NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predicting Self-Regulated Learning Skills using Learning Analytics in Moodle

Towards Precision Education

Mariana Rodrigues Cabral

Master Thesis

presented as a partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Predicting Self-Regulated Learning Skills using Learning Analytics in Moodle

Towards Precision Education

by

Mariana Rodrigues Cabral

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Roberto Henriques, Phd and Ricardo Santos, MSc

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 15th, 2025

DEDICATION

Para o meu irmão mais novo, Duarte,

que ainda tem o mundo inteiro por descobrir.

Segue os teus sonhos e nunca deixes de acreditar em ti.

Estarei sempre aqui para te apoiar e celebrar as tuas vitorias, tal como tu celebras as minhas.

ACKNOWLEDGEMENTS

Esta dissertação simboliza não apenas o trabalho desenvolvido nos últimos meses, mas também o reflexo dos desafios superados ao longo de todo o meu percurso académico.

"Alone we can do so little, together we can do so much." - Helen Keller

Começo por agradecer a todos os que contribuíram mais diretamente para esta dissertação:

Ao Leon, ao André, à Sofia e à Susana, colegas de mestrado, pelos trabalhos em conjunto, pelas ideias partilhadas e pelo constante companheirismo ao longo desta etapa.

À Mariana, amiga de longa data, pelo apoio e, em especial, por ter revisto esta dissertação.

Aos meus orientadores, Professor Roberto Henriques e Professor Ricardo Santos, pelo desafio proposto e pela orientação exigente e dedicada.

Quero também deixar um agradecimento especial àqueles que, mesmo não tendo contribuído diretamente para esta dissertação, tiveram um papel determinante no meu percurso:

Aos meus amigos, por me ajudarem a manter o equilíbrio nos momentos mais exigentes.

Ao meu namorado, Paulo, por me ter acompanhado nesta fase diariamente, apoiando-me mesmo nos dias longos e silenciosos e nos momentos de maior cansaço.

À minha família, e em particular à minha mãe, professora por vocação e por coração, que desde cedo me ensinou o valor da perseverança e da curiosidade. És quem mais me compreende e quem sempre me desafiou a ser a melhor versão de mim mesma.

A todos vós, o meu sincero obrigada.

ABSTRACT

Traditional “one-size-fits-all” education paradigm often produces generalized predictive models that overlook different learning environments, affecting academic success prediction. Precision education has emerged as a data-driven approach to personalized learning. Learning Management Systems (LMSs) offer rich behavioral data, but this data often lacks a theoretical basis. Self-regulated learning (SRL) theory provides a framework to address this, but it is usually measured through self-reports, which are biased and difficult to scale. Building on prior research that used clickstream data as an objective proxy for SRL's time management subscale, this study broadens the scope at NOVA Information Management School. It investigates whether Moodle LMS clickstream data can predict students’ SRL skills in three subscales (time management, effort regulation and peer learning) with course-specific models. Data were collected from two graduate courses (Course A and Course B) in the first semester of 2024/2025. SRL targets were extracted from the Motivated Strategies for Learning Questionnaire. Each dataset–target pair was processed through a pipeline with ten configurations and seven algorithms, combining feature selection, dimensionality reduction, and data augmentation techniques. The top three models were shortlisted, tuned, and the best model selected. For each SRL subscale, the most effective dataset was identified by comparing final models. Parametric models generally outperformed non-parametric ones. The best-performing models often showed moderate predictive performance, with models from Course B outperforming those from Course A in two of three subscales. Among the subscales, effort regulation achieved the lowest mean absolute error on the test (MAE = 0.73), followed by time management (MAE = 0.83). In contrast, peer learning was the most challenging subscale to predict (MAE = 1.23), likely due to its offline and social characteristics. Overall, feature–target relationships were weak, and most final models used only the minimum features, indicating limited signal. Additional limitations included sample imbalance, static feature design and reliance on subjective self-report measures. Still, this study offers an initial step toward SRL prediction through behavioral learning analytics. Future research should expand the dataset, adopt time-dependent features, and improve theoretical alignment to support more robust and interpretable SRL prediction in precision education.

KEYWORDS

Clickstream data; Learning analytics; Moodle; Precision education; Self-regulated learning

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AUCROC	Area Under the Receiver Operating Characteristic
COPES	Conditions, Operations, Products, Evaluations, Standards
CS	Course Specific dataset: CS-A (Course A), CS-B (Course B)
CTGAN	Conditional Tabular Generative Adversarial Network
DT	Decision Tree
EDA	Exploratory Data Analysis
GDPR	General Data Protection Regulation GPA
GPA	Grade Point Average
kNN	*k*-Nearest Neighbors
KPCA	Kernel Principal Component Analysis
LMS	Learning Management System
MAE	Mean Absolute Error
MI	Mutual Information
ML	Machine Learning
MOODLE	Modular Object-Oriented Dynamic Learning Environment
MSLQ	Motivated Strategies for Learning Questionnaire
NOVA IMS	Nova Information Management School
PCA	Principal Component Analysis
PQ	Practical Quiz
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SRL	Self-Regulated Learning

TBL	Team-Based Learning
VIF	Variance Inflation Factor
URL	Uniform Resource Locator

1. INTRODUCTION

1.1. CONTEXT

The success of learning analytics models heavily relies on institutional design (Gašević et al., 2016). However, the traditional "one-size-fits-all" paradigm often neglects this factor, resulting in predictive models that fail to perform effectively across diverse learning environments and course contexts. In response, precision education has emerged as a data-informed method to personalized learning. As noted by Luan & Tsai (2021), it integrates several research purposes, including diagnosing and profiling individual learner differences, predicting academic performance, and designing timely and personalized prevention strategies.

While learning analytics and machine learning (ML) models have been recognized as valuable tools for precision education research, the majority of work in this domain has focused on predicting academic performance (van Sluijs & Matzat, 2024). These studies frequently rely on Learning Management Systems (LMS), such as Moodle[1], as they track diverse student activities and provide a wealth of digital behavioral data. Nevertheless, the full potential of LMS data remains underexplored due to a significant methodological gap: a lack of strong theoretical grounding to explain how digital behaviors relate to meaningful learning concepts. Few studies in learning analytics explicitly ground the selection of predictive variables based on theoretical reasoning (Conijn et al., 2017). However, some researchers used the theory of self-regulated learning (SRL), recognizing its potential for bridging this gap (Gašević et al., 2016).

SRL represents a key factor in precision education, as it helps profile students by recognizing their learning differences. Self-regulation involves cognitive, metacognitive, and motivational components that enable students to manage their learning (Pintrich & De Groot, 1990; Pintrich et al., 1993). Students who often regulate their learning tend to be consistently associated with improved academic performance (Zimmerman, 2000). Thus, identifying SRL profiles can help instructors deliver more targeted and effective pedagogical interventions, provide a well-established framework for interpreting individual learning processes, and support personalized interventions.

Despite its importance, SRL is often assessed through self-report instruments (Schellings & Van Hout-Wolters, 2011), such as the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1991). These traditional tools raise concerns about their reliability and validity due to potential biases and the time required for administration (Schellings & Van Hout-Wolters, 2011).

[1] https://moodle.org

LMS log data, also known as clickstream data, can be an alternative method for assessing students' SRL skills. It captures detailed, time-sensitive information about student interactions, allowing for large-scale and real-time analysis of learning behaviors. For example, clickstream data such as resource access timestamps, task submissions, and navigation patterns can be used to objectively measure SRL skills, including time management, goal setting, and effort regulation (Baker et al., 2020). Supporting this, recent research has shown that clickstream data can directly predict self-reported time management skills, highlighting their potential as proxies for traditionally subjective constructs (van Sluijs & Matzat, 2024).

1.2. PURPOSE: OBJECTIVE AND RELEVANCE

This study builds on the prior work of van Sluijs & Matzat (2024) by exploring the predictive capabilities of Moodle data from NOVA Information Management School (NOVA IMS) regarding students' SRL skills. Guiding this research is a key question, framed as follows: Is it possible to accurately predict students' SRL skills using NOVA IMS Moodle LMS clickstream data?

To answer this question, the study explores two key dimensions:

> **RQ1:** How does predictive performance vary across different SRL subscales?
>
> **RQ2:** How can course context and student interaction patterns influence predictive performance across courses that share a similar structure?

To address these questions, the following objectives were defined:

1. Collect SRL skills scores of NOVA IMS students using the MSLQ.
2. Identify which MSLQ subscales are most frequently targeted in SRL studies using clickstream data and evaluate their suitability based on the behavioral traces available in Moodle.
3. Extract and preprocess Moodle clickstream data, focusing on behavioral features theoretically linked to SRL dimensions.
4. Train ML regression models to assess the relationship between LMS clickstream data and SRL subscale scores.
5. Evaluate model performance and feature importance, comparing results across SRL subscales and course contexts.

The significance of this study lies in its ability to establish a connection between clickstream data and SRL skills at NOVA IMS, thereby providing an alternative approach to measuring theoretical SRL variables that overcomes the limitations of traditional self-report methods. This is particularly significant because it could benefit various research objectives within

precision education. For instance, it could create opportunities to improve the efficiency of performance prediction models by studying the impact of SRL on academic outcomes and integrating predicted SRL variables directly into LMS-based performance prediction frameworks.

1.3. STUDY OUTLINE

Following this introductory chapter, the remainder of the work is structured as follows: the next chapter provides a literature review, progressing from a broad overview to a focused discussion of relevant works that cover the relationship between LMS clickstream data and SRL. The third chapter details the methodology guiding the investigation to address the research questions. The fourth chapter presents the results obtained from applying the outlined methodology and highlights the key findings. The fifth and final chapter discusses the main limitations of the research and proposes future work directions.

2. LITERATURE REVIEW

A summary of the methodology that guided the literature review is presented in Figure 2.1. The process followed three steps: (1) identification of research domains from the thesis title, (2) definition of guiding questions per domain, and (3) establishment of a systematic strategy for sourcing relevant literature.

From the title, we identified three core domains and ordered them by scope (from the most general to the most directly related to the research objectives): Precision education, SRL and LMSs. For each domain, standard questions were formulated to clarify key concepts, summarize key findings and identify research gaps and recommendations for future work. The search process began with the inspirational study (van Sluijs & Matzat, 2024), expanded through its references and related works, and, when needed, was broadened using systematic reviews.

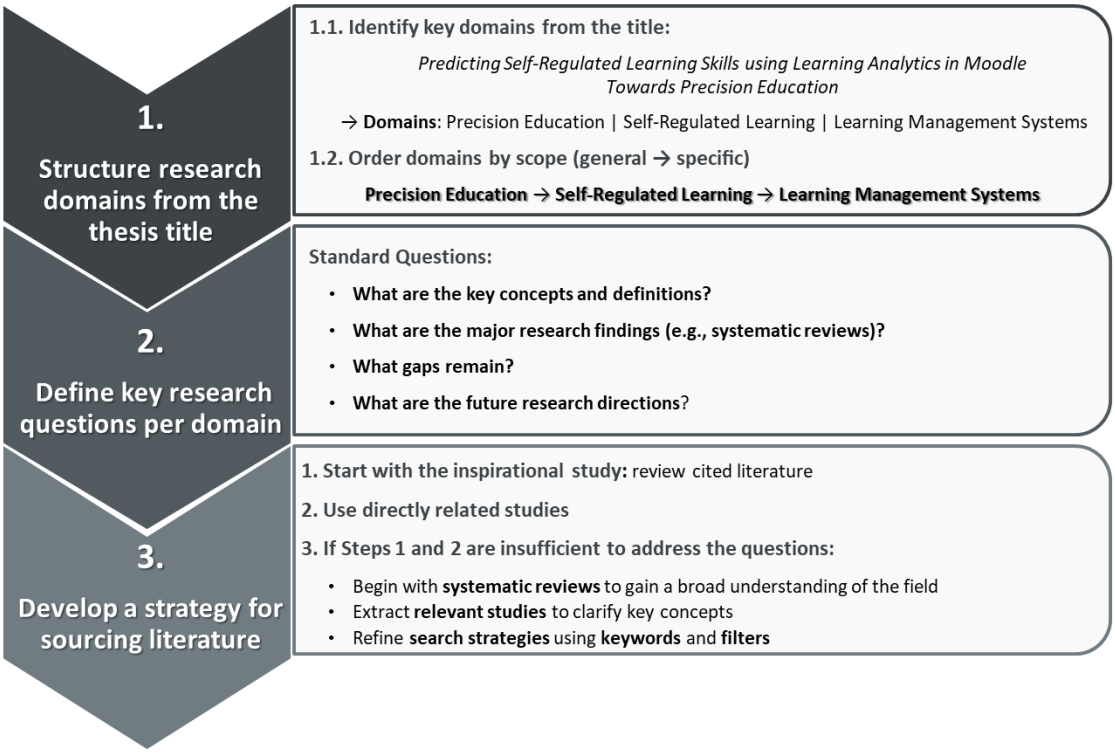


Figure 2.1 – Literature review process schema

As a result of the methodology employed, the literature review is organized into four main sections aligned with the identified research domains. It begins with the broader topic of precision education, emphasizing its focus on data-driven personalization and its reliance on learning analytics and ML to enhance learning outcomes. The second section concentrates on SRL, specifically on assessing SRL skills. It reviews traditional self-report tools, particularly the MSLQ, highlighting their limitations and the growing interest in more dynamic, data-driven alternatives. The third section examines the role of LMSs in monitoring student online

behavior, with particular attention to how clickstream data has been used to predict academic performance. It also identifies key research gaps, including weak theoretical foundations, data context dependence, and limited model portability. The final section narrows the focus to the aims of this study, connecting clickstream data with SRL. A literature review table of studies closely aligned with our research objectives is provided in Appendix A, Table A.1. The subsequent sections offer a detailed examination of each of these domains.

2.1. PRECISION EDUCATION

Precision education is a modern approach that replaces the traditional "one-size-fits-all" paradigm with tailored learning experiences (Luan & Tsai, 2021). Inspired by precision medicine, it aims to improve learning outcomes by adapting to individual cognitive, behavioral and environmental differences through targeted evidence-based strategies (Hart, 2016). It is a technology-driven approach (Williamson, 2019) that uses learning analytics to collect data on student activities and applies ML models to predict academic performance (Luan & Tsai, 2021).

A systematic review (Luan & Tsai, 2021) revealed that most research in precision education using ML has been focusing on predicting academic performance. These studies primarily involved university students in online or blended courses in the areas of science, technology, engineering, and mathematics. The primary data sources included learning logs from massive open online courses, institutional records and surveys. Frequently used algorithms included regression, decision trees, support vector machines, and neural networks, with cross-validation being the most frequently used method for model evaluation. A particular example was Azcona et al. (2019), who provided weekly predictions of assignment failure in programming courses by combining LMS behavioral logs and demographic data to deliver personalized feedback and improve academic outcomes. Similarly, Delen et al. (2020) addressed dropout prediction by applying Bayesian networks to model the causal relationships between various student features and the risk of dropout.

While these studies offer valuable insights, they also expose significant gaps, such as the limited diversity of research across educational levels and disciplines, and a tendency to prioritize data-driven approaches over solid theoretical frameworks, often neglecting classic learning theories. Future research should address these gaps by exploring broader research contexts and combining pedagogical frameworks with advanced technologies (Luan & Tsai, 2021).

A suitable candidate for that is SRL theory (Gašević et al., 2016), which offers a well-established framework for understanding individual learning differences, enhancing model interpretability, and enabling more effective theory-informed applications of learning analytics. Building on this, Baker et al. (2020) explored how LMS clickstream data could reveal SRL-related behaviors and discussed the methodological challenges involved. The following

section provides an in-depth exploration of SRL, focusing on its conceptual foundations and assessment.

2.2. SELF-REGULATED LEARNING SKILLS ASSESSMENT: SELF-REPORTED INSTRUMENTS

SRL refers to the ability of students to manage their learning by employing cognitive, metacognitive and motivational strategies (Pintrich & De Groot, 1990; Pintrich et al., 1993). Since the late 1990s, various models and definitions have been discussed (Pintrich, 2000; P. H. Winne, 1996; Zimmerman, 2000). This study focuses on Pintrich's model (2000), which serves as the foundation for the MSLQ, a widely used self-report instrument to assess SRL skills (Broadbent & Poon, 2015) and the source from which the SRL scores used in this research were derived.

According to Pintrich's model (2000, 2004), SRL is a cyclical process structured into four distinct phases: (1) forethought, planning and activation, which involves goal setting, task analysis and the activation of prior knowledge along with motivational beliefs; (2) monitoring, by self-tracking comprehension, effort, and motivation during learning; (3) control where learners select and adjust cognitive and metacognitive strategies; and (4) reaction and reflection that involve evaluating results, identifying reasons for success or failure, and adjusting strategies for the future. These phases operate across four domains: cognition, motivation, behavior and context. The purpose of SRL is to develop learners' autonomy, helping them become more strategic, reflective and adaptable. Students who actively regulate their learning tend to be more motivated, better at managing their time, and more resilient when facing academic challenges, which is consistently associated with improved academic performance (Zimmerman, 2000).

Self-reported instruments, such as questionnaires, are traditional methods in educational research used to assess students' perceptions, thoughts, and behaviors, offering valuable insights into SRL skills (Jo et al., 2016; Kaya et al., 2012; Li et al., 2020; J. Xu et al., 2013). In addition to the widely used MSLQ, other examples include the Self-Regulated Learning Interview Schedule (Zimmerman & Martinez-Pons, 1986) and the Online Self-Regulated Learning Questionnaire (Barnard et al., 2009). Their ease of administration and minimal disruption to learning contribute to their popularity (Schellings & Van Hout-Wolters, 2011). However, they are limited by inaccuracies in recall (Veenman, 2005), potential biases and incomplete representation of behaviors. Addressing these requires refining constructs, tailoring measurements and adopting multi-method approaches to enhance validity and reliability (Schellings & Van Hout-Wolters, 2011).

The MSLQ (Pintrich et al., 1991) assesses college students' SRL skills within a specific course context. It comprises 81 items and 15 scales divided into two sections: motivation and learning strategies. The motivation section assesses students' reasons for learning and perceptions of their ability to succeed, including goal orientation (intrinsic/extrinsic), task value, self-efficacy,

and test anxiety. The learning strategies section evaluates cognitive and metacognitive strategies (e.g., rehearsal, elaboration, organization, critical thinking, and self-regulation) and resource management skills (e.g., time management, effort regulation, peer learning, and help-seeking). Its modular design allows scales to be used independently or together. Items are scored on a 7-point Likert scale (Likert, 1932), where higher scores (e.g., 4–7) indicate greater motivation or strategy use. For negatively worded items, scores are reversed during analysis to align with the positive scales (e.g., a score of 2 is recalculated as 6 by subtracting 2 from 8). Each scale's score is calculated as the average of its item scores and compared to a reference group, such as the class. Scores in the top 25% indicate higher motivation or strategy use, the middle 50% suggest similarity to peers, and the bottom 25% reflect lower levels. Generally, scoring below 3 on more than six scales suggests that seeking support is needed.

A following study (Pintrich et al., 1993) confirmed the reliability of the MSLQ in consistently measuring motivation and learning strategies. It also demonstrated strong predictive validity, with scales such as self-efficacy showing significant correlations with academic success indicators, including GPA. Additionally, the MSLQ proved to be generalizable, performing consistently across diverse student samples.

As a self-reported instrument, the MSLQ has notable limitations (Schellings & Van Hout-Wolters, 2011). A meta-analytic review (Credé & Phillips, 2011) highlighted its reliability but pointed to moderate predictive validity for academic performance. Although some subscales exhibited moderate to strong correlations with grades, such as self-efficacy, effort regulation, and time and study environment, most showed weaker correlations, including help-seeking and peer learning. Additionally, unvalidated context-specific assumptions and redundancy between constructs, such as time and study environment regulation with effort regulation, underscore the need for improvement.

While improving the MSLQ remains an open issue, this study focuses on exploring LMS clickstream data as an alternative for assessing SRL behaviors. Recent evidence suggests that clickstream data can offer a more accurate and dynamic understanding of self-regulated learning and its impact on student performance (Li et al., 2020).

2.3. LEARNING MANAGEMENT SYSTEMS

The integration of information and communication technologies into higher education has transformed the traditional education model, redesigning face-to-face courses into blended or entirely online formats (Hoic-Bozic et al., 2009). This change has led to the adoption of LMSs to support these new learning models (Conijn et al., 2017).

LMSs, also known as Managed Learning Environments or Integrated Learning Systems, are e-learning platforms designed to manage, deliver and evaluate educational programs. They provide a digital space to share course materials, facilitate communication through forums,

track progress via quizzes and assignments, and offer feedback to students and educators (Holmes & Gardner, 2006). Besides Moodle, well-known examples include Blackboard[2] and Canvas[3].

Beyond administrative functions, LMS clickstream data are powerful tools for revealing patterns of student motivation and learning activities (Baker et al., 2020). However, while this data is automatically recorded, simplifying the data collection process, it remains raw. Therefore, it is often disconnected from established theoretical frameworks, resulting in mixed results across contexts (Conijn et al., 2017). This emphasizes the importance of theory-grounded features to support the development of learning analytics (van Sluijs & Matzat, 2024).

According to Baker et al. (2020), there are two primary data analysis strategies for extracting insights from clickstream data: aggregate non-temporal analysis and time-dependent analysis. The first approach aggregates clickstream data over time for each student, summarizing their counts of different activity types into a static multivariate representation of the course (e.g., total clicks on a lecture video). It is suitable for statistical methods like regression or clustering, but it loses temporal details about behavior. The second retains time-dependent information, allowing for the detection of patterns over time (e.g., daily click count), but is often more complex to analyze.

2.3.1. Clickstream data: predicting academic performance

The focus of research in learning analytics has been the use of clickstream data to predict academic performance (Conijn et al., 2017; van Sluijs & Matzat, 2024). Initial studies started to show its effectiveness in specific course contexts. For instance, Calvo-Flores et al. (2006) achieved 80.2% accuracy in predicting pass/fail outcomes using artificial neural networks, and Macfadyen and Dawson (2010) identified 80.9% of at-risk students with an overall accuracy of 73.7% using logistic regression.

Additionally, research has also explored models trained on data from multiple courses. Studies by Conijn et al. (2017) and Gašević et al. (2016) evaluated the effectiveness of these models, concluding that they often perform worse than models tailored to individual courses, which highlights the inconsistency of findings and the portability issues. This underperformance is attributed to the variability in the predictive power of key metrics, including logins, time spent online, and discussion posts. These metrics can be highly context-dependent, influenced by differences in LMS platforms, course formats (e.g., blended vs. online) and the characteristics of study populations (Conijn et al., 2017; Gašević et al., 2016; van Sluijs & Matzat, 2024).

[2] https://www.anthology.com/products/teaching-and-learning/learning-effectiveness/blackboard
[3] https://www.instructure.com/canvas

Despite these advancements, significant challenges remain in using clickstream data for predictive purposes, including the tendency to make late predictions in courses, which reduces the effectiveness of early interventions, and methodological inconsistencies, such as variations in calculating time-on-task, which compromise the reliability of the findings (Conijn et al., 2017). Additionally, the majority of research focuses on identifying at-risk students, with limited exploration of high-achieving learners or institution-wide applications (Riestra-González et al., 2021; Santos & Henriques, 2023b).

A significant limitation is the lack of strong theoretical frameworks to understand behavioral indicators of learning (Conijn et al., 2017), which compromises the reliability of such metrics across diverse contexts. While SRL theory has shown promise in addressing this issue (Gašević et al., 2016), it is also important to acknowledge the limitations associated with traditional self-reported instruments used to assess SRL (Schellings & Van Hout-Wolters, 2011). In response, recent research has begun to investigate clickstream data as a more dynamic and objective means of assessing SRL skills (Li et al., 2020). The following section builds on this emerging perspective by reviewing studies that integrate SRL theory with behavioral data, thereby laying the groundwork for the present study.

2.3.2. Clickstream data: self-regulated learning skills

Using learning analytics to assess students' digital footprints within LMSs often provides a more objective and accurate approach to measure SRL skills than self-reported methods (Baker et al., 2020). Accurate measurement of SRL skills is crucial, as subskills such as time management, effort regulation, metacognition and critical thinking are consistently linked to improved performance in online courses (Broadbent & Poon, 2015). Particularly, time management stands out as a predictor of academic performance (Baker et al., 2018; Cicchinelli et al., 2018; Credé & Phillips, 2011). Additionally, since LMS features are not explicitly designed to measure SRL skills, and time management is the subskill most aligned with observable behaviors in these systems, most research using LMS clickstream data has focused on it (Baker et al., 2020).

Research on time management has employed two main approaches: questionnaire-based methods (Kaya et al., 2012; J. Xu et al., 2013) and data-driven methods using learning analytics. The latter approach involved studying three types of behaviors: studying on time, studying in advance and spacing (Li et al., 2020). Studying on time refers to reviewing course materials or completing assignments before deadlines, for instance, by tracking how often students in blended courses access resources related to face-to-face meetings before those meetings occur (Cicchinelli et al., 2018). Studying in advance involves starting tasks rather than waiting until deadlines approach (Baker et al., 2018), which is measured by analyzing how far in advance students begin or submit assignments in fully online courses (Crossley et al., 2016). Spacing refers to spreading study time over an extended period, rather than concentrating

work into short periods. This is often measured by analyzing how work sessions are distributed throughout the week (Baker et al., 2018).

Relevant contributions to the field included the study of Ahmad Uzir et al. (2020), which analyzed session data and the number of actions within a session in a flipped classroom context, identifying distinct time management strategies such as preparing, working ahead, revisiting, and catching up. However, their findings were limited to highly structured courses. Additionally, Li et al. (2020) used clickstream data to analyze students' access to materials (in advance vs. on time) and regular engagement, finding moderate correlations between these metrics and self-reported skills after a course, but not before. They also highlighted that clickstream data is a better predictor of course performance than surveys. Finally, Jo et al. (2016) further explored the relationship between clickstream behaviors (e.g., total login time, login regularity, and frequency) and self-reported time management scores, showing these behaviors mediated the relationship between self-reported scores and academic performance, but did not directly predict time management skills, suggesting the need for further investigation into this complex relationship.

Building on prior research, Van Sluijs and Matzat (2024) proposed a study that used LMS clickstream data to directly predict students' time management skills, moving beyond the investigation of correlations and mediation. Their study addressed the possibility of predicting time management skills in blended courses without a specific predefined structure. Moreover, they also combined learning analytics with MSLQ data to validate predictions. The results showed that features such as mean session time, session interval, login irregularity, number of files accessed, and average daily start time predicted time management in more than one course. These features are less tied to specific schedules, indicating their potential in less structured learning contexts. However, the effectiveness of linear and multilevel regression models varied, accounting for 14% to 71% of the variance in skills across five courses. This inconsistency underscores the limited reliability and transferability of these models, influenced by differences in instructional design and content. Although clickstream data accurately reflects online behaviors, its failure to capture offline activities and cognitive processes highlights the need for additional data sources to gain a fuller understanding of SRL strategies.

Additional challenges were outlined in a systematic review by Alhazbi et al. (2024) regarding indicators and metrics for measuring SRL in higher education. The findings indicated that most studies focused on indicators of time management skills, such as engagement, regularity, and anti-procrastination, which often fail to capture the full complexity of SRL. For instance, engagement metrics such as total logins may reflect cramming before exams, rather than consistent study habits, which indicates poor time management (Asarta & Schmidt, 2013). Therefore, further research should explore diverse types of engagement and develop more accurate indicators of SRL. Moreover, regularity has been measured using various statistical

methods, including averages, variance, standard deviation, and entropy. However, these methods come with limitations. For example, averages can be misleading, masking access patterns, while smaller entropy values suggest steadier engagement but fail to account for adaptive behaviors, a critical aspect of SRL.

Alhazbi et al. (2024) also concluded that few studies have examined help-seeking behaviors, usually relying on discussion forum data, which often have low engagement and are affected by student-instructor relationships, course content, and delivery mode. Additionally, most research depends on data from single LMS courses, limiting the capture of SRL's context-sensitive nature. To address this, future research should analyze student behavior across all enrolled courses. Finally, many existing indicators lack validation against established theoretical models, raising concerns about their relevance. Combining clickstream data with self-reports, such as the MSLQ (van Sluijs & Matzat, 2024), has improved validity, but challenges persist as these methods measure different aspects of SRL. While self-reports capture SRL as a stable trait, clickstream data reflect dynamic, context-specific behaviors (Fan et al., 2022, as cited by Alhazbi et al., 2024). Thus, future research should track behaviors across multiple subjects and semesters, helping to understand how learning adapts over time, beyond what self-reports alone can reveal (Alhazbi et al., 2024).

A relevant future recommendation is promoting collaboration between researchers and educators to develop robust and practical metrics that are theoretically based (Alhazbi et al., 2024). Supporting this, Cristea et al. (2024) developed an unobtrusive, scalable, and portable approach to assess SRL behaviors in online learning environments, using only LMS data. Grounded in the COPES model of SRL (Conditions, Operations, Products, Evaluations, and Standards; P. Winne & Hadwin, 1998), they designed multidimensional behavioral indicators corresponding to key SRL phases: task definition, goal-setting, enactment, and adaptation. These phases reflect how learners interpret tasks, set goals, engage with content, and adjust their strategies in response to feedback. Their findings demonstrated that these clickstream-based measures outperformed traditional self-reported surveys in predicting academic performance, particularly for phases like enactment and adaptation.

However, the study highlighted challenges in effectively capturing reflective phases such as goal-setting and task definition due to offline and cognitive processes that are not observable in LMS data. Furthermore, the researchers found a lack of correlation between the clickstream-based scales and the MSLQ-SRL subscale, indicating that these two approaches may measure distinct dimensions of SRL. These findings highlight the difficulty of fully assessing all SRL phases using clickstream data alone, emphasizing the need for further research to improve the scales and combine multiple data sources for a more complete understanding.

Why investigate whether LMS data can be used as an alternative instrument to assess SRL behaviors, initially measured by MSLQ, when innovative and unobtrusive methods like Cristea

et al. (2024) already use LMS data directly and demonstrate scalability and portability? While their study highlights the effectiveness and portability of clickstream-based metrics for general SRL assessment, our research targets specific SRL skills critical for tailored interventions. Therefore, our work complements their broader approach, providing actionable insights into how students regulate their learning processes and bridging the gap between theoretical assessments and practical applications to advance precision education through actionable insights.

3. METHODOLOGY

This chapter outlines the methodology employed in this study, designed to address the research questions and objectives. Our approach was structured into four sequential phases: population definition, data collection, data preparation, and ML pipeline. A schematic representation of our experimental approach is in Figure 3.1.

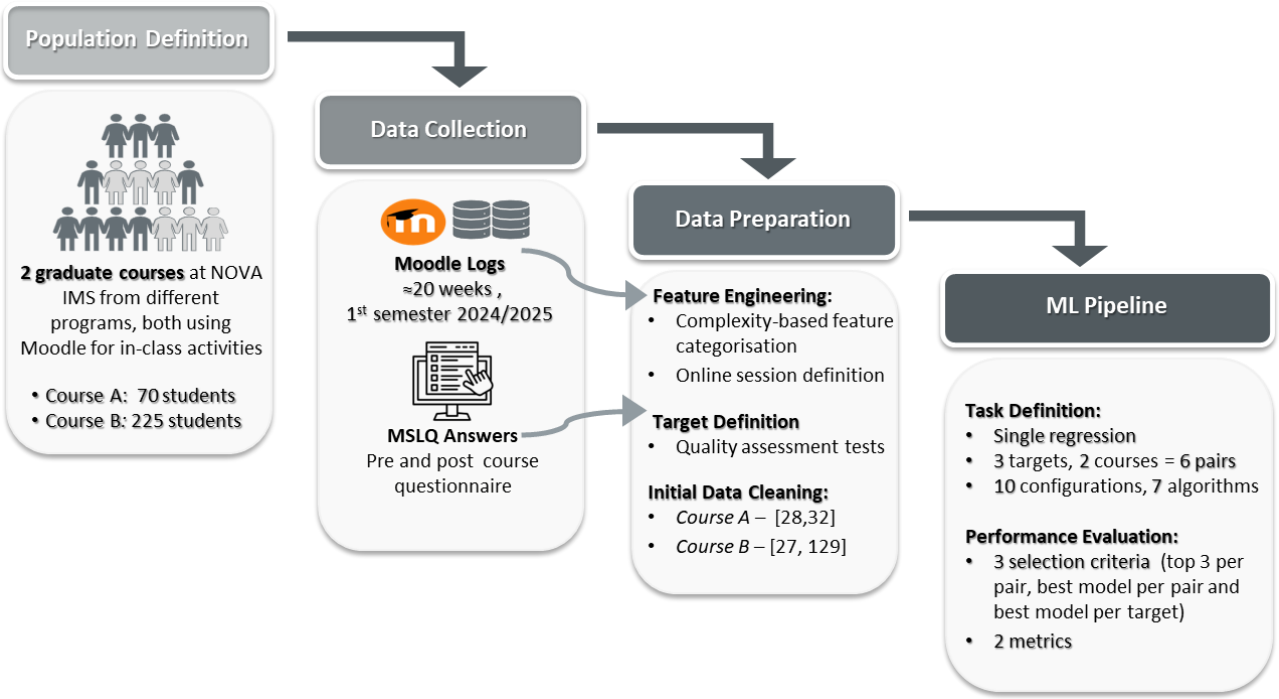


Figure 3.1 – Overview of the experimental approach adopted

In the first phase, two graduate-level courses at NOVA IMS were selected for their similar structure, particularly their consistent use of the LMS to support in-class activities. This ensured the availability of LMS logs for extracting clickstream data and allowed a focused comparison of how course-specific factors might influence predictive performance. While structurally similar, the courses differ in content, class frequency, and their emphasis on individual versus group work, which could influence LMS interaction patterns and potentially impact the prediction accuracy.

During the second phase, data collection, we addressed research objective 1 by administering the MSLQ. Additionally, we began working on objective 3 by using Moodle LMS to extract the logs. Aligning these two data sources defined the study sample, with the MSLQ subscale scores serving as target candidates and the LMS logs as potential features for modelling.

The third phase, data preparation, focused on achieving objectives 2 and 3. It involved defining the target variables and features, and initially exploring and cleaning the data prior to modelling. Target selection was based on prior literature, course context and data quality tests (Shapiro & Wilk, 1965; Wilcoxon, 1945), leading to the choice of three SRL subscales: time

management, effort regulation, and peer learning. The features were extracted from clickstream data, utilizing a complexity-based feature framework (Santos & Henriques, 2023a) and session definitions adapted from prior research (Conijn et al., 2017; Zacharis, 2015).

The final phase aimed to address research objectives 4 and 5 by developing a structured ML pipeline designed to train predictive models and generate results that directly answer our research questions. Each SRL subscale was treated as an independent regression task and modelled separately for each course dataset. The pipeline tested seven algorithms across ten configurations, incorporating data augmentation, dimensionality reduction, and hyperparameter tuning. Model performance was evaluated through cross-validation using both mean absolute error (MAE) and root mean square error (RMSE). To identify the best-performing models for each subscale and course, three predefined selection criteria were applied, enabling comparative analysis across SRL subscales (RQ1) and course contexts (RQ2).

Before proceeding to the subsequent sections, which provide a detailed description of each phase, it is essential to outline two methodological considerations. First, all data manipulation and analysis were performed using Python (McKinney, 2018) and Scikit-learn (Pedregosa et al., 2011) unless stated otherwise. Second, all student data were anonymized in compliance with the General Data Protection Regulation (GDPR) and the institution's ethical guidelines. This project received approval from the university's Ethics Committee and Institutional Review Board under the reference code DSCI2024-11-185925 (Appendix B).

3.1. POPULATION DEFINITION

The student sample considered in this study corresponded to two courses from different graduate-level programs at NOVA IMS during the fall quarter of the 2024/2025 academic year. *Data Mining I*, from the Master's Degree program in Information Management (hereafter referred to as Course A), and *Machine Learning*, from the Master's Degree program in Data Science and Advanced Analytics (corresponding to Course B).

These courses were mainly selected because they included Team-Based Learning (TBL) sessions as an evaluation element. The purpose of these sessions was to evaluate students' preparation for lectures and promote collaborative learning. Before each session, students had access to preparatory materials on the Moodle platform. During the session, students first completed an individual quiz (I) and then repeated it as a team (T), with the final score being a weighted average of both. All these activities took place in class and were tracked through Moodle, producing structured, time-stamped data useful for examining students' online behavior.

Additionally, the courses shared several structural features that facilitated meaningful cross-course comparisons. They employed a blended delivery mode, shared teaching staff, and the program lasted 14 weeks, consistently utilizing Moodle for materials and forums.

Despite this, the courses differ in the total number of students enrolled, with 70 in Course A and 225 in Course B, and in content, with Course A focused on unsupervised learning and Course B on supervised learning. Additional differences include class frequency, with Course A meeting once a week and alternating between lectures and practical labs, while Course B has two sessions each week, one lecture and one lab. Finally, they differ in their emphasis on individual versus group work, as evident in their evaluation schemes, shown in Table 3.1. Course A has six TBL sessions, worth 20% of the grade, while Course B distributes this across five TBL sessions and five practical quizzes (PQs) conducted during lab sessions on Moodle to assess individual problem-solving skills.

Table 3.1 – Course evaluation structure as proposed in the course syllabus

Evaluation Element	Description	Course A	Course B
Team-Based Learning (TBL)	Multiple-choice quiz to assess students' preparation for the theoretical class. Initially taken individually **(I)**, followed by a team repetition **(T)**. Final score for each TBL is calculated as 50% **I** + 50% **T**. The total TBL grade is the sum across all sessions.	**20%** Distributed across 6 (Each:1.67%**I**+1.67%)	**10%** Distributed across 5 (Each:1.00%**I**+ 1.00%**T**)
Handout (H)	Group assignment involving practical exercises.	**10%**	**10%**
Practical Quiz (PQ)	Individual quiz to test students' coding or problem-solving skills.	-	**10%** Distributed over 5
Final Project (FP)	Course group project.	**30%**	**30%**
Exam (E)	Individual formal assessment covering the entire course syllabus.	**40%**	**40%**
Final Grade 1st Attempt		**0.2*TBL +0.1*H+0.3*P+0.4*E**	**0.1*TBL+0.1*H +0.3*P+0.4*E**

3.2. DATA COLLECTION

In this study, two data sources were considered: the MSLQ and Moodle log data. The number of valid questionnaire responses determined the sample size and the amount of Moodle logs under analysis. The questionnaire data was used to extract the target variable, while the clickstream data was employed to define the features of our models.

First, we administered the full version of the MSLQ (Pintrich et al., 1991) using 7-point Likert scale items. Data collection was conducted in person during theoretical classes in the 12th week out of 14. A total of 174 valid responses were received from a target population of 295 enrolled students, resulting in a response rate of approximately 59%. This timing, near the end of the lectures and course completion, was chosen to minimize the gap between students' perceptions and their actual learning experiences. The underlying hypothesis was that self-reported methods assess SRL perceptions, while Moodle logs provide evidence of SRL behavior. Therefore, to approximate these metrics, we assumed that by the end of the course, students have a clearer perception of their actual learning skills and their impact on learning outcomes. As a further step to investigate this assumption and evaluate the quality of the responses collected, we also administered a pre-course MSLQ, following a similar approach to that of Li et al. (2020). Although this questionnaire achieved a 95% response rate, only students who participated in both questionnaires were included in the comparative analysis.

After the course concluded, we extracted the Moodle logs for the defined population and applied filtering to ensure relevance. Following an approach similar to van Sluijs & Matzat (2024), we included logs from one week before the courses started up to the day before the first exam of the first examination period, including class weeks, breaks, and holidays.

3.3. DATA PREPARATION

3.3.1. Target definition

After collecting the MSLQ responses, general preprocessing was performed to obtain only the response items and then allocate them into predefined subscales. The selection of the target SRL subscales was guided by three main criteria: relevance of the study, prior literature and course characteristics.

First, we considered the established link between SRL skills and academic performance (Zimmerman, 2000), which supports the relevance of our study. Based on this, the focus was on subscales identified in the literature as being correlated with academic performance, such as effort regulation, self-efficacy, time and study environment management, and metacognitive strategies (Broadbent & Poon, 2015; Credé & Phillips, 2011). Then we narrowed the selection, prioritizing subscales that were widely referenced in the literature or considered accessible for LMS-based assessment, ending up with time management and effort regulation (Ahmad Uzir et al., 2020; Baker et al., 2020; Jo et al., 2016; Li et al., 2020; van Sluijs & Matzat, 2024). Finally, as an experimental addition, we also included the peer learning subscale, considering the structure of the courses under analysis. Both had multiple TBL recorded on the Moodle platform as described in Table 3.1.

Thus, three SRL subscales were selected as the target in this study: peer learning, time management, and effort regulation. The questionnaire items corresponding to each subscale

are presented in Appendix C (Tables C.1, C.2, and C.3). The final target computation involved additional data cleaning procedures to exclude unauthorized, duplicate, or suspiciously fast responses. This resulted in a final dataset of 172 valid entries, approximately 58% of the enrolled student population. For each selected subscale, a mean score was calculated by averaging responses to the relevant items, with proper adjustments for reverse items. These mean subscale scores served as the three continuous target variables in the modelling phase.

3.3.1.1. Quality test

Before proceeding, we decided to test the quality of the responses of the computed targets. We began by analyzing the data distribution to determine the most appropriate statistical methods, using Shapiro's test (Shapiro & Wilk, 1965) to check for normality. The results indicated that none of the targets were normally distributed.

Additionally, we explored how students' self-perceptions evolved over the course. Therefore, we applied the same preprocessing steps to the pre-course MSLQ data and then used the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945) to compare paired pre-course and post-course responses from students who completed both assessments (N = 159). Test results are summarized in Table D.1 of Appendix D. Since all p-values were below 0.05, the results suggest that students' perceptions of their SRL skills changed significantly from the beginning to the end of the course. To infer in what way they changed, we also compared mean scores. The results presented in Table D.2 of Appendix D indicate a decline in self-reported SRL skills across all three subscales: time management ($\Delta = -0.75$), effort regulation ($\Delta = -0.56$), and peer learning ($\Delta = -0.27$). These findings support our hypothesis that self-report instruments accurately reflect students' perceptions of their SRL skills, which can evolve over time. By the course's end, we expected students to have a better understanding of their learning behaviors and their connection to academic outcomes. The drop in self-reported scores likely reflects a shift from initial overconfidence to more realistic self-assessments based on experience.

Finally, to assess the internal consistency of each subscale, we computed Cronbach's alpha (Cronbach, 1951) for the post-course responses. Peer learning showed the highest reliability coefficient (0.72), followed by time management (0.64) and effort regulation (0.57). These values suggested a moderately acceptable reliability for most scales, though the effort regulation subscale ($\alpha = 0.57$) falls below the commonly accepted threshold of 0.60–0.70, indicating lower internal consistency in this case. We decided to keep it since it is close to the lower bound of the considered threshold.

3.3.2. Feature engineering

After obtaining valid questionnaire responses, we filtered the Moodle logs to include only students who had responded to the final MSLQ and defined the time frame of interest. The

resulting sample included 346,395 interaction logs, involving 161 students (32 from Course A and 129 from Course B).

Before the feature extraction, we decided to gain deeper insights into engagement on Moodle. Therefore, we analyzed the average student's interactions on the Moodle LMS weekly (Figure 3.2) and daily (Figure 3.3) for each course.

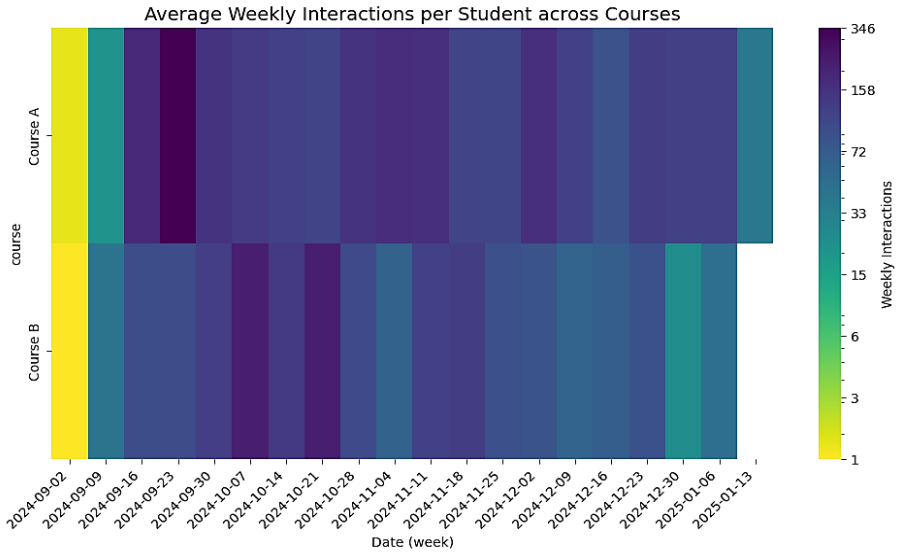


Figure 3.2 – Average weekly Moodle LMS interactions per student across courses

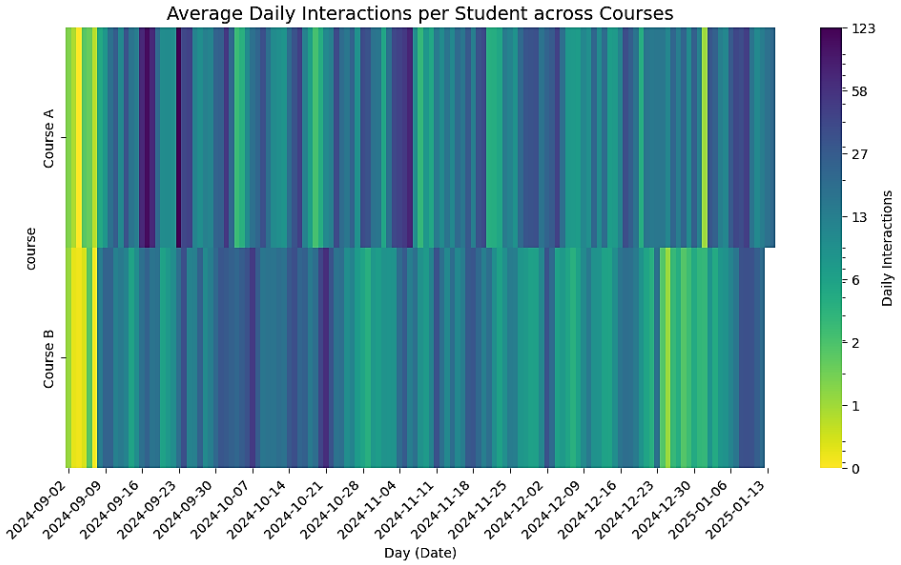


Figure 3.3 – Average daily Moodle LMS interactions per student across courses

The figures highlighted the importance of investigating how course context and student interaction patterns influence predictive performance. While they reinforced the information in Table 3.1 and the course schedules, indicating that the courses were overall similarly structured, they also revealed the need to treat the courses separately due to distinct Moodle usage patterns. Both courses showed significant peaks in interaction around key LMS

activities, such as TBL sessions and practical classes. Engagement decreased during breaks and holidays but increased before exams, although not as much as during the semester, suggesting a shift toward offline study. However, Course A sustained higher engagement, peaking at 346 clicks per student per week in Week 3, whereas Course B was more variable, reaching its highest interaction in Week 7, with 221 clicks per student per week. Overall, Course B showed lower engagement in Moodle LMS than Course A, with more pronounced drops during breaks.

Continuing our analysis of the interactions, we examined the log table in detail and identified specific entries for key activities such as assignments, quizzes, and TBLs. Comparing these with the evaluation schema (Table 3.1) and course schedules, we noted that Course B had 4 TBLs and 4 quizzes instead of the proposed 5.

To convert Moodle logs into predictive features, we began by collecting variables used in prior studies aligned with our research objectives, as mentioned in the literature review section (Calvo-Flores et al., 2006; Conijn et al., 2017; Cristea et al., 2024; Gašević et al., 2016; Jo et al., 2016; Li et al., 2020; Macfadyen & Dawson, 2010; Santos & Henriques, 2023a, 2023b; van Sluijs & Matzat, 2024; Zacharis, 2015). Then, we filtered the features that were more suitable for our study context by using the same or a comparable proxy.

Furthermore, to support feature engineering, we categorized the features based on their complexity, following the methodology proposed by Santos & Henriques (2023a), which classifies features into three perspectives: raw activity, representing the number of times a student performs a specific action on the LMS; time-on-task, capturing the time spent on the LMS; and frequency, measuring how often and when students access the LMS.

Finally, to session-based features, we needed a definition of an online session in LMS. Therefore, we adopted the definition proposed by Conijn et al. (2017), inspired by Zacharis (2015), where a session begins with the first click after login and ends with the last click before logout or after 40 minutes of inactivity. It requires at least two clicks, and their time difference measures the duration. After identifying periods of interaction without logging, we adapted the definition to consider a session as a continuous period of activity, where interactions are grouped based on a threshold of inactivity, set to 40 minutes.

3.3.3. Initial data cleaning

From the feature engineering strategy, 51 feature candidates were extracted, highlighting the need for initial data cleaning to ensure the quality and relevance of our data before modelling. It involved two phases, each with two steps: exploratory data analysis (EDA) (where the number of features remained unchanged) and preprocessing (where the features were modified, leading to a change in their number).

In the first phase, steps were taken for each category of features (raw, time-on-task, and frequency) across courses. From the EDA steps, we highlight the study of correlations using

the Spearman (Spearman, 1904) and Pearson (Pearson, 1895) methods. It directly impacted the preprocessing by eliminating redundant, constant, or irrelevant features and converting features into a binary format to enhance their interpretability. At this stage, only perfect correlation scores were considered redundant, with the removal criteria favoring weekly activity features over absolute counts, as the course is structured weekly. Additionally, percentage-based features were prioritized over absolute values to ensure better comparability across students by normalizing against course-wide totals.

As a result of the first phase of preprocessing, two course-specific datasets were extracted: Course A (CS-A), containing 32 students and 32 features, and Course B (CS-B), with 129 students and 33 features. Each row corresponds to a student who completed the questionnaire and provided a valid response for analysis, while each column represents a candidate behavioral feature.

In the second phase, the EDA was conducted using the previously defined datasets. This analysis involved a deeper exploration of the relationships between features and between features and targets. For the feature-target analysis, we further investigate the high correlation using a threshold score of 0.8. For the feature-feature analysis, we evaluated multicollinearity using the variance inflation factor (VIF)(Kutner et al., 2005) (following the implementation[4]). During preprocessing, variables with a correlation coefficient of 0.9 or higher were identified as candidates for removal due to high collinearity. The removal criteria were based on the last phase preprocessing criteria combined with feature correlation with targets (preferring higher correlations), its correlation with other features (preferring lower correlations), and its VIF score (aiming for the lowest possible value). Ultimately, the number of features in each dataset was updated to 28 for CS-A and 27 for CS-B.

Tables 3.2, 3.3 and 3.4 list all features retained across datasets, grouped by unit for each of the three predefined categories. All listed features are included across datasets (CS-A and CS-B), unless marked otherwise: (*) indicates exclusion from CS-A, and (**) from CS-B.

[4]https://www.statsmodels.org/dev/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html

Table 3.2 – Raw activity feature set across datasets

Unit (short-term)	**Feature Name**
Binary	File download
	Forum discussion post read
	TBL completed
	PQ completed*
Number (n)	Assignments viewed
	TBL
	TBL resources clicks
	TBL out of the schedule
	URLs viewed
Number as a percentage of the course mean (% of the course mean)	Announcements read
	Forum discussion posts read
	Resources view
Number as a percentage of the course total (% of the total course)	Course clicks
	Submissions
Number per week (n/week)	Course clicks
	Folder clicks**

Features excluded from: *CS-A, ** CS-B

Table 3.3 – Time-on-task feature set across datasets

Unit (short-term)	**Feature Name**
Hours (h)	Handout submission delay
	Project submission delay
Minutes (min)	Average duration of online sessions
	Time spent online before the course start
	Total time online
Minutes per week (min/week)	Variation of online session duration
Number per online session (n/session)	Course clicks

Table 3.4 Frequency feature set across datasets

Unit (short-term)	**Feature Name**
Days (d)	Largest inactivity period
Hours (h)	Average spacing between online sessions
Hours per week (h/week)	Variation of spacing irregularity
Minutes (min)	Online Session duration irregularity
Minutes per week (min/week)	Variation of online session duration irregularity
Number as a percentage of course duration in days (% of total course days)	Days with no interaction **

Features excluded from: *CS-A, ** CS-B

Overall, raw activity features (Table 3.2) captured absolute counts for each student's interactions (e.g., URLs viewed), weekly-normalized counts aligned with the course structure (e.g., folder clicks per week), percentage metrics relative to course-wide means or totals (e.g., forum discussion posts read as a percentage of the course mean) and binary indicators (e.g., whether a student read at least one forum post). Additionally, time-on-task features (Table 3.3) included session-related features: timing (e.g., time spent online before the course start), duration (e.g., average duration of online sessions) and variation in duration (e.g., variation in online session duration). Finally, frequency features (Table 3.4) reflected inactivity levels (e.g., days with no interaction as a proportion of the total course duration in days), session spacing behaviors (e.g., average spacing between online sessions) and variation in irregularity (e.g., variation in online session duration irregularity). Tables E.1, E.2 and E.3 in Appendix E provide a detailed description of each feature for a particular category.

3.4. ML PIPELINE

This section presents the final phase of the methodology, in which we designed a structured ML pipeline to model each of the three SRL subscales (peer learning, time management, and effort regulation) across two course-specific datasets (CS-A and CS-B). Each subscale was treated as a separate regression task, and the pipeline was applied independently to each dataset–target pair. We tested seven algorithms with ten configurations involving data preprocessing, dimensionality reduction and oversampling.

Model selection followed a three-stage process: (1) selection of the top three models per dataset–target pair, (2) identification of the best model for each pair, and (3) selection of the final model per SRL subscale. Between (1) and (2), we also tested the impact of hyperparameter tuning, comparing the tuned and untuned versions of the top 3 models. Tuned models were retained only if they improved or maintained overall test performance

(defined as the sum of selected metrics, MAE and RMSE) without exceeding a fixed 0.1 train–test gap.

3.4.1. Configurations

Table 3.5 lists the ten configurations used. All strategies build upon a baseline with skewness correction, using Yeo-Johnson transformation (Yeo & Johnson, 2000) for features with skewness above 0.5 and feature standardization.

Table 3.5 Configurations

Strategy	Configuration	Description
Baseline	**B**	Feature scaling and skewness correction
Feature Selection	**FS**	Feature selection only
	FS+RO	Feature selection with random oversampling
	FS+CO	Feature selection with CTGAN oversampling
PCA	**PCA**	PCA only
	PCA+RO	PCA with random oversampling
	PCA+CO	PCA with CTGAN oversampling
KPCA	**KPCA**	KPCA only
	KPCA+RO	KPCA with random oversampling
	KPCA+CO	KPCA with CTGAN oversampling

Due to the high feature-to-observation ratio, particularly in the CS-A dataset, feature selection and dimensionality reduction methods were selected to mitigate the curse of dimensionality (Bellman, 1961). Thus, we employed Principal Component Analysis (PCA)(Pearson, 1901) as a linear technique and Kernel Principal Component Analysis (KPCA) (Schölkopf et al., 1998) as its non-linear counterpart. Additionally, to address data imbalance, two oversampling strategies were included: random oversampling and Conditional Tabular Generative Adversarial Network (CTGAN) (L. Xu et al., 2019) (following the implementation[5]). These methods were selected due to their complementary effectiveness: random oversampling balances the dataset by replicating existing minority class samples, while CTGAN generates synthetic samples by learning the underlying data distribution, thus potentially producing more diverse and realistic examples. Each of these steps is described in detail in the subsections below.

[5] https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer

3.4.1.1. Feature selection

For all configurations except the baseline and those with dimensionality reduction (PCA or KPCA) a robust feature selection strategy was adopted, combining five methods within a voting-based framework. Features selected by at least three methods were retained. Thresholds and constraints were empirically defined, with warnings triggered when all features were selected, enabling dynamic adjustments to preserve selectivity and robustness.

The selected techniques grounded in prior literature were: Recursive Feature Elimination (RFE) (Guyon et al., 2002) (with a Decision Tree regressor, DT), feature importance from a Random Forest (RF) (Breiman, 2001), Mutual Information (MI) scores (Vergara & Estévez, 2014), and coefficient-based selection from Lasso (Tibshirani, 1996) and Ridge (Hoerl & Kennard, 2000) regressions. Table 3.6 summarizes these methods along with the empirical parameter choices used in this study.

Table 3.6 Feature selection methods

Configuration	Selection criterion
RFE	Retain 3 to 80% features, optimized by the lowest RMSE
RF	Feature importance > 1 / total number of features
MI	MI score > 80th percentile
Lasso	Retain features with non-zero coefficients
Ridge	Absolute coefficient > 80th percentile

The selection process was applied only to the training data in each cross-validation fold. Features selected in each iteration were recorded, and those chosen in at least 70% of folds were kept for the final model. An adaptive mechanism was implemented to ensure robustness: if fewer than three features met the voting criteria, the highest-voted features were added until at least three were selected.

3.4.1.2. PCA and KPCA

PCA retained 70% of the variance, capturing the essential linear components. In contrast, KPCA used 70% of input features, a conservative heuristic that ensures substantial representation of the original data. It employed a Radial Basis Function kernel with a gamma of 0.1.

3.4.1.3. Oversampling

Although Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling technique, it was not applied in this study for two reasons. First, it was traditionally designed

for classification tasks (Chawla et al., 2002) and does not natively support regression problems with continuous target variables. Second, even adaptations for regression (Branco et al., 2017) may generate unrealistic samples near the edges of the distribution, potentially introducing noise. Instead, we decided to use random oversampling as a simple baseline technique and CTGAN to generate more realistic synthetic data in underrepresented regions of the target distribution.

Random oversampling increased the frequency of samples from extreme target regions, defined as below the 20th or above the 80th percentile. CTGAN oversampling generated new synthetic samples from the same extreme regions. The generative model was trained on minority subsets to create realistic samples, enhancing the learning signal in underrepresented areas of the target distribution.

3.4.2. Algorithms

The seven algorithms used ranged from simple baselines to more complex ensemble methods. As a baseline algorithm for model performance, we included a simple algorithm that always predicts the mean value of the target variable, regardless of the input data. We also tested regularized linear models such as Lasso (Tibshirani, 1996), Ridge (Hoerl & Kennard, 2000), and ElasticNet (Zou & Hastie, 2005). Despite this, and acknowledging that the true relationship between variables is often unknown in real-world data, we also included flexible non-parametric methods to account for potential non-linear patterns that linear models may fail to capture: *k*-Nearest Neighbors (kNN) (Cover & Hart, 1967) and DT (Breiman et al., 1984). Additionally, we further included RF (Breiman, 2001), an ensemble method that combines multiple decision trees to reduce overfitting and improve accuracy, particularly in noisy or complex data settings (James et al., 2013).

Before hyperparameter tuning, all algorithms were evaluated using default settings, except for DT and RF, where the maximum depth of trees was specified to limit tree growth and reduce the risk of overfitting. For the assessment, we used cross-validation with five folds repeated three times, to ensure sufficient data per fold and to support robust statistical comparisons.

3.4.3. Performance evaluation strategy

3.4.3.1. Metrics

Performance evaluation was based on two complementary metrics: MAE to assess the magnitude of errors for each subscale due to its simplicity and straightforward interpretation, and RMSE to account for the impact of larger errors (Chai & Draxler, 2014). Metrics were computed per fold for training and test sets, and references to them in the following sections (e.g., MAE train or MAE test) typically refer to the average across all folds.

The mathematical definitions of MAE and RMSE are presented in the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

3.4.3.2. Selection criteria

We established performance-based selection criteria to support decision-making at three key stages in our pipeline. These criteria were designed to promote generalization, reduce overfitting, and ensure robustness in model comparison across datasets and target variables.

Selection criterion 1 – top 3 candidate models per dataset–target pair

Given the large number of initial models tested for each target variable, we applied a multi-stage model selection strategy to reduce the hypothesis space and focus subsequent optimization efforts on the most promising candidates. The process described in Figure 3.4 was guided by five sequential filtering phases: (1) extraction of valid candidates, (2) MAE filter, (3) RMSE filter, (4) final ranking, and (5) top 3 selection. If fewer than three models passed at any stage, a fallback sorting process was applied on the remaining candidates.

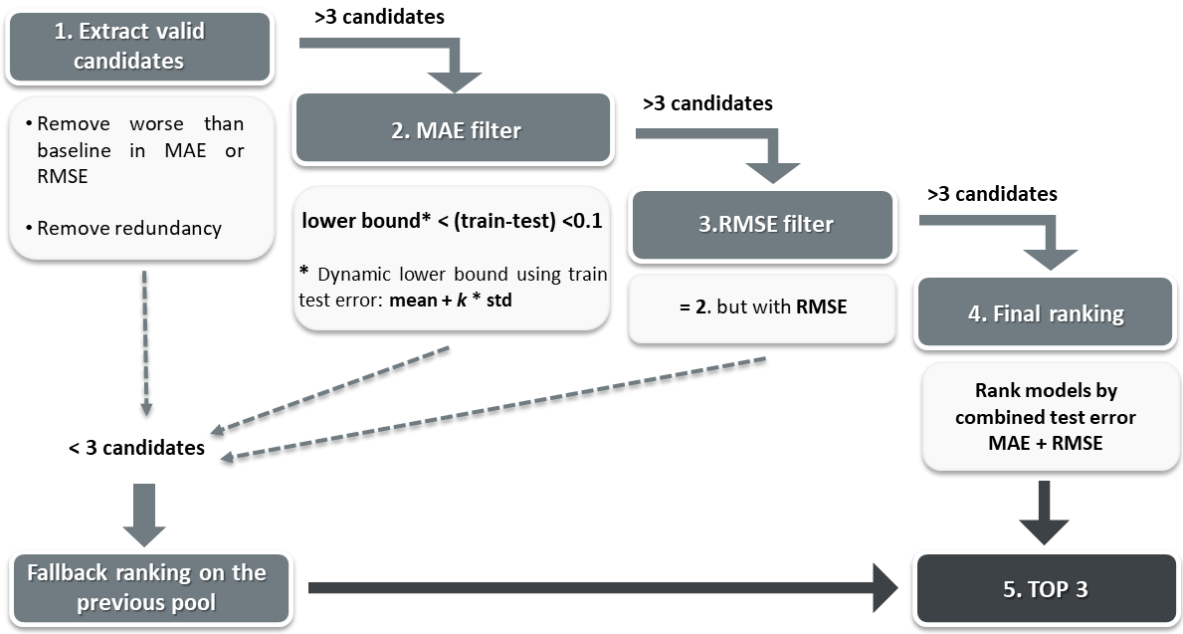


Figure 3.4 – Selection criteria 1: multistage model filtering

The process began by identifying valid candidates through the removal of underperforming and redundant models. A model was considered underperforming if it performed worse than the baseline algorithm in its baseline configuration in at least one of two evaluation metrics. Redundancies were identified by comparing fold-level MAE and RMSE values, retaining only one instance of each duplicate.

After identifying the valid candidates, we applied the MAE filter, retaining only those instances whose train-test difference fell within a set range. This range was composed of a fixed upper limit, set at 0.1, to mitigate overfitting, and a dynamic lower limit to cover possible statistical perturbations. The lower threshold was calculated as the average difference between test and train errors plus *k* times the standard deviation, with *k* set to 1.5. If more than three models passed this filter, a similar RMSE filter was applied. If models still exceeded the limit, a final tie-breaker ranked them by their combined test error (MAE + RMSE), selecting the top three.

If fewer than three models passed at any stage, a fallback sorting process was applied on the remaining candidates. It prioritized the MAE train-test absolute distance first, followed by a similar rank for the RMSE distance to resolve ties, and finally by the combined test error as the ultimate tiebreaker.

As a note, we emphasize that the process was stopped early each time exactly three models remained at any stage. These top three models were then kept for further tuning and analysis.

Selection criterion 2 – best model selection per dataset–target pair

The best model was selected from the revised top three candidates using a two-stage procedure that combined statistical significance testing and a structured fallback strategy.

The selected test was the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945), as it does not assume normality, thereby being more robust for data that may not follow a known distribution, as in this study. We used the signed-rank test because it accounts for paired predictions across the same folds. The rank-sum test assumes independence. All models, despite different training strategies, were evaluated on the same data splits, justifying paired comparisons.

The process involved comparing the performance of different models using the Wilcoxon signed-rank test on fold-level MAE scores, with a significance level set at 0.10. This threshold was chosen because our repeated cross-validation with a few folds resulted in small sample sizes for the tests. According to Demšar (2006) non-parametric tests like Wilcoxon can have low statistical power in such cases, increasing the risk of missing real differences (Type II errors). Increasing the significance level slightly can make it easier to detect meaningful differences. Furthermore, according to Fisher (1950) the conventional 5% threshold isn't absolute. In exploratory analyses like ours, where the goal is to identify promising patterns rather than definitive conclusions, a more relaxed threshold, such as 10%, may be appropriate.

A significant difference indicated that the model with the lower average MAE was the winner. Wins were counted for each model, and the one with the most wins was selected. In case of ties, the test was applied to RMSE scores among the tied models, and the model with the most RMSE wins was chosen. If a tie persisted after both MAE and RMSE tests or if no model had any significant wins, a fallback sorting strategy was applied equal to the one in the selection criterion 1. Finally, the selected model was then labelled with a reason for selection, indicating whether it was chosen based on Wilcoxon MAE, Wilcoxon RMSE, or the fallback sorting strategy.

Selection criterion 3 – final model selection per target

Once the best model for each pair was identified, we compared the two candidates (CS-A and CS-B) for each target variable to determine which was most suitable for representing the corresponding SRL subscale in the final interpretation and discussion. Only candidates with statistically significant results were considered eligible for selection. However, if no significant models were found, all candidates were eligible. Among the eligible models, the one with the lowest combined test score was selected as the final representative.

3.4.4. Hyperparameter tuning

For each dataset-target pair corresponding to the top three selected models, a random search using 60 hyperparameter configurations was conducted (Bergstra & Bengio, 2012), maintaining each model's original strategy and selected features where applicable. Each configuration was evaluated using 5-fold cross-validation repeated 3 times, and the one with the lowest average score was selected, with MAE chosen as the scoring metric. The complete hyperparameter search space for each model is detailed in Appendix F, Table F.1.

4. RESULTS AND DISCUSSION

This chapter presents the key findings of our study. It begins with an EDA of the final datasets obtained during the data preparation phase. Next, it summarizes the ML pipeline modeling results through a stepwise evaluation process: from an initial ranking of modeling strategies to the identification of the most robust and interpretable models. Finally, it assesses the prediction quality of the final selected models and revisits the research questions to synthesize the implications of the findings.

4.1. FINAL EDA

The final EDA was structured in two parts. First, we examined descriptive statistics of the final feature set. Then, we explored the relationships between individual features and targets using three complementary techniques: MI (Vergara & Estévez, 2014) to capture non-linear dependencies, F-score(Fisher, 1950) to assess linear explanatory power of each feature, and Pearson correlation (Pearson, 1895) to measure the strength and direction of linear associations. These analyses were conducted separately for each course-specific dataset (CS-A and CS-B) using the original non-standardized data.

4.1.1. Descriptive statistics

For this analysis, only features shared across all datasets were considered, and they were grouped by complexity level. Table 4.1 reports raw activity, Table 4.2 time-on-task, and Table 4.3 frequency features. Each table presents the mean and standard deviation, in parentheses, for each feature. For binary features, only the count and the percentage of occurrences (i.e., value = 1) are shown, with the percentage also reported in parentheses.

Table 4.1 – Descriptive statistics of raw activity features across datasets

Feature (unit)	**CS-A**	**CS-B**
Announcements read (% of the course mean)	1 (1.32)	1 (1.54)
Assignments viewed (n)	5.5 (4.9)	5.29 (5.4)
Course clicks (% of the total course)	0.03 (0.03)	0.01 (0.01)
Course clicks (n/week)	7 (4.33)	8.84 (5.05)
File download (binary)	29 (90.62%)	7 (5.43%)
Forum discussion posts read (% of the course mean)	1 (1.49)	1 (2.18)
Forum discussion post read (binary)	19 (59.38%)	59 (45.74%)
Resources view (% of the course mean)	1 (0.63)	1 (0.4)
Submissions (% of the total course)	0.03 (0.04)	0.01 (0.01)
TBL (n)	13.16 (10.33)	8.43 (3.93)
TBL completed (binary)	15 (46.88%)	83 (64.34%)
TBL resources clicks (n)	207.72 (129.21)	153.21 (68.18)
TBL out of the schedule (n)	3.31 (4.37)	1.18 (2.12)
URLs viewed (n)	1.69 (1.93)	9.25 (6.89)

TBL features show greater student engagement in Course A than in Course B. They attempted more TBL sessions on average (13 vs. 8) and clicked more on preparation materials (208 vs. 153), likely reflecting the higher number of mandatory TBL sessions in Course A (6 vs. 4). Despite this, students in Course A had a lower completion rate (47% vs. 64%) and engaged in more out-of-schedule TBL (3 vs. 1), suggesting that many were completed independently outside of class, possibly affecting the quality of the learning experience.

Focusing on other absolute activity metrics, in mean, students in Course B accessed more URLs (9 vs. 2) and recorded slightly higher weekly course clicks (9 vs. 7). This may reflect their course structure, which included separate lecture and lab sessions, unlike Course A*'s* single weekly class. In contrast, students in Course A read more forum discussions (60% vs. 46%) and

download more files than those from Course B (91% vs. 5%), indicating distinct engagement strategies.

Relative activity metrics, such as announcements read, forum posts, and resource views, showed average values around 1 across both courses, indicating that students generally behaved in line with their peers. However, greater variability in Course B, especially in forum and announcement engagement, points to a few students with significantly different interaction patterns. Finally, the consistently low percentages for course clicks and submissions (1%–3%) across all groups suggest minimal additional engagement beyond the core tasks.

Table 4.2 – Descriptive statistics of time-on-task features across datasets

Feature (unit)	**CS-A**	**CS-B**
Average duration of online sessions (min)	20.43 (5.78)	19.3 (4.14)
Course clicks (n/session)	21.34 (13.03)	14.5 (5.99)
Handout submission delay (h)	19.53 (5.83)	9.98 (3.98)
Project submission delay (h)	35.06 (2.22)	2.37 (32.1)
Time spent online before the course start (min)	10.34 (18.08)	11.47 (19.87)
Total time online (min)	2197.03 (1361.89)	2016.53 (1116.95)
Variation of online session duration (min/week)	-0.1 (0.58)	0.03 (0.54)

When comparing the course means, evidence suggests that course design may influence differences in student profiles. On average, students in Course A spent approximately 180 minutes more (over three hours) online than students in Course B. They also clicked considerably more per session (21 clicks vs. 15 clicks), with higher variability in Course A, indicating more diverse interaction patterns. Regarding assignment management, students in Course A delayed their handout submissions almost twice as long as those in Course B (19.5 hours vs. 10 hours). Conversely, for the final project submission, Course B exhibited more significant variability, suggesting that although the average delay was lower, submission behaviors were more inconsistent among students. Pre-course activity levels were similarly low across both groups, while the variation in session duration over time was minimal, indicating relatively stable engagement throughout the course.

Table 4.3 – Descriptive statistics of frequency features across datasets

Feature (unit)	**CS-A**	**CS-B**
Average spacing between online sessions (h)	36.62 (17.82)	34.02 (13.15)
Largest inactivity period (d)	9.74 (4.5)	11.34 (3.97)
Online Session duration irregularity (min)	25.06 (8.23)	23.6 (5.23)
Variation of online session duration irregularity (min/week)	-0.18 (0.73)	0.04 (0.54)
Variation of spacing irregularity (h/week)	1.16 (1.79)	1.64 (1.51)

Table 4.3 shows moderate Moodle engagement in both courses, with sessions spaced over a day apart and offline periods reaching up to two working weeks, suggesting erratic, irregular usage patterns. However, this was expected given the blended format of the courses and the relatively light in-class schedule (typically one or two sessions per week). While we cannot determine with certainty the reason for these offline periods, it is plausible that students were studying or participating in other academic activities outside the LMS.

Despite similarities, some differences emerge. Students in Course A, on average, space sessions slightly further apart (36.6 hours vs. 34 hours) and show more variability in session duration (25.1 minutes vs. 23.6 minutes). However, this slight irregularity decreases over time (–0.18 minutes). Their spacing irregularity increases moderately (1.16 hours per week), which is less than in Course B (1.64 hours per week). Students in Course B, while engaging slightly more frequently, exhibit longer inactivity periods (11.3 vs. 9.7 days) and growing irregularity in both session duration (0.04 minutes per week) and spacing, suggesting a more fragmented engagement pattern despite a denser class schedule.

4.1.2. Target-feature relationships

This section highlights only the key findings from the analysis. The complete results for each SRL subscale are presented in Appendix G (Tables G.1–G.3). Each table displays the top three ranked features for MI, F-score (F) and Pearson correlation (r) across both course-specific datasets (CS-A and CS-B), enabling side-by-side comparison and facilitating the identification of the most informative and consistent features.

For peer learning (Table G.1, Appendix G), the most consistent features across metrics were related to forum engagement, especially the percentage of forum discussions read relative to the course mean. Due to the course structure, features likely associated with preparing TBL activities, such as the number of URLs viewed, also emerged as relevant. While this aligns with the collaborative nature of peer learning described in SRL literature (Pintrich et al., 1991),

correlation results revealed negative associations for these variables. This may suggest that students who spend more time reading forum posts or accessing links engage more in individual online study than in offline collaboration. The strongest relationships were observed in CS-A, both in correlation (|r| = [0.26, 0.44]) and MI ([0.16, 0.21]). In contrast, CS-B displayed weaker and less consistent associations, with no feature surpassing a correlation magnitude of 0.2 and MI scores remaining negligible.

For time management (Table G.2, Appendix G), the most relevant features across metrics were related to online session patterns, particularly average session duration and spacing between sessions. These were especially strong in CS-A, with high F-scores (3.37 and 2.92) and meaningful correlations (r = 0.29 and r = –0.26), suggesting that students with stronger time management tend to access the platform more regularly and remain online longer per session. In CS-B, features related to content interaction and task completion also emerged as relevant. Notable examples include forum discussions as a percentage of the course mean (F = 2.09), assignment views (r = 0.24), and project submission delay (F = 1.44). Although these relationships were generally weaker than those observed in CS-A, they suggest that task-related engagement may serve as a more prominent behavioral signal of time management in Course B.

Effort regulation findings (Table G.3, Appendix G) closely align with time management, reflecting their conceptual proximity within SRL theory (Pintrich, 2000). In CS-A, effort regulation was most strongly associated with consistent platform engagement: percentage of course clicks (MI = 0.18), session duration irregularity (MI = 0.14), and spacing between sessions (F = 3.69), with strong negative correlations for spacing (r = –0.54) and for days without interaction (r = –0.45). These results suggest that students with higher effort regulation engage more steadily with the platform over time. CS-B highlighted different behavioral signals, with a focus on content interaction and task completion. Key features included TBL resource clicks (F = 2.22, r = 0.18), assignment views (MI = 0.11), and submission rate (F = 1.99, r = 0.20).

Although the associations found were modest overall, the analysis revealed several insights. CS-A consistently showed the strongest relationships across all three SRL subscales, especially in the correlation analysis, aligning with earlier findings (Figure 3.2) and indicating more regular LMS use despite fewer weekly classes. Effort regulation was most strongly linked to LMS activity, followed by time management. Both subscales were more clearly linked to behavioral regularity in CS-A, while CS-B revealed stronger signals related to task-driven interaction. Peer learning had the weakest associations, mostly tied to forum activity, underscoring the LMS's limited ability to capture offline or socially driven collaborative behaviors.

4.2. ML PIPELINE RESULTS

The findings of this section are structured around the three selection criteria defined in the methodology: (1) identification of the top three candidate models per dataset–target pair, (2) selection of the best model per dataset-target pair, and (3) selection of the best model overall per target.

4.2.1. Top 3 candidate models per dataset–target pair

Tables 4.4, 4.5, and 4.6 present the top three models for each SRL subscale across the two course-specific datasets (CS-A and CS-B), based on selection criterion 1. Each table includes the algorithm (from the seven tested), the modelling strategy (from the ten configurations), whether the model was selected before or after hyperparameter tuning, and the mean training and testing errors (MAE and RMSE) across folds.

Table 4.4 – Top 3 peer learning candidate models across datasets

Dataset	Algorithm	Strategy	Default (D) Tunned (T)	MAE		RMSE	
				Train	Test	Train	Test
CS-A	ElasticNet	FS+RO	D	1.18	1.23	1.236	1.40
	ElasticNet	FS	D	1.15	1.23	1.35	1.42
	ElasticNet	B	D	1.14	1.24	1.35	1.42
CS-B	Lasso	B	T	1.20	1.21	1.42	1.43
	Lasso	KPCA+CO	D	1.30	1.21	1.57	1.42
	kNN	KPCA	D	1.03	1.21	1.24	1.46

Table 4.5 – Top 3 time management candidate models across datasets

Dataset	Algorithm	Strategy	Default (D) Tunned (T)	MAE		RMSE	
				Train	Test	Train	Test
CS-A	Lasso	FS+RO	D	1.02	0.95	1.19	1.08
	ElasticNet	FS+RO	D	1.02	0.96	1.18	1.08
	Lasso	PCA+RO	D	1.02	0.96	1.18	1.08
CS-B	Lasso	FS+RO	T	0.81	0.83	0.98	1.01
	ElasticNet	PCA+RO	D	0.93	0.84	1.10	1.01
	Lasso	FS+CO	T	0.81	0.83	0.98	1.01

Table 4.6 – Top 3 effort regulation candidate models across datasets

Dataset	Algorithm	Strategy	Default (D) Tunned (T)	MAE		RMSE	
				Train	Test	Train	Test
CS-A	Lasso	FS+RO	D	0.84	0.75	1.05	0.96
	ElasticNet	FS+RO	D	0.81	0.76	1.00	0.95
	Lasso	B	T	0.75	0.77	0.95	0.97
CS-B	Ridge	KPCA+CO	T	0.64	0.73	0.82	0.94
	Lasso	B	T	0.74	0.75	0.95	0.96
	ElasticNet	B	T	0.74	0.75	0.95	0.96

Tables 4.4–4.6 reveal a clear dominance of parametric models across all dataset–target pairs, except kNN in CS-B for peer learning. This trend likely reflects that the relationships between features and targets were either largely linear or weak and sparse, as supported by the final EDA section. In such low-signal and small-sample settings, regularized linear models are typically more effective, as they offer a favorable bias–variance trade-off and generalize better than more flexible non-parametric alternatives (James et al., 2013).

Additional insights emerge from the impact of different ML techniques. Feature selection combined with random oversampling (FS+RO) was the most prevalent, particularly in CS-A, suggesting its suitability for smaller samples. In contrast, configurations using PCA/KPCA and CTGAN oversampling were more prominent in CS-B, likely due to its larger and more heterogeneous dataset. Hyperparameter tuning had a greater impact on effort regulation. In contrast, default models were often well-suited for peer learning and time management, indicating that tuning provided limited or target-dependent benefits.

Examining the three tables individually reveals important differences in predictive performance across SRL subscales. Peer learning consistently emerged as the most challenging dimension to predict, with test MAEs ranging from 1.21 to 1.24, values that represent substantial errors considering the 1–7 scale of the target variable. In contrast, time management models achieved lower error rates, with test MAEs typically between 0.83 and 0.96, while effort regulation models performed slightly better, with MAEs ranging from 0.73 to 0.77. These findings confirmed what was observed in the target-feature relationship analysis section, suggesting that the behavioral features extracted from LMS data may be more informative and reliable for capturing effort and time-related self-regulatory behaviors than for peer learning. Those behaviors might depend on more social or offline interactions not captured in the log data, as observed by prior literature (Cristea et al., 2024; van Sluijs & Matzat, 2024).

4.2.2. Best model per dataset–target pair and final best model per target

Tables 4.7, 4.8, and 4.9 present the best model for each target across the two course-specific datasets (CS-A and CS-B), based on the application of selection criterion 2 to the previously identified top three candidates. The structure mirrors the previous section, adding a "Selection Reason" column indicating whether the model was selected based on statistical significance or retained as a fallback without such evidence. The dataset showing the best overall fit for each target (according to selection criterion 3) is highlighted in bold, supporting the interpretation of how predictive performance varies across modelling perspectives. Statistical significance levels are indicated by the number of asterisks in the table footnotes.

Table 4.7 – Best model for peer learning across datasets

Dataset	Algorithm	Strategy	Selection Reason	MAE		RMSE	
				Train	Test	Train	Test
CS-A	**ElasticNet**	**FS**	**Wilcoxson MAE *****	**1.15**	**1.23**	**1.35**	**1.42**
CS-B	Lasso	B	Fallback	1.20	1.21	1.42	1.43

Selection criterion 3, significance win: *p < .1, ** p < .05, *** p < .01

Table 4.8 – Best model for time management across datasets

Dataset	Algorithm	Strategy	Selection Reason	MAE		RMSE	
				Train	Test	Train	Test
CS-A	Lasso	FS+RO	Wilcoxon MAE*	1.02	0.95	1.19	1.08
CS-B	**Lasso**	**FS+RO**	**Wilcoxon RMSE***	**0.81**	**0.83**	**0.98**	**1.01**

Selection criterion 3, significance win: *p < .1, ** p < .05, ***p < .01

Table 4.9 – Best model for effort regulation across datasets

Dataset	Algorithm	Strategy	Selection Reason	MAE		RMSE	
				Train	Test	Train	Test
CS-A	Lasso	FS+RO	Wilcoxon MAE*	0.84	0.75	1.05	0.96
CS-B	**Ridge**	**KPCA+CO**	**Wilcoxon MAE****	**0.64**	**0.73**	**0.82**	**0.94**

Selection criterion 3, significance win: *p < .1, ** p < .05, ***p < .01

The tables above reinforce earlier findings, confirming the consistent superiority of parametric models. All best-performing configurations across dataset–target pairs relied exclusively on regularized linear models, with no non-parametric alternatives retained. Among the SRL subscales, peer learning had already emerged as the most challenging to predict, consistently yielding the highest test errors. This difficulty is further supported by the absence of statistically significant improvements in most datasets. Models in the top three often performed similarly, and even in the only exception (CS-A), showed the highest train–test gaps (ΔMAE = 0.08; ΔRMSE = 0.07), suggesting lower generalization.

The models highlighted in bold represent the best overall fit for each target. For peer learning (Table 4.7), the top model was ElasticNet with feature selection, trained on CS-A, achieving a test MAE of 1.23 and RMSE of 1.42. For time management (Table 4.8), Lasso, combined with feature selection and random oversampling, trained on CS-B, yielded the best performance, with a test MAE of 0.83, RMSE of 1.01, and minimal train–test gaps of 0.02 and 0.03 for MAE and RMSE, respectively. Lastly, for effort regulation (Table 4.9), Ridge with KPCA and CTGAN oversampling, also trained on CS-B, achieved the lowest test error, with a test MAE of 0.73, RMSE of 0.94, and train-test gaps of 0.09 and 0.12 on MAE and RMSE, respectively. Although all final selections combined statistical support with the lowest combined test scores, this sometimes came at the cost of slightly higher overfitting, as observed in the effort regulation model.

A key pattern across these results is that, in cases where more than one dataset yielded significant models (e.g., time management and effort regulation), the final selection was based on the model with the lowest combined test error. In both instances, this favored Course B, which had a larger sample size (n = 129) compared to Course A (n = 32), reinforcing the impact of sample size on model stability and generalizability. In these cases, we also observed that the best-performing configurations employed feature selection or dimensionality reduction in combination with oversampling. This suggests that managing the high feature-to-observation ratio through dimensionality control and data augmentation was essential to achieving robust performance.

4.3. PREDICTIONS QUALITY ANALYSIS

To conclude this chapter, we discuss the predictive performance of the final models selected according to the third selection criterion for each SRL subscale. For each case, we conducted an error analysis to examine the distribution of predictions, followed by a feature interpretation using SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017) to understand which behavioral variables most influenced the model's output.

4.3.1. Peer learning

Figure 4.1 illustrates the distribution of true and predicted values, along with the absolute errors for the selected final model.

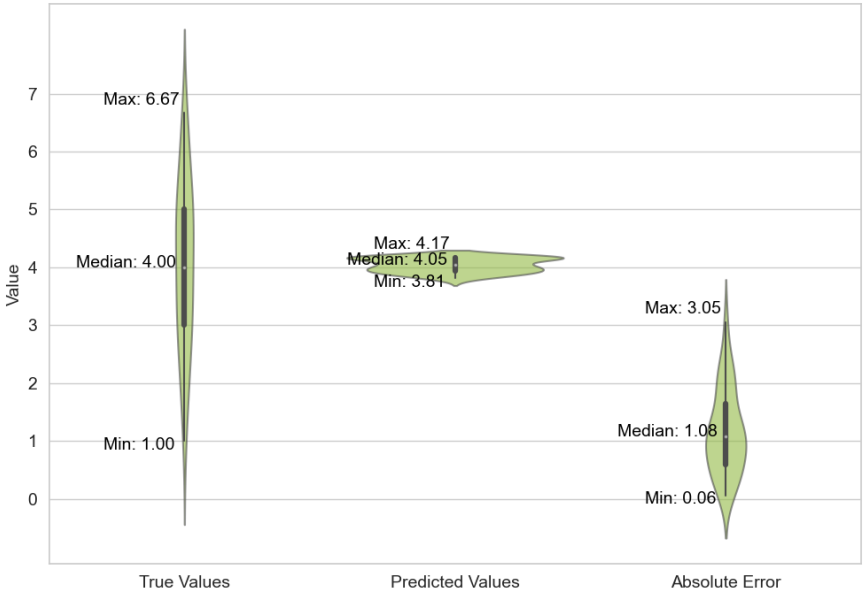


Figure 4.1 – Distribution of true and predicted values along with absolute error for the best peer learning model

The true peer learning values range from 1.00 to 6.67 (median = 4.00), showing substantial variability across student responses. In contrast, the predicted values are narrowly concentrated between 3.81 and 4.17 (median = 4.05), indicating the model's inability to capture the target's true variance and its tendency to produce median predictions. The absolute errors, ranging from 0.06 to 3.05 (median = 1.08), are high relative to the 1–7 scale and confirm the relatively challenging predictive performance already discussed in the previous section (test MAE = 1.23).

To explore whether the model struggles more with certain student profiles, an error analysis by final grade group was conducted (Table 4.10). Students were categorized according to the institutional grading system[6]. For each group (of size n), the table reports the mean final grade, true and predicted peer learning scores, prediction bias (predicted – true), and the mean and standard deviation of absolute errors.

[6] The ECTS grading scale and its conversion table for NOVA IMS can be found in the official institutional guide: NOVA Information Management School. *Grading System – ECTS Guide*. Retrieved from: https://www.novaims.unl.pt/media/3hlfngn0/ects.pdf

Table 4.10 – Error analysis by student group for the best peer learning model

Student Group	Mean				Absolute Error	
	Final Grade	True Value	Predicted Value	Bias	Mean	Std.
A (n=3)	18.00	5.56	4.06	-1.49	1.49	0.58
B (n=11)	16.45	3.91	4.03	0.12	1.20	0.79
C (n=3)	15.00	4.22	4.07	-0.15	0.94	0.29
D (n=8)	13.50	4.08	4.06	-0.03	1.43	0.81
F (n=7)	6.00	3.48	4.02	0.54	0.74	0.71

The true values of most groups follow the expected trend, where higher final grades are associated with stronger SRL skills, particularly among excellent (group A) and failing students (group F). The remaining show similar but ordered scores, except for group B, which deviates from this pattern.

The results confirm the model's tendency to regress predictions toward the median, underestimating true high scores (e.g., groups A, C, and D) and overestimating true low ones (e.g., groups F and B). This pattern aligns with the previously observed low variance in predictions, limiting the model's ability to capture finer distinctions across student groups.

These findings highlight that although the model is statistically significant, its practical utility for predicting peer learning is limited. This may be due to a weak relationship between the features and the target, a small sample size (n = 32), and the social, offline nature of peer interactions, which is not fully captured in LMS log data.

To better understand how these limitations influence predictions, Figure 4.2. presents a SHAP values summary plot (following the standard implementation[7]). Each feature is listed from top to bottom in order of importance. Along each line, every dot represents an individual prediction. The horizontal position of the dot indicates the SHAP value, reflecting the feature's impact on that specific prediction. The color gradient represents the original feature value, ranging from blue (low) to pink (high). Additionally, we complement this analysis by presenting SHAP mean absolute values and linear model coefficients to understand the overall magnitude and direction of each feature's contribution to predictions.

[7] https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html

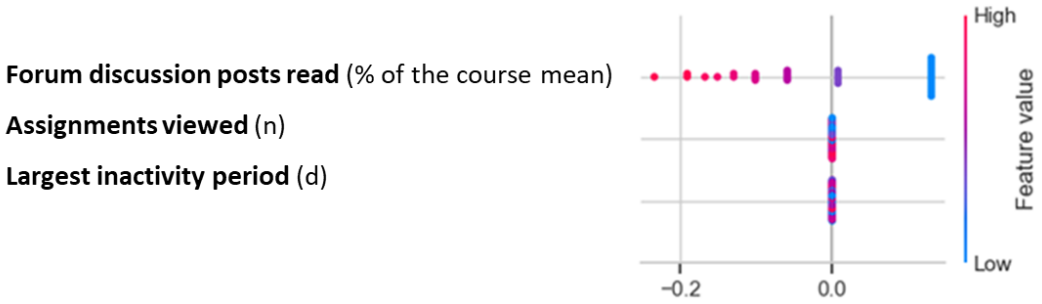


Figure 4.2 – SHAP summary plot of feature importance for the best peer learning model

The final model retains the minimal features, with only the percentage of forum posts read showing a moderate impact on predictions. Higher values are associated with lower predicted peer learning scores, as reflected by the ElasticNet coefficient (–0.12). The mean absolute SHAP value (0.12) further confirms its strong and consistently negative contribution to the model's predictions. This aligns with the results in Section 4.1.2, where this feature was found to correlate negatively with peer learning. It does not invalidate prior literature findings. While previous studies link forum activity with collaborative learning behaviors (Alhazbi et al., 2024), this metric captures passive engagement (reading) rather than active participation (e.g., posting). Therefore, its negative contribution may indicate that students who read more forum content engage less in offline activities, potentially replacing collaborative dialogue with individual online strategies.

4.3.2. Time management

Figure 4.3 shows the distribution of the true and predicted values, along with the absolute errors for the selected final model.

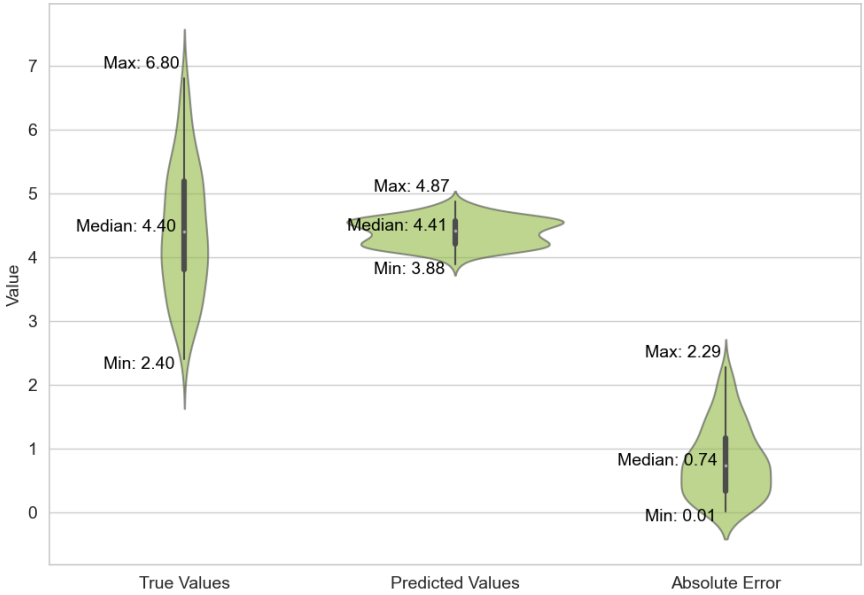


Figure 4.3 – Distribution of true and predicted values along with absolute error for the best time management model

The true time management values range from 2.40 to 6.80 (median = 4.40), generally higher than those observed for peer learning, though not directly comparable due to differences in the underlying samples (Course B vs Course A). The variability across students is moderate, but predicted values remain narrowly concentrated between 3.88 and 4.87 (median = 4.41). That highlights the model's tendency to regress toward the median, however, in this case, with a slightly wider spread than in peer learning. Absolute errors range from 0.01 to 2.29 (median = 0.74), reflecting a model with more potential, as also supported by the lower test MAE (0.83), though the overall predictive power remains limited.

A detailed error analysis was also conducted for the time management model (Table 4.11), following the same approach applied to peer learning (Table 4.10).

Table 4.11 – Error analysis by student group for the best time management model

Student Group	Mean				Absolute Error	
	Final Grade	True Value	Predicted Value	Bias	Mean	Std.
A (n=24)	18.42	4.82	4.50	-0.32	0.97	0.65
B (n=21)	17.00	4.30	439	0.08	0.76	0.56
C (n=46)	15.72	4.42	4.39	-0.03	0.81	0.59
D (n=24)	13.54	3.98	4.37	0.40	0.67	0.49
E (n=7)	11.29	4.49	4.30	-0.19	0.98	0.36
F (n=7)	8.00	4.40	4.28	-0.12	0.67	0.40

A pattern observed in the peer learning analysis is more evident here: true mean values do not increase with final grade as expected and fall within a narrower range (3.98–4.82). Most group means deviate from the expected order, except for groups A, E, and F. This suggests that final grades are weakly aligned with self-reported time management, likely due to the subjective and biased nature of questionnaire data. This not only limits the reliability of the target but also makes it harder to predict from objective LMS behaviors, as the two may not reflect the same underlying construct.

The predicted means generally increase with final grade, but the model produces relatively narrow estimates across groups (ranging from 4.28 to 4.50), reinforcing the earlier observation of regression toward the median. Bias values are smaller than those observed in peer learning, with the most notable cases being the excellent students (group A: –0.32) and the satisfactory group (D: 0.40).

The MAE values are consistently below 1.00, with the best performance observed in groups D and F (MAE = 0.67) and the highest errors in groups A and E (0.97 and 0.98, respectively). Although direct comparison with the peer learning model is limited by differences in sample size and composition, the lower error levels may suggest comparatively better predictive performance. This may be due to several factors: the larger sample size (n = 129) improving model stability; time management behaviors being better captured by LMS logs, as evidenced by the literature (Baker et al., 2020); and limited variance in responses, which favors models predicting near the median. Therefore, this apparent predictive advantage should be interpreted with caution, as it does not necessarily reflect an accurate understanding of individual behavioral differences.

Figure 4.4 presents the SHAP feature importance plot. As in the peer learning model, the final model retains only three features. The most influential feature, based on the highest mean absolute SHAP value (1.03), is the average spacing between online sessions. The lasso regression coefficient (-0.10) supports the distribution observed in the summary plot, where shorter intervals between sessions (i.e., more regular engagement) contribute positively to predictions, likely reflecting a more organized study routine. This finding align with prior literature findings Li et al. (2020), who highlight temporal engagement patterns, such as frequency and regularity of access, as important behavioral indicators of time management skills, and van Sluijs & Matzat (2024), who also found metrics like session intervals predictive of time management in blended learning settings.

Additionally, the number of assignments viewed also shows a significant impact. Although it appears at the top of the SHAP summary plot due to high variability in individual contributions, its overall effect is reflected in a lower mean absolute SHAP value of 0.62. The positive regression coefficient (0.14) further confirms that students who engage more frequently with assignment-related content tend to show stronger time management skills.

Lastly, the number of clicks on the course page per week shows a more limited positive influence, with a mean absolute SHAP value of 0.07 and a coefficient of 0.02. This supports the behavioral patterns identified by the other two features. The significance of content interaction features aligns with the findings in Section 4.1.2 for Course B, where such metrics are more strongly associated with time management.

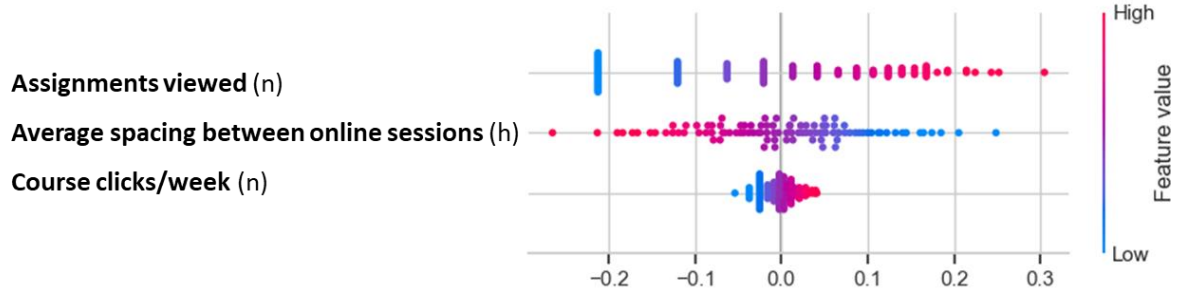


Figure 4.4 – SHAP summary plot of feature importance for the best time management model

4.3.3. Effort regulation

Figure 4.5 shows the distribution of the true and predicted values, along with the absolute errors for the selected final model.

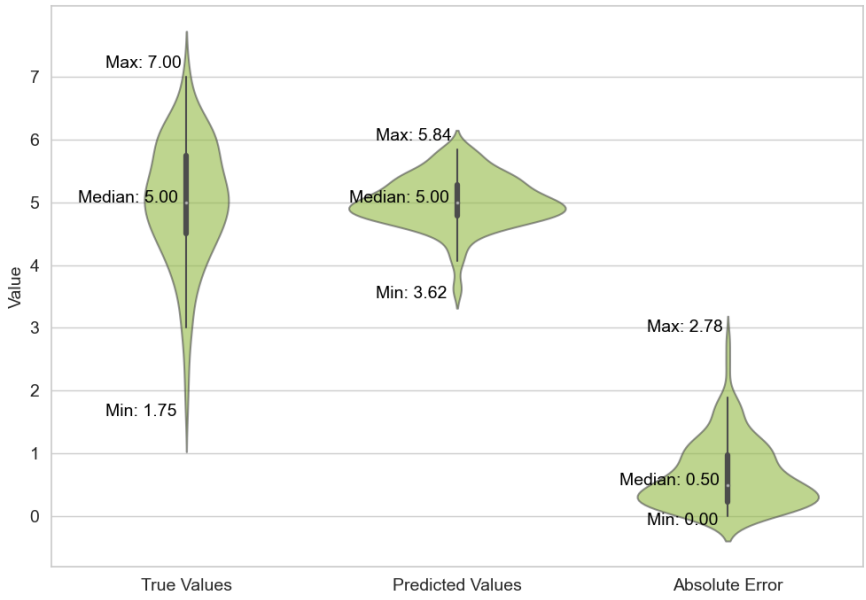


Figure 4.5 – Distribution of true and predicted values along with absolute error for the best effort regulation model

The true values for effort regulation range from 1.75 to 7.00 (median = 5.00), indicating substantial variability among students' responses, greater than time management but slightly less than peer learning, though not directly comparable. Therefore, it is the subscale with the highest median. Predicted values, while slightly more dispersed (ranging from 3.62 to 5.84, median = 5.00), remain concentrated around the center, reflecting the model's limited sensitivity to extreme cases.

Absolute errors range from 0.00 to 2.78, with a median of 0.50, the lowest among the three subscales (peer learning: median = 1.08; time management: median = 0.74). This may indicate superior performance in terms of error magnitude, which is further supported by the test MAE of 0.73, making effort regulation the best-predicted subscale with the available data. This outcome may be due to factors similar to those observed in the time management model, as well as the use of kernel-based dimensionality reduction, which likely helped the model capture more complex structures and generalize more effectively.

Nevertheless, the regression to the median and the limited spread of predicted values indicate that the model still struggles to distinguish students with particularly high or low effort regulation, highlighting a limited explanatory capacity despite better overall performance.

To explore whether the model struggles more with certain student profiles, an error analysis by final grade group was conducted (Table 4.12), using the same approach applied to the other SRL subscales.

Table 4.12 – Error analysis by student group for the best effort regulation model

Student Group	Mean				Absolute Error	
	Final Grade	True Value	Predicted Value	Bias	Mean	Std.
A (n=24)	18.42	5.26	5.02	-0.24	0.69	0.45
B (n=21)	17.00	5.23	5.11	-0.12	0.46	0.43
C (n=46)	15.72	4.97	4.99	0.02	0.66	0.53
D (n=24)	13.54	4.92	4.98	0.07	0.73	0.59
E (n=7)	11.29	4.71	5.00	0.28	0.82	0.89
F (n=7)	8.00	4.43	4.95	0.52	0.59	0.43

The true values show a general trend of decreasing effort regulation as final grades decrease, particularly evident between group A (mean true = 5.26) and group F (mean true = 4.43). However, the differences between groups are relatively small, which may limit the model's ability to learn distinctive patterns. This is reflected in the predicted means, which range narrowly from 4.95 to 5.11 and do not consistently follow the expected grade order (e.g., groups B and E). These findings reinforce the model's tendency to regress toward the median, as seen before in the other SRL subscales.

In terms of prediction bias, the model slightly underestimates scores for high-performing students (e.g., group A: –0.24) and overestimates them for lower-performing ones (e.g., group F: +0.52), reflecting the general pattern across subscales. However, these biases are more moderate than those observed in peer learning (e.g., group A: –1.49; group F: +0.54) and comparable or smaller than those in time management (e.g., group A: –0.32; group D: +0.40). This suggests that the effort regulation model maintains a more balanced bias across student profiles, contributing to its comparatively better overall performance.

The absolute error analysis shows relatively consistent performance across groups, with mean errors below 1.00 in all cases, similar to time management. The lowest error was observed in group B (MAE = 0.46), and the highest in group E (MAE = 0.82), though the small sample size

in group E (n = 7) may contribute to this variability. This further supports the comparatively better predictive quality of this model.

The SHAP analysis was not conducted for this target because KPCA was used, which transforms the input features into a nonlinear space, thereby preventing the direct interpretation of individual feature contributions.

4.3.4. Revisiting the research questions

Across all SRL subscales, the final models exhibited a common limitation: predictions regressed toward the median, underestimating high scores and overestimating low ones. While results should be interpreted with caution due to differences in sample sizes, features, and modelling strategies, they offer useful insights for addressing the research questions.

RQ1: How does predictive performance vary across SRL subscales?

Treating each target as an independent regression task allows us to compare performances across subscales. We demonstrated that effort regulation was predicted with the lowest test error (MAE = 0.73), followed by time management (MAE = 0.83). In contrast, peer learning was the most difficult to predict (MAE = 1.23), likely due to its offline, socially driven nature, which is less visible in LMS clickstream data.

RQ2: How does course context influence predictive performance?

Although previous studies have highlighted the role of course context in SRL prediction (Conijn et al., 2017; Gašević et al., 2016; van Sluijs & Matzat, 2024), in this study, its effect is difficult to separate from sample imbalance. Course B models often outperformed those of Course A, possibly due to its larger sample size (n = 129 vs. 32), despite students in Course A being more active on the LMS, displaying stronger SRL behavior associations. However, the most effective peer learning model came from Course A, which included more TBL activities that were linked to collaborative learning.

5. CONCLUSIONS AND FUTURE WORKS

Precision education has emerged as an alternative to the “one-size-fits-all” paradigm, using LMS data to support personalized learning experiences. Among its research applications, performance prediction stands out. However, its theoretical foundation remains limited, making it challenging to interpret how LMS-derived behaviors relate to meaningful learning processes.

SRL theory provides a framework to address this gap. However, self-report instruments, commonly used to measure SRL, are limited by issues of bias, subjectivity, and practicality. Recent research (van Sluijs & Matzat, 2024) suggests that LMS clickstream data may serve as more objective proxies for certain SRL dimensions, such as time management, thereby offering a viable alternative to traditional self-report methods. Building on this, our study investigates whether clickstream data from the Moodle LMS at NOVA IMS can predict students’ SRL skills, as measured by the MSLQ. Specifically, it examines how predictive performance varies across SRL subscales and how course context and interaction patterns influence model effectiveness, even when courses are similarly structured. More than just an alternative to traditional assessment methods, this approach opens new directions in precision education as it may improve performance prediction by incorporating inferred SRL dimensions.

We collected data from two graduate courses at NOVA IMS (Course A and Course B) during the first semester of the 2024/2025 academic year and assessed the predictive power of Moodle clickstream data across three SRL subscales (peer learning, time management, and effort regulation) using two course-specific datasets (CS-A and CS-B). A structured ML pipeline was applied to all six dataset–target pairs, testing ten configurations that combined feature selection, dimensionality reduction, and data augmentation. Seven algorithms were evaluated per pair, generating 70 models. Under defined selection criteria, the top three were shortlisted, tuned, and the best model was selected. For each SRL subscale, we identified the most effective dataset by comparing the best models.

Can SRL skills be accurately predicted using NOVA IMS Moodle clickstream data? Results showed moderate predictive performance, with models consistently regressing toward the median, a pattern further confirmed by error analyses across student grade groups. Parametric models consistently outperformed non-parametric models, suggesting either low signal strength or predominantly linear relationships between LMS clickstream data and SRL subscales. Among the best-performing models, Course B dataset yielded more significant results, possibly due to its larger sample size. Among the subscales, effort regulation showed the best predictive performance according to the mean absolute error on test (MAE = 0.73), followed by time management (MAE = 0.83), while peer learning was the most challenging to model (MAE = 1.23), likely due to its social and offline nature.

Despite its contributions, this study represents an initial exploration and involves several methodological simplifications. As such, the findings should be interpreted with caution. Several limitations should be considered when assessing the generalizability of the results.

First, although previous studies have highlighted the role of course context in SRL prediction (Conijn et al., 2017; Gašević et al., 2016; van Sluijs & Matzat, 2024), it remains unclear whether the performance differences observed in this study truly reflect contextual effects or are partly driven by sample imbalance. In our case, only 161 out of 295 students presented a valid response to the questionnaire, with just 32 from Course A. This imbalance may help explain why Course B often outperformed Course A in model comparisons, despite Course A student's being more active on the LMS (see Figures 3.2 and 3.3) and showing stronger SRL associations (Section 4.1.2).

Second, relying solely on LMS data excludes offline and collaborative aspects of SRL, such as those involved in peer learning. Prior studies have noted the limitations of clickstream data in capturing these dimensions (Cristea et al., 2024; van Sluijs & Matzat, 2024). This limitation was also evident in our results. Although the peer learning subscale demonstrated the highest internal consistency among the assessed dimensions (Cronbach's $\alpha = 0.72$), its predictions based on LMS-derived features were considerably weaker. This suggests that, despite the reliability of the self-report data, the behavioral traces available in the LMS may be insufficient to capture the collaborative nature of peer learning.

Third, the relationship between features and targets was consistently weak. The final EDA revealed low MI and correlation scores across modelling perspectives, and most final models retained only the minimum three features despite the use of a robust voting-based selection strategy. This points to limited predictive value in the LMS-derived variables. A likely explanation is the static nature of the feature representation, which fails to reflect the temporal and adaptive dynamics central to SRL.

Fourth, a known limitation from the literature, also evident in our study, is the lack of theoretical validation for many LMS-derived indicators. Although we combined behavioral data with MSLQ responses to strengthen construct validity, the two data sources inherently capture distinct measures of SRL: stable perceptions versus dynamic behaviors (Fan et al., 2022). This misalignment may help explain the limited predictive performance observed.

Finally, the use of self-reported targets introduces bias and measurement error, as students' perceptions may not align with their actual behaviors (Credé & Phillips, 2011; Schellings & Van Hout-Wolters, 2011).

Future research should build on this initial exploration by expanding the sample to include more courses and academic terms, thereby improving the generalizability of results. To better capture the complexity of SRL, studies should also incorporate complementary data sources

(Cristea et al., 2024; van Sluijs & Matzat, 2024), such as forum activity, assignment metadata, and group structures, that account for social and offline behaviors. Additionally, using time-dependent feature representations and tracking students across semesters (Alhazbi et al., 2024) may offer a more accurate view of SRL as a dynamic and adaptive process. Finally, strengthening the theoretical alignment of LMS-based indicators and validating them across diverse contexts (Alhazbi et al., 2024; van Sluijs & Matzat, 2024) will be essential to developing robust, interpretable, and transferable SRL models for precision education.

BIBLIOGRAPHICAL REFERENCES

Ahmad Uzir, N., Gašević, D., Matcha, W., Jovanović, J., & Pardo, A. (2020). Analytics of time management strategies in a flipped classroom. *Journal of Computer Assisted Learning*, *36*(1), 70–88. https://doi.org/10.1111/jcal.12392

Alhazbi, S., Al-ali, A., Tabassum, A., Al-Ali, A., Al-Emadi, A., Khattab, T., & Hasan, M. A. (2024). Using learning analytics to measure self-regulated learning: A systematic review of empirical studies in higher education. *Journal of Computer Assisted Learning*, *40*(4), 1658–1674. https://doi.org/10.1111/jcal.12982

Asarta, C. J., & Schmidt, J. R. (2013). Access Patterns of Online Materials in a Blended Course. *Decision Sciences Journal of Innovative Education*, *11*(1), 107–123. https://doi.org/10.1111/j.1540-4609.2012.00366.x

Azcona, D., Hsiao, I.-H., & Smeaton, A. F. (2019). Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. *User Modeling and User-Adapted Interaction*, *29*(4), 759–788. https://doi.org/10.1007/s11257-019-09234-7

Baker, R., Evans, B., Li, Q., & Cung, B. (2018). Does Inducing Students to Schedule Lecture Watching in Online Classes Improve Their Academic Performance? An Experimental Analysis of a Time Management Intervention. *Research in Higher Education*, *60*(4), 521–552. https://doi.org/10.1007/s11162-018-9521-3

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes.

International Journal of Educational Technology in Higher Education, *17*(1), 13. https://doi.org/10.1186/s41239-020-00187-1

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education*, *12*(1), 1–6. https://doi.org/10.1016/j.iheduc.2008.10.005

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press. https://www.jstor.org/stable/j.ctt183ph6v

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(10), 281–305.

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: A Pre-processing Approach for Imbalanced Regression. *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 36–50. https://proceedings.mlr.press/v74/branco17a.html

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees* (https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-olshen-charles-stone; 1st ed.). Chapman and Hall / Wadsworth. https://doi.org/10.1201/9781315139470

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review.

The Internet and Higher Education, *27*, 1–13. https://doi.org/10.1016/j.iheduc.2015.04.007

Calvo-Flores, M., Gibaja, E., Pegalajar Jiménez, M. del C., & Pérez, O. (2006). Predicting Students' Marks from Moodle Logs using Neural Network Models. *Current Developments in Technology-Assisted Education*.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., & Lindstädt, S. (2018). Finding traces of self-regulated learning in activity streams. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 191–200. https://doi.org/10.1145/3170358.3170381

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Credé, M., & Phillips, L. A. (2011). A meta-analytic review of the Motivated Strategies for Learning Questionnaire. *Learning and Individual Differences*, *21*(4), 337–346. https://doi.org/10.1016/j.lindif.2011.03.002

Cristea, T., Snijders, C., Matzat, U., & Kleingeld, A. (2024). Unobtrusive measurement of self-regulated learning: A clickstream-based multi-dimensional scale. *Education and Information Technologies*, *29*(11), 13465–13494. https://doi.org/10.1007/s10639-023-12372-6

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 6–14. https://doi.org/10.1145/2883851.2883931

Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research*, *281*(3), 575–587. https://doi.org/10.1016/j.ejor.2019.03.037

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, *7*, 1–30.

Fan, Y., Van Der Graaf, J., Lim, L., Raković, M., Singh, S., Kilgour, J., Moore, J., Molenaar, I., Bannert, M., & Gašević, D. (2022). Towards investigating the validity of measurement of self-regulated learning based on trace data. *Metacognition and Learning*, *17*(3), 949–987. https://doi.org/10.1007/s11409-022-09291-1

Fisher, Ronald. A. (1950). *Statistical Methods for Research Workers* (https://ia601406.us.archive.org/2/items/in.ernet.dli.2015.137901/2015.137901.Statistical-Methods-For-Research-Workers-Thirteenth-Edition_text.pdf; 13th ed.). Oliver and Boyd.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46 (1-3)*, 389–422. https://doi.org/10.1023/A:1012487302797

Hart, S. A. (2016). Precision Education Initiative: Moving Toward Personalized Education. *Mind, Brain, and Education*, *10*(4), 209–211. https://doi.org/10.1111/mbe.12109

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.2307/1271436

Hoic-Bozic, N., Mornar, V., & Boticki, I. (2009). A Blended Learning Approach to Course Design and Implementation. *IEEE Transactions on Education*, *52*(1), 19–30. IEEE Transactions on Education. https://doi.org/10.1109/TE.2007.914945

Holmes, B., & Gardner, J. (2006). *E-learning: Concepts and practice*. SAGE Publications.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jo, I.-H., Park, Y., Yoon, M., & Sung, H. (2016). Evaluation of Online Log Variables that Estimate Learners' Time Management in a Korean Online Learning Context. *The International Review of Research in Open and Distributed Learning*, *17*(1). https://doi.org/10.19173/irrodl.v17i1.2176

Kaya, H., Kaya, N., Palloş, A. Ö., & Küçük, L. (2012). Assessing time-management skills in terms of age, gender, and anxiety levels: A study on nursing and midwifery students in

Turkey. *Nurse Education in Practice*, *12*(5), 284–288. https://doi.org/10.1016/j.nepr.2012.06.002

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences*, *9*, 3093. https://doi.org/10.3390/app9153093

Li, Q., Baker, R., & Warschauer, M. (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, *45*, 100727. https://doi.org/10.1016/j.iheduc.2020.100727

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55–55.

Luan, H., & Tsai, C.-C. (2021). A Review of Using Machine Learning Approaches for Precision Education. *Educational Technology & Society*, *24*(1), 250–266.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

McKinney, W. (2018). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media, Inc.

Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation* (pp. 451–502). Academic Press. https://doi.org/10.1016/B978-012109890-2/50043-3

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology. *Journal of Educational Psychology*, *82*(1), 33–40. https://doi.org/10.1037/0022-0663.82.1.33

Pintrich, P. R., Smith, D. A., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement - EDUC PSYCHOL MEAS*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Pintrich, P. R., Smith, D., Duncan, T., & Mckeachie, W. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). *Ann Arbor. Michigan*, *48109*, 1259.

Riestra-González, M., Paule-Ruíz, M. D. P., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, 104108. https://doi.org/10.1016/j.compedu.2020.104108

Santos, R., & Henriques, R. (2023a). *GROUPING BACHELOR'S STUDENTS ACCORDING TO THEIR MOODLE INTERACTION PROFILES: A K-MEANS CLUSTERING APPROACH*. 7383–7389. https://doi.org/10.21125/edulearn.2023.1920

Santos, R., & Henriques, R. (2023b). *PREDICTING STUDENT PERFORMANCE FROM MOODLE LOGS IN HIGHER EDUCATION: A COURSE-AGNOSTIC APPROACH*. International Conference on Education and New Developments. https://doi.org/10.36315/2023v2end016

Schellings, G., & Van Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments: Theoretical and empirical considerations. *Metacognition and Learning*, *6*(2), 83–90. https://doi.org/10.1007/s11409-011-9081-9

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, *10*(5), 1299–1319. https://doi.org/10.1162/089976698300017467

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611. https://doi.org/10.2307/2333709

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van Sluijs, M., & Matzat, U. (2024). Predicting time-management skills from learning analytics. *Journal of Computer Assisted Learning*, *40*(2), 525–537. https://doi.org/10.1111/jcal.12893

Veenman, M. V. J. (2005). *The assessment of metacognitive skills: What can be learned from multi-method designs?* https://dare.uva.nl/search?identifier=5e20ffb2-a4e6-49a5-a67f-18a75dc112ec

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, *24*(1), 175–186. https://doi.org/10.1007/s00521-013-1368-0

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Williamson, B. (2019). Digital policy sociology: Software and science in data-intensive precision education. *Critical Studies in Education*, *62*, 1–17. https://doi.org/10.1080/17508487.2019.1691030

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, *8*(4), 327–353. https://doi.org/10.1016/S1041-6080(96)90022-9

Winne, P., & Hadwin, A. (1998). Studying as Self-Regulated Learning. In *Metacognition in Educational Theory and Practice* (Vol. 93, pp. 277–304).

Xu, J., Du, J., & Fan, X. (2013). “Finding our time”: Predicting students’ time management in online collaborative groupwork. *Computers & Education*, *69*, 139–147. https://doi.org/10.1016/j.compedu.2013.07.012

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling Tabular data using Conditional GAN* (No. arXiv:1907.00503). arXiv. http://arxiv.org/abs/1907.00503

Yeo, I., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, *27*, 44–53. https://doi.org/10.1016/j.iheduc.2015.05.002

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a Structured Interview for Assessing Student Use of Self-Regulated Learning Strategies. *American Educational Research Journal*, *23*, 614–628. https://doi.org/10.3102/00028312023004614

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

APPENDIX A. LITERATURE REVIEW TABLE – RELATED WORK

This section of the appendices presents Table A.1 that summarizes a subset of the studies discussed in Section 2.3.2 of the literature review. The selection focuses on works most closely aligned with our research objectives and methodological approach. Broader systematic reviews and studies with limited applicability to our context were excluded to maintain relevance and coherence.

Table A.1 – Literature review: relevant studies

Reference	Objective(s)	Participants	Data Source(s)	Features	Methodology	Key Findings	Limitations
Ahmad Uzir et al. (2020)	• Identify **time management strategies** from **trace data** • Investigate their **association** with **academic performance**	1,134 students over 3 years in 1 computer engineering course	• **Trace data** from flipped classrooms online preparatory learning activities • **Midterm** and final **exam scores**	• **Time of activity completion:** preparing, ahead, revisiting, and catching up • **Learning session data** encoded into sequences of **learning modes**	• **Hierarchical clustering:** group learning mode sequences into time management tactics and strategies • **Markov chains**: analyze transitions in learning modes • **Inferential statistics**: associate time management strategies with academic performance	• **4 time management tactics:** mixed and short, revisiting, short preparing, and long preparing • **3 strategy groups**: comprehensive and active; selective and active and limited activity • Comprehensive and Active strategy is linked to higher performance	• **Limit generalizability:** focuses only on flipped classroom settings • **Not account** for **external factors** (e.g., personal, social, or technical challenges) affecting time management

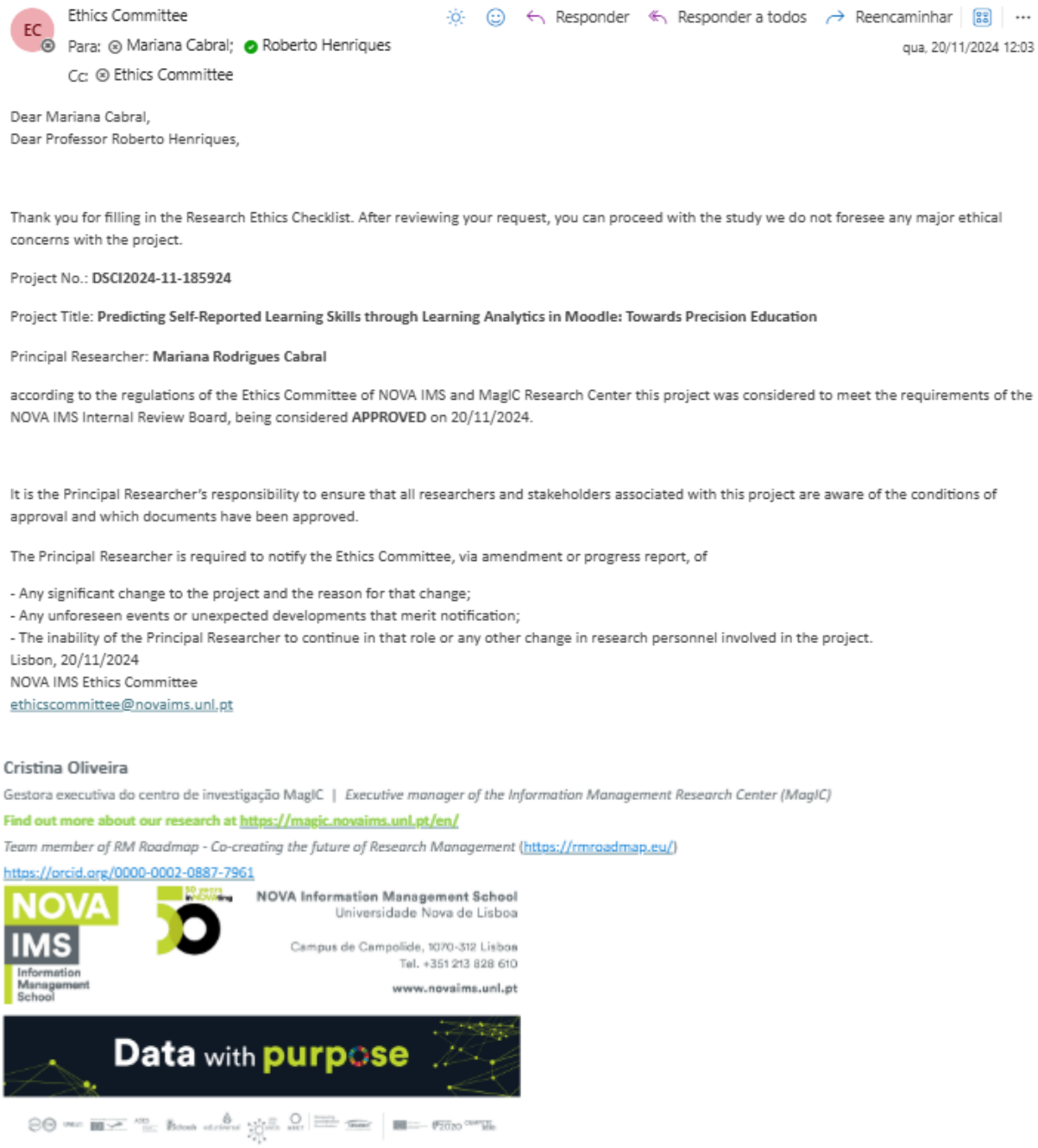
Cristea et al. (2024)	• Develop **clickstream-based scales** for measuring **SRL phases** in online learning • Create a **reliable**, **valid**, and **portable scale**	757 students, 4 courses at a Dutch university	• **LMS clickstream data** • **MSLQ SRL scale** • **Final grades**	LMS indicators related to the 4 COPES phases: **task definition**, **goal-setting**, **enactment**, and **adaptation** (e.g. total clicks, time spent on resources)	• **Mapping LMS** indicators into the **4 COPES phases** • **PCA** and **Cronbach's alpha:** reliability and validity • **Portability test:** scales across similar and dissimilar courses • **Correlation Analysis:** scales, final grades, and SRL survey results	• **4 reliable SRL scales** • **Strong portability** for **enactment** and **adaptation** • **Enactment** phase correlated with final grades (r = 0.38) • **Goal-setting** showed some **inconsistencies** and **lower portability** • **Clickstream-based scales** provided **better predictive power** for **academic performance** compared to traditional self-reported surveys	• **Subjectivity interpreting COPES model** phases and assigning **indicators** • **Lack of external validation** or convergence with **established SRL scales**
Li et al (2020)	• Examine the **validity** of **clickstream data** for measuring **time management** and **effort regulation**	238 students enrolled in a 10 week fully online chemistry course at a public university	• **Clickstream data:** LMS Canvas	• **Time management:** study on time (% units accessed before deadlines); study in advance (time difference between 1st access and deadlines); spacing (distribution of study sessions) • **Effort regulation:** change in time on task • **MSLQ metrics**	• **Correlation Analysis**: assess the relationship between survey and clickstream measures	• **Pre-course self-reported measures not correlate** with **clickstream measures** or **predict performance**	• **Limited generalizability** • **Clickstream data:** not measure internal cognitive process

	• **Improve performance prediction** compared to self-reported measures		• **Adapted MSLQ** to measure **time management** and **effort regulation**: pre- and post-course • **Performance data:** final exam scores, and grades from other courses	• **Grades**	• **Regression Analysis:** predictive power of clickstream and survey for course performance • Pre- and post-course survey comparison with clickstream for validity and prediction improvement	• **Post-course survey** measures showed **moderate alignment with clickstream data** and were **predictive** of **performance** • **Clickstream data improved predictions of course performance over survey data alone,** especially for studying on time, studying in advance, and changing time on task	• **Pre-course surveys:** may suffer from overconfidence and inaccuracy
Jo et al. (2016)	• Evaluate the **relationship** between time and study environment management **(TSEM)** and **online behavioral patterns** • Investigate the **impact** of that on **academic performance**	124 students enrolled in a management statistics course at a private university in Seoul, Korea	• **Adaptation of MSLQ focusing on TSEM:** 8 questions scored on a five-point Likert scale • **LMS log data:** extracted weekly • **Course grading system:** course Scores	• **Dependent: l**ogin frequency, regularity, total login time • **Independent:** TSEM scores from the MSLQ survey • **Final course scores**	• **Structural Equation Modelling:** assess relationships between TSEM, LMS log data, and final scores • **Sobel Test:** evaluate mediation effects of online behavioral patterns	• **TSEM significantly influenced online behavioral patterns:** higher TSEM correlated with longer login times and higher frequency • **Online behavioral patterns mediated the relationship between TSEM and final scores**	• **Limited Population:** only female and 1 online course • **Limited SRL behaviors analyzed**: only on TSEM

					• **Descriptive statistics:** ensure data normality	• **TSEM alone:** no direct effect on final scores	
van Sluijs & Matzat (2024)	• Use **LMS trace data** to **predict self-reported time management skills** • **Evaluate the portability** of the models across different courses	679 students from 3 bachelor's and 4 master's programs at a Dutch technical university	• **Clickstream data:** LMS Canvas • **Adapted MSLQ** to measure **time management**	• **Survey features** • **Clickstream Features:** – Number of clicks, sessions, and files accessed/downloaded. – Session intervals, login irregularity, and forum activity. – Start times of sessions and session length	• **Linear Regression Models:** individual courses and the overall dataset • **Multi-Level Regression Models:** examined model portability across courses	• **LMS data successfully predicted self-reported time-management skills in some courses** • **Predictive power varied significantly between courses:** explained variances ranging from 14% to 71% • **Key features:** login irregularity, session intervals, and files accessed	• **Lack of model portability:** course-specific factors heavily influenced model effectiveness • Data did **not capture offline activities** or cognitive processes, limiting insights into full learning behaviors

APPENDIX B. ETHICS APPROVAL EMAIL

This appendix includes Figure B.1, which shows a screenshot of the email confirming the ethical approval of this project, granted under reference code DSCI2024-11-185925.



Ethics Committee

Responder | Responder a todos | Reencaminhar

Para: Mariana Cabral; Roberto Henriques

Cc: Ethics Committee

qua, 20/11/2024 12:03

Dear Mariana Cabral,
Dear Professor Roberto Henriques,

Thank you for filling in the Research Ethics Checklist. After reviewing your request, you can proceed with the study we do not foresee any major ethical concerns with the project.

Project No.: **DSCI2024-11-185924**

Project Title: **Predicting Self-Reported Learning Skills through Learning Analytics in Moodle: Towards Precision Education**

Principal Researcher: **Mariana Rodrigues Cabral**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 20/11/2024.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 20/11/2024
NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

Cristina Oliveira
Gestora executiva do centro de investigação MagIC | *Executive manager of the Information Management Research Center (MagIC)*
Find out more about our research at https://magic.novaims.unl.pt/en/
Team member of RM Roadmap - Co-creating the future of Research Management (https://rmroadmap.eu/)
https://orcid.org/0000-0002-0887-7961
NOVA Information Management School
Universidade Nova de Lisboa
Campus de Campolide, 1070-312 Lisboa
Tel. +351 213 828 610
www.novaims.unl.pt

Data with purpose

Figure B.1 – Screenshot of the email confirming ethical approval of the thesis

APPENDIX C. MSLQ ITEMS FOR THE SELECTED SRL SUBSCALES

This section presents the MSLQ items used in our study to assess the selected SRL subscales: peer learning (Table C.1), time management (Table C.2), and effort regulation (Table C.3).

Table C.1 – Peer learning subscale used in the survey

Peer learning subscale items
When studying, I often try to explain the material to a classmate or friend.
I try to work with other students to complete the course assignments.
When studying, I often set aside time to discuss course material with a group of students from the class.

Table C.2 – Time management subscale used in the survey

Time management subscale items
I make good use of my study time.
I find it hard to stick to a study schedule. *
I make sure I keep up with the weekly readings and assignments.
I often find that I don't spend very much time studying because of other activities. *
I rarely find time to review my notes or readings before an exam. *

***Negatively worded questions**

Table C.3 – Effort regulation subscale used in the survey

Effort regulation subscale items
I often feel so lazy or bored when I study that I quit before I finish what I planned to do. *
I work hard to do well even if I don't like what we are doing.
When course work is difficult, I either give up or only study the easy parts. *
Even when course materials are dull and uninteresting, I manage to keep working until I finish.

***Negatively worded questions**

APPENDIX D. WILCOXON SIGNED-RANK TEST RESULTS

In this appendix, we showcase the results of the Wilcoxon Signed-Rank Test applied to paired pre- and post-course MSLQ responses from students who completed both assessments (N = 159). Table D.1 summarizes the test statistic and p-value for each SRL subscale.

Table D.1 – Wilcoxon signed-rank test results

SRL subscale	Statistic	p-value
Time management	1617	0.00
Effort regulation	1783	0.00
Peer learning	3889	0.01

All p-values fell below the 5% significance threshold, providing statistical evidence that students' self-reported SRL skills changed over the course. To examine the direction of these changes, Table D.2 reports the pre- and post-course mean scores for each subscale, along with the observed mean variation (Δ= pos – pre).

Table D.2 – SRL subscale mean variation: pre-course vs post-course

SRL subscale	Pre-Course Mean	Pos-Course Mean	Mean variation (Δ)
Time management	5.17	4.42	-0.75
Effort regulation	5.54	4.98	-0.56
Peer learning	4.48	4.21	-0.27

APPENDIX E. FEATURES

This section provides an overview of the final feature set extracted from the Nova IMS Moodle logs. The features were grouped into three categories: raw activity (Table E.1), time-on-task (Table E.2), and frequency (Table E.3). Each table includes the feature name, a brief description, and references to prior studies that used the same or a comparable proxy. All listed features are included across datasets (CS-A and CS-B), unless marked otherwise: **(*)** indicates exclusion from CS-A, and **(**)** from CS-B.

Table E.1 – Raw activity features

Feature (unit)	Description	References
Announcements read (% of the course mean)	Number of announcements read as a percentage of the course mean.	Adapted from: Cristea et al. (2024)
Assignments viewed (n)	Number of assignment page views.	Conijn et al. (2017); Gašević et al. (2016); Macfadyen & Dawson (2010); Santos & Henriques (2023a, 2023b)
Course clicks (% of the total course)	Number of clicks made in the course as a percentage of the total course.	Santos & Henriques (2023a, 2023b)
Course clicks (n/week)	Number of clicks made in the course divided by the number of course weeks.	van Sluijs & Matzat (2024)
File download (binary)	Binary indicator for whether the student downloaded at least one file.	-
Folder clicks** (n/week)	Number of folder clicks divided by the number of course weeks.	Adapted from: van Sluijs & Matzat (2024)
Forum discussion posts read (% of the course mean)	Number of forum discussion posts read as a percentage of the course mean.	Adapted from: Conijn et al. (2017); Macfadyen & Dawson (2010); Santos & Henriques (2023a, 2023b); Zacharis (2015)
Forum discussion post read (binary)	Binary indicator for whether the student read at least one forum discussion post.	-

Resources view (% of the course mean)	Number of resources viewed as a percentage of the course mean.	Adapted from: Calvo-Flores et al. (2006); Conijn et al. (2017); Gašević et al. (2016); Santos & Henriques (2023a, 2023b); Zacharis, 2015);
Submissions (% of the total course)	Number of submissions made in the course as a percentage of the total course.	Santos & Henriques (2023a, 2023b)
TBL (n)	Total number of TBL activities accessed.	-
TBL completed (binary)	Binary indicator for whether the student completed all TBLs proposed by the course.	-
TBL resources clicks (n)	Number of clicks on TBL page resources.	-
TBL out of the schedule (n)	Number of TBL activities accessed outside the official course schedule.	-
PQ completed* (binary)	Binary indicator for whether the student completed all course PQs.	-
URLs viewed (n)	Number of clicks on external links.	Conijn et al. (2017); Macfadyen & Dawson (2010); Santos & Henriques (2023a, 2023b); Zacharis, 2015)

Features excluded from: *CS-A, ** CS-B

Table E.2 – Time-on-task features

Feature (unit)	Description	References
Average duration of online sessions (min)	Total time online in minutes divided by the number of online sessions.	Conijn et al. (2017); Santos & Henriques (2023a, 2023b)
Course clicks (n/session)	Number of clicks divided by the number of online sessions.	Santos & Henriques (2023a, 2023b)
Handout submission delay (h)	Time difference in hours between the handout deadline and the submission moment.	Adapted from: Li et al. (2020)
Project submission delay (h)	Time difference in hours between the project deadline and the submission moment.	Adapted from: Li et al. (2020)
Time spent online before the course start (min)	Sum of the duration in minutes of all online sessions of a student before the beginning of the classes.	-
Total time online (min)	Sum of the duration in minutes of all online sessions of a student.	Conijn et al. (2017); Jo et al. (2016); Macfadyen & Dawson (2010); Santos & Henriques (2023a, 2023b)
Variation of online session duration (min/week)	Slope of a linear regression estimating change in average online session duration (in minutes) over time (weeks). A positive value indicates that the student is spending more time per session. Adapted from time-on-task indicators.	Adapted from: Li et al. (2020)

Table E.3 – Frequency features

Feature (unit)	Description	References
Average spacing between online sessions (h)	Sum of time in hours between each user's consecutive sessions divided by the number of session intervals (i.e., total sessions – 1).	Adapted from: van Sluijs & Matzat (2024)
Days with no interaction ** (% of total course days)	Number of days without any interaction with the Moodle LMS, expressed as a percentage of the total number of course days.	Adapted from: Santos & Henriques (2023a, 2023b)
Largest inactivity period (d)	Maximum number of consecutive days without any interaction with the Moodle LMS.	Adapted from: Conijn et al. (2017); Santos & Henriques (2023a, 2023b); van Sluijs & Matzat (2024)
Online Session duration irregularity (min)	Standard deviation of session durations (in minutes) across the course timeline.	Adapted from: Conijn et al. (2017); van Sluijs & Matzat (2024)
Variation of online session duration irregularity (min/week)	Slope of a linear regression estimating change in the irregularity of session durations (standard deviation in minutes) over time (weeks). A positive value indicates growing inconsistency.	Adapted from: van Sluijs & Matzat (2024)
Variation of spacing irregularity (h/week)	Slope of a linear regression estimating change in the irregularity of spacing between sessions (standard deviation in hours) over time (weeks). A positive value indicates increasingly inconsistent session pacing.	Adapted from: van Sluijs & Matzat (2024)

Features excluded from: *CS-A, ** CS-B

APPENDIX F. HYPERPARAMETERS

This appendix complements Section 3.4.3.4 by detailing the hyperparameter search spaces used during the random search procedure applied to the top three models selected for each dataset–target combination. Table F.1 presents the full configuration space defined for each ML algorithm. All random searches were run with 60 sampled configurations and evaluated using 15-fold cross-validation, with MAE as the primary scoring metric.

Table F.1 – Hyperparameter space for random search of each ML algorithm

ML algorithm	Space
DT	'criterion': ['squared_error', 'friedman_mse','absolute_error'] 'splitter': ['best', 'random'] 'max_depth': [2, 4, 6, 8, 10, 20] 'min_samples_split': [2, 5, 0.1] 'min_samples_leaf': [1, 2, 5, 0.05] 'max_features': [None, 'sqrt', 'log2'] 'ccp_alpha': [0.0, 0.01, 0.05] 'random_state': [3]
Elasticnet	'alpha': np.unique(np.concatenate([np.logspace(-4, 1, 50), [1.0]])) 'max_iter': [100, 500, 1000, 3000] 'tol': [1e-4, 1e-3, 1e-2] 'l1_ratio': np.linspace(0.1, 0.9, 5) 'random_state':[3] 'selection': ['cyclic', 'random']
kNN	'n_neighbors': [3, 5, 7, 9] 'weights': ['uniform', 'distance'] 'p': [1, 2] 'leaf_size': [10, 20, 30, 50] 'n_jobs': [-1]

Lasso	'alpha': np.unique(np.concatenate([np.logspace(-4, 1, 50), [1.0]])) 'max_iter': [100, 500, 1000, 3000] 'tol': [1e-4, 1e-3, 1e-2] 'random_state':[3] 'selection': ['cyclic', 'random']
RF	'n_estimators': [50, 100, 300] 'criterion': ['squared_error', 'friedman_mse', 'absolute_error'] 'max_depth': [2, 4, 6, 8, 10, 20] 'min_samples_split': [2, 5, 0.1] 'min_samples_leaf': [1, 2, 5, 0.05] 'ccp_alpha': [0.0, 0.01, 0.05] 'random_state': [3] 'bootstrap': [True, False] 'n_jobs': [-1]
Ridge	'alpha': np.unique(np.concatenate([np.logspace(-4, 1, 50), [1.0]])) 'max_iter': [100, 500, 1000, 3000] 'tol': [1e-4, 1e-3, 1e-2] 'random_state':[3]

APPENDIX G. RESULTS – TARGET-FEATURE RELATIONSHIP

This appendix presents the full results referenced in Section 4.1.2. Tables G.1, G.2, and G.3 correspond to individual target analysis: peer learning, time management, and effort regulation, respectively. Each table presents the top 3 ranked features for each metric (MI, F-score, and Correlation) across all datasets (CS-A and CS-B) side by side. All listed features are included across datasets, unless marked otherwise: **(*)** indicates exclusion from CS-A, and **(**)** from CS-B.

Table G.1 – Peer learning: top 3 ranked features by metric and dataset

Feature (unit)	MI		F-score		Correlation	
	CS-A	CS-B	CS-A	CS-B	CS-A	CS-B
Announcements read (% of the course mean)				1.09		
Average spacing between online sessions (h)			2.51			
Course clicks (n/week)	0.18					
File download (binary)					0.26	
Forum discussion post read (binary)	0.16				-0.44	
Forum discussion posts read (% of the course mean)	0.21			1.25	-0.35	-0.17
Handout submission delay (h)						0.12
Project submission delay (h)		0.07				
Submissions (% of the total course)			1.64	1.11		
TBL (n)		0.06				
Variation of spacing irregularity (h/week)		0.08				
URLs viewed (n)			2.31			-0.14

Table G.2 – Time management: top 3 ranked features by metric and dataset

Feature (unit)	MI		F-score		Correlation	
	CS-A	CS-B	CS-A	CS-B	CS-A	CS-B
Assignments viewed (n)						0.24
Average duration of online sessions (min)			3.37		0.29	
Average spacing between online sessions (h)			2.92		-0.26	-0.22
Course clicks (n/session)	0.11					
Course clicks (% of the total course)	0.15					
Course clicks (n/week)	0.09					
Forum discussion post read (binary)		0.04				
Forum discussion posts read (% of the course mean)				2.09		
Handout submission delay (h)					-0.24	
Online Session duration irregularity (min)			2.91			
Project submission delay (h)				1.44		
Resources view (% of the course mean)		0.10				
Submissions (% of the total course)						0.20
TBL completed (binary)		0.04				
Variation of online session duration (min/week)				1.46		

Table G.3 – Effort Regulation: top 3 ranked features by metric and dataset

Feature (unit)	MI		F-score		Correlation	
	CS-A	CS-B	CS-A	CS-B	CS-A	CS-B
Assignments viewed (n)		0.11				
Average duration of online sessions (min)		0.08				
Average spacing between online sessions (h)			3.69		-0.54	-0.21
Course clicks (% of the total course)	0.18					
Days with no interaction (% of total course days) **	0.13				-0.45	
Forum discussion post read (binary)		0.06				
Online Session duration irregularity (min)	0.14					
Project submission delay (h)			1.70			
Submissions (% of the total course)				1.99		0.20
TBL completed (binary)					-0.36	
TBL resources clicks (n)			1.56	2.22		0.18

Features excluded from: *CS-A, ** CS-B

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa