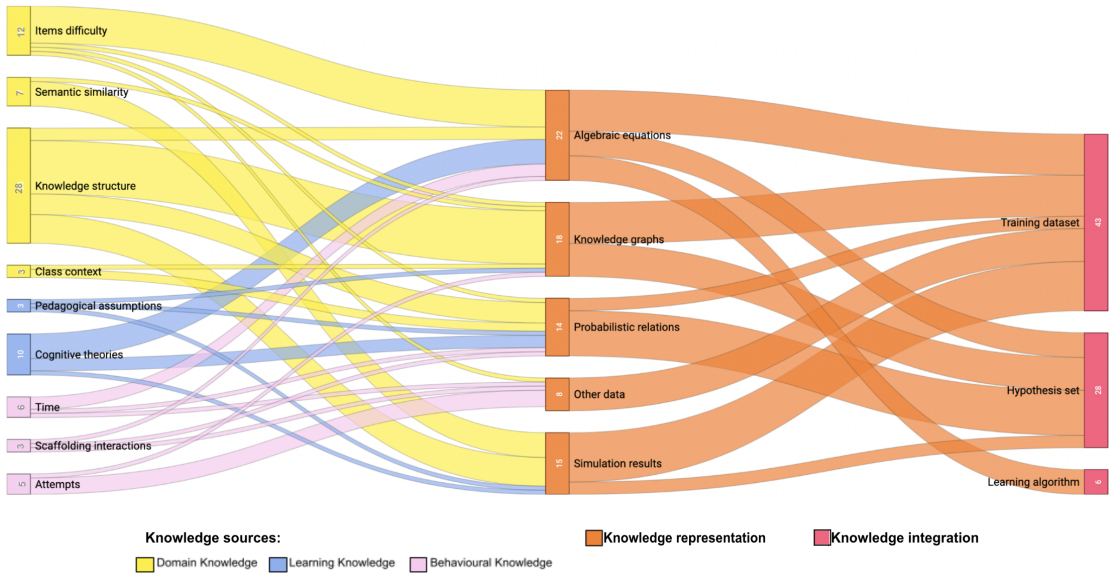

Figure 5.4: **Paths for integrating prior knowledge in a knowledge tracing model**. The nodes on the left represent the spectrum of knowledge sources distilled from our iterative literature survey; we used three colors to distinguish the three main classes (yellow for domain knowledge, light blue for learning knowledge, and pink for behavioral knowledge). Central nodes cover the forms of knowledge representations. The right nodes are for the types of knowledge integration. The paths among nodes represent different approaches to integrating prior knowledge into the ML pipeline.

et al. in [139] introduce 4 ways to integrate exercises learning dependencies in their KT model. As a knowledge source, they exploit knowledge structure from experts, the semantic similarity between items, class context leveraging common behaviours in the class, or class context retrieved by studying the correlation among items answered correctly by many students in the class. They represent the first three knowledge sources using knowledge graphs, while the last is tackled through probabilistic relations. In each path, the integration occurs in the hypothesis set step. Hence, we have 4 triads and considered 4 paths in our labeling counting process.

The quantitative overview of this analysis is summarized through the Sankey diagram in Figure 5.4. This visualization format depicts a flow from one set of nodes to another. The paths connect the elements across the three dimensions of the taxonomy and illustrate the approaches we found in the analyzed papers. The height of each node is proportionate to its absolute frequency, which is also expressed in number. The thickness of the links depends on the absolute frequency with which the path connecting two nodes has been recorded.

Let us point out three main pieces of evidence from the Sankey diagram. Firstly, domain knowledge sources are the most exploited, with 40 paths out of 77 which use them. Specifically, in 28 cases, the prior knowledge to integrate concerns the knowledge structure, i.e. more than a third of the paths.

Secondly, the training dataset is the privileged integration path (43 cases out of 77) for all forms of representation except for probabilistic relations. The latter case is often connected to the choice of the Bayesian network as inspiring architecture for the model. Therefore, it is brought back to the hypothesis set class, i.e. the form chosen for the objective function of the predictor.

Thirdly, we have some cases of exclusive inbound or outbound paths. As expected, when knowledge is represented by exploiting other data, we have a unique outbound path to the training dataset, i.e. these new data sources are used to increase the features considered for a more rich knowledge representation. Moreover, the knowledge integration into the learning algorithm step occurs only with an inbound path from algebraic forms of knowledge representation. All the papers that present this approach introduce a regularization term in the loss function, which is optimized during the model training phase.

In addition, the diagram suggests that some representation and integration approaches (paths from knowledge sources to knowledge representations and paths from knowledge representations to knowledge integrations) are more frequent than others, i.e. some paths are more common. A measure for the relevance of each approach is expressed by computing its conditional probability, i.e. the probability that a path ends in a certain node B knowing that the source node is A. In other words, we aim to interpret the Sankey diagram in Figure 5.4 as a weighted 3-partite direct graph. The weights of the links are defined through conditional probabilities; the three sets of independent nodes correspond to the three dimensions of the taxonomy (knowledge sources, knowledge representations, and knowledge integrations); the paths' direction is from left to right.

We define the conditional probability $p(B = b_j|A = a_i)$ as

$$p(B = b_j|A = a_i) = \frac{f_{ij}}{f_{i\cdot}} \tag{5.1}$$

where A and B are two variables, a_i and b_j stand for one of the modalities respectively of A and B, f_{ij} is the absolute bivariate frequency (i.e. how many times a_i and b_j occur together), and $f_{i\cdot}$ is the absolute marginal frequency (i.e. the number of total occurrences of a_i). The weights determine the relevance of the different approaches.

For instance, we assume A as the variable for knowledge sources and B as the variable for knowledge integration; $A \in \{a_1, a_2, ..., a_9\}$, where a_i denotes the i-th modality for knowledge source from the top in Figure 5.4; similarly, $B \in$

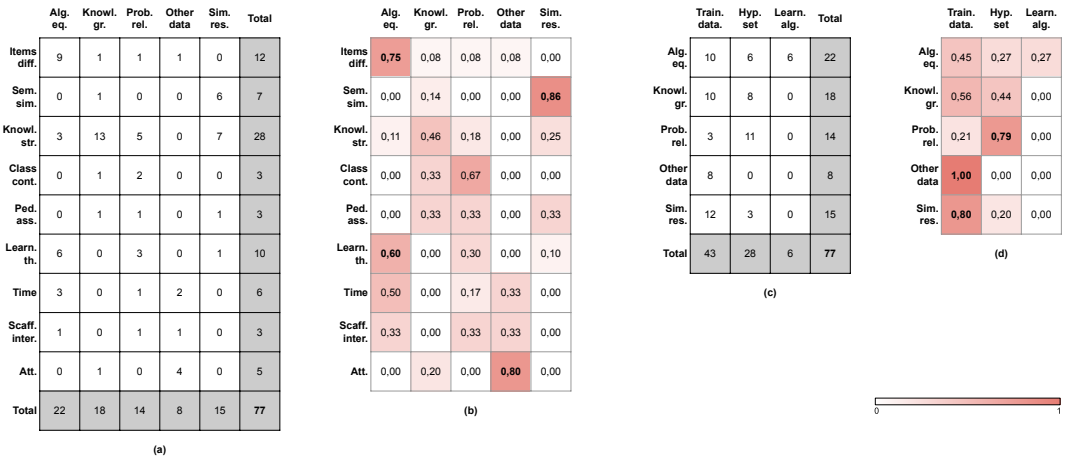

Figure 5.5: **Contingency Tables and Adjacency Matrices of the graph for Informed Machine Learning Taxonomy for Knowledge Tracing**. (a) is the contingency table for all the possible combinations among knowledge sources and knowledge representations modalities (left paths in Figure 5.4). (b) is the adjacency matrix of the left fold of the graph defined on the Sankey diagram in Figure 5.4. (c) and (d) are respectively the contingency table and the adjacency matrix among knowledge representations and knowledge integrations (right fold in Figure 5.4). We have bolded in the adjacency matrices the weights associated with the most relevant approaches in the prior knowledge integration pipeline according to the criteria described in the final part of section 5.4.

$\{b_1, b_2, ..., b_5\}$, b_j is the j-th modality for knowledge representation from the top in the same figure. Hence, we have a_1 = 'items difficulty', b_1 = 'algebraic equations'. The co-occurencies of a_1 and b_1 is $f_{11} = 9$; the marginal frequency for a_1n is $f_{1\cdot} = 12$. Thus, according to definition 5.1, we have $p(B = b_1|A = a_1) = 0.75$. In practice, 75% of the paths outgoing from the item difficulty source are integrated into the model through an algebraic representation.

The tables in Figure 5.5 summarize the results for all the possible combinations which define the paths in the Sankey diagram. In particular, we present the contingency tables for the frequencies of each possible path and the adjacency matrices which describe the graph defined on the Sankey diagram. The elements of the adjacency matrices are the conditional probabilities computed according to definition 5.1 as weights for the links. More details about the tables are provided in the description of the figure.

From the results presented, we can highlight the following relevant approaches among the paths from knowledge sources to knowledge representations: (i) items difficulty - algebraic equations; (ii) semantic similarity - simulation results; (iii)

cognitive theories - algebraic equations; (iv) attempts - other data. As for the paths from knowledge representations to knowledge integration, the relevant approaches are (i) probabilistic relations - hypothesis set; (ii) other data - training dataset (already mentioned as an exclusive outbound path); (iii) semantic similarity - training dataset. The relevant approaches have been identified by applying the following criteria: the weight associated with the path is greater than or equal to 60%; both the marginal frequencies and the joint frequency are greater than 5% of the total number of paths, i.e. $f_{ij}, f_{i\cdot}, f_{\cdot j} \geq 4$.

5.4.3 Application of IML taxonomy for KT to a real case study

Before discussing our results for our RQs, we want to show a case of the application of our taxonomy. We aim to make clear what it means to integrate prior knowledge sources in a KT model, following the full path in one of the reviewed papers. We choose the paper by Wang et al. [141] because it is the only one that considers the most frequent category in each dimension of our taxonomy (knowledge structure as knowledge source, algebraic equations as representation form, and training dataset as integration step).

They present a model based on a Deep Knowledge Tracing (DKT) architecture. In figure 5.6, we present its diagram. The grey part concerns the DKT in its purely data-driven fashion: the student past question-answer sequence feeds an embedding layer; then passes through the sequential layer RNN-based; finally, the feedforward layer predicts the student's future answer to each question.

The red part encodes the prior knowledge injection flows. The first knowledge source consists of question-question relations. These relations are based on their skills and concepts similarity, i.e. two questions are the more similar the more they test the same skills and deal with the same concepts. This knowledge source is represented through a knowledge graph with adjacency matrix A. The adjacency matrix is a square matrix encoding whether pairs of vertices are adjacent or not in the graph. Its integration takes place at the level of training data: the question-question knowledge graph is used as input for the embedding layer of the architecture.

The second knowledge source is the intuition that if a pair of questions requires similar skills or involves similar concepts, students are expected to perform similarly. The knowledge representation here occurs in the form of a regularization term $\mathfrak{L}_{\mathfrak{r}}$, i.e. the algebraic expression $p^T Lp$, where $p(i)$ indicates the probability that the student can answer the i-th question correctly and L is the Laplacian matrix associated to A. The loss function in the model is designed to capture this information, and it is defined by two additional terms: $\mathfrak{L}_{\mathfrak{p}}$, the cross-entropy loss,

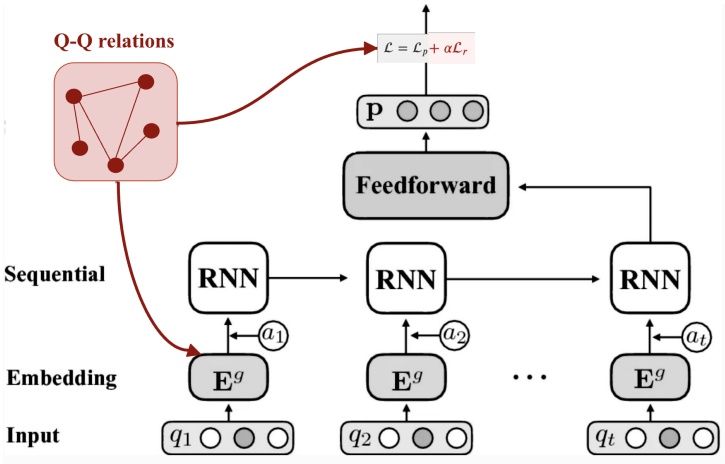

Figure 5.6: **Architecture for DKT with prior knowledge integration**. The figure is adapted from the original paper [141]. In gray there is the standard architecture for DKT, and in red the prior knowledge that is integrated. Questions are given as input to an embedding layer. The RNN-based sequential layer is fed with the embedding output and a_i which encodes whether the student answered the question q_i correctly. Prior knowledge is the similarity relationships between questions. There are two integration paths. Firstly, it is represented through a knowledge graph given as input to the embedding layer. Thus it is integrated into the training data. Secondly, it is included in a regularization term added to the binary cross-entry loss. Thus there is an algebraic re-shaping of the learning algorithm.

and $\mathfrak{L}_r$ the regularization term. Thus the integration occurs at the level of the learning algorithm.

5.5 Discussion

We now discuss our results, pointing out how they answer to our RQs. We keep two main focuses. Firstly, we highlight some remarks on our distilled taxonomy for IML applied to KT. Given our starting point in von Rueden et al.'s taxonomy [7], it is worthwhile to compare them, to show both points of contact and divergence, and assessing the effectiveness of IML for KT. In this way, we can stress strengths and limitations in our proposal, which can be interpreted as possible future research avenues.

Secondly, we draw some considerations from the quantitative analysis. We mostly exploited the quantitative results to emphasize relevant and widespread

IML approaches among the results. This supports our response to the RQs. We present this discussion in three subsections, one for each RQ.

5.5.1 Knowledge sources for Knowledge Tracing

As for the knowledge sources which can be integrated into the ML pipeline to address the KT problem (**RQ4.1**), we distilled a two-level taxonomy which is schematized in Figure 5.2. Comparing this to the taxonomy for the IML, there is a basic difference in the type of labels that have been searched. In our case, we have identified a label for each type of content or information that is integrated, e.g. under the class "domain knowledge" we have four types of content taken as prior knowledge (i.e. item difficulty, semantic similarity, knowledge structure, and class context). von Rueden et al., on the other hand, defined the labels for the knowledge sources by identifying who holds the prior knowledge, i.e. a scientific theory, a human heritage, or the experts in a specific field.

The strength of our choice is the higher granularity and, thus, its descriptive power. The list of identified classes represents a reference of valuable knowledge sources which may integrate student performance data in KT tools. They can be considered factors that influence the KT, improving the models' performances; hence, researcher may consider them while developing their models. Also, they can be seen as elements which enhance the models' explainability and interpretability; that is, they are factors with a high semantic load, making it easier to attribute meaning to the weights or components of the architectures of the designed models (explainability) or favoring the identification of causal relationships between the input and the output of the models (interpretability). In our opinion, this is the first possible line of research that has not yet been sufficiently explored.

On the other hand, the drawback of this improvement in granularity is a loss of generality. Our taxonomy is closely linked to the educational context, with specific reference to the KT problem, while the one for IML applies to domains that may be very different. Furthermore, our taxonomy for knowledge sources is non-exhaustive because it refers to the types of prior knowledge encountered in this literature survey. This does not exclude that there may be other types of relevant content or information to consider. Our result outlines a picture of what exists in the state of the art and can be taken as a starting point subject to updating as a result of new research.

The quantitative analysis highlights one main point: the predominance of domain knowledge as the prior source. Within this class, some sub-classes are more exploited than others: knowledge structure is highly considered, while little attention is paid to the class context. The high consideration of the knowledge structure may be due to its easy availability. In fact, in ITS and AEHS it is often necessary

to provide this type of structure to organize the contents of the course. At the same time, the lack of consideration of the class context is not surprising. Most KT systems are applied to online asynchronous learning contexts, where the class context is non-existent or has little influence. However, it would be interesting, also in light of the teaching experiences that have characterized the recent years of the COVID-19 pandemic, to investigate if and how these systems can be integrated into mixed schooling contexts and estimate which is in this setting the weight of the class context.

As regards other knowledge sources, there are few attempts to consider learning theories, perhaps also due to the difficulty of modeling and representing this kind of knowledge. Most papers in which this occurs refer to cognitive science models of learning curves or forgetting, which are easily representable through algebraic equations. However there are many other psychological and cognitive factors studied in the literature. For example, a little-considered aspect is the influence of emotions on learning, which has great relevance according to educational experts [172]. This is a further gap to explore in IML for KT.

5.5.2 Knowledge representations for Knowledge Tracing

How the prior knowledge is represented (**RQ4.2**) to attain its integration into the ML pipeline is depicted by the second dimension of our taxonomy (see the left part in the schema of Figure 5.3). As already mentioned, we have a subset of the labels used by von Rueden et al. (see Table 5.1), and we add a new label, i.e. other data. The knowledge representations forms never met in our literature survey are differential equations, spatial invariances, logic rules, and human feedback. The first three forms of representation fit better to fields where mathematical modeling of a phenomenon is among the best strategies for its description and study. In the literature survey by von Rueden et al., neither of them is used to represent expert knowledge sources, which is the case for almost all of our knowledge source labels (except semantic similarity, which they include in the class of world or general knowledge). Hence, it is not surprising that they are missing.

On the other hand, we expected human feedback, as defined by von Rueden et al., among the knowledge representation forms of our taxonomy. In the IML taxonomy, human feedback "refers to technologies that transform knowledge via direct interfaces between users and machines. [...] Typical modalities include the keyboard, mouse, and touchscreen, followed by speech and computer vision, e.g., tracking devices for motion capturing. [...] This often occurs in areas of reinforcement learning, or interactive learning combined with visual analytics" [7]. In other words, human feedback, as knowledge representation form, occurs when the human user intuitively and informally expresses a preference or a relevant

opinion concerning the output of the automatic model and this is used to enhance the model's performance.

We envision the human feedback representation as useful for personalized learning tools, where KT is the base to suggest to students resources based on their individual needs, and content which is predicted to be too easy or too hard can be skipped or delayed. For instance, camera devices can be exploited to integrate the learner's emotions during the learning process, i.e. the learner facial expressions are assumed in the form of informal human feedback [173]. Another example concerns the interactions among peers in face-to-face lessons, which could be detected by recording audio or asking for explicit feedback from the teacher in the classroom. The main obstacle to this information is the technological equipment normally available to monitor learning, which is connected to a well-known problem of multimodal learning analytics [174]. In other words, this representation also relies on the hardware technology's availability, which affects its effective use. Both examples refer to knowledge sources we have already stated as underconsidered in the papers selected for this systematic literature review. This could justify why human feedback as a knowledge representation form is unused. However, in dealing with a research question about which learning theories may be integrated into KT models, there is an issue on which forms of representations fit better. We believe that an informal representation through human feedback could have an interesting role here to be investigated.

The quantitative analysis has highlighted that the form of representation is often linked to the type of knowledge source: (i) items difficulty and cognitive theories are often represented in algebraic form; (ii) knowledge structure is almost represented through graphs; (iii) the semantic aspects are represented through simulation results (usually intended as embedding layers, see section 5.4, the subsection on knowledge representation). Except for the representation form "other data", which is used in only 8 out of 77 paths, the other modalities are distributed evenly. This, in our view, reinforces the need to investigate alternative forms of representation to valorise all knowledge sources that may be relevant to KT.

5.5.3 Knowledge integration for Knowledge Tracing

Dealing with **RQ4.3**, the labels in our taxonomy are a subset of the one for general IML and are shown in the right part of Figure 5.3. Our literature survey has no cases for integration in the final hypothesis step. According to von Rueden et al. definition, the integration in the final hypothesis step occurs when the output of a learning pipeline is "benchmarked or validated against existing knowledge". This sounds like ensuring trustworthiness and reliability to the output of the models through a comparison with an authoritative apriori knowledge, i.e. a scientific

theory or formal constraints. Such an approach to tackle the KT problem seems unlikely.

The quantitative analysis points out that the training dataset is the privileged step where prior knowledge is integrated into the ML pipeline. In many cases, prior knowledge is used to find effective representations of either the students learning or the items with which they interact. Thus, the tailored representation is a way to arrange differently or augment the training dataset to fully exploit its potential. This trend is not a surprise because it can be led to a characteristic of the phenomenon under study. Specifically, the KT problem is strongly connected with assessing students' learning.

Referring to Højgaard [175], assessment is modeled as a three-step process: characterizing, identifying, and judging. It is impossible to measure students learning directly; it is necessary to characterize what you are looking for, identify the extent to which it is present in the situations involved in the assessment, and then judge the identified. In other words, when dealing with the assessment of students learning, there is an intrinsic problem of identifying some indicators that need to be interpreted in some way. Automating this process, which is usually managed by the teacher, means integrating it into the model. Characterization and identification precede the judgment phase and, in some way, are the premises on which the judgment can be formulated. KT models, according to the definition we presented in section 5.2, automate the learning judgment phase and use it to predict students' performances on new items. Therefore, studying adequate representations becomes the way to manage the preliminary operation of characterization and identification. It foregoes the model's training, which is oriented to learning how to judge, thus involving mainly the training dataset step.

The quantitative analysis has also brought out three main approaches to knowledge integration based on the knowledge representation kind: (i) the probabilistic relations form is often integrated into the hypothesis set step because probabilistic reasoning is often handled with Bayesian networks (chosen as the form for the objective function in the ML pipeline); (ii) when the prior knowledge is represented through new data, this is always integrated into the training dataset, becoming an extra input source for the model; (iii) simulation results as knowledge representation form is almost integrated into the training dataset. This last point is in line with two observations already stated in this paper: simulation results often occur as embedding layers, thus connected to the problem of representing the input for the models properly; the representation problem is a necessary pre-training phase which enable the model to learn how to predict future learners' performances, thus it is handled in the first phase of the ML pipeline.

5.6 Chapter Conclusion

To conclude, we summarise the main findings of our systematic literature review and some final remarks.

To answer the three RQs on integrating prior knowledge in KT models, we obtained a three-dimensional taxonomy (knowledge source, knowledge representation, knowledge integration) as main result of a qualitative analysis (see Figures 5.2 and 5.3). This taxonomy has been benchmarked with the one proposed by von Rueden et al. [7] for the IML, taking into account the specific focus on KT. Through a quantitative analysis, some common integration approaches were also identified, which can be deduced interpreting the sankey diagram in Figure 5.4. The analysis displays the state of the art at the moment in which the papers involved in the systematic literature review were selected.

Discussing our results in section 5.5, we have emphasized some gaps in IML for KT which outline future research directions. We summarize them by posing a new set of research questions (NRQs). They are the result either of strengths that we have found in our taxonomy (e.g. its high granularity), of the assumptions we have made to justify why some prior knowledge injection approaches are more widespread than others, or of some gaps with respect to the literature on learning theories (e.g. the neglect of emotional aspects on KT). In this sense, they represent open issues to be investigated and verified. We formulate six questions:

NRQ1 How the integrated prior knowledge sources impact in terms of explainability and interpretability of KT models?

NRQ2 Which prior knowledge sources were not considered in the papers selected for the systematic literature review and could expand the proposed taxonomy?

NRQ3 To what extent KT can be applied in contexts that include face-to-face teaching?

NRQ4 Which role does the class context play as a prior knowledge source in face-to-face (or mixed) teaching settings?

NRQ5 Which cognitive, psychological, or pedagogical theories have relevance in KT (e.g. theory of emotions impact on learning)?

NRQ6 What forms of representation can be used to integrate these theories? (e.g. can we exploit the under-considered human feedback form for knowledge representation?)

Furthermore, we want to stress that despite the selected papers present hybrid machine learning models, most approaches to KT are still purely data-driven.

However, all the papers considered in this systematic literature review claim that their results are comparable to or better than those of traditional ML methods. This encourages further research in this direction. The three RQs posed in section 5.3 can be a trace for researchers to identify which prior knowledge sources should be considered, how to represent them, and where to integrate them during model development. Our taxonomy can be a tool to use in the exploratory phase to determine what to consider. Moreover, the good performances achieved by these models can be evaluated with respect to the bias issues that characterize AI applications in education [176]. There are different levels at which bias may affect the ML pipeline, e.g. in the data collection process, the data annotation step, the learning algorithm choice, or the performance metrics selection. Integrating prior knowledge can be a new source of bias or, conversely, act as a mitigating effect. This was out of the scope of this work, but future research could investigate the bias challenges of IML for KT.

As a final remark, we point out that in this study we didn't consider the implication for practice, i.e. how the results obtained can support teaching and learning. This was outside the aim of the systematic literature review, whose focus was more methodological. However, both to validate the utility of these hybrid machine learning approaches to enhance personalized learning and to investigate some aspects proposed in the NRQs, i.e. models' interpretability or the use of human feedback, it is important to develop research focused on the implication for practice. We explicitly mention the link with the interpretability of the model or the use of human feedback, because these are aspects that directly involve teachers and learners; therefore, a global study of the benefits on teaching and learning is needed.

The chapter makes a significant contribution to achieving Goal 3 of this thesis. Specifically, this study has outlined the framework for a new form of student encoding for the initial case study presented in Chapter 1. In the next chapter, we will introduce this new encoding strategy, aiming to enhance the previous predictive models for low achievement.

Chapter 6

Students' Low Achievement: generalizability via IML

In this chapter, I introduce a novel extension of the predictive models discussed in Chapter 2. The aim is to assess whether a new form of student encoding, designed according to the Informed Machine Learning (IML) framework introduced in Chapter 5, can enhance the performance of models for predicting the risk of low academic achievement. Specifically, I will focus on how this approach can improve the models' predictive capabilities, considering both performance metrics (such as accuracy and recall) and explainability.

The new form of student representation employs basic graph theory techniques. It was developed in collaboration with Dr. Balzan, Dr. Ebli, Prof. Gabbrielli, and Dr. Zingaro. We first introduced this concept in a position paper [9], presented at a workshop on Responsible Knowledge Discovery in Education (RKDE 2023). This work was further expanded in Dr. Ebli's master's thesis, where I served as a co-advisor. In the thesis, we presented various graph-based techniques evaluated through different clustering methods to assess their effectiveness in characterizing diverse student groups. Here, I will summarize the essential aspects of this work relevant to the advancement of my thesis, introducing the most promising graph-based encoding. Moreover, we provide a semantic interpretation of the features that have been identified as significant for characterizing the graphs and clusters of student encodings. By doing so, we gain deeper insights into the underlying factors that influence learning processes and outcomes.

Furthermore, I will evaluate the impact of this enriched representation on both the predictive power and the explainability of the Random Forest model previously introduced. By comparing these results with prior approaches, the research aims to demonstrate how integrating diverse knowledge sources can lead to more effi-

cient and comprehensible predictive models in educational settings. This chapter, therefore, not only advances the theoretical understanding of student representation but also offers practical implications for the development of more effective AI tools in education.

The chapter concludes with an epistemological reflection, focusing on the knowledge sources utilized, how this knowledge is processed, and the type of knowledge generated when employing these predictive models. The chapter's case study exemplifies a methodological framework that we introduce to address some limitations identified in previous chapters when developing predictive models in educational settings.

6.1 Background and Motivation

The work presented thus far confirms the potential of artificial intelligence, particularly machine learning models, to enhance the education system [177, 178]. Previous chapters have introduced research studies on predictive models for the risks of low student achievement and academic dropout, as well as examined the literature on knowledge tracing. These efforts provide concrete examples of the success of AI-based predictive models in various educational tasks.

It is well-documented in the literature, supported by our cases, that sub-symbolic models often outperform standard machine learning techniques. However, despite their ability to handle large volumes of data and recognize intricate patterns, sub-symbolic ML models struggle to offer detailed explanations for their predictions [179]. This limitation significantly impacts stakeholders, including educators and policymakers, by hampering their understanding of the rationale behind AI decisions, which ultimately leads to reduced adoption rates [180, 181].

Within this context, we would like to refocus our attention on the case study regarding the risk of low achievement. To frame the contribution of this chapter, we would like to highlight a few crucial points in the development of this thesis, which will be revisited later:

1. In Chapter 2, when we introduced the initial models for predicting the risk of low achievement in mathematics, we faced the challenge of generalizing the use of these models to make them applicable across different cohorts of students. This led us to work on engineering the available dataset, specifically the INVALSI standardized test dataset, to create a student representation that spans various cohorts. Our initial approach involved encoding student learning through a concatenation of scores related to different thematic areas or skills in mathematics. While the results were satisfactory, there remains room for improvement.

2. In the same chapter, a preliminary study on the explainability of the Random Forest models was introduced through the analysis of feature importance, which can be supplemented by other post-hoc explainability techniques introduced later in Chapter 4, particularly SHAP. The preliminary study indicated that the most important features for determining the model's predictions include certain socio-economic-demographic context features and the overall test score.

3. The process of student encoding has, in fact, been an initial, somewhat unconscious approach to Informed Machine Learning (IML). The feature extraction process integrated a domain knowledge source, namely the taxonomy used by mathematics education experts for constructing and classifying the items that comprise the tests. This taxonomy considers various dimensions: areas of knowledge, specific skills, and macro-processes (see Section 2.4). However, a simplification was made that overlooks the relationships that can exist between these different dimensions of mathematical learning. The literature has shown that students who struggle to achieve the expected outcomes often have difficulties in making connections between different areas of mathematical knowledge or competency [182, 183].

In light of these three points, we aim to investigate alternative strategies that can enhance model exploitation. Specifically, as previously mentioned, we will contribute further to the topic of generalizability across different cohorts by proposing a new form of student encoding based on certain graph theory techniques, which accommodates an interconnected view of the various dimensions of mathematical learning. More details will be provided in the following sections.

The representation introduced here, as will be elaborated further, explicitly utilizes certain metrics used to describe the topological characteristics of a graph. To maintain the interpretability desired by stakeholders, we have sought to provide a possible semantic interpretation of the features used for student encoding, with the support of a group of italian maths teachers.

There are two methodological choices I would like to clarify upfront, as they enable a more informed understanding of this chapter's contribution. First, the experiments conducted in this chapter were performed on a subsample of approximately 2,000 students. This decision was primarily motivated by the computational cost associated with calculating a graph for each student's representation and the recognition that, at this stage, we are more interested in understanding the mechanisms by which knowledge can be included and generated through our analysis, rather than applying the model in a real-world context.

Additionally, we chose to consider only features related to students' mathematical learning in the student encoding, focusing on granular rather than summative

information. This means we excluded data on socio-economic context and composite scores, such as the overall test score or grades assigned by schools. This choice aligns with the specific interest of many stakeholders, including myself as a secondary school teacher, in exploring which teaching and learning aspects are most relevant and connected to the problem of low achievement, and on which schools can directly intervene through instructional and educational actions. Therefore, unlike previous chapters, the main focus here is as follows: an epistemological reflection and a comparison of the use of standard machine learning methods versus Informed Machine Learning (IML) in education for data analysis.

The chapter is organized as follows. Section 6.2 introduces the novel student encoding. A graph-based representation is presented, along with some metrics useful for describing them. Section 6.3 introduces semantics for the selected graph metrics, helping to provide interpretability to the representation through domain expertise. The research questions investigated in this chapter are also explicitly outlined at the end of the section. In Section 6.4, the ML techniques used to conduct the experiment are revisited, with appropriate references to parts of this thesis where they are described more comprehensively. Section 6.5 reports the results in terms of both predictive metrics and the explainability of the models. In Section 6.6, these results are discussed in comparison with the models introduced earlier in Chapter 2. The chapter concludes with Section 6.7, summarizing the chapter's contribution to the overall goals of the thesis.

6.2 Graph-based student encoding

In this section, we introduce an alternative method for student encoding, distinct from that presented in Section 2.4. As previously discussed, the exploration of alternative encoding strategies aligns with our objective of achieving generalizability in the predictive models developed for assessing low achievement risk. This aims to integrate prior knowledge sources from the mathematical education domain, in accordance with the principles of IML.

6.2.1 Methodology for the student graph-based encoding

The proposed encoding utilizes a representation grounded in specific metrics that describe spanning trees associated with each student. The selection of this representation is founded on a systematic research approach encompassing several pivotal steps. Firstly, we posited the hypothesis that accurately capturing a student's learning status in mathematics necessitates consideration of the interconnections between various topics and skills, as delineated in the prior section.

This perspective aligns with the taxonomy introduced in Chapter 5, categorizing this as a form of knowledge source designated as *learning knowledge*, underpinned by pedagogical assumptions drawn from the relevant literature in math education [182, 183].

Secondly, we resolved to encapsulate this knowledge in a graph structure, implemented directly on the training dataset. For each student, we constructed a graph following this procedure:

1. Test items are represented as nodes.
2. Two nodes are connected by a directed link if they share at least one classification dimension (such as area, process, or macro-process).
3. Each link is assigned a weight ranging from 1 to 3, corresponding to the number of classification dimensions shared by the nodes.
4. If both items connected by an edge have been answered correctly, the edge is considered bidirectional.
5. An item representing a correct answer that cannot be linked to any others, according to the aforementioned criteria, appears as an isolated node.
6. Items that have not been answered correctly and lack connections to other items based on step (2) are excluded from the graph.

Figure 6.1 shows an example of student graph encoding, taking a student from the cohort 2013/14.

Thirdly, we computed the spanning tree for each graph. The spanning tree of a connected graph is the minimal subset of edges that forms a tree connecting all the graph's nodes, while avoiding cycles. This representation is designed to eliminate any cycles present in the previous configuration, while retaining the original criteria of edge directionality, weight, and node inclusion. During the spanning tree construction, whenever a node in the graph is explored, all child nodes are incrementally added to the tree. A critical aspect of this process is the strategy for selecting the next node to explore, which prioritizes child nodes based on their difficulty levels. Specifically, the easiest node—defined as the one that has received the highest number of correct answers relative to the reference dataset—is selected. The criterion of item difficulty is a foundational element frequently incorporated into student modeling, as elaborated in Section 5.4.

This methodology generates a cycle-free representation; however, to maintain the integrity of correctness indicators for each answered question, any node representing a correct answer without child nodes is augmented with a fictitious child node. This approach preserves the structural properties of the tree while ensuring

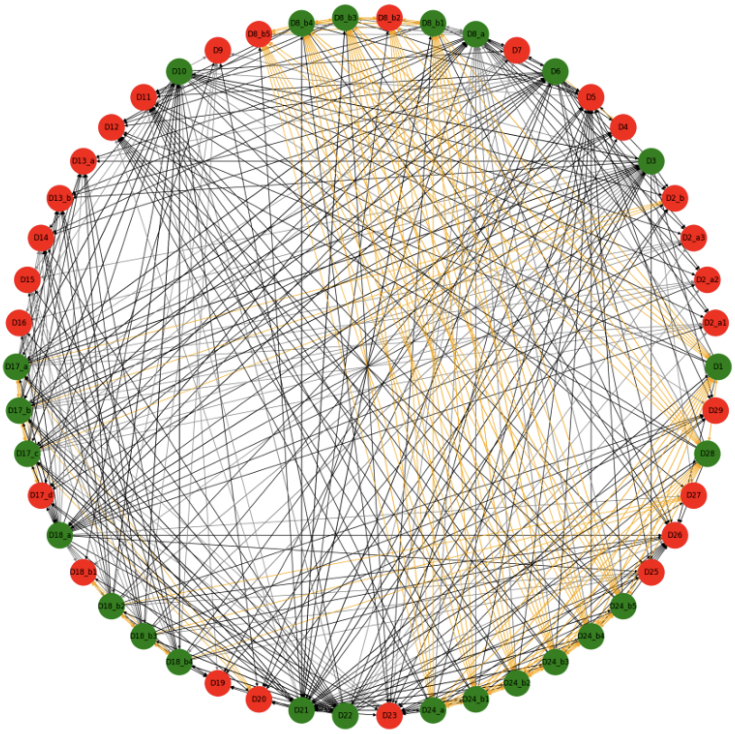

Figure 6.1: Illustration of a student graph, where green nodes signify correct answers and red nodes indicate incorrect answers. The black edges are weighted at 1, the grey edges at 2, and the orange edges at 3.

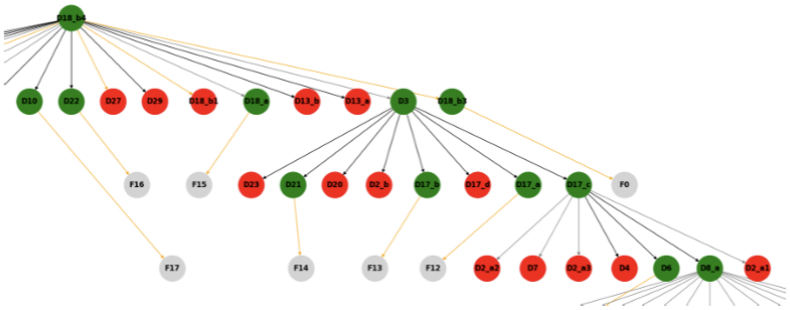

Figure 6.2: Central portion of an example student spanning tree, where green nodes represent correct answers, red nodes denote incorrect answers, and light grey nodes indicate fictitious nodes. The edge weights are as follows: black edges have a weight of 1, grey edges a weight of 2, and orange edges a weight of 3.

that essential information regarding answer correctness is retained. In Figure 6.2, we present a detailed view of the spanning tree representation of a student, with a comprehensive example provided in Figure 6.3.

Following the construction of the spanning trees, we analyzed a set of global metrics pertinent to the proposed representations. The decision to employ metrics arises from the need for effective comparison of different student encodings. The selected metrics, which will be briefly discussed in the next section, facilitate these comparisons and, with the engagement of domain experts, have been utilized to interpret the characteristics that emerge from the spanning tree encodings (see Section 6.3).

6.3 Semantic for the selected graph features

The final representation we consider involves calculating a set of global metrics for the spanning tree representation constructed from the graph associated with individual students. In tables 6.1 and 6.2 we sum up all the metrics used to build this student encoding.

According to our understanding, we aim to propose a potential interpretation for this set of metrics in the context of student learning encoding. Recognizing how to interpret these metrics may yield valuable insights into a student's learning characteristics and comprehension of the subject matter. It is important to note

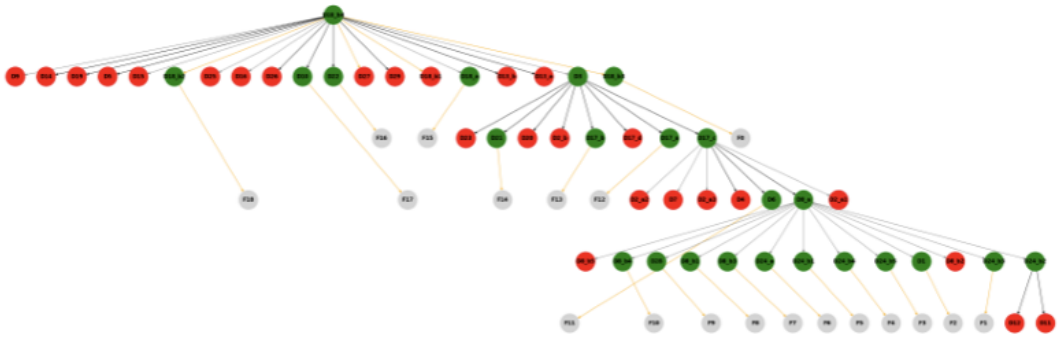

Figure 6.3: Example of a student shallow spanning tree. Colors of nodes and edges are consistent with those in figure 6.2.

that our interpretation is not comprehensive; rather, it represents a foundational perspective developed through collaborative discussions with a team of secondary school mathematics teachers. This proposal serves as a starting point for interdisciplinary research that integrates domain experts into the learning feedback loop. A more detailed exploration of this approach will be presented in the final section of this chapter, where we will introduce the concepts of human-in-the-loop [184, 185], or more appropriately, machine-in-the-loop.

The *average out-degree* serves as an important indicator of the student's ability to connect different mathematical concepts. A higher average out-degree suggests that the student is integrating various dimensions of their knowledge, indicative of a comprehensive understanding of the material. Conversely, a lower out-degree may reveal a fragmented conceptual framework, where the student struggles to link different topics effectively.

The *density* of the graph reflects the overall connectivity of the student's responses. A high density indicates that the student has successfully linked their correct answers to multiple dimensions, pointing towards a robust grasp of the content. In contrast, a low density may signify a prevalence of incorrect responses and limited ability to connect concepts.

The *number of isolated nodes* is critical for identifying specific areas of misunderstanding. Isolated nodes represent topics where the student demonstrates a lack of connections to the broader context, suggesting targeted areas for instructional intervention.

Similarly, the *number of weakly connected components* provides insight into the organization of the student's knowledge. A higher count of these components may indicate an inability to integrate separate topics, highlighting potential gaps in the student's understanding that could be addressed through focused teaching

strategies.

The *S-metric* assesses the robustness of the student's knowledge by evaluating the interconnections within their responses. A larger S-metric indicates a well-structured understanding with key concepts that create strong links throughout the graph, suggesting a more comprehensive learning experience.

In terms of spanning tree-specific metrics, the *size* of the tree, quantified by the sum of the weights of the edges, signifies the extent of the student's engagement with the material. An expansive size suggests that the student has navigated a wide range of related concepts with proficiency, whereas a smaller size may reflect limited engagement or understanding.

The *breadth*, determined by the maximum out-degree of any node, reveals the student's capacity to answer questions across various dimensions. High breadth indicates a well-rounded command of the subject, while low breadth may reflect difficulties in engaging with broader mathematical topics.

The *height* of the spanning tree, representing the longest path from the root to the leaves, offers insight into the complexity of the student's understanding. Limited height may indicate that incorrect answers restrict deeper exploration of content, while significant height could imply successful navigation through complex topics.

Finally, the *load balance* metric assesses the equilibrium between incoming and outgoing edges in the spanning tree. A balanced load suggests that the student effectively connects concepts, while an imbalance may highlight areas where the student is over-relying on certain topics at the expense of others.

After this excursus summarizing some relevant aspects of alternative forms of student encoding, we now shift our focus to the effects of its use on one of the models previously presented in Chapter 2. To facilitate this discussion, we will introduce several research questions that will be addressed through the methodology outlined in the following section.

RQ5.1: What effects does a student encoding designed according to the Informed Machine Learning (IML) paradigm have on model predictive performance?

RQ5.2: How does the integration of a data-driven method with a theory-driven approach alter the explainability of the model?

6.4 Methods

The experiment was conducted on a subset of the INVALSI dataset, specifically focusing on a sample of 1,733 students from the 2013/14 cohort. These students were selected based on their province of residence. This selection was necessary due to the high computational cost associated with generating the graph and the

spanning tree for each student. The primary aim of this exploratory study was to evaluate the potential benefits of employing the Informed Machine Learning (IML) paradigm in addressing the previously examined issue of low achievement risk prediction.

The dataset was processed using the procedures outlined in Section 6.2. Since the focus of this study is to assess which intrinsic characteristics of learning possess greater predictive power concerning the risk of low achievement, only features related to this aspect were considered. Consequently, socio-economic, cultural, or demographic features were excluded from the analysis.

For the modeling phase, a Random Forest algorithm was employed (see Section 2.4). Interpretation of the model was facilitated through techniques previously illustrated in this thesis, specifically using Permutation Feature Importance and SHAP (SHapley Additive exPlanations), applied in both global and local perspectives (see Section 4.3). To provide a comparative benchmark with the student learning encoding presented in Chapter 2, the model was trained and tested on the same subset using the two forms of learning encoding proposed in this thesis.

In contrast to previous experiments, and given the limited sample size, an Oversampling technique was utilized to address the issue of class imbalance within the dataset, which exhibited approximately 80% instances of the False class and 20% of the True class (i.e., Low Achievement). The Synthetic Minority Oversampling Technique (SMOTE) was employed for this purpose. SMOTE functions by generating synthetic examples of the minority class based on a combination of existing instances, effectively balancing the dataset without simply duplicating existing samples [186].

Regarding the experimental setup, it is worth noting that the dataset was split into training and test sets in an 85-15 ratio. This partitioning ensures that the model is adequately trained while retaining a robust test set for evaluation.

6.5 Results

The results of the experiment are summarized in Table 6.3. They highlight the performance of the two student encoding approaches: graph-based encoding and base encoding. For clarity, graph encoding refers to the methodology introduced in this chapter, while base encoding pertains to the method discussed in Chapter 2, limited to features directly related to learning.

These results were obtained using the following set of hyperparameters, determined through grid search: for graph encoding, the parameters were 'max_depth': 11, 'max_features': 0.8, 'max_samples': 0.6, and 'n_estimators': 50; for base encoding, the parameters were 'max_depth': 11, 'max_features': 0.5, 'max_samples': 0.5, and 'n_estimators': 60.

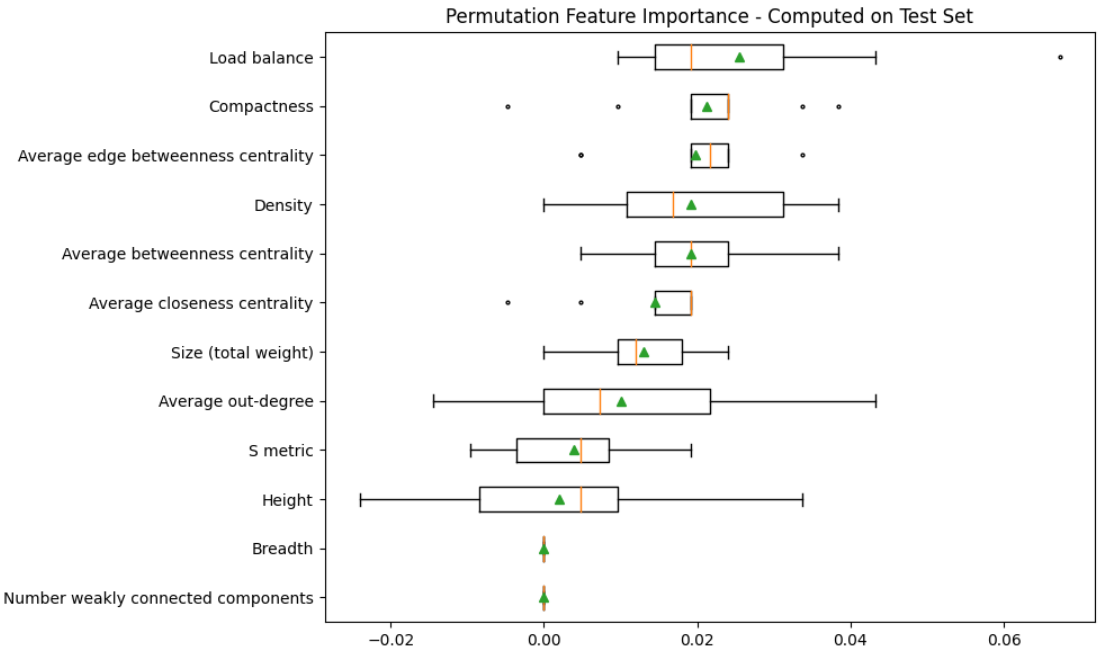

Figure 6.4: **Feature Importance with PFI for RF model trained with graph-based enconding of data.**. The legend for the box-plot is the following: the central rectangle in each line shows the second and the third quartiles together, where the orange line represents the 50% threshold; the green triangle is the average value for the importance of the selected feature on all the decision trees in the RF model; dots represent outliers.

In terms of explainability, we present the analysis performed on the model trained with the graph-based student learning encoding. Figure 6.4 illustrates the Permutation Feature Importance (PFI) analysis, contributing to global post-hoc explainability. Notably, only four features exhibited a median PFI value exceeding 2%, namely Load Balance, Compactness, Average Edge Betweenness Centrality, and Density. Among these, Compactness and Average Edge Betweenness Centrality showed lower variance, indicating less dispersion in their contributions.

Furthermore, two of these features were also identified as important by SHAP, as depicted in Figure 6.5, namely density and load balance. The newly selected features in the top four by SHAP are Average Closeness Centrality and height, highlighted as the most significant features overall. It is also evident from Figure 6.5 that low values of height and density are positively correlated with the risk of low achievement. Conversely, low values of Average Closeness Centrality and Load Balance frequently characterize students who do not experience low achievement.

Another relevant aspect emerges regarding the feature size; cases of low achieve-

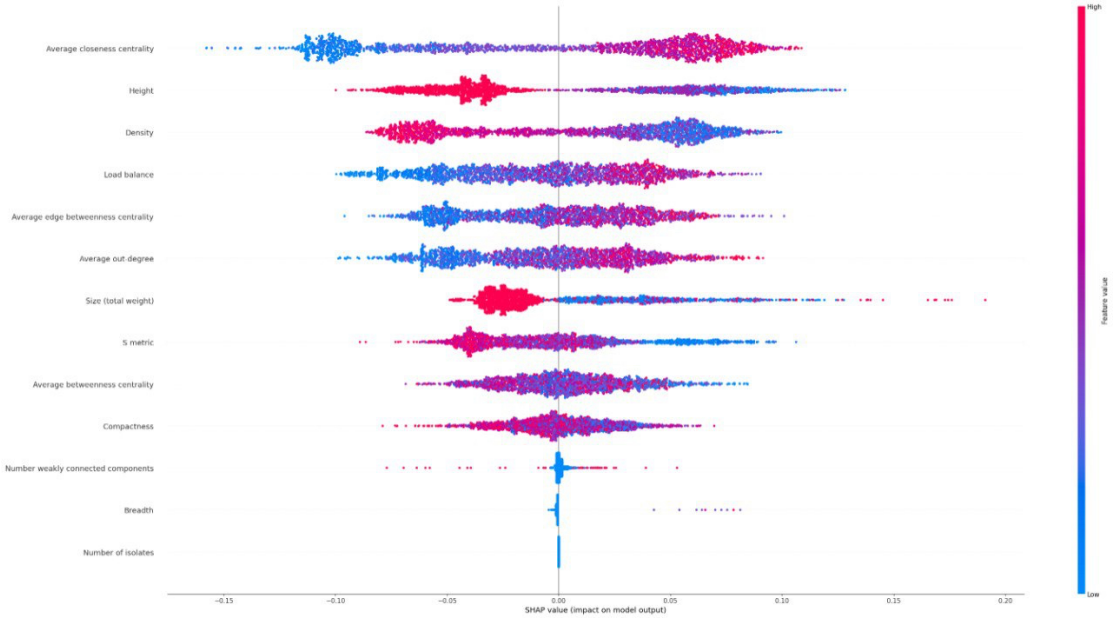

Figure 6.5: **SHAP global explanations for RF model trained with graph-based enconding of data.** The beeswarm plot shows features ordered by average SHAP value. Each dot represents an instance, positioned by SHAP value; colors indicate numeric feature values.

ment are related to low values of this feature, which cluster in a very dense area, indicating that these students are similar to one another in their learning profiles.

For completeness, Figure 6.6 provides an example of SHAP utilized locally on a student to explain the model's prediction.

6.6 Discussion

The results obtained from the experiment provide interesting insights into the efficacy of different student encoding methods —namely, graph encoding versus base encoding—in predicting risks associated with low achievement among students. However, they also come with certain limitations. This discussion will be presented in two subsections: one focusing on the analysis of predictive performance and the other on the analysis of explainability.

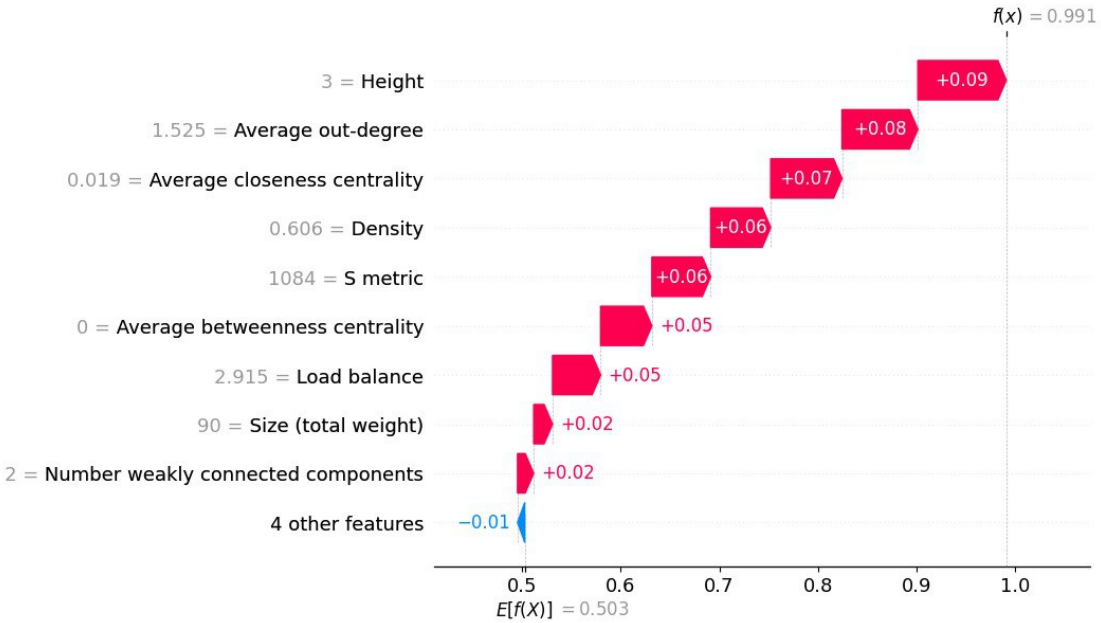

Figure 6.6: **SHAP Local explanations for RF model trained with graph-based enconding of data.**. Each line shows the main features impacting the predicted dropout risk for a student, with bar lengths proportional to their SHAP values. Pink bars indicate features that increase dropout risk, while blue bars indicate features that decrease it. The combined contributions determine the predicted value.

6.6.1 Predictive Performance

The comparison of performance metrics indicates that graph encoding demonstrates a higher recall of 0.60 compared to the base encoding's recall of 0.46. This suggests that the graph-based approach is more effective at identifying students at risk of low achievement, thereby highlighting its potential utility in early intervention strategies. The ability of the graph encoding to capture intricate relationships and interactions among various learning dimensions likely contributes to this observed performance. This result contributes to a positive answer to research question RQ5.1 of this chapter regarding the benefits of integrating a data-driven approach with a knowledge source from the application domain.

However, while these findings are encouraging, there are clear limitations. While both encoding methods achieved reasonably high accuracy, with graph encoding achieving 0.78 and base encoding at 0.82, these values must be viewed in light of the other metrics. The graph encoding's F1-score of 0.48 indicates room for improvement in balancing precision and recall, emphasizing the potential need for further refinement of the feature set or modeling approach to enhance both metrics.

Some of these limitations can be addressed by enhancing the data and computational sources used. The dataset used for this experiment is relatively small, comprising only 1,733 students, which could affect the robustness of the results. Furthermore, the exclusion of demographic and socio-economic variables from the feature selection process may overlook significant contextual factors influencing student performance. When comparing the results of base encoding with those presented in Chapter 2, based on a dataset on 706733 students, it becomes evident that while accuracy remains consistent, recall and precision differ by 21 and 16 percentage points, respectively. This discrepancy raises important questions regarding the relevance of contextual factors in learning outcomes.

On the other hand, the new graph-learning encoding demonstrates promising results in recall, falling short by only 7 percentage points with respect to the model introduced in chapter 2, despite the great difference in the dataset size. Therefore, it suggests that with an adequately sized dataset, this gap could potentially diminish further. In summary, graph encoding appears to be more informative than the previous representation in terms of intrinsic learning aspects relevant to low achievement risk, outlining a viable direction for future work.

6.6.2 Explainability

The results from the explainability analysis provide important insights into the features driving the predictions of the model trained with graph-based student learning encoding. The Permutation Feature Importance (PFI) analysis identified

Load Balance, Compactness, Average Edge Betweenness Centrality, and Density as key metrics, with Compactness and Average Edge Betweenness Centrality exhibiting lower variance in their contributions. This suggests a consistent influence of these features on model predictions.

Notably, the SHAP analysis corroborates the significance of Density and Load Balance, while highlighting Average Closeness Centrality and Height as additional critical features. The observations indicate that low values of Height and Density correlate positively with the risk of low achievement, implying that students with less intricate learning connections may face greater challenges. Conversely, lower values of Average Closeness Centrality and Load Balance tend to characterize students who do not experience low achievement, suggesting that these students effectively integrate their knowledge across concepts.

Moreover, the finding regarding Size is particularly striking. Instances of low achievement are associated with low values of this feature, clustered in a dense area, indicating that these students share similar learning characteristics. This underscores the importance of understanding the intricacies of student profiles, as a nuanced approach to analyzing these metrics could potentially inform targeted interventions to support student learning.

Overall, these results emphasize the potential of graph-based student learning encoding to provide valuable insights into the factors that influence student achievement. While the findings offer a promising direction for future research, it is essential to approach these interpretations with caution, acknowledging the complexity of educational contexts and the need for ongoing validation of these metrics. Thus, it is essential to consolidate the interpretations that can be assigned to these metrics within the contextual domain. This appears to necessitate collaboration with domain experts and support from a qualitative analysis of the results, as will be further elaborated in the concluding chapter of this thesis.

Moreover, regarding the use of local explainability techniques, as illustrated in Figure 6.6, an ablative study would be beneficial, comparing various prediction cases—some correct and some incorrect—to gain a better understanding of the areas for improvement in what the model has learned. This outlines other directions for future work.

6.7 Chapter Conclusion

This chapter has introduced a novel approach to student learning encoding through the implementation of graph-based encoding methods. By comparing this new encoding with the base encoding presented in Chapter 2, the findings provide valuable insights into the predictive performance and explainability of the models employed. The results suggest that graph encoding is more effective in

identifying students at risk of low achievement, thereby highlighting its potential utility in early intervention strategies.

In response to research question RQ5.1, which investigates the effects of a student encoding designed according to the Informed Machine Learning (IML) paradigm on model predictive performance, the results indicate that graph encoding achieves a higher recall compared to base encoding. Despite some limitations due to the reduced dataset size and the exclusion of demographic and socio-economic features, the graph encoding demonstrates a promising ability to capture intricate relationships among learning dimensions that are crucial for effective prediction.

Regarding research question RQ5.2, which explores how integrating a data-driven method with a theory-driven approach alters the explainability of the model, the analysis reveals that key metrics such as Density and Load Balance contribute significantly to the model's predictions. The identification of these features through both PFI and SHAP confirms their relevance in understanding student performance. Interpreting Density and Load Balance in relation to our domain context, both metrics appear to be associated with the level of interconnection between different areas and mathematical skills. However, to confirm this interpretation and enhance the transparency and reliability of the explainability analysis, a dedicated sector study is required.

This work contributes significantly to the overarching goals of the thesis. First, the exploration of explainability through metrics relevant to student learning encoding directly supports the goal G2, emphasizing the importance of transparent and understandable models in educational settings. By highlighting learning key features that influence predictions, this chapter contributes to fostering trust among stakeholders and enhancing user comprehension, which are essential for effective implementation of predictive analytics in education.

Additionally, the emphasis on developing robust student representations addresses the issue of *generalizability*. By leveraging graph-based encoding, this chapter lays the groundwork for future applications of predictive models across diverse student cohorts, developing a framework applicable beyond the specific context of the study.

To conclude, this chapter presents a work that culminates in a doctoral journey that has opened new avenues and challenges. This work is not exhaustive and complete; rather, it serves as a foundational step towards delineating fresh perspectives beyond traditional purely data-driven and conventional machine learning approaches. This exploration prompts reflections on the value of mixed methods and the role of human expertise in the implementation of AI solutions in education. These themes will be explored in greater detail in the conclusion of this thesis.

Metrics for Directed Graphs		
Metric	**Definition**	**Formal Definition**
Average Out-degree	It calculates the average number of edges that leave each node in the graph.	$\frac{\sum_{i=1}^{N} \text{Out-degree}(v_i)}{N}$
Density	It indicates how dense a graph is by comparing the actual number of edges to the maximum possible number of edges.	$\frac{2E}{N(N-1)}$
Number of Isolated Nodes	It counts the number of nodes in the graph that have no connections to any other nodes.	$\lvert\{v \in V : \text{Degree}(v) = 0\}\rvert$
Number of Weakly Connected Components	It calculates the number of groups of nodes where each group is connected by paths but may not connect to other groups.	
Average Betweenness Centrality	It measures the average fraction of shortest paths that pass through a given node.	Average Betweenness Centrality $= \frac{1}{N(N-1)} \sum_{s \neq t} \sum_{v \in V}$
Average Closeness Centrality	It calculates the mean reciprocal of the shortest path distances from a node to all other reachable nodes.	Average Closeness Centrality $= \frac{1}{\lvert V\rvert - 1} \sum_{u \in V} \frac{1}{d(v,u)}$
Average Edge Betweenness Centrality	It evaluates the average fraction of shortest paths that pass through each edge.	Average Edge Betweenness $= \frac{1}{N} \sum_{u \neq v} \frac{\sigma_{uv}(e)}{\sigma_{uv}}$
S-metric	It assesses the robustness of a graph by evaluating connections between nodes.	$S = \sum_{(v_i,v_j) \in E} \text{degree}(v_i) \cdot \text{degree}(v_j)$

Table 6.1: Summary of Metrics for Directed Graphs

Metrics for Spanning Trees		
Metric	**Definition**	**Formal Definition**
Size	It is the sum of the weights of the edges in the spanning tree.	$\sum_{(v_i,v_j)\in E} \text{Weight}(v_i, v_j)$
Breadth	It measures the maximum number of outgoing edges from any single node in the spanning tree.	$\max_{v\in V} \text{Out-degree}(v)$
Height	It calculates the longest path from the root of the tree to its leaves.	
Load Balance	It assesses the average difference between incoming and outgoing edges, accounting for edge weights.	$\frac{1}{\lvert V\rvert}\sum_{v\in V}(\text{In-degree}(v) - \text{Out-degree}(v))$

Table 6.2: Summary of Metrics for Spanning Trees

Metric	**Graph Encoding**	**Base Encoding**
Recall	0.60	0.46
Precision	0.40	0.46
Accuracy	0.78	0.82
F1-Score	0.48	0.46

Table 6.3: Performance Metrics for Different Student Encoding Approaches

Chapter 7

Conclusion

7.1 A two-steps methodology

In this section, we will introduce a framework for the implementation of AI systems to support decision-making processes, with a specific focus on the educational context. This framework is general in nature; however, we will present it using our foundational case study, which addresses the prediction of low achievement risk among students. The framework was first introduced in the position paper referenced earlier [9].

The goal of this framework is to provide a structured approach that integrates machine learning and data-driven methodologies with educational expertise. By doing so, we aim to enhance the effectiveness of AI solutions in addressing the challenges faced in educational settings. Through our case study, we will illustrate how this framework can be operationalized to support stakeholders in making informed decisions that contribute to improved educational outcomes.

In our proposal, achieving inherently explicable and successful AI depends on distinguishing and sequencing its exploratory and exploitative uses, where epistemic (exploratory) processes precede pragmatic play (exploitation) [187]. The exploratory phase, characterized as epistemic, can be viewed as a cycle of refinement that continues until a state of saturation is reached. In other words, knowledge is input into the system through the data utilized during the training phase and the integration of additional knowledge sources. Through machine learning techniques, knowledge is produced by processing this input; however, this knowledge must be interpreted, verified, and compared.

The analysis of results, both predictive and explanatory, can suggest areas for improvement. The nature of these modifications can vary: for instance, it may be necessary to augment the dataset if it is not sufficiently representative; to

consider other pedagogical, cognitive, or sociological theories during the machine learning phases; or to modify the architecture of the chosen model, among other considerations. The quality of the knowledge obtained at the end of each cycle is evaluated through domain expertise, underscoring the need to incorporate human expertise into the loop. Only when the cycle reaches saturation, indicating that there are no longer significant changes in the knowledge generated by the predictive model, can we transition to the actual implementation of the developed tool in practice.

Following this exploratory phase, we transition into the exploitative, pragmatic phase. This phase emphasizes the application of the knowledge gained during exploration to implement actionable strategies aimed at improving educational outcomes. In this context, exploitation involves utilizing AI systems to make informed decisions, create personalized learning experiences, and ultimately refine teaching practices based on the insights derived from the exploratory phase.

Thus, a comprehensive understanding of the educational system is not merely a preliminary step; it is a prerequisite guided by the epistemic cycle. By emphasizing the epistemic phase before entering the pragmatic stage, we ensure that the AI solutions deployed are informed by a rich context, which enhances their effectiveness and acceptability in educational settings.

Here, we primarily focus on the epistemic phase, through the main case study of this thesis. Our goal is not (only) to improve the data gathering process used by INVALSI (left in figure 7.1); instead, we follow an approach similar to [27],using the dataset as a valuable source of information about students' learning states and corresponding achievement rates. In Figure 7.1, we illustrate the general schema of our proposal, applied to the casestudy presented in chapters 2 and 6, based on the dataset extracted from the INVALSI test administration.

In a nutshell, the INVALSI dataset has been engineered to create a form of student encoding that has been utilized for model training. The outputs are then subjected to an explainability analysis. At this stage, the results are compared with a domain expert (researchers or industry professionals), who can highlight potential critical points. This represents an iteration of the epistemic cycle, corresponding to the work presented in Chapter 2. The outcome of this phase suggests that it is appropriate to enrich the dataset with a pedagogical hypothesis derived from mathematical education, which takes into account the interconnections between different mathematical skills and knowledge. A second iteration of the cycle then begins, where the representation is adjusted with the graph-based hypothesis (of which the previous chapter represents a naive attempt at application).

In this framework, the role of the domain expert is crucial, leading us to refer to the paradigm known in the literature as human-in-the-loop [184, 185]. However, we would also like to introduce another perspective. The starting point of the

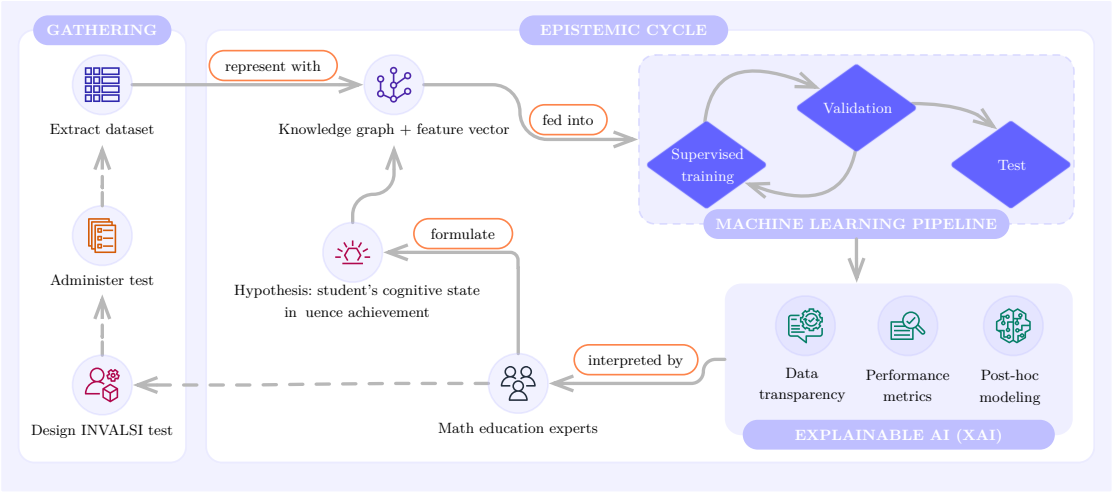

Figure 7.1: The epistemic cycle involving the ML pipeline and the XAI module, leading tomodel scaling and use for pragmatic purposes after achieving consensus among domainexperts.

entire framework is a data collection process centered around the human element, both because the educational context is strongly human-centered, and because the very design of the data collection derives from pedagogical and cognitive hypotheses upon which the assessment tests are built. The machine learning process is also mediated by human expertise; furthermore, the implications of this complete process (including the exploitation phase) return to the real educational context as a decision-support tool for relevant stakeholders [55], rather than as a replacement for their role. Thus, we could say that this thoughtfully designed process embodies the machine-in-the-loop concept, continuing to acknowledge the central and irreplaceable role of individuals at various levels, either in the field or in policy, who are engaged in education.

7.2 Final Remarks

This thesis encompasses a coherent journey through the application of Data Science and Artificial Intelligence (AI) in educational contexts, culminating in the development of predictive models aimed at improving student outcomes. Each chapter has contributed uniquely to the overarching themes of transferability, explainability, and generalizability, effectively shaping the road to addressing the associated challenges.

Chapter 2 presented the initial case study focused on predicting the risk of low achievement among students using machine learning methodologies. This chapter

contributed significantly to the theme of generalizability by establishing a foundation for applying predictive models across different cohorts, demonstrating how these models can adapt to diverse educational contexts.

In Chapter 3, the exploration shifted toward academic dropout prediction. By assessing the effectiveness of various machine learning models, particularly the Feature Tokenizer Transformer (FTT) compared to traditional models like Random Forest, this chapter emphasized the theme of transferability. The results illustrated how methodologies can be effectively transferred to different educational settings, fostering a deeper understanding of the factors contributing to student success.

Chapter 4 focused on the explainability of predictive models, analyzing the significance of key features in the context of low achievement and dropout prediction. By employing methods such as Permutation Feature Importance (PFI) and SHAP, this chapter reinforced the relevance of explainability in educational AI applications, contributing directly to the goal of enhancing transparency and understanding among stakeholders.

Chapter 5 provided a systematic literature review that opened up epistemological questions regarding the integration of prior knowledge into knowledge tracing models. This chapter not only laid the groundwork for addressing the challenges of explainability and generalizability in a new way, but also introduced the basis for the two-step methodology framework that underscores the importance of integrating domain expertise into the model-building process (see section 7.1).

Finally, Chapter 6 introduced a novel student learning encoding through graph-based methods, demonstrating its potential benefits in both predictive performance and explainability. The insights gained from this chapter further solidified the contributions to generalizability and explainability by highlighting how well-structured student representations enhance the effectiveness of predictive modeling.

In summary, the contributions of this thesis have collectively advanced our understanding of how AI tools can be effectively implemented in educational settings. The findings highlight the importance of addressing the interrelated themes of transferability, explainability, and generalizability, each contributing to the development of a framework that guides the implementation of AI systems for educational decision-making.

This work represents a foundational step in a doctoral journey that opens new avenues for research and practice, suggesting the need for further exploration beyond traditional data-driven approaches. Moving forward, it invites consideration of mixed methods [188] and emphasizes the central role of human expertise in the application of AI solutions in education.

Bibliography

[1] T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder, "Data mining and machine learning retention models in higher education," *Journal of College Student Retention: Research, Theory & Practice*, p. 1521025120964920, 2020.

[2] OECD, *PISA 2022 Results (Volume I)*. 2023.

[3] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Student low achievement prediction," in *International Conference on Artificial Intelligence in Education*, pp. 737–742, Springer, 2022.

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[5] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[6] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Machine learning for academic dropout risk assessment, with explanations," *IEEE Transactions on Learning Technologies*, in press.

[7] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[8] A. Zanellati, D. Di Mitri, M. Gabbrielli, and O. Levrini, "Hybrid models for knowledge tracing: a systematic literature review," *IEEE Transactions on Learning Technologies*, 2024.

[9] F. Balzan, A. Zanellati, S. P. Zingaro, M. Gabbrielli, *et al.*, "A 2-step methodology for xai in education," in *Proceedings of the 1st International Tutorial*

and Workshop on Responsible Knowledge Discovery in Education, pp. 1–7, 2023.

[10] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, vol. 4, no. 4, pp. 253–278, 1994.

[11] W. Pietsch, "Aspects of theory-ladenness in data-intensive science," *Philosophy of Science*, vol. 82, no. 5, pp. 905–916, 2015.

[12] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[13] Oecd, "Who and where are the low-performing students?," pp. 33–59, 2016.

[14] D. D. Curtis and J. McMillan, "School non-completers: Profiles and initial destinations," 2008.

[15] S. Flisi, V. Goglio, E. C. Meroni, and E. Vera-Toscano, "School-to-work transition of young individuals: what can the elet and neet indicators tell us," *Luxembourg: Publications Office of the European Union, EUR-Scientific and Technical Research Reports*, 2015.

[16] K. L. Alexander, D. R. Entwisle, and L. S. Olson, "Schools, achievement, and inequality: A seasonal perspective," *Educational evaluation and policy analysis*, vol. 23, no. 2, pp. 171–191, 2001.

[17] S. J. Ingels, T. R. Curtin, P. Kaufman, M. N. Alt, X. Chen, and J. A. Owings, "Coming of age in the 1990s: The eighth-grade class of 1988 12 years later. initial results from the fourth follow-up to the national educational longitudinal study of 1988. statistical analysis report.," 2002.

[18] R. Ricci, "La dispersione scolastica implicita," 2019.

[19] S. Sava, *Needs analysis and programme planning in adult education*. Verlag Barbara Budrich, 2012.

[20] Y. Abd Algani and J. Eshan, "Reasons and suggested solutions for low-level academic achievement in mathematics," *International e-Journal of Educational Studies*, vol. 3, no. 6, pp. 181–190, 2019.

[21] INVALSI, "Rapporto invalsi 2023," tech. rep., INVALSI, 2023.

[22] S. Aljawarneh and J. A. Lara, "Data science for analyzing and improving educational processes," *Journal of Computing in Higher Education*, vol. 33, no. 3, pp. 545–550, 2021.

[23] S. Vincent-Lancrin and R. van der Vlies, "Trustworthy artificial intelligence (ai) in education," no. 218, 2020.

[24] M. Ekowo and I. Palmer, "The promise and peril of predictive analytics in higher education: A landscape analysis.," *New America*, 2016.

[25] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[26] R. Cassen and G. Kingdon, *Tackling low educational achievement*. Joseph Rowntree Foundation, 2007.

[27] D. C. Geary, "Consequences, characteristics, and causes of mathematical learning disabilities and persistent low achievement in mathematics," *Journal of developmental and behavioral pediatrics: JDBP*, vol. 32, no. 3, p. 250, 2011.

[28] J. L. Rastrollo-Guerrero, J. A. Gomez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied sciences*, vol. 10, no. 3, p. 1042, 2020.

[29] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student'performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[30] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students'performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.

[31] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final gpa using decision trees: a case study," *International journal of information and education technology*, vol. 6, no. 7, p. 528, 2016.

[32] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression," in *21st Annual SAS Malaysia Forum, 5th September*, 2007.

[33] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.

[34] C.-J. Villagrá-Arnedo, F. J. Gallego-Durán, P. Compañ, F. Llorens Largo, R. Molina-Carmona, *et al.*, "Predicting academic performance from behavioural and learning data," 2016.

[35] S. Sultana, S. Khan, and M. A. Abbas, "Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts," *International Journal of Electrical Engineering Education*, vol. 54, no. 2, pp. 105–118, 2017.

[36] L. K. Baartman and E. De Bruijn, "Integrating knowledge, skills and attitudes: Conceptualising learning processes towards vocational competence," *Educational Research Review*, vol. 6, no. 2, pp. 125–134, 2011.

[37] B. Ertl, F. G. Hartmann, and J.-H. Heine, "Analyzing large-scale studies: Benefits and challenges," *Frontiers in Psychology*, vol. 11, 2020.

[38] G. E. Fischman, A. M. Topper, I. Silova, J. Goebel, and J. L. Holloway, "Examining the influence of international large-scale assessments on national education policies," *Journal of education policy*, vol. 34, no. 4, pp. 470–499, 2019.

[39] O. Publishing, *Equity in education: Breaking down barriers to social mobility*. Organisation for Economic Co-operation and Development OECD, 2018.

[40] L. Branchetti, F. Ferretti, A. Lemmo, A. Maffia, F. Martignone, M. Matteucci, and S. Mignani, "A longitudinal analysis of the italian national standardized mathematics tests," in *CERME 9-Ninth Congress of the European Society for Research in Mathematics Education*, pp. 1695–1701, 2015.

[41] H. Blossfeld, H. Roßbach, and J. von Maurice, "The german national educational panel study (neps)," *Zeitschrift für Erziehungswissenschaft: Sonderheft*, vol. 14, 2011.

[42] A. Pejić, P. S. Molcer, and K. Gulači, "Math proficiency prediction in computer-based international large-scale assessments using a multi-class machine learning model," in *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 49–54, Ieee, 2021.

[43] M. Saarela, B. Yener, M. J. Zaki, and T. Kärkkäinen, "Predicting math performance from raw large-scale educational assessments data: a machine learning approach," in *JMLR Workshop and Conference Proceedings; 48*, Jmlr, 2016.

[44] S. Thomson, "Achievement at school and socioeconomic background–an educational perspective," 2018.

[45] G. Reimann, M. Stoecklin, K. Lavallee, J. Gut, M.-C. Frischknecht, and A. Grob, "Cognitive and motivational profile shape predicts mathematical skills over and above profile level," *Psychology in the Schools*, vol. 50, no. 1, pp. 37–56, 2013.

[46] P. Lerman, "Fitting segmented regression models by grid search," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, no. 1, pp. 77–84, 1980.

[47] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation.," *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[48] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Advances in neural information processing systems*, vol. 26, 2013.

[49] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[50] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches to educational data mining," *Complexity*, 2019.

[51] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," 2021.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[53] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[54] V. Shah and S. R. Konda, "Neural networks and explainable ai: Bridging the gap between models and interpretability," *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, vol. 5, no. 2, pp. 163–176, 2021.

[55] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. Zingaro, "Student dropout prediction," in *Artificial Intelligence in Education*, (Cham), pp. 129–140, Springer International Publishing, 2020.

[56] V. Tinto, "Dropout from Higher Education: A Theoretical Synthesis of Recent Research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.

[57] E. Commision, "Europe 2020 a strategy for smart, sustainable and inclusive growth," mar 2010.

[58] Istat, "Livelli di istruzione e occupazione nazionali," jul 2020.

[59] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and major using random forests," *Research in Higher Education*, vol. 60, pp. 1048–1064, 2019.

[60] R. Z. Pek, S. T. Özyer, T. Elhage, T. ÖZYER, and R. Alhajj, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure," *IEEE Access*, vol. 11, pp. 1224–1243, 2023.

[61] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Comput. Surv.*, vol. 53, may 2020.

[62] K. Fahd, S. Venkatraman, S. J. Miah, and K. Ahmed, "Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature," *Education and Information Technologies*, vol. 27, p. 3743–3775, apr 2022.

[63] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks," *IEEE Access*, vol. 9, pp. 140731–140746, 2021.

[64] V. Anand, S. A. Rahiman, E. B. George, and A. Huda, "Recursive clustering technique for students' performance evaluation in programming courses," in *2018 Majan International Conference (MIC)*, pp. 1–5, IEEE, 2018.

[65] M. Albán and D. Mauricio, "Neural networks to predict dropout at the universities," *International Journal of Machine Learning and Computing*, 2019.

[66] M. Baranyi, M. Nagy, and R. Molontay, "Interpretable deep learning for university dropout prediction," in *Proceedings of the 21st Annual Conference on Information Technology Education*, SIGITE '20, (New York, NY, USA), p. 13–19, Association for Computing Machinery, 2020.

[67] T. Tang, J. Hou, T. Guo, X. Bai, X. Tian, and A. Noori Hoshyar, "KIDNet: A Knowledge-Aware Neural Network Model for Academic Performance Prediction," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 37–44, 2021.

[68] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18932–18943, 2021.

[69] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain*, pp. 41–50, 2009.

[70] B. Kiss, M. Nagy, R. Molontay, and B. Csabay, "Predicting dropout using high school and first-semester academic achievement measures," in *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pp. 383–389, 2019.

[71] J. Jayaraman, "Predicting student dropout by mining advisor notes," in *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, pp. 629–632, 2020.

[72] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting student academic performance at higher education using data mining: A systematic review.," *Applied Computational Intelligence & Soft Computing*, vol. 2022, 2022.

[73] M. A. U. Alam, "College student retention risk analysis from educational database using multi-task multi-modal neural fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12689–12697, 2022.

[74] Z. Zheng, Y. Li, and Y. Cai, "Oversampling Method for Imbalanced Classification," *Comput. Informatics*, vol. 34, pp. 1017–1037, 2015.

[75] C. Chen and L. Breiman, "Using Random Forest to Learn Imbalanced Data," *University of California, Berkeley*, 01 2004.

[76] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *FAT*'19: PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY*, pp. 220–229, Assoc Comp Machinery, 2019. ACM Conference on Fairness, Accountability, and Transparency (FAT), Atlanta, GA, JAN 29-31, 2019.

[77] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), p. 3323–3331, Curran Associates Inc., 2016.

[78] J. Vasquez Verdugo, X. Gitiaux, C. Ortega, and H. Rangwala, "Faired: A systematic fairness analysis approach applied in a higher educational context," in *LAK22: 12th international learning analytics and knowledge conference*, pp. 271–281, 2022.

[79] C. Pan and Z. Zhang, "Examining the algorithmic fairness in predicting high school dropouts," in *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 262–269, 2024.

[80] S. Zingaro, A. Del Zozzo, F. Del Bonifro, and M. Gabbrielli, "Predictive models for effective policy making against university dropout," *Form re-Open Journal per la formazione in rete*, vol. 20, no. 3, pp. 165–175, 2020.

[81] D. Hooshyar and Y. Yang, "Neural-symbolic computing: A Step Toward Interpretable AI in Education," *Bulletin of the Technical Committee on Learning Technology (ISSN: 2306-0212)*, vol. 21, no. 4, pp. 2–6, 2021.

[82] K. Fiok, F. V. Farahani, W. Karwowski, and T. Ahram, "Explainable artificial intelligence for education and training," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 2, pp. 133–144, 2022.

[83] L. Cohausz, "Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science.," *International Educational Data Mining Society*, 2022.

[84] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[85] M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni, "Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques," *Studies in Higher Education*, vol. 47, no. 9, pp. 1935–1956, 2022.

[86] M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards xai-based personalized intervention," *International Journal of Artificial Intelligence in Education*, pp. 1–27, 2023.

[87] D. Delen, B. Davazdahemami, and E. Rasouli Dezfouli, "Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework," *Information Systems Frontiers*, pp. 1–22, 2023.

[88] L. Plagwitz, A. Brenner, M. Fujarski, and J. Varghese, "Supporting AI-Explainability by Analyzing Feature Subsets in a Machine Learning Model," in *Challenges of Trustable AI and Added-Value on Health*, pp. 109–113, IOS Press, 2022.

[89] B. Škrlj, S. Džeroski, N. Lavrač, and M. Petkovič, "Feature importance estimation with self-attention networks," *arXiv preprint arXiv:2002.04464*, 2020.

[90] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[91] Q. Au, J. Herbinger, C. Stachl, B. Bischl, and G. Casalicchio, "Grouped feature importance and combined features effect plot," *Data Mining and Knowledge Discovery*, vol. 36, no. 4, pp. 1401–1450, 2022.

[92] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, pp. 1340–1347, 04 2010.

[93] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[94] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Student Low Achievement Prediction," in *Artificial Intelligence in Education* (M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, eds.), (Cham), pp. 737–742, Springer International Publishing, 2022.

[95] J. R. Anderson, C. F. Boyle, and B. J. Reiser, "Intelligent tutoring systems," *Science*, vol. 228, no. 4698, pp. 456–462, 1985.

[96] V. J. Shute and D. Zapata-Rivera, "Adaptive educational systems," *Adaptive technologies for training and education*, vol. 7, no. 27, pp. 1–35, 2012.

[97] P. A. Cohen, J. A. Kulik, and C.-L. C. Kulik, "Educational outcomes of tutoring: A meta-analysis of findings," *American educational research journal*, vol. 19, no. 2, pp. 237–248, 1982.

[98] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational researcher*, vol. 13, no. 6, pp. 4–16, 1984.

[99] V. Vagale and L. Niedrite, "Learner model's utilization in the e-learning environments.," in *DB&Local Proceedings*, pp. 162–174, Citeseer, 2012.

[100] L. D. Kurup, A. Joshi, and N. Shekhokar, "A review on student modeling approaches in its," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 2513–2517, 2016.

[101] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of item response theory*, vol. 2. Sage, 1991.

[102] J. Templin, R. A. Henson, *et al.*, *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, 2010.

[103] A. Abyaa, M. Khalidi Idrissi, and S. Bennani, "Learner modelling: systematic review of the literature from the last 5 years," *Educational Technology Research and Development*, vol. 67, no. 5, pp. 1105–1143, 2019.

[104] S. Minn, "Ai-assisted knowledge assessment techniques for adaptive learning environments," *Computers and Education: Artificial Intelligence*, p. 100050, 2022.

[105] A. Tato and R. Nkambou, "Some Improvements of Deep Knowledge Tracing," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, (Portland, OR, USA), pp. 1520–1524, IEEE, Nov. 2019.

[106] C. Conati, *Bayesian Student Modeling*, pp. 281–299. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[107] P. Pavlik Jr, H. Cen, and K. Koedinger, "Performance factors analysis - a new alternative to knowledge tracing," vol. 200, pp. 531–538, 01 2009.

[108] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis – a general method for cognitive model evaluation and improvement," in *Intelligent Tutoring Systems* (M. Ikeda, K. D. Ashley, and T.-W. Chan, eds.), (Berlin, Heidelberg), pp. 164–175, Springer Berlin Heidelberg, 2006.

[109] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," *Advances in neural information processing systems*, vol. 28, 2015.

[110] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *Proceedings of the 26th international conference on World Wide Web*, pp. 765–774, 2017.

[111] C.-K. Yeung and D.-Y. Yeung, "Addressing two problems in deep knowledge tracing via prediction-consistent regularization," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, (London United Kingdom), pp. 1–10, ACM, June 2018.

[112] X. Sun, X. Zhao, B. Li, Y. Ma, R. Sutcliffe, and J. Feng, "Dynamic Key-Value Memory Networks With Rich Features for Knowledge Tracing," *IEEE Transactions on Cybernetics*, vol. 52, pp. 8239–8245, Aug. 2022.

[113] J. R. Anderson, L. M. Reder, and H. A. Simon, "Situated learning and education," *Educational researcher*, vol. 25, no. 4, pp. 5–11, 1996.

[114] R. Nkambou, R. Mizoguchi, and J. Bourdeau, *Advances in intelligent tutoring systems*, vol. 308. Springer Science Business Media, 2010.

[115] J. Vainshtein, R. Esin, and G. Tsibulsky, "Constructing domain model based on logical and epistemological analysis," in *CEUR Workshop Proceedings*, pp. 140–146, 2020.

[116] A. Zanellati, S. P. Zingaro, F. Del Bonifro, M. Gabbrielli, O. Levrini, and C. Panciroli, "Informing predictive models against students dropout," *Atti Convegno Nazionale*, p. 18, 2021.

[117] A. Tato and R. Nkambou, "Infusing Expert Knowledge Into a Deep Neural Network Using Attention Mechanism for Personalized Learning Environments," *Frontiers in Artificial Intelligence*, vol. 5, p. 921476, June 2022.

[118] S. Cheng, Q. Liu, E. Chen, K. Zhang, Z. Huang, Y. Yin, X. Huang, and Y. Su, "AdaptKT: A Domain Adaptable Method for Knowledge Tracing," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, (Virtual Event AZ USA), pp. 123–131, ACM, Feb. 2022.

[119] W. Zhang, K. Qu, Y. Han, and L. Tan, "A Novel Knowledge Tracing Model Based on Collaborative Multi-Head Attention," in *2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI)*, (Guangzhou China), pp. 210–215, ACM, Mar. 2022.

[120] G. G. Towell and J. W. Shavlik, "Knowledge-based artificial neural networks," *Artificial intelligence*, vol. 70, no. 1-2, pp. 119–165, 1994.

[121] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021.

[122] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*, vol. 4. AMLBook New York, 2012.

[123] P. I. Pavlik Jr, K. Brawner, A. Olney, and A. Mitrovic, "Tutoring systems," *Design Recommendations for Intelligent Tutoring Systems: Volume 1-Learner Modeling*, vol. 1, p. 39, 2013.

[124] W. Gan, Y. Sun, and Y. Sun, "Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent tutoring systems," *Neurocomputing*, vol. 488, pp. 36–53, June 2022.

[125] Z. Song, S. Huang, and Y. Zhou, "A Deep Knowledge Tracking Model Integrating Difficulty Factors," in *The 2nd International Conference on Computing and Data Science*, (Stanford CA USA), pp. 1–5, ACM, jan 2021.

[126] H. Dai, Y. Zhang, Y. Yun, and X. Shang, "An improved deep model for knowledge tracing and question difficulty discovery," in *PRICAI 2021: Trends in Artificial Intelligence* (D. N. Pham, T. Theeramunkong, G. Governatori, and F. Liu, eds.), vol. 13032, pp. 362–375, Cham: Springer International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[127] L. He, "Integrating Performance and Side Factors into Embeddings for Deep-Learning Based Knowledge Tracing," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, (Shenzhen, China), pp. 1–6, IEEE, July 2021.

[128] X. Zhang, J. Zhang, N. Lin, and X. Yang, "Sequential Self-Attentive Model for Knowledge Tracing," in *Artificial Neural Networks and Machine Learning – ICANN 2021* (I. Farkaš, P. Masulli, S. Otte, and S. Wermter, eds.), vol. 12891, pp. 318–330, Cham: Springer International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[129] A. Ghosh, N. Heffernan, and A. S. Lan, "Context-Aware Attentive Knowledge Tracing," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (Virtual Event CA USA), pp. 2330–2339, ACM, Aug. 2020.

[130] W. Gan, Y. Sun, X. Peng, and Y. Sun, "Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing," *Applied Intelligence*, vol. 50, pp. 3894–3912, Nov. 2020.

[131] Z. A. Pardos and N. T. Heffernan, "Kt-idem: Introducing item difficulty to the knowledge tracing model," in *International conference on user modeling, adaptation, and personalization*, pp. 243–254, Springer, 2011.

[132] S. Oeda and K. Asai, "Student modeling method integrating knowledge tracing and irt with decay effect.," in *EKM@ EKAW*, pp. 19–26, 2016.

[133] R. Krishnan, J. Singh, M. Sato, Q. Zhang, and T. Ohkuma, "Incorporating wide context information for deep knowledge tracing using attentional bi-interaction.," in *L2D@ WSDM*, pp. 1–13, 2021.

[134] R. Xiao, R. Zheng, Y. Xiao, Y. Zhang, B. Sun, and J. He, "Deep Knowledge Tracking Based on Exercise Semantic Information," in *Emerging Technologies for Education* (W. Jia, Y. Tang, R. S. T. Lee, M. Herzog, H. Zhang, T. Hao, and T. Wang, eds.), vol. 13089, pp. 278–289, Cham: Springer International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[135] Q. Liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu, "EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, pp. 100–115, Jan. 2021.

[136] N. Zhang and L. Li, "Knowledge Tracing with Exercise-Enhanced Key-Value Memory Networks," in *Knowledge Science, Engineering and Management* (H. Qiu, C. Zhang, Z. Fei, M. Qiu, and S.-Y. Kung, eds.), vol. 12815, pp. 566–577, Cham: Springer International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[137] H. Tong, Y. Zhou, and Z. Wang, "Exercise Hierarchical Feature Enhanced Knowledge Tracing," in *Artificial Intelligence in Education* (I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, eds.), vol. 12164, pp. 324–328, Cham: Springer International Publishing, 2020. Series Title: Lecture Notes in Computer Science.

[138] S. Pandey and J. Srivastava, "RKT: Relation-Aware Self-Attention for Knowledge Tracing," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, (Virtual Event Ireland), pp. 1205–1214, ACM, Oct. 2020.

[139] H. Tong, Z. Wang, Y. Zhou, S. Tong, W. Han, and Q. Liu, "Introducing Problem Schema with Hierarchical Exercise Graph for Knowledge Tracing," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Madrid Spain), pp. 405–415, ACM, July 2022.

[140] Y. Liang, W. Wu, L. Wu, and M. Wang, "Inferring How Novice Students Learn to Code: Integrating Automated Program Repair with Cognitive Model," in *Big Data* (H. Jin, X. Lin, X. Cheng, X. Shi, N. Xiao, and Y. Huang, eds.), vol. 1120, pp. 46–56, Singapore: Springer Singapore, 2019. Series Title: Communications in Computer and Information Science.

[141] Z. Wang, X. Feng, J. Tang, G. Y. Huang, and Z. Liu, "Deep Knowledge Tracing with Side Information," in *Artificial Intelligence in Education* (S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, and R. Luckin, eds.), vol. 11626, pp. 303–308, Cham: Springer International Publishing, 2019. Series Title: Lecture Notes in Computer Science.

[142] J. Lee and D.-Y. Yeung, "Knowledge Query Network for Knowledge Tracing: How Knowledge Interacts with Skills," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, (Tempe AZ USA), pp. 491–500, ACM, Mar. 2019.

[143] Y. Ma, P. Han, H. Qiao, C. Cui, Y. Yin, and D. Yu, "SPAKT: A Self-Supervised Pre-TrAining Method for Knowledge Tracing," *IEEE Access*, vol. 10, pp. 72145–72154, 2022.

[144] C. Liu and X. Li, "Memory Attentive Cognitive Diagnosis for Student Performance Prediction," in *Web and Big Data. APWeb-WAIM 2021 International Workshops* (Y. Gao, A. Liu, X. Tao, and J. Chen, eds.), vol. 1505, pp. 79–90, Singapore: Springer Singapore, 2021. Series Title: Communications in Computer and Information Science.

[145] S. Li, L. Xu, Y. Wang, and L. Xu, "Self-learning Tags and Hybrid Responses for Deep Knowledge Tracing," in *Web Information Systems and Applications* (C. Xing, X. Fu, Y. Zhang, G. Zhang, and C. Borjigin, eds.), vol. 12999, pp. 121–132, Cham: Springer International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[146] Y. Yang, J. Shen, Y. Qu, Y. Liu, K. Wang, Y. Zhu, W. Zhang, and Y. Yu, "GIKT: A Graph-Based Interaction Model for Knowledge Tracing," in *Machine Learning and Knowledge Discovery in Databases* (F. Hutter, K. Kersting, J. Lijffijt, and I. Valera, eds.), vol. 12457, pp. 299–315, Cham: Springer

International Publishing, 2021. Series Title: Lecture Notes in Computer Science.

[147] J. Zhang, Y. Mo, C. Chen, and X. He, "GKT-CD: Make Cognitive Diagnosis Model Enhanced by Graph-based Knowledge Tracing," in *2021 International Joint Conference on Neural Networks (IJCNN)*, (Shenzhen, China), pp. 1–8, IEEE, July 2021.

[148] H. Nakagawa, Y. Iwasawa, and Y. Matsuo, "Graph-based knowledge tracing: Modeling student proficiency using graph neural networks," *Web Intelligence*, vol. 19, pp. 87–102, Dec. 2021.

[149] C. JIANG, W. GANb, G. SUa, Y. SUNb, and Y. SUNa, "Improving knowledge tracing through embedding based on metapath," *Proceedings of the 29th International Conference on Computers in Education*, 2021.

[150] H. Liu, T. Zhang, F. Li, Y. Gu, and G. Yu, "Tracking Knowledge Structures and Proficiencies of Students With Learning Transfer," *IEEE Access*, vol. 9, pp. 55413–55421, 2021.

[151] Y. Sun, L. Wang, Q. Xie, Y. Dong, and X. Lin, "Online Programming Education Modeling and Knowledge Tracing," in *Knowledge Science, Engineering and Management* (G. Li, H. T. Shen, Y. Yuan, X. Wang, H. Liu, and X. Zhao, eds.), vol. 12274, pp. 259–270, Cham: Springer International Publishing, 2020. Series Title: Lecture Notes in Computer Science.

[152] S. Tong, Q. Liu, W. Huang, Z. Hunag, E. Chen, C. Liu, H. Ma, and S. Wang, "Structure-Based Knowledge Tracing: An Influence Propagation View," in *2020 IEEE International Conference on Data Mining (ICDM)*, (Sorrento, Italy), pp. 541–550, IEEE, Nov. 2020.

[153] F. Ai, Y. Chen, Y. Guo, Y. Zhao, Z. Wang, G. Fu, and G. Wang, "Concept-aware deep knowledge tracing and exercise recommendation in an online learning system.," *International Educational Data Mining Society*, 2019.

[154] H. N. M. Ferreira, T. Brant-Ribeiro, R. D. Araujo, F. A. Dorca, and R. G. Cattelan, "An Automatic and Dynamic Knowledge Assessment Module for Adaptive Educational Systems," in *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, (Timisoara, Romania), pp. 517–521, IEEE, July 2017.

[155] P. Nedungadi and M. S. Remya, "Predicting students' performance on intelligent tutoring system — Personalized clustered BKT (PC-BKT)

model," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, (Madrid, Spain), pp. 1–6, IEEE, Oct. 2014.

[156] T. Käser, S. Klingler, A. G. Schwing, and M. Gross, "Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks," in *Intelligent Tutoring Systems* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, eds.), vol. 8474, pp. 188–198, Cham: Springer International Publishing, 2014. Series Title: Lecture Notes in Computer Science.

[157] M. Sao Pedro, R. Baker, and J. Gobert, "Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition," in *Educational Data Mining 2013*, 2013.

[158] Y. Wang and J. Beck, "Class vs. Student in a Bayesian Network Student Model," in *Artificial Intelligence in Education* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, eds.), vol. 7926, pp. 151–160, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.

[159] Q. Pan and T. Tezuka, "Prior knowledge on the dynamics of skill acquisition improves deep knowledge tracing," in *Proceedings of the Proceedings of the 29th International Conference on Computers in Education, Bangkok, Thailand*, pp. 22–26, 2021.

[160] W. Lee, J. Chun, Y. Lee, K. Park, and S. Park, "Contrastive Learning for Knowledge Tracing," in *Proceedings of the ACM Web Conference 2022*, (Virtual Event, Lyon France), pp. 2330–2338, ACM, Apr. 2022.

[161] R. S. Baker, S. M. Gowda, and E. Salamin, "Modeling the learning that takes place between online assessments," in *Proceedings of the 26th international conference on computers in education*, pp. 21–28, 2018.

[162] S. Liu, J. Yu, Q. Li, R. Liang, Y. Zhang, X. Shen, and J. Sun, "Ability boosted knowledge tracing," *Information Sciences*, vol. 596, pp. 567–587, June 2022.

[163] T. Long, Y. Liu, J. Shen, W. Zhang, and Y. Yu, "Tracing Knowledge State with Individual Cognition and Acquisition Estimation," in *Proceedings of the*

44th International ACM SIGIR Conference on Research and Development in Information Retrieval, (Virtual Event Canada), pp. 173–182, ACM, July 2021.

[164] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian Knowledge Tracing Models," in *Artificial Intelligence in Education* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, eds.), vol. 7926, pp. 171–180, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.

[165] L. Wei, B. Li, Y. Li, and Y. Zhu, "Time Interval Aware Self -Attention approach for Knowledge Tracing," *Computers and Electrical Engineering*, vol. 102, p. 108179, Sept. 2022.

[166] D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. Choi, "SAINT+: Integrating Temporal Features for EdNet Correctness Prediction," in *LAK21: 11th International Learning Analytics and Knowledge Conference*, (Irvine CA USA), pp. 490–496, ACM, Apr. 2021.

[167] K. Nagatani, Q. Zhang, M. Sato, Y.-Y. Chen, F. Chen, and T. Ohkuma, "Augmenting Knowledge Tracing by Considering Forgetting Behavior," in *The World Wide Web Conference*, (San Francisco CA USA), pp. 3101–3107, ACM, May 2019.

[168] X. Sun, X. Zhao, Y. Ma, X. Yuan, F. He, and J. Feng, "Muti-behavior features based knowledge tracking using decision tree improved DKVMN," in *Proceedings of the ACM Turing Celebration Conference - China*, (Chengdu China), pp. 1–6, ACM, May 2019.

[169] A. Asselman, M. Khaldi, and S. Aammou, "Evaluating the impact of prior required scaffolding items on the improvement of student performance prediction," *Education and Information Technologies*, vol. 25, pp. 3227–3249, July 2020.

[170] J.-J. Vie and H. Kashima, "Knowledge tracing machines: Factorization machines for knowledge tracing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 750–757, 2019.

[171] M. Birenbaum, A. E. Kelly, and K. K. Tatsuoka, "Diagnosing knowledge states in algebra using the rule-space model," *Journal for Research in Mathematics Education*, vol. 24, no. 5, pp. 442–459, 1993.

[172] C. M. Tyng, H. U. Amin, M. N. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in psychology*, vol. 8, p. 1454, 2017.

[173] F. Böttger, U. Cetinkaya, D. Di Mitri, S. Gombert, K. Shingjergji, D. Iren, and R. Klemke, "Privacy-preserving and scalable affect detection in online synchronous learning," in *European Conference on Technology Enhanced Learning*, pp. 45–58, Springer, 2022.

[174] P. Blikstein and M. Worsley, "Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks," *Journal of Learning Analytics*, vol. 3, no. 2, pp. 220–238, 2016.

[175] T. Højgaard, "Competencies, skills and assessment," in *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia*, vol. 1, pp. 225–231, Citeseer, 2009.

[176] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, pp. 1–41, 2021.

[177] A. Alam, "Should robots replace teachers? mobilisation of AI and learning analytics in education," in *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*, pp. 1–12, Dec. 2021.

[178] A. Alam, "A digital game based learning approach for effective curriculum transaction for Teaching-Learning of artificial intelligence and machine learning," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 69–74, Apr. 2022.

[179] D. Castelvecchi, "Can we open the black box of AI?," *Nature*, vol. 538, pp. 20–23, Oct. 2016.

[180] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data Soc.*, vol. 3, p. 205395171562251, Jan. 2016.

[181] D. Hooshyar and Y. Yang, "Predicting course grade through comprehensive modelling of students' learning behavioral pattern," *Complexity*, vol. 2021, pp. 1–12, May 2021.

[182] G. Egodawatte and D. Stoilescu, "Grade 11 students' interconnected use of conceptual knowledge, procedural skills, and strategic competence in algebra: A mixed method study of error analysis.," *European journal of science and mathematics education*, vol. 3, no. 3, pp. 289–305, 2015.

[183] H. Retnawati, E. Apino, A. Santoso, *et al.*, "High school students' difficulties in making mathematical connections when solving problems," *International Journal of Learning, Teaching and Educational Research*, vol. 19, no. 8, pp. 255–277, 2020.

[184] F. M. Zanzotto, "Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, vol. 64, pp. 243–252, 2019.

[185] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022.

[186] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[187] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, Mar. 2022.

[188] R. B. Johnson and A. J. Onwuegbuzie, "Mixed methods research: A research paradigm whose time has come," *Educational researcher*, vol. 33, no. 7, pp. 14–26, 2004.