

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AI-Powered Learning Analytics Dashboard for Monitoring Student Progress

Luís Tiago Trindade Cabral



Mestrado em Engenharia Informática e Computação

Supervisor: Rui Pedro Ferreira Pinto

July 28, 2025

AI-Powered Learning Analytics Dashboard for Monitoring Student Progress

Luís Tiago Trindade Cabral

Mestrado em Engenharia Informática e Computação

Approved by:

President: António Soares

Examiner: Constantino Martins

July 28, 2025

Resumo

Nas últimas décadas, a transformação digital no ensino e no setor da educação em geral tem vindo a redefinir a forma como docentes e estudantes interagem com os conteúdos, com os ambientes educacionais e com o próprio processo de educativo. O crescimento dos Sistemas de Gestão de Aprendizagem (SGA) como Moodle e Canvas e a sua crescente adoção em instituições de ensino permitiu uma gestão mais eficiente dos recursos pedagógicos, promovendo flexibilidade, acessibilidade e autonomia para todos os intervenientes no processo de ensino-aprendizagem.

Contudo, apesar do acesso a uma vasta quantidade de dados, os SGA tradicionais apresentam limitações significativas. Para os educadores, a dificuldade de detetar atempadamente padrões de desmotivação ou risco de insucesso entre os alunos, especialmente em turmas numerosas aliada ao possível reduzido conhecimento técnico cria um ambiente de incerteza e ineficácia pedagógica, levando à tomada de decisões avaliativas com base em intuições ou avaliações retrospectivas, ao invés de dados objetivos e atualizados. Por outro lado, para os estudantes, a inexistência de uma visão clara e acessível sobre o seu próprio progresso bem como a falta de feedback personalizado e em tempo útil compromete a autorregulação e o envolvimento com a aprendizagem, podendo levar à desmotivação gradual que, em limite, resulta no insucesso académico.

Neste contexto, a Análise de Aprendizagem e a Inteligência Artificial (IA) surgem como ferramentas com grande potencial para transformar dados em conhecimento útil, contribuindo para ambientes de aprendizagem mais personalizados, preditivos e focados na melhoria de resultados educacionais dos docentes, e dos estudantes.

É neste enquadramento que esta dissertação apresenta o desenvolvimento do *CABRAL*, um Painel de Análise de Aprendizagem potenciado por IA, com o objetivo de melhorar a monitorização do desempenho dos estudantes e apoiar a tomada de decisões pedagógicas. O sistema propõe-se a integrar-se com SGA existentes e a utilizar algoritmos de previsão para identificar padrões de desempenho e motivação, fornecendo insights explicáveis e personalizados, tanto para docentes como para estudantes. A interface do *CABRAL* foi desenhada com base em princípios de design centrado no utilizador e esteticamente otimizada para um ambiente de elevada carga cognitiva, promovendo uma experiência agradável e intuitiva, que seja um garante de acessibilidade e usabilidade.

As avaliações empíricas da capacidade preditiva bem como da usabilidade do sistema demonstram que o *CABRAL* é eficaz no apoio à tomada de decisão dos docentes e na melhoria dos mecanismos de apoio aos estudantes e os resultados delas obtidos validam a hipótese de que painéis de análise de aprendizagem baseados em IA podem contribuir significativamente para ambientes de aprendizagem mais proativos, personalizados e orientados por dados.

Com esta abordagem, o *CABRAL* representa um contributo relevante para a construção de um ensino mais inteligente, equitativo e centrado nas necessidades reais dos seus intervenientes.

Palavras-Chave: Inteligência Artificial, Análise de Aprendizagem, Painel de Análise de Aprendizagem, Sistemas de Gestão de Aprendizagem, Previsão de Desempenho Académico, Motivação Estudantil, IA Explicável, IA Generativa, Engajamento Estudantil

Abstract

In recent decades, the digital transformation of the education sector has redefined the way teachers and students interact with content and the learning process itself. The growth of Learning Management Systems (LMS) such as Moodle and Canvas their increasing adoption in educational institutions has enabled more efficient management of teaching resources, promoting flexibility, accessibility and autonomy for all those involved in the teaching-learning process.

However, despite access to a vast amount of data, traditional LMSs have significant limitations. For educators, the difficulty of detecting early patterns of demotivation or risk of underachievement among students, especially in large classes, coupled with possible limited technical knowledge, creates an environment of uncertainty and pedagogical ineffectiveness, leading to evaluative decisions based on intuition or retrospective assessments, rather than objective and up-to-date data. On the other hand, for students, the lack of a clear and accessible view of their own progress, as well as the lack of personalized and timely feedback, compromises self-regulation and engagement with learning and can lead to gradual demotivation, which ultimately results in academic failure.

In this context, Learning Analytics and Artificial Intelligence (AI) are emerging as tools with great potential for transforming data into useful knowledge, contributing to more personalized, predictive learning environments focused on improving educational outcomes for teachers and students.

It is within this framework that this dissertation presents the development of *CABRAL*, an AI-powered Learning Analytics Dashboard (LAD), with the aim of improving student performance monitoring and supporting pedagogical decision-making. The system aims to integrate with existing LMS and use predictive algorithms to identify patterns of performance and motivation, providing explainable and personalized insights for both teachers and students. The *CABRAL* interface was designed based on user-centered design principles and aesthetically optimized for a high cognitive load environment, promoting a pleasant and intuitive experience that is a guarantee of accessibility and usability.

Empirical evaluations of the predictive capacity and usability of the system show that *CABRAL* is effective in supporting teachers' decision-making and improving student support mechanisms, and the results obtained validate the hypothesis that AI-based LADs can contribute significantly to more proactive, personalized and data-driven learning environments.

With this approach, *CABRAL* makes a significant contribution to building a more intelligent, equitable education system centered on the real needs of those involved.

Keywords: Artificial Intelligence, Learning Analytics, Learning Analytics Dashboard, Learning Management Systems, Academic Performance Prediction, Student Motivation, Explainable AI, Generative AI, Student Engagement

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

Among these objectives, a few closely align with this project's goals, focusing mostly on health, education, and equal access to resources:

The specific Sustainable Development Goals mentioned have the following names:

SDG 3 Good Health and Well-Being

SDG 4 Quality Education

SDG 10 Reduced Inequalities

SGD	Target	Contribution	Performance Indicators and Metrics
3	3.1	Empowering educational stakeholders with a resource capable of improving their learning outcomes and enhance their productivity allows them to achieve a good work-life balance, thus promoting a healthier lifestyle	Educators' intent of using the system, Educators' and students' qualitative feedback on usability; System Usability Scale (SUS); Engagement rates; Time-to-task completion.
4	4.1	Providing students with a tool that promotes bidirectional communication with educators transforms the classroom into a more connected environment that enhances the likelihood of motivation	Single Ease Question (SEQ) results; Feedback frequency; Motivation prediction model accuracy; Usability metrics.
	4.2	Providing educators with a tool capable of flagging at-risk students in the beginning stages of the semester prevents early student disengagement	Predictive Evaluation Results (MAE, RMSE, R^2 , MHSP)
10	10.1	Offering a tool that provides personalized strategies for each student in a class regardless of their academic background can help minimize inequalities and empower students with more difficulties.	Insight coverage across performance bands. Accessibility and usage statistics;

Acknowledgments

Although signed by me, and marked as an individual accomplishment, I do not see this document as the work of one person. This thesis, in one way or another, is the culmination of the efforts of multiple people who helped place me in a position to succeed.

First and foremost, I want to express my gratitude towards Professor Rui Pinto, for providing me with the necessary guidance to improve day-in and day-out and for giving me the constant reassurance that my effort and hard work would, indeed, be rewarded.

Secondly, I want to thank my girlfriend, for being my emotional rock through these ten challenging months and constantly reminding me of my value as an engineer, as a writer and as a person.

Last, but most definitely not least, I want to thank my parents in particular, because even without understanding a single line of code, or knowing what an AI model consists of, they were, and always are, this thesis' biggest fans. I want to thank my mother for calling me every day and offering help, even with things she had no idea how to do, and I want to thank my father, for giving me my love of education that would inevitably lead me down the path to these very thesis. These two people gave me the role models of a lifetime, and for that, I am forever grateful.

Though this signature is mine. . . This work transcends me.

Obrigado.

Luís Cabral

*“We don’t get to choose how we start in this life. Real greatness is what you do with the hand
you’re dealt.”*

Victor Sullivan, Uncharted 3: Drake’s Deception

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Problem	2
1.3	Objectives	3
1.4	Document Structure	4
2	Background	5
2.1	Learning Management Systems	5
2.2	Learning Analytics	6
2.3	Learning Analytics Dashboards	9
2.4	XAI	11
2.5	Generative AI	13
3	Literature Review	15
3.1	Related Work	15
3.2	Search Strategy and Methodology	15
3.2.1	First Iteration	16
3.2.2	Second Iteration	17
3.2.3	Final Iteration	17
3.2.4	Methodology	17
3.2.5	Eligibility Criteria	19
3.2.6	Title and Abstract Screening	19
3.2.7	Final Selection	20
3.3	Literature Review Results Analysis	21
3.4	Literature Review Results Content Analysis	25
3.4.1	AI-Powered Data Analysis Based Categorization	25
3.4.2	Data Retrieval Based Categorization	36
3.4.3	Data Visualization Based Categorization	37
3.4.4	GUI-Based Categorization	39
3.5	Literature Findings	42
3.5.1	AI-Powered Data Analysis	42
3.5.2	Data Retrieval	43
3.5.3	Data Visualization Techniques	43
3.5.4	GUI	43
3.5.5	Limitations and Gaps	44
3.5.6	Conceptual Framework	46

4	Proposed Solution	49
4.1	Approach Description	49
4.1.1	Solution Goals	50
4.1.2	System Architecture	51
4.2	Methods	53
4.2.1	Research Design	53
4.2.2	Data Collection	56
4.2.3	User Insights and Design Motivation	57
4.2.4	Limitations	59
5	Developed System	60
5.1	Technological Stack	60
5.1.1	Backend and Data Handling	60
5.1.2	Machine Learning and AI Components	61
5.1.3	Frontend and Interface Design	61
5.1.4	Deployment and Integration	61
5.2	Data Retrieval and Management	61
5.2.1	<i>Prediction</i>	62
5.2.2	<i>Course</i>	62
5.2.3	<i>Student</i>	63
5.2.4	<i>StudentCourseEvent</i>	63
5.3	Grade and Motivation Predictions	63
5.3.1	Grade Prediction	63
5.3.2	Motivation Prediction	66
5.4	Actionable Insights	69
5.4.1	Database Interaction	70
5.4.2	Analytical Engine	70
5.4.3	Insight Generation Engine	72
5.4.4	Orchestrator	73
5.5	User Interface	73
5.6	Messaging Generation System	77
6	Evaluation and Discussion	81
6.0.1	Usability evaluation	81
6.0.2	Predictive evaluation	85
6.1	Evaluation Results	89
6.1.1	Usability Results	89
6.1.2	Predictive Results	110
6.1.3	Results Interpretation	121
6.1.4	General Findings	127
6.1.5	Research Questions Reflections	128
6.2	Ethical Considerations in <i>CABRAL</i>	130
7	Conclusions and Future Work	132
7.1	Limitations	133
7.2	Future Work	134
A	Usability Evaluation Task Durations	151

B Preliminary Survey Responses	156
C Usability Survey Responses	176

List of Figures

2.1	2017 LMS Usage Distribution across US Schools	6
2.2	Adaptation of Romero and Ventura’s Cycle of Applying DM to Educational Relationships	8
3.1	PRISMA Flow Diagram of the Selection Process	18
3.2	Publication Count per Year - Pie Chart and Bar Chart Distribution	22
3.3	Publication Count - Geographical Distribution	23
3.4	Article Abstract Word Cloud	25
3.5	Example of a DrillDownTree by Tableau Exchange [1]	29
3.6	DV Techniques Across the SLR Results (Sorted by Publications)	38
3.7	GUI Navigation Design Distribution	41
3.8	App Environment Distribution	41
3.9	Conceptual Framework of an AI-powered LAD based on SLR Results and Findings	46
4.1	System’s Main Capabilities	51
4.2	Proposed Solution’s Architecture	52
4.3	Student-Oriented Use Case Diagram	53
4.4	Educator-Oriented Use Case Diagram	54
4.5	Student Motivation to Use LMSs - Response Distribution	57
4.6	Student Regularity when Accessing LMS - Response Distribution	58
5.1	<i>CABRAL</i> ’s Grade Prediction Pipeline	64
5.2	Prediction Buttons in <i>CABRAL</i> ’s UI	66
5.3	<i>CABRAL</i> ’s Motivation Prediction Pipeline	66
5.4	Emojis employed in <i>CABRAL</i>	69
5.5	Actionable Insights Pipeline	70
5.6	<i>CABRAL</i> ’s main page UI	74
5.7	Traffic Light System employed to convey Student Predictions	75
5.8	Examples of <i>CABRAL</i> ’s Visualization Techniques	76
5.9	At-Risk Courses Help Modal	77
5.10	Message Generation System	78
6.1	<i>CABRAL</i> ’s Home Student Page in the Developed Figma Prototype	84
6.2	Example of User Interaction with the Maze Questionnaire	85
6.3	Progression of Educator Responses throughout <i>CABRAL</i> ’s Usability Evaluation’s Experimental Phase	91
6.4	Progression of Student Responses throughout <i>CABRAL</i> ’s Usability Evaluation Experimental Phase	92
6.5	Success Rate Progression	93

6.6	Hotspot Visualization of User Clicks on a <i>CABRAL</i> Screen	94
6.7	Misclick Rate Progression	95
6.8	Average Duration to Necessary Screen Ratio Progression	97
6.9	Average SEQ Value Progression for each Task	99
6.10	SUS Scores Distribution	100
6.11	Educator SUS Values Histogram	101
6.12	Student SUS Values Histogram	101
6.13	Success Rate vs. SEQ	106
6.14	Misclick Rate vs. SEQ	106
6.15	Average Task Duration vs. SEQ	107
6.16	Necessary Screens vs. SEQ	108
6.17	# Comments vs. Success Rate	108
6.18	# Comments vs. Misclick Rate	109
6.19	# Comments vs. Average Task Duration	110
6.20	LGP - MAE Distribution	111
6.21	LGP - RMSE Distribution	112
6.22	LGP - R^2 Distribution	113
6.23	LGP - MHSP Distribution	114
6.24	PROG - MAE Distribution	115
6.25	PROG - RMSE Distribution	115
6.26	PROG - R^2 Distribution	116
6.27	PROG - MHSP Distribution	117
6.28	SCI - MAE Distribution	119
6.29	SCI - RMSE Distribution	119
6.30	SCI - R^2 Distribution	120
6.31	SCI - MHSP Distribution	121
A.1	Task 1 - Educator Durations	151
A.2	Task 2 - Educator Durations	152
A.3	Task 3 - Educator Durations	152
A.4	Task 4 - Educator Durations	153
A.5	Task 5 - Educator Durations	153
A.6	Task 1 - Student Durations	154
A.7	Task 2 - Student Durations	154
A.8	Task 3 - Student Durations	155
A.9	Task 4 - Student Durations	155
B.1	Preliminary Survey - Q1 Response Distribution	156
B.2	Preliminary Survey - Q2 Response Distribution	156
B.3	Preliminary Survey - Q3 Response Distribution	157
B.4	Preliminary Survey - Q4 Response Distribution	157
B.5	Preliminary Survey - Q5 Response Distribution	158
B.6	Preliminary Survey - Q6 Response Distribution	158
B.7	Preliminary Survey - Q7 Response Distribution	159
B.8	Preliminary Survey - Q8 Response Distribution	159
B.9	Preliminary Survey - Q9 Response Distribution	160
B.10	Preliminary Survey - Q10 Response Distribution	160
B.11	Preliminary Survey - Q11 Response Distribution	161
B.12	Preliminary Survey - Q12 Response Distribution	161

B.13 Preliminary Survey - Q13 Response Distribution	162
B.14 Preliminary Survey - Q14 Response Distribution	162
B.15 Preliminary Survey - Q15 Response Distribution	163
B.16 Preliminary Survey - Q16 Response Distribution	163
B.17 Preliminary Survey - Q17 Response Distribution	164
B.18 Preliminary Survey - Q18 Response Distribution	164
B.19 Preliminary Survey - Q19 Response Distribution	165
B.20 Preliminary Survey - Q20 Response Distribution	165
B.21 Preliminary Survey - Q21 Response Distribution	166
B.22 Preliminary Survey - Q22 Response Distribution	166
B.23 Preliminary Survey - Q23 Response Distribution	167
B.24 Preliminary Survey - Q24 Response Distribution	167
B.25 Preliminary Survey - Q25 Response Distribution	168
B.26 Preliminary Survey - Q26 Response Distribution	168
B.27 Preliminary Survey - Q27 Response Distribution	169
B.28 Preliminary Survey - Q28 Response Distribution	169
B.29 Preliminary Survey - Q29 Response Distribution	170
B.30 Preliminary Survey - Q30 Response Distribution	170
C.1 Usability Questionnaire - Course - Student Response Distribution	176
C.2 Usability Questionnaire - Degree - Student Response Distribution	177
C.3 Usability Questionnaire - "Are you a mobility student?" - Student Response Distribution	177
C.4 Usability Questionnaire - "Are / Were you enrolled in any Lifelong learning course?" - Student Response Distribution	178
C.5 Usability Questionnaire - Age - Student Response Distribution	178
C.6 Usability Questionnaire - "The courses you usually teach are framed in what topics?" - Educator Response Distribution	180
C.7 Usability Questionnaire - "Have you had any experience with a Learning Analytics Dashboard? (Learning Management Systems such as Moodle and Canvas do not apply)" - Educator Response Distribution	181
C.8 Task 1 - Student Performance - Maze Breakdown	182
C.9 Task 2 - Student Performance - Maze Breakdown	183
C.10 Task 3 - Student Performance - Maze Breakdown	184
C.11 Task 4 - Student Performance - Maze Breakdown	185
C.12 Task 1 - Educator Performance - Maze Breakdown	185
C.13 Task 2 - Educator Performance - Maze Breakdown	186
C.14 Task 3 - Educator Performance - Maze Breakdown	187
C.15 Task 4 - Educator Performance - Maze Breakdown	187
C.16 Task 5 - Educator Performance - Maze Breakdown	188

List of Tables

2.1	Comparison between EDM and LA	9
2.2	Khosravi <i>et al.</i> 's Criteria to labeling XAI Approaches	12
2.3	Approaches and their Explanations in Explainable AI, according to Khosravi <i>et al.</i> [2]	12
3.1	Related studies - Sorted by Date	16
3.2	Search Query Breakdown	19
3.3	Exclusion Criteria for Study Selection	20
3.4	Selected Documents	21
3.5	Distribution of Publications by Country (Sorted by Number of Publications) . . .	23
3.6	Stopwords categorized by Exclusion Reason	24
3.7	Overview of selected studies on AI-powered Learning Analytics Dashboards . .	26
3.8	Summary of Articles Reviewed based on their AI Application	28
3.9	Nguyen <i>et al.</i> 's Results of calculating Correlation Coefficient in Predicted Interactive Activities over the Weeks	32
3.10	Majumdar <i>et al.</i> 's Performance Scores	33
3.11	Ramaswami <i>et al.</i> 's Performance Scores of Classifiers with Standard Deviations .	34
3.12	GUI Navigation Process Divisions	40
3.13	LADs Deployment Environment and Tech Stack	42
3.14	Critical Comparison of Existing LADs and their Limitations	44
5.1	Activity Weights	67
5.2	Feature Set used in Motivation Prediction Model	68
5.3	Grade Band Labels	71
5.4	Activity Labels	71
6.1	Summary of Course Data	86
6.2	Feature Set used in "Simple" Grade Prediction Models	87
6.3	Feature Set used in "Smart" Grade Prediction Models	88
6.4	Sample Size Distribution across <i>CABRAL</i> 's Evaluation Procedure Phases	90
6.5	Progression of Student and Educator Responses throughout <i>CABRAL</i> 's Usability Evaluation Experimental Phase	91
6.6	Educator and Student Success Rates throughout each Task	93
6.7	Educator and Student Misclick Rates throughout each Task	95
6.8	Educator and Student Average Duration for each Task	96
6.9	Average SEQ Values for each Task	98
6.10	Linear Regression Performance across Datasets	124
6.11	Random Forest Performance across Datasets	124
6.12	AdaBoost Performance across Datasets	125

6.13	CatBoost Performance across Datasets	126
B.1	Preliminary Survey - Q31 Response Distribution (Part A)	171
B.2	Preliminary Survey - Q31 Response Distribution (Part B)	172
B.3	Preliminary Survey - Q31 Response Distribution (Part C)	173
B.4	Preliminary Survey - Q32 Response Distribution (Part A)	174
B.5	Preliminary Survey - Q32 Response Distribution (Part B)	175
C.1	Usability Questionnaire - "Could a prototype like this help you in your academic life?" - Student Response Distribution	179
C.2	Usability Questionnaire - "Could a prototype like this help you or your students in your / their academic life?" - Educator Response Distribution	181

Abbreviations

AI	Artificial Intelligence
LMS	Learning Management System
LAD	Learning Analytics Dashboard
SLR	Systematic Literature Review
LA	Learning Analytics
EDM	Educational Data Mining
DM	Data Mining
CS	Course Signals
AOL	Asynchronous Online Learning
HCI	Human-Computer Interaction
GenAI	Generative AI
LLM	Large Language Model
PRISMA	Preferred Reporting Items for Systematic Review and Meta-Analysis
EX-LAD	Explainable Learning Analytics Dashboard
MAE	Mean Absolute Error
MSE	Mean Squared Error
AID	Automated Insightful Drill-Down
BKT	Bayesian Knowledge Tracing
DR	Data Retrieval
DV	Data Visualization
IO	Information Overload
GDPR	General Data Protection Regulation
SUS	System Usability Scale
SEQ	Single Ease Question
CABRAL	Clarifying AI for Better Results in Academic Learning
RMSE	Root Mean Squared Error
MHSP	Mean Hassanat Similarity Percentage
LGP	Project Management Laboratory (Laboratório de Gestão de Projetos)
SCI	Cyberphysical Systems and Internet of Things (Sistemas Ciberfísicos e Internet das Coisas)
PROG	Programming (Programação)
FEUP	Faculty of Engineering of the University of Porto

Chapter 1

Introduction

In an increasingly digital world where the advent of Artificial Intelligence (AI) is not only unavoidable but a little intimidating to the well-being of crucial jobs, it becomes a necessity, nay, a duty, to try to come up with new and innovative ways to use the latest available technologies to speed up or enhance otherwise tedious and repetitive tasks' most professionals must endure in a day-to-day basis.

As defended by Nogueira *et al.* [3], teaching is a job that requires constant “fine-tuning” and evolution from both the educators and the methods used by them to convey a better message to the alumni. Educating, as expected, isn't, nor should be, unbothered by the frequent AI-related changes. AI-powered tools like Eduaide.AI [4] for lesson development, efficient presentation slide makers like SlidesGo [5], and helpful ChatGPT [6] prompts prove to be making teaching a little easier and more effective through the controlled use of AI.

However, it is hard to not agree that the art of teaching not only applies to the specific act of providing useful information to the students but also to keeping a good track of their learning processes and employing personalized methods depending on the students at hand, both crucial factors in academic success, as argued by Shin *et al.* [7].

Furthermore, in the view of Mallocci *et al.* [8], the increasing complexity and volume of educational data, cause traditional methods of monitoring student progress to fall short in providing timely and actionable insights for educators, thus compromising the student's learning process.

As pointed out by Dobre *et al.* [9], one way educators are using new technology to their advantage is through Learning Management Systems (LMS) [10]. While being software platforms designed to facilitate the administration, documentation, tracking, and delivery of educational courses or training programs, they also enable educators and organizations to create, manage, and deliver content, as well as provide feedback, often in an online or hybrid setting. By streamlining the learning process with a centralization of resources and flexibility, they are able to improve collaboration and communication. However, since they are mostly used to organize student's information and documents, they tend to leave a lot on the table when it comes to student progress tracking.

1.1 Motivation

To better understand the motivation for this research work, it is crucial to analyze the current paradigm in the educational world. It is undeniable that, in the current day, and predictably for the foreseeable future, most educational practices tend to heavily rely on the use of digital technology to, on the one hand manage large amounts of data, while also to, on the other hand, conduct communication between both students and educators.

LMSs like Moodle [11] or Canvas [12] are finding a bigger role than ever in today's courses, giving educators new platforms to handle student submissions, provide class materials, and even manage assignments. These systems however, Meißner *et al.* [13] point out, while a great tool for organizing and managing courses as a whole, often lack the ability to provide a comprehensive and intuitive way to analyze student progress, either because of lecturers' technological inexperience, or of LMS' often complex interfaces that require a steep learning curve, not always possible to take on by educators.

In practical terms, many teachers struggle to locate and interpret relevant information across different modules of an LMS. For example, to identify students at risk of disengagement, an educator might need to manually check submission dates, attendance logs, forum activity, and grade trends, often spread across multiple dashboards and pages. This fragmented structure not only consumes time but also introduces human error and subjectivity, delaying effective interventions.

Furthermore, another factor to take into account is that the educational process, as a whole, mostly depends on a periodical assessment and retrospective analysis of the student's progress, thus solidifying the need for a more intuitive and comprehensive way to analyze their development. With the application of data-driven decision-making, some of these more complex tasks could be sped up and, more importantly, heavily empowered. What's more, through an intuitive analytics dashboard, the data that goes into making learning processes complex and hard to predict can be easily scattered through a graph or table, thus massively simplifying the tracking process from the educators' side, leaving them with more time for managing classes and students, producing a win-win situation for both parties.

1.2 Problem

Despite the widespread adoption of LMSs in educational institutions, most of these platforms still lack robust, intuitive tools to support real-time monitoring and meaningful interpretation of student performance data. Educators often face difficulties in identifying students who are at risk of underperforming or disengaging, especially in large classes where individualized attention is difficult to maintain.

This issue is twofold. Firstly, LMSs typically focus on course content delivery and administrative functionalities, offering limited support for data analytics or intelligent insights. Secondly, many educators lack the time, training, or technical background required to manually extract and

interpret complex data from these systems. As a result, crucial patterns in student learning behavior may go unnoticed until it is too late to intervene effectively.

Moreover, traditional approaches to performance monitoring rely heavily on periodic assessments, which provide only a retrospective view of student progress. Without continuous and predictive feedback mechanisms, opportunities for timely support and pedagogical adjustments are often missed.

Therefore, there is a clear need for a solution that can:

- Automatically process and analyze educational data in real-time;
- Translate this data into actionable insights for educators;
- Be accessible and usable by non-technical teaching staff;
- Help anticipate learning difficulties before they materialize into poor outcomes.

Addressing this problem is essential to empower educators with the tools needed to adopt a more proactive, data-driven approach to teaching, ultimately contributing to more personalized, effective, and inclusive educational experiences.

1.3 Objectives

The hypothesis of this thesis is the following:

An AI-powered Learning Analytics Dashboard (LAD) that integrates real-time data visualization and advanced AI algorithms to analyze student performance will significantly enhance educators' ability to make data-driven decisions, leading to improved teaching strategies and better student learning outcomes.

This thesis' objectives can be summarized as follows:

- Develop an AI-powered LAD that monitors and visualizes student progress in real-time.
- Seamlessly integrate the dashboard with existing LMS.
- Provide educators with actionable insights to enhance teaching strategies and student support.
- Improve student learning outcomes through data-driven decision-making.

To correctly address the problem previously pointed out, it is important to analyze how an AI-Powered LAD can benefit both students' and educators' learning outcomes. Thus, the following research questions will be aimed to be addressed through the entirety of this thesis' work.

- **RQ1: Can an AI-Powered LAD help students and educators improve their learning outcomes?**
- **RQ2: What are the main capabilities of existing AI-Powered LADs?**

- **RQ3: How can advanced AI techniques, such as predictive and prescriptive analytics, improve the accuracy and utility of insights provided by LADs?**

1.4 Document Structure

To ensure a clear and systematic exposition of the research work carried out throughout this thesis, the document is organized into seven chapters, each with a distinct purpose and logical progression. Following this introductory chapter, which outlines the motivation, problem statement, research objectives, and questions, Chapter 2 provides the necessary background on key topics that sets the stage for a deeper understanding of the challenges and opportunities explored in this work.

Chapter 3 presents a comprehensive literature review, grounded in a Systematic Literature Review (SLR) methodology, that maps the current landscape of AI-powered LADs and critically analyzes the existing body of knowledge in the field. Chapter 4 introduces the proposed solution, detailing its approach, main capabilities, and system architecture. The subsequent Chapter 5 elaborates on the developed system, including data management, predictive models, insights generation pipeline, user interface, and messaging components. Chapter 6 evaluates the system through usability and predictive performance analyses and discusses the resulting findings. Finally, Chapter 7 concludes the thesis by summarizing key contributions and outlining possible directions for future research.

Chapter 2

Background

This chapter provides a background and brief explanation on some topics explored in this work, as well as related work.

2.1 Learning Management Systems

Despite the uneven balance between the demands of the digital age and the still relevant and applied traditional teaching methods, it is hard to see a more concrete application of digitization in the educational environment than in the use of LMS. Either with Moodle, Blackboard [14] or Canvas, educational institutions tend to prefer centralizing their students' learning process in a digitalized platform rather than in old traditional methods like binders or portfolios.

Although in the latest decade it would be extremely hard to find a higher education institution without either one, or more LMS integrated into their system, assuming LMS's dominance in the classroom has happened since they first appeared would be a considerable mistake, as the first recollection of a LMS being employed in a classroom setting goes as back as 1926, with the invention of the Pressey Teaching Machine, a simple mechanism with four keys designed to reward correct responses in multiple answer questions. [15] In spite of, many years later, being perceived as a precursor to modern computer-based learning systems, Pressey's work was, in the 1920s, as argued by [16], nothing more than an absolute commercial failure, either due to resistance from educators, economic or practical barriers, or simply technological backlash. However, despite being a powerful invention brought to life in the wrong era, it is doubtful that Pressey's machine would, in the modern time, fall under the definition of a LMS due to the simplicity of its functionality as an answer verifier.

As a result, it becomes important to first define what a modern LMS really consists of, and only then dive into its evolution in the last three decades.

Although a long time ago the definition of an LMS might be a complex question, most modern authors tend to agree on what, nowadays, constitutes this type of systems. As such, today, an LMS is a software or web-based application that can be used to either plan, implement, aid, handle, or assess a specific learning process. [17] [18]

Driven by the growth in personal computers and more modern operating systems, these modern LMS started to appear in 1993 with the birth of Blackboard, and were mostly focused on tracking message delivery between the system and the users. [19] In 2007, exactly ten years after the emergence of LMS as a viable profitable solution in the educational world, the adoption of LMS was a much different phenomena. According to Falvo *et al.*'s study from the same year [20], from 100 random universities in the USA, 79 were already employing some kind of LMS to aid their classes' activities and overall educational process, with most of them preferring Blackboard. Although undeniably larger than ten years prior, the amount of higher educational institutions that would not use LMS as a means to help the learning process was still, compared to modern statistics, staggeringly high at 21 percent. Ten years later, however, the paradigm would be considerably changed, as in 2017, according to Edutechnica's "LMS Data - Spring 2017 Updates" [21], seen on Figure 2.1, on a study on over 3900 schools, only 7 percent of them weren't using any LMS to their advantage, cutting this number to a third in only ten years. In 2024, we have a concrete way to prove the LMS adoption and investability trend was here to stay from its first breath, since, as pointed out by Bouchrika [22], the number of current LMS users in the world is already upwards of 73 million, and, by the end of 2025, the market is expected to rise above 28 billion dollars.

	ANGEL	BbLearn	Canvas	D2L	Moodle	Sakai	Pearson	Other	None
Institutions	34	1185	713	360	678	107	105	494	252
	1%	33%	19.8%	10%	18.9%	3%	2.9%	13.7%	7%
Enrollments	124,679	7,383,086	4,773,367	2,314,816	2,611,762	757,643	353,688	1,420,744	248,488
	0.7%	43.5%	28.1%	13.6%	15.4%	4.5%	2.1%	8.4%	1.5%
Average Size	3667	6246	6704	6430	3852	7081	3368	2876	990
Median Size	1597	3365	3295	3568	1899	2828	1022	1001	722

Spring 2017, 500+ FTE students, www.edutechnica.com



Figure 2.1: 2017 LMS Usage Distribution across US Schools

2.2 Learning Analytics

With the increasingly digitization in the education process as a whole, the sheer volume of educational data available made it necessary to have an entire field dedicated to “the measurement, collection, analysis and reporting of data about learners and their contexts” [23] : Learning Analytics (LA). The emergence of LA as a critical field lies at the intersection of education, data science, and technology, and its ability to leverage large-scale data generated from various educational environments, such as LMS, proves to enhance teaching and learning experiences for

both the learners and the educators involved. However, LA, as a field, was not always of critical importance in education.

While it can be hard to consider the use of data in education without the existence of both LMS and Online Education, teachers and educators have, for a very long time, leveraged the use of data to both aid and empower their day-to-day jobs, even without calling it LA. As Ye argues [24], the collection and analysis of educational data is closely related to the behaviorist philosophy educators have been employing since the start of the profession. Although behaviorism is a complex topic dating back to the late 1800s and early 1900s with Thorndike's doctoral dissertation *Animal Intelligence: An Experimental Study of the Associative Processes in Animals* [25] and John B. Watson's *Psychologist as the behaviorist views it* [26] it can be defined, in simple terms, as an approach that aims to understand human and animal behavior. Given Watson's unprecedented view in utilizing data to both predict and influence human behavior, his work needs a deeper analysis.

According to the previously mentioned Watson address, for a behaviorist, not only is the theoretical goal of psychology the prediction and control of behavior, but, in his view, "in a system of psychology completely worked out, given the response the stimuli can be predicted; given the stimuli the response can be predicted" thus highlighting the belief in the deterministic nature of behavior, allowing data to become the main enabler of precise predictions and control. Despite behaviorism being a largely utilized and regarded theory up until the half of the 20th century, its inability to view internal mental states as explanations for observable behavior deemed it outdated and propelled its replacement by cognitive psychology, an approach that resulted from the cognitive revolution of the late 1950s. Although considering a 1910s address on an already outdated psychological school of thought as basis for the use of data in educational environments may seem farfetched, Watson's view is, to the best of the available knowledge on the topic, the first true pioneer of utilizing data to predict human and student behavior. As the psychologist puts it: "*we find out by experimentation whether a series of stanzas may be acquired more readily if the whole is learned at once, or whether it is more advantageous to learn each stanza separately and then pass to the succeeding. We do not attempt to apply our findings. The application of this principle is purely voluntary on the part of the teacher*" and "*If psychology would follow the plan I suggest, **the educator**, the physician, the jurist and the business man could utilize our data in a practical way, as soon as we are able, experimentally, to obtain them*". As brilliantly stated by Dan Ye [24], the shift to behavioral learning outcomes marks a vital step towards dealing with quantitative learning data and, as a result, towards modern LA.

When analyzing the growth of LA as a field, especially through the late 1990s it becomes extremely difficult to not mention Educational Data Mining (EDM). EDM is, as defined by the International Educational Data Mining Society as "an emerging discipline, concerned with developing methods for exploring increasingly large-scale data that come from educational settings and using those methods to better understand students" [27], however, its process is mostly defined as the conversion of raw data from educational systems into useful information aiming to help educational research and practice [28], meaning the application of Data Mining (DM) techniques to educational data [29]. Although the application of DM techniques in education has been target of

research since the late 1990s [30], it is fairly consensual in the academic community that Cristobal Romero and Sebastian Ventura are the two main pioneers when it comes to EDM with their book *Data mining in e-learning* [31] and their *Educational data mining: A survey from 1995 to 2005* article [32] being the beginning of their long list research works on the topic. In the latterly mentioned article, the authors argue that EDM can be valuable to three different stakeholders: students, educators and academics responsible and administrators. To Romero and Ventura, applying DM in educational systems can be represented by a simple cycle, like the slightly simpler adaptation shown in Figure 2.2, where the students, who use the educational system produce data that in turn is analyzed through DM techniques. Those techniques can both help advise students through actionable insights or show patterns to both educators and academics responsible who can use it to adapt the educational system, whether it being the actual classroom or a more complex structure, like an e-learning system or a LMS.

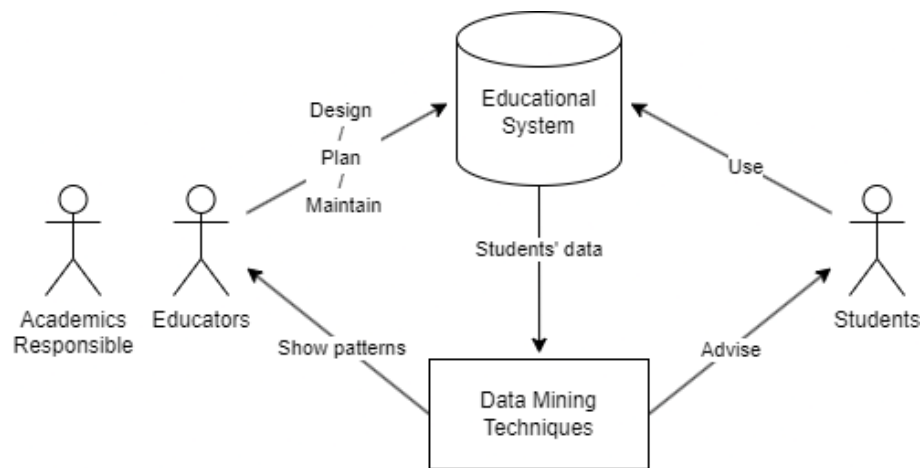


Figure 2.2: Adaptation of Romero and Ventura's Cycle of Applying DM to Educational Relationships

However, in spite of being intimately related, LA and EDM are not one and the same. According to Siemens and Baker 2012 study [33], there are key differences between LA (or LAK as the authors put it) and EDM, with the most notables being in the type of discovery prioritized, adaptation and personalization and the way both reduce (or do not reduce) systems. To the authors, when it comes to the type of discovery prioritized, while LA values highly the leveraging in human judgement while only looking as automated discovery as a means to this end, EDM does the opposite, considering automated discovery as the final goal and leveraging human judgement as a tool to accomplish it. As for the adaptation and personalization, Siemens and Baker point out that, whilst EDM focuses highly on automated adaptation with no need for a human-in-the-loop, LA aims to inform and empower educators and learners, not undervaluing their presence in the system. Last, but not least, whereas LA emphasizes understanding systems as wholes without partitioning them, EDM attempts to reduce systems to components and analyses them individually, as well as the relationships between them. An overview of these differences can be seen in Table 2.1.

Table 2.1: Comparison between EDM and LA

Aspect	EDM	LA
Discovery Prioritization	Automated discovery is the final goal, with human judgment leveraged as a tool to achieve it.	Leverages human judgment highly, with automated discovery as a means to this end.
Adaptation and Personalization	Focuses on automated adaptation without the need for human-in-the-loop.	Aims to inform and empower educators and learners, emphasizing the importance of their presence in the system.
System Analysis	Reduces systems to components, analyzing them individually and their relationships.	Emphasizes understanding systems as wholes without partitioning them.

Nowadays, it would be hard to mention LA as a field without mentioning its use of data analysis to generate insights through three main types of analytics: Descriptive, Predictive, and Prescriptive. In broad terms, Descriptive Analytics places its focus on summarizing past and current data to provide a clear understanding of what has happened, allowing users to view patterns in key areas such as student engagement, performance trends, and resource usage. [34] Predictive Analytics, on the other hand takes it a step further by using historical data and statistical modeling to forecast future outcomes, such as, for example, identifying students who are at risk of underperforming or dropping out. Prescriptive Analytics aims to leverage the insights from both descriptive and predictive analyses to recommend actionable strategies both for educators and students, enabling them to make informed decisions to improve learning outcomes, personalize instruction, and optimize educational practices. These types of analytics collectively empower stakeholders to create data-informed strategies that enhance the overall efficacy of the education process.

2.3 Learning Analytics Dashboards

To synthesize, organize, and provide relevant insights from the data analysis produced by LA, LADs have received a lot of attention as tools that can capitalize on its power, and add to the user experience that data can sometimes lack. [35] In short, LADs are, by definition digital interfaces designed to visualize data either related to learners' performance, engagement, or progress in educational settings. These tools provide the ability to aggregate data, offer actionable insights for educators, students, and administrators, and to present key metrics like grades, participation rates, time spent on tasks, and skill mastery in an intuitive format. Furthermore, LADs can help Educators identify struggling or at-risk students, tailor interventions to specific learners, and enhance curriculum design, while still allowing students to monitor their own progress and set goals, but throughout history, LADs did not always conform to this definition. As the exact concept is fairly modern, it is difficult to pinpoint the date of birth of the first LAD, but a reasonable estimate of it could fall on 2007, with the appearance of Course Signals (CS). [36] [37]

According to a 2009 article written by Steve Tally published in the Purdue University Website [38], CS was piloted in the 2006-07 and 2007-08 academic years with nearly 2,000 university students, and was no stranger to great results right from its arrival, with positive feedback from both students and educators. As argued by Pistilli and Arnold [39], CS went a “*step beyond the traditional models*” as it, unlike them, did not need to rely on midterm grades reported by faculty members that, due to the stage of the semester on which they were made available, usually left students with not enough time to rectify the issue with the course. Another way it differentiated itself from the competition is the way CS was built. Through the use of empirical data in every stage, ensuring the most predictive student-success algorithm was an easier step to take, allowing CS to be built “*from the ground up*”. Furthermore, CS was, by all means, a behaviorally based model, placing more emphasis on the time the student interacted with the course management system rather than on single points of assessment like midterms and single tests, since, as the authors put it “*a student of average intelligence that works hard is just as likely to get a good grade as a student that has above-average intelligence but does not exert any effort; hence, the inclusion of effort in the model*”. A big advantage of CS lied on the fact that, the identification of an at-risk or struggling student directly prompted an intervention set up by instructors, allowing students to receive direct feedback from their tutors, resulting in an increase in motivation and success. Additionally, CS also presented a pioneering way of classifying and categorizing students that is still used today: Traffic Lights. As explained in the 2011 YouTube video “*Course Signals Explanation*” [40], students would be able to check their progress and projected grade in one of three colors: “*If you see a green light, you’re on track to do well in the course. If you see a red light, Course Signals predicts that you need to make serious changes to do well in the course. If you see a yellow light, there is room for improvement.*”, this presented students with an intuitive and revolutionary way of, in the blink of an eye, find out if their progress was sufficient to complete the course with a passing grade, a technique still vastly employed today. Its success as a precursor to LADs rendered CS a partnership with SunGard Higher education in October 2010 and as of April 2012, more than 24,000 students had been impacted by the project [41], showing the need for LADs was present from the start, just not known.

In the modern day, as pointed out by Lee and Kim [42], LADs’ potential to enhance student engagement, specifically in Asynchronous Online Learning (AOL), has been an unanimous point in the research community. Although the rise of AOL might be seen as an unavoidable evolution propelled by either the increase in the global tertiary education enrolment, as according to UNESCO [43] in the last two decades that number more than doubled, or the equally unavoidable development of new and improved technology, better catered to educational institutions’ needs and therefore of extreme usefulness. Despite these factors affect in the evolution to AOL-centered courses, one major catalyzer in the transition to AOL is often swept under the rug when discussing the topic, COVID-19. According to UNESCO’s 2020 article entitled “*Education: From COVID-19 school closures to recovery*” [44], the COVID-19 pandemic affected more than 1.6 billion students and youth around the world, leaving educators and entire institutions wondering how to replace the otherwise reliable face-to-face interaction between students and teachers in extremely short time

periods. Now through the sheer need to adapt to unprecedentedly adverse conditions, Remote Education finally had the perfect conditions to flourish as the employed method of both teaching and learning, giving education stakeholders the opportunity and resources to perform their regular day-to-day activities, without completely compromising them. Although a simple concept, Remote Education can be divided into two categories, specifically when considering the temporal dimension, as argued by Hung *et al.* [45]: Synchronous and Asynchronous, with synchronous learning requiring both learners and educators to communicate in real-time, and asynchronous not needing schedule from both parts synchronization. As explained by Hung *et al.*, when it comes to its effects, AOL can be seen as something of a double-edged sword, as while it provides time flexibility to both students and educators it does require a higher degree of self-discipline due to the limit on teacher-student communication regarding course material management, learning progress awareness and performance evaluation. The combination of these factors allied with the increasingly powerful LA techniques researched through the time created a paradigm that drove the adoption of LADs to unparalleled heights, setting a new trend in education environments.

2.4 XAI

As LA systems increasingly adopt advanced AI techniques for prediction and insight generation, ensuring the transparency and interpretability of those models becomes critical, especially in sensitive contexts like education. Educators, who are often non-technical stakeholders, need to understand and trust the recommendations produced by such systems in order to act upon them effectively and, as pointed out by Storey *et al.* in *Explainable AI* [46], machine learning techniques lack of explicit human-understandable specification of the rules that go into creating a predictive model together with “black box” AI techniques’ complexity create a problem that is becoming increasingly alarming. To solve this problem, XAI appeared. According to Khosravi *et al.*’s [2] research, XAI focuses on the promotion of the use of methods that can enable human users to both understand, trust and manage artificial intelligence [47] by developing post-hoc explainability techniques that intend to explain the output of a specific model without “elucidating the structure of the model” itself. Education is not untouched by this trend. To the research team, the role and need for XAI in education is not dissimilar to its broader use. On the other hand, while a large number of insights and practices are drawn from the field of AI and of Human-Computer Interaction (HCI), there are distinctive needs grounded in theories from both cognitive and learning sciences, the authors argue. To Khosravi *et al.*, the potential benefits of XAI boil down to four concepts: Agency, meaning facilitating conversation among different stakeholders thus encouraging co-design and value co-creation; Student-teacher interactions, with explanations of AI propelling dialogue between students and educators allowing the challenging and improvement of those same explanations; AI Literacy, both among teachers and learners in order to decide when and how AI should be used in the classroom setting; and Accountability and Trust, in particular in Higher Education institutions given the constant pressure to provide evidence of education quality, both for assurance and competition.

In Khosravi *et al.*'s view, XAI approaches are usually labeled taking into account two main criteria, shown in Table 2.2.

Table 2.2: Khosravi *et al.*'s Criteria to labeling XAI Approaches

Dimension	Approach explains the entire model
	Approach explains an individual prediction
Explainability	Model is self-explainable and has a single structure
	Explanation is post-hoc and focuses on how the model produces the predictions without elucidating its structure

On the other hand, the six concrete examples of used approaches in XAI are shown in Table 2.3 as well as their basic definition.

Table 2.3: Approaches and their Explanations in Explainable AI, according to Khosravi *et al.* [2]

Approach	Explanation
Global Explanations	This approach focuses on explaining how different features/variables affect predictions. It is generally model-specific and works best for models that are interpretable by design. For less interpretable models, a global surrogate model can approximate predictions of a black-box model. Care must be taken as the surrogate might not generalize across all subsets.
Feature Relevance	Presents the computed relevance of each feature in terms of a score, which shows the importance granted by the model to each feature. Scores can be computed using models such as linear regression or Shapley values. This method is used in predictive learning analytics to highlight important features affecting learners' "at-riskness."
Example-based	Selects particular instances to explain the model, providing both local and global explanations. This model-agnostic approach helps in interpreting boundaries by visualizing instances and highlighting examples.
Local Explanations	Focuses on explaining a particular instance independently of the model. Works well with rule-based models and decision trees. Local surrogate models can approximate predictions locally for less interpretable models. This method is common in student-facing systems to inform students of their progress.
Comparison	Selects particular instances to locally explain other instances. Best suited for non-parametric models like K nearest neighbors. Often used in plagiarism detection to compare instances and provide evidence for predictions.
Counterfactual Explanation	Describes causal situations like "If X had not occurred, Y would not have occurred." Demonstrates the smallest feature change required to alter a prediction. This method can show the importance of features but is challenged by identifying instances and potential conflicting "stories" from multiple counterfactuals.

Additionally, the researchers also provide specific design processes that can lead or help AI-powered educational systems mainly focusing on User Experience, User-Centered Design, Top-down theory driven design, Participatory Design and Co-Design and HCI and Interaction Design all techniques aimed at putting the end-user in the center of the process, therefore helping to create systems adapted to their actual target.

2.5 Generative AI

While explainability focuses on improving human understanding of AI-driven predictions, another emerging challenge in LADs lies in how these insights are communicated to end-users. Educators and students often receive results in technical formats such as charts, metrics, or classifications that, although informative, may not always be immediately actionable or engaging. To bridge this gap, recent advancements in Generative AI (GenAI) have introduced new opportunities for transforming analytical outputs into natural language feedback.

As explained by Feuerriegel *et al.* [48], GenAI “refers to computational techniques (...) capable of generating seemingly new, meaningful content such as text, images or audio from training data” while a GenAI model “refers to generative modeling that is instantiated with a machine learning architecture (...) and, therefore, can create new data samples based on learned patterns”.

GenAI’s rapid growth profoundly accelerated the development and adoption of AI technologies in several different environments as it shed light on a new world of technology otherwise hidden to non-tech-savvy users. By enabling machines to produce human-like text, images, code, and even music, GenAI has not only made AI more accessible and impactful across industries but has also demonstrated the practical, creative, and problem-solving potential of AI, captured public imagination and drove unprecedented investment and innovation. This leap has not only simplified complex AI concepts for broader audiences but also unlocked transformative possibilities in different fields, specifically education, propelling AI into mainstream consciousness like never before.

Despite the first reactions to the advent of GenAI in educational institutions being mixed and sparking larger debates, with the gradually increasing knowledge acquisition on the topic, a lot of institutions learned to coexist with GenAI viewing it as a tool, rather than a threat [49]. While multiple tools like DALL-E [50] or even Turnitin [51] play an important role in a GenAI-assisted education environment, it is impossible to not mention ChatGPT as the main propeller of generative AI in both education and other diverse fields.

As described by Su and Yang [52], ChatGPT is a type of Large Language Model (LLM) designed to generate human-like text based on large volumes of data. In the authors’ views, in education, the use of ChatGPT can provide various benefits, ranging from providing a more personalized learning experience for students, encouraging educators to reflect on their educational content or helping them in answering various student questions. However, since it is trained to answer any question an user might have, an interesting approach to take when trying to understand how ChatGPT can help educators or students fulfill their goals or complete their tasks, is ask it.

ChatGPT, when asked the question: “How can you help students and/or educators in an educational system reach their goals or perform their tasks?”, answers the following:

“I can support students and educators by simplifying learning, enhancing teaching, and streamlining tasks. For students, I explain complex topics, help with homework, create study guides, and prepare them for exams. I can also assist with coding, language learning, and building critical thinking skills. My support includes personalized resources and strategies to match their learning styles, as well as tools to stay motivated and manage time effectively.

For educators, I provide help with lesson planning, creating engaging materials, and offering strategies for inclusive teaching. I can simplify difficult concepts for diverse learners, suggest innovative classroom activities, and assist with grading or providing feedback. Additionally, I support professional development by summarizing new research, integrating technology into teaching, and offering solutions to common challenges. My goal is to empower both students and educators to reach their full potential efficiently and effectively.”¹

Despite the constantly increasing adoption of GenAI in classrooms and educational settings, there are still some important challenges and limitations associated with its use. Su and Yang [52] underline, as main concerns, the ethical and safety issues stemmed from the use of specifically ChatGPT and the possible students’ disregard of the knowledge acquisition process since allowing students to complete their homework using GenAI tools may lead to detrimental results.

Nowadays, GenAI has emerged as a transformative force, propelling AI into mainstream adoption and redefining its role across industries, particularly in education, with tools like ChatGPT demonstrating immense potential to enhance learning experiences and teaching methodologies. However, while the benefits of GenAI are substantial, its integration into educational systems must be approached thoughtfully, addressing ethical concerns, safety issues, and the potential impact on students’ learning processes. Balancing innovation with responsibility will be key to leveraging the full potential of GenAI in education and beyond.

¹This response was generated in a regular ChatGPT chat on January 17th 2025, using a non-premium account with no previous use. The model used to perform this operation was GPT-4, available, at the time, for a limited number of requests. Since the first answer the LLM provided was quite extensive and not written in plain text but in topics, the LLM was asked to refine the answer through the following prompt: “Can you summarize so it fits in, at most, two paragraphs and is written in plain text?”. The response generated after this prompt was copied and pasted verbatim to the document.

Chapter 3

Literature Review

The intent of this phase trended towards performing a SLR on AI-powered LADs and the results that stemmed from their use in the educational context. This chapter aims to provide a detailed overview of the current developed work surrounding LADs and its main effect on students' learning outcomes. This chapter also intends to provide a comparison between already existing studies and this thesis' work.

3.1 Related Work

In the 21st century, the fields of LA, LADs and AI, specifically in education have definitely seen a considerable growth in terms of both widespread adoption and concept recognition with many different research works focusing on the topics. However, despite this advancement and the complementary nature of the three topics, there is still a gap in the literature when it comes to the intersection of LADs and AI, with most studies either focusing on LA and AI or on LA and LADs, as can be seen in Table 3.1. This study attempts to build on these advancements by presenting a comprehensive approach that combines LA, LAD, and AI to push the boundaries of educational technology and aid both students and educators in improving their learning outcomes.

3.2 Search Strategy and Methodology

During the initial development stages of this research phase, a really strong focus was placed on establishing iterative goals that would guide the process as a whole. As such, the first main goal was to understand the current state of the art in LADs, their integrability with LMS, and their applications in the educational context. To do so, it would be foolish to not also research the current state of the art in LA and LMS, as they are the main components of the proposed solution. This meant that the research process would have to revolve around these three main topics, which would prove to be difficult.

Table 3.1: Related studies - Sorted by Date

Work	Authors	Year	LA	LAD	AI
[53]	Park et. al	2015	x	✓	x
[54]	Matcha et. al	2020	✓	✓	x
[55]	Chaudy et. al	2021	✓	x	x
[56]	Valle et. al	2021	✓	✓	x
[57]	Williamson et. al	2022	x	✓	x
[58]	Salas et. al	2022	✓	x	✓
[59]	Peláez et. al	2023	✓	x	x
[60]	Nadimpalli et. al	2023	x	x	✓
[61]	Alfredo et. al	2024	✓	x	✓
[62]	Kaliisa et. al	2024	✓	✓	x
[63]	Valle Torre et. al	2024	✓	x	x
[64]	Paulsen et. al	2024	x	✓	x
[65]	Ouyang et. al	2024	✓	x	✓
[66]	Yan et. al	2024	✓	x	✓
This study	Cabral & Pinto	2025	✓	✓	✓

3.2.1 First Iteration

As expected, the vast amount of information available on the subject deemed it necessary to conduct a exploratory and less formal review to better acknowledge the existing SLRs on the area, as they are the most reliable way of understanding the current state of the art in a specific topic. This preliminary review was conducted by searching for reviews on the three topics of LA, LMS, and LAD. From this process, a major problem arose: there are not enough studies acknowledging all three topics at the same time: LA, LMS, and LAD, as well as its outcomes in the educational context. While LA and LAD appeared to be widely covered by the same articles, most papers analyzed either solely focused on the development of LADs, with little mentions of the integration of these tools with LMS, or did not refer to LMS at all.

Another problem was the fact that most of the articles did not place a big emphasis on the specific topic of LA, but rather on its applications in the development of LADs. This meant it would be harder to rely on a large number of documents to guide the development of a research plan that focused on all the three referred topics at the same time. After evaluating these results, it was clear that the initial goal of understanding the current state of the art in LA, LADs, their integrability with LMS, and their applications in the educational context was not feasible. It would be needed to refine the plan of research. Since LA is not the main focus of this research, and is more of a supporting component to the development of LADs, it could easily be discarded from the research plan, as studies focusing on LADs would already cover the main aspects of LA. This new perspective forced a decision to be made: either to focus on the integrability of LAD with LMS, or to focus on the impact of LADs on students' learning outcomes.

Although the integrability of LADs with LMS is a crucial factor in the development of the proposed solution, it does not pose much importance in the context of the research, as it is more

of a functional requirement than a research goal. After establishing that the integrability of LADs with LMS was not as crucial a factor in the proposed solution as the impact of LADs on students' learning outcomes, the main goal of this phase shifted towards understanding the current state of the art in LADs and the results that stemmed from their use in the educational context.

3.2.2 Second Iteration

The same process was carried out in order to fulfill this new goal. A new exploratory review was conducted, this time focusing solely on work that underlined the impact of LADs on students' learning outcomes. After an exploratory review to find the current SLRs on the topic, a specific article proved to be of great value to the research process, as it provided a detailed, comprehensive and recent overview of the current state of the art in LADs and its impact on students' learning outcomes [62]. As brilliantly noted by Kaliisa *et al.* [62] through a deep systematic review, LADs, in their current state, show little impact on students' performance outcomes, with the authors believing not to exist enough evidence to correlate the use of LADs with better student results. On the other hand, while not an enormous improvement, the article pointed out that the use of LADs did show a slight improvement in student engagement, a crucial factor in the learning process, as well as in student interaction, specifically in online learning environments. As shown, this research work practically answered the goal defined at the end of the first iteration, as it managed to provide a detailed overview of the current state of the art in LADs' impact on students' learning outcomes. This signaled the need to refine the research goals once again, in order to expand on the already existing knowledge on the topic. Despite the article's brilliance and it being fairly recent (March 2024), Kaliisa *et al.* [62] decided not to focus on the impact of employing AI-based algorithms in LADs as a way to enhance their possible outcomes, which is the hypothesis this thesis aims to prove. This meant that there was still a large gap in the current state of the art in the area and the proposed solution. So, after refining the research goal once again, the main objective now turned towards understanding the main differences between AI-powered LADs and traditional LADs, as well as the impact of AI-powered LADs on both students and educators, to better understand the potential of the proposed solution in the educational context.

3.2.3 Final Iteration

After an exploratory review of the current studies addressing AI in LADs, as no systematic or non-systematic review was found that focused on the impact of AI-powered LADs on educators' or students' outcomes, the main goal of the Literature Review phase was now clearly defined: Perform a SLR on the impact of AI in LADs' learning outcomes, both for students and educators.

3.2.4 Methodology

To perform the SLR, the guidelines outlined in the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) [67] statement were followed and the flow chart describing the Literature Review different phases is shown in Figure 3.1.

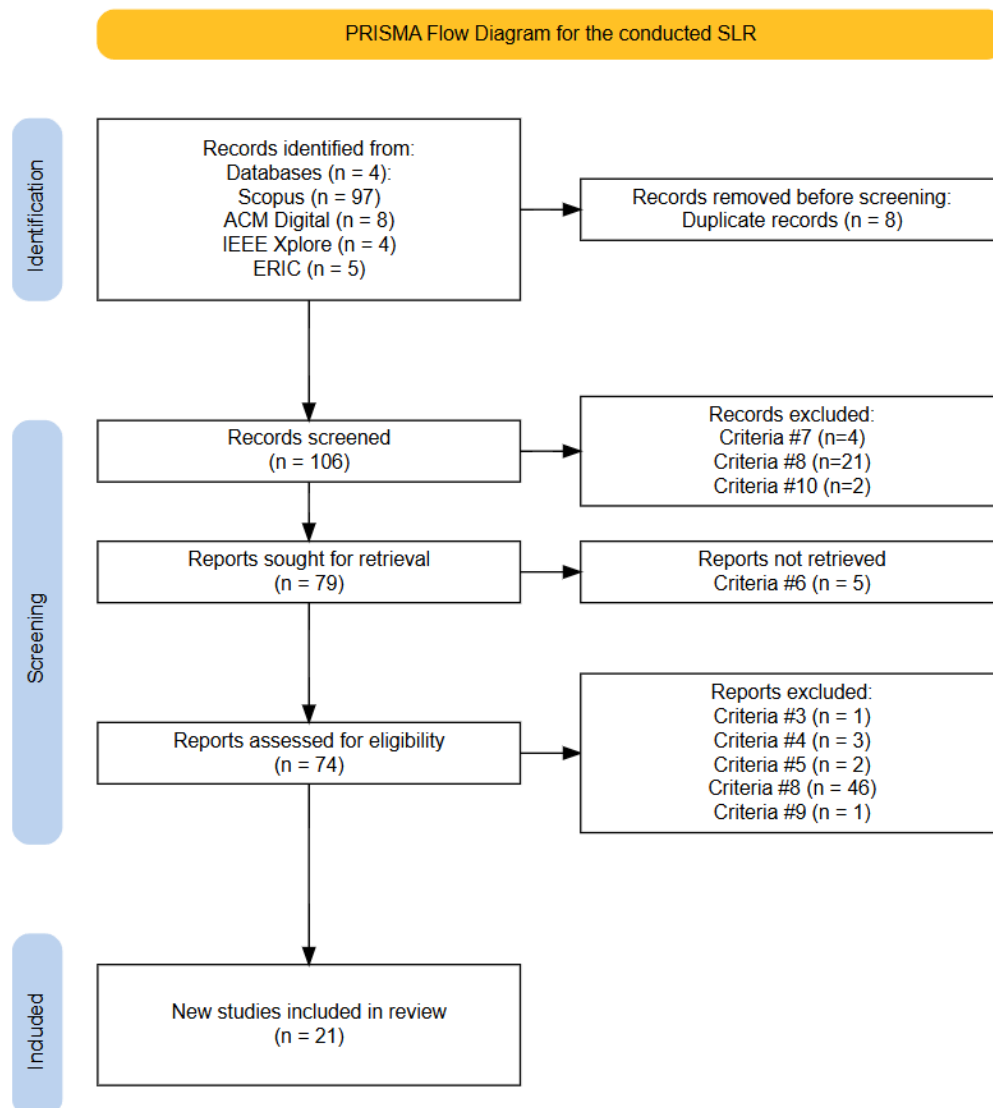


Figure 3.1: PRISMA Flow Diagram of the Selection Process

The search strategy developed aimed to ensure the selection of the most relevant documents for this research. The search query was the result of a trial-and-error process and suffered several modifications to ensure the retrieval of the most relevant documents to the research, as it was designed to retrieve documents that were either directly related to the development of AI-powered LADs, or that heavily contributed to the process of developing one. The full breakdown of the Search Queries¹ utilized in each database are shown in Table 3.2.

The (*AI algorithm OR artificial intelligence OR machine learning*) clause was used to ensure the documents retrieved from the search would present strong correlations with automated data

¹The search query explained is adapted to the specific standard of the Scopus database. The LAD acronym was not used in the search query, as it erroneously returned documents related to the expression "Logic Analysis of Data", which is not related to the topic of this research.

Table 3.2: Search Query Breakdown

Database	Search Query
SCOPUS	(AI algorithm OR artificial intelligence OR machine learning) AND TITLE-ABS(learning analytics dashboard)
ACM Digital	AllField:("AI algorithm" OR "artificial intelligence" OR "machine learning" OR "AI") AND (Title:("learning analytics dashboard") OR Abstract:("learning analytics dashboard"))
IEEE Xplore	("All Metadata": "AI algorithm" OR "artificial intelligence" OR "machine learning" OR "AI") AND ("Document Title": "learning analytics dashboard" OR "Abstract": "learning analytics dashboard")
ERIC	("AI algorithm" OR "artificial intelligence" OR "machine learning") AND (title: "learning analytics dashboard" OR "abstract": "learning analytics dashboard")

analysis or AI-algorithms that could help expand LADs ability to help students and educators, which are key components of the proposed solution. The (*learning analytics dashboard*) clause was used to ensure the documents retrieved were directly related to the main topic of this research: LADs.

The use of the TITLE-ABS operator in the last two clauses was used to ensure the documents retrieved had the term directly mentioned in the title or the abstract, which are the most relevant parts of a document to determine its content.

For the purpose of this systematic review, the databases used were Scopus [68], ACM Digital [69], IEEE Xplore [70] and ERIC [71]. This choice was made due to the databases' comprehensive coverage of computer science and educational technology literature.

3.2.5 Eligibility Criteria

To guide the selection process, the exclusion criteria were defined and are shown in Table 3.3.

Of this reviewing process must result studies directly related to the development of AI-powered learning tools that utilize LA to provide insights to both students and educators, or to already existing solutions in the area. Studies not inside the described scope will be excluded.

3.2.6 Title and Abstract Screening

As reflected in Figure 3.1, of the 114 initial documents, 8 were duplicates. After removing them, 106 documents remained. Of these, after performing a title and abstract review, 27 records were

Table 3.3: Exclusion Criteria for Study Selection

ID	Exclusion Criteria
#1	Release date before 2010
#2	Opinion articles or editorials
#3	Work-in-progress articles, retracted publications, or summaries from workshops, conferences, or seminars
#4	Proceedings
#5	Short papers
#6	Studies not available in full text
#7	Studies not published in English
#8	Studies not focusing on AI-powered LADs or not directly related to the development of the tool
#9	Studies that do not focus on the Education environment
#10	SLRs or articles that do not propose a solution

excluded due to the fact that they were either completely off-topic, or only briefly mentioned LADs, not making it the center of the research². After seeking the full-text of the remaining 79, 5 more documents were excluded due to the fact that its full-text was not available. After analyzing the remaining 74 documents, 46 of them were excluded since they made absolutely no reference to AI as a way to enhance LADs, or to AI algorithms, thus being irrelevant to the research proposed and 6 more documents were excluded, 3 being proceedings, 2 being short papers, and 1 being a work-in-progress article.

3.2.7 Final Selection

After the selection process, 21 documents were selected to be included in the systematic review³. The selected documents are listed in Table 3.4.

²Most of the documents excluded for this reason were merely mentioning the use of LAD in the educational context in order to fulfill a different need, but did not focus on the development of the tool itself, hence being irrelevant

³From the SLR, two pairs of articles, while not verbatimly equal, showed very strong inner similarities in both the research team and the work content itself. Therefore, from the pairs [72] - [73] and [74] - [75], only the most recent ones will be deeply analyzed, them being [72] and [75]

Table 3.4: Selected Documents

Document	Research Objective
[76]	Investigate the impact of a social-comparison dashboard on motivation and learning in university courses
[72]	Develop an Explainable Learning Analytics Dashboard (EX-LAD) for teachers and students
[77]	Evaluate AI-supported Smart Learning Environments for regulating behavior and improving learning
[74]	Explore current Learning Analytics Dashboard approaches and their challenges
[78]	Study the effects of dashboard information design on cognitive load and performance
[79]	Assess how explainable recommendations affect learning performance and self-regulation
[80]	Create a human-in-the-loop AI system to help educators identify student interventions
[75]	Measure the SensEnablr dashboard's impact on LMS engagement
[73]	Build an EX-LAD to assess performance, engagement, and perseverance
[81]	Design a dashboard providing actionable feedback to students
[82]	Develop Moodle Analytics Dashboard to detect at-risk students with machine learning
[83]	Explore how dashboards promote self-regulated learning in higher education
[84]	Create a dashboard to improve programming education diversity and outcomes
[85]	Design a data flow system for analyzing student behavior with integrated tools
[86]	Analyze dashboards' effectiveness combined with weekly personalized emails
[87]	Predict student outcomes and provide progress insights for educators
[88]	Study the relationship between learning activity and performance in Moodle
[89]	Investigate how instructors use dashboards to inform teaching practices
[90]	Propose a user-centered approach for adaptable dashboards
[91]	Develop a prototype dashboard for educators to analyze learning and teaching
[92]	Study professional learners' use of dashboards for exam preparation

3.3 Literature Review Results Analysis

By analyzing the graphs in Figure 3.2, it is possible to notice and identify a few patterns.

Firstly, when looking at the Pie Chart, it becomes clear that the majority (62%) of all articles were published during or after 2022. To understand or even explain this phenomenon, it is important to delve deeper into the first year of the range, 2022. As pointed out by García-Peñalvo and Vázquez-Ingelmo [93], despite AI having experienced an increase in popularity throughout the whole 2010s decade, the year of 2022 and, specifically, the later stages of it were the witnesses of the biggest explosive growth the concept had ever seen. Despite Gen AI not being entirely related to AI-Powered LADs or LADs in general, its growth may greatly influence the research on broader AI concepts or algorithms, hence the increase noted in the publication years after the

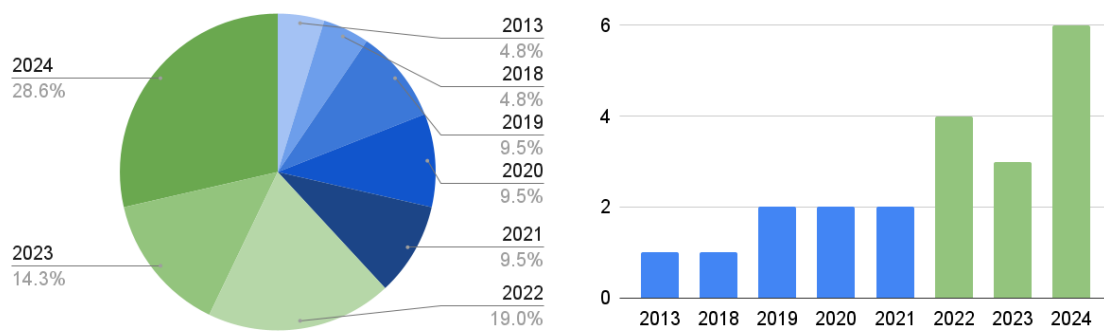


Figure 3.2: Publication Count per Year - Pie Chart and Bar Chart Distribution

“GenAI boom”. Another interesting pattern is the steadily increasing publication count through time, most likely propelled by the development of AI and AI-powered technology in recent years.

Although assuming the previously stated hypothesis may seem a little far-fetched, according to Mustafa *et al.*’s recent systematic review of AI specifically in education [94], doing so may have some viability in it. As pointed out by the authors, the landscape of AI in education in 2024 has reached a pivotal moment, both due to the widespread adoption of AI or AI-powered tools across different educational sectors and to the innovative advancements in AI both as a concept and a tool with major educational platforms integrating advanced AI functionalities that have the ability to both personalize learning experiences and facilitate real-time adjustments based on a student’s performance metrics and profile. This unprecedented paradigm can be a viable explanation for the increased academic interest in the last few years as the integration of AI-powered tools in the classroom has shifted upwards in the last decade, indicating a trend that will likely be sustained over the entire 2020s decade.

When analyzing the geographical distribution of the articles in the set, one thing becomes certain, the topic of AI-powered LADs is, more than anything else, a global research topic.⁴

As can be seen in Figure 3.3 and Table 3.5, despite Europe having produced the biggest number of articles (10 of the 21) the research in AI-powered LADs has undoubtedly reached a global status, as all five continents have been the birthplace of one or more of the research works this systematic review encapsulates.

⁴Each article’s birthplace was assumed to be the country of the institution or research lab where the study was conducted. In the few cases where it was not available, the birthplace was assumed to be the article’s first author country of origin.

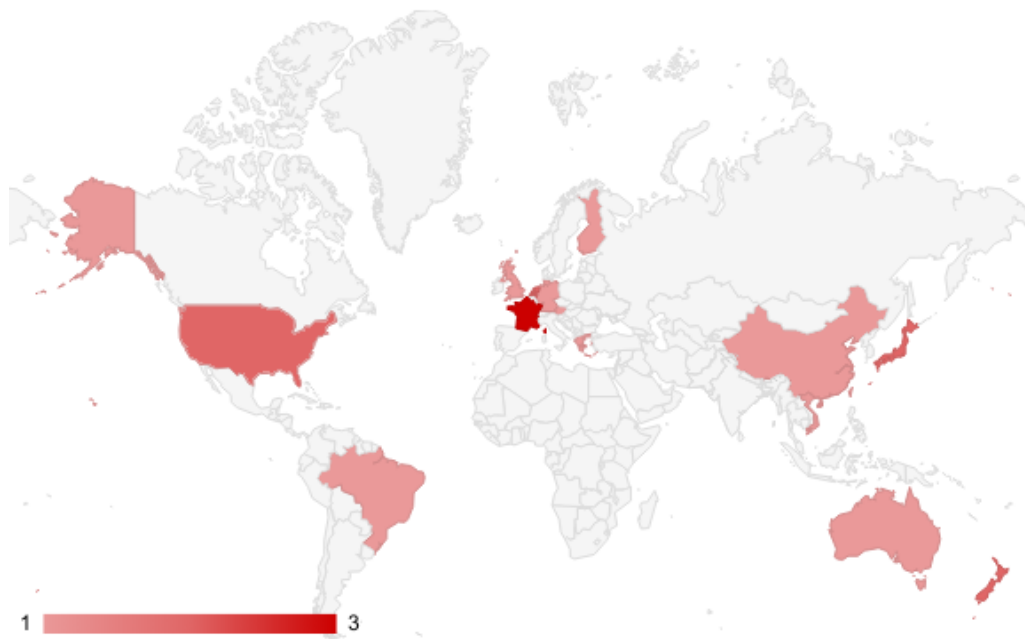


Figure 3.3: Publication Count - Geographical Distribution

Another interesting conclusion to draw from the geographical distribution lies on the fact that multiple countries produced more than one of the articles with the most notable example being France with 3 of the 21 articles.

Table 3.5: Distribution of Publications by Country (Sorted by Number of Publications)

Country	# Publications	Percentage
France	3	13.6%
Japan	2	9.1%
New Zealand	2	9.1%
The Netherlands	2	9.1%
United States of America	2	9.1%
Australia	1	4.5%
Austria	1	4.5%
Brazil	1	4.5%
China	1	4.5%
Finland	1	4.5%
Germany	1	4.5%
Greece	1	4.5%
Taiwan	1	4.5%
United Kingdom	1	4.5%
Vietnam	1	4.5%

To better understand the central themes and recurring terminologies across the selected literature, a word cloud analysis on the reviewed articles' abstracts was performed. This method's result is a visual representation of the most frequently occurring words, therefore offering insights into the dominant topics and research focus within the field. A limitation of this method, however, is the ubiquitous non-informative words such as "and", "the", "is" and "of" that, while frequent in natural language, do not convey any meaningful or conceptual information, posing the risk of overshadowing domain-specific vocabulary that reflects the core focus of the reviewed literature. Another group of words that can be detrimental to the word cloud's visualization accuracy are contextual words and verbs or adjectives that, while related to the study's domain, are overly common in the abstracts and do not add meaningful context or provide relevant insights like "learning", "course", "analytics" or "LAD".

Table 3.6: Stopwords categorized by Exclusion Reason

Category	Stopwords	Reason for Exclusion
Common Words	and, or, the, of, in, to, with, on, for, as, is, was, by, an, at, from, this, that, it, be, are, can, we, has, have, been, not, but, their	Frequently occurring words in English that lack thematic relevance.
Contextual Words	learning, course, analytic, analytics, LAD, study, results, model, approach, design, dashboard, research	Overused terms in the field, adding noise to the analysis.
Verbs and Adjectives	use, analyze, developed, found, using, significant, important, provides, effective, improve	Generic descriptors or actions that don't define themes specifically.

To ensure a more accurate visualization of the key trends and topics present in the literature and the abstracts, a list of stop words was constructed and can be seen in Table 3.6.

After constructing the afore mentioned list, a word cloud was generated using the WordCloud Python package [95] and can be seen in Figure 3.4.

By analyzing the word cloud, some trends emerge. Firstly, the most dominant words being student, data, performance, learner, and dashboard highlight the primary focus of the reviewed literature: analyzing educational data to assess and improve student performance through the use of LADs. Additionally, the prominence of predictive, system, and process suggests a strong emphasis on the technical and analytical underpinnings of LADs, particularly their ability to forecast student outcomes and support decision-making. The term online, paired with environment and support, reflects the prevalence of studies exploring LADs in digital or asynchronous learning contexts, which has become increasingly relevant post-pandemic.

On the other hand, secondary terms such as insight, behavior, feedback, and engagement indicate a growing interest in not just raw performance metrics but also in the interpretation and

Table 3.7: Overview of selected studies on AI-powered Learning Analytics Dashboards

Author (Year)	Objective	AI	DR	DV	Method	Evaluation	Key Findings
Fleur <i>et al.</i> (2020) [76]	Motivation via social comparison LAD	X	LMS	✓	Dashboard	Motivation, Grades	Increased engagement and outcomes
Khelifi <i>et al.</i> (2024) [72]	Explainable LAD for early support	X	LMS	✓	Usability-focused	Self-assessment, Engagement	Effective for student monitoring
Hu <i>et al.</i> (2022) [77]	AI-driven feedback using emotion data	✓	External	X	Behavior Recognition	Not Reported	Enables feedback, raises privacy issues
Susnjak <i>et al.</i> (2022) [74]	Overview of LADs and analytics depth	✓	LMS	✓	CatBoost	Prediction Metrics	Descriptive dominates; predictive underused
Cheng <i>et al.</i> (2024) [78]	Dashboard design vs. cognitive load	X	LMS	✓	LAD variants	Cognitive Load, Grades	Explanatory LAD boosts understanding
Majumdar <i>et al.</i> (2023) [79]	Explainable quiz recommendations	✓	LMS	X	BKT	ANOVA	Improves motivation, limited outcome effect
Khosravi <i>et al.</i> (2021) [80]	Human-in-loop LAD filters via AID	✓	LMS	✓	Drill-down Trees	Insight Quality	Enables subgroup analysis and filtering
Hellings <i>et al.</i> (2022) [86]	LAD + personalized email intervention	✓	N/A	✓	AdaBoost + Stump	$F1, R^2$	Behavioral changes, no final grade impact
Ramaswami <i>et al.</i> (2023) [75]	Identify at-risk students via ML	✓	LMS	✓	CatBoost	AUC, Accuracy, F1	High performance, interpretable insights
Khelifi <i>et al.</i> (2024) [73]	Assessing engagement via EX-LAD	X	LMS	✓	Usability-focused	Performance, Engagement	Teacher insight improved; at-risk flagged
Duan <i>et al.</i> (2022) [81]	Compare ML models for outcome prediction	✓	LMS	X	RF, LR, DT	AUC, F1, Precision	RF best; co-designed with instructors
Cechinel <i>et al.</i> (2021) [82]	Predicting success using LMS logs	✓	LMS	✓	AdaBoost	Accuracy	Engagement is key predictor
Kleimola <i>et al.</i> (2024) [83]	Promote SRL with dashboard support	X	SIS + Interviews	✓	Qualitative	Grades, Interviews	Facilitates planning and reflection
Groher <i>et al.</i> (2024) [84]	Support diversity in programming ed.	X	Assignment Logs	✓	Data Analysis	Tasks, Gender Info	Gender gaps and submission trends
Liu <i>et al.</i> (2020) [85]	Enhance LAD via xAPI integration	X	SCROLL + xAPI	✓	System Integration	View Logs, Quizzes	Personalization via behavior logs
Nguyen <i>et al.</i> (2018) [87]	Forecast outcomes with regression	✓	LMS	X	K-means + Regr.	MAE ⁵ , MSE, R^{26}	Accuracy improves over weeks
Kotsiantis <i>et al.</i> (2013) [88]	Analyze Moodle activity and learning performance	X	LMS	X	C4.5 Decision Tree	Accuracy	Optional activity and perceptions predict performance
Wise <i>et al.</i> (2019) [89]	Instructor use of LAD for pedagogy	X	LMS	✓	Observational	N/A	Translating data into action is complex
Ines <i>et al.</i> (2019) [90]	UCD approach for adaptive LADs	X	Digital Traces	✓	User-Centered Design	Visual Indicators	Captures needs, supports decisions
Kodumuru <i>et al.</i> (2022) [91]	Predict dropout with LAD	✓	Activity Logs	✓	LogReg + PCA	Team Data	Prediction supports team monitoring
Sapsai <i>et al.</i> (2023) [92]	Dashboard for teacher awareness	X	LMS	✓	Adaptive AI Framework	Interaction Logs	Encourages teaching reflection

Legend: ✓ = Used; X = Not Used; DR = Data Retrieval (LMS, External, SCROLL, SIS); DV = Data Visualization; AI = Use of ML/AI

Methods: RF = Random Forest, LR = Logistic Regression, DT = Decision Tree, BKT = Bayesian Knowledge Tracing, AID = Automated Insightful Drill-down

just cited studies that focused on it. As AI / ML is a crucial component of this work, this set will not be extensively described in this subsection. The second set of articles, however, provides extensive knowledge on AI / ML techniques specifically employed in either predictive or prescriptive analysis, as well as instructions on how they were used in each article, thus, will be extensively analyzed. A summarized description of every article in this set can be seen on Table 3.8.

Despite not focusing to a great extent AI / ML techniques to predict students' learning outcomes, Kotsiantis *et al.*'s article [88] does point out interesting techniques that could be extremely useful when developing an LAD, and actively employs a ML algorithm no other study in the group mentions, therefore deeming it worthy of being examined in detail. The article's goal was, according to the authors, threefold: study whether students' learning performance is affected by their activity in an LMS, in this case Moodle, identify what types of collected data help more in predicting students' learning outcomes, and understand if the knowledge obtained during the investigation could inform the design of a LAD. Like in previously reviewed articles, Kotsiantis *et al.* also started by collecting the data from the LMS, pre-processing it, using algorithms such as the C4.5 decision tree [96] to perform the predictions and then interpreting the results derived from the method. To collect the needed data and pre-process the log files, the authors used a tool named Moodle Parser, specifically developed by one of the authors. To the best of the knowledge available on the subject, this tool was not further developed, nor published anywhere online, therefore deeming it irrelevant for future researches. After retrieving 19 different variables related to both the students' perceptions about the Moodle LMS, (opinions collected after lab completion via survey) as well as their activity, the authors employed a C4.5 decision tree algorithm to identify which of the variables could help predict students' grades, as well as their overall course completion. After analyzing the results of the algorithm's execution, as well as a varplot produced using the improve criteria reported by Hastie *et al.* [97], the authors concluded that the students' perception of Moodle materials' usefulness as well as the number of times each student interacted with the LMS strongly influenced their final grade, which is expected. From these conclusions, the research team extrapolated that, in general, failure in the studied course is closely associated with the negative attitudes and perceptions of students towards Moodle, while excellence in it is associated with increased usage. Additionally, Kotsiantis *et al.* also point out that the use of specific LMS services such as forums and glossaries can indicate students' motivation to learn and can be a successful learning strategy. As a final step, the authors decided to cluster the students using the k-Means algorithm [98], mentioning clustering as a helpful technique to identify learners with different expected performance and prompt collaboration groups reorganization, accordingly, thus aiding educators. Despite not developing it and only setting it as a future goal, the authors mentioned that, to further help educators adapt their teaching in a more objective manner without the need for extensive technical savviness, a visualization tool would reduce the effort needed to deploy LA techniques in a real-world setting, increasing the likelihood of the tool's adoption.

The second article in the set, contrary to the other ones, approached the employment of AI / ML in a less conventional way, and therefore should be treated individually. Hu *et al.* [77], envisions the integration of AI technologies in LADs as, not only a way to improve a learner's outcomes by

Table 3.8: Summary of Articles Reviewed based on their AI Application

Author(s) and Year	Objective	Methodology	Metrics	Findings
Hu <i>et al.</i> (2022) [77]	Integration of AI to improve learning outcomes and provide self-reflective insights	Facial, image, and behavior recognition integrated into visual dashboards	None explicitly stated; focuses on potential application	Highlights privacy concerns and potential for real-time feedback.
Khosravi <i>et al.</i> (2021) [80]	Development of human-in-the-loop AI for LADs	Automated Insightful Drill-Down using drill-down trees and scoring functions	Insight quality (qualitative)	Proposes a unique way to filter and analyze student subpopulations.
Duan <i>et al.</i> (2022) [81]	Performance evaluation of ML models for student outcome prediction	Comparison of logistic regression, random forest, decision trees, and others	Accuracy, AUC, F-measure, precision, recall	Random forest performed best; emphasizes instructor collaboration.
Nguyen <i>et al.</i> (2018) [87]	Creation of a forecast model for learning outcomes with different ML techniques	Multi-module system using K-means clustering and regression models	R^2 , MSE, MAE over weeks	Accuracy increases over time; detailed time interval-based analysis.
Majumdar <i>et al.</i> (2023) [79]	Development of explainable AI tools (EXAIT) for math quiz recommendations	Bayesian Knowledge Tracing (BKT) and LMS activity logs	Performance difference between groups (e.g., ANOVA)	Explanations improve motivation but limited effect on outcomes.
Ramaswami <i>et al.</i> (2023) [75] [74]	Identification of at-risk students with course-agnostic features	CatBoost algorithm with feature importance analysis	F-measure, accuracy, AUC	CatBoost outperformed other algorithms in accuracy and interpretability.
Cechinel <i>et al.</i> (2021) [82]	Prediction of student success using LMS interaction data	AdaBoost with interaction count metrics	Accuracy (not explicitly stated)	Highlights engagement as a key predictor in distance learning.
Kotsiantis <i>et al.</i> (2013) [88]	Investigate the relationship between students' learning performance and their activity in a Learning Content Management System, specifically Moodle, in the context of a blended learning academic course	Data mining and statistical analysis methods to analyze students' performance prediction data. C4.5 Decision Tree Algorithm	Accuracy (not explicitly stated)	Importance of students' perceptions and non-obligatory activities in predicting learning outcomes
Hellings <i>et al.</i> (2022) [86]	Analyze the effectiveness of a LAD with personalized weekly emails	Decision Stump combined with AdaBoost to predict the chance of success in the course	F-Measure and R^2	The dashboard had positive effects on student online behavior, but did not significantly impact the final exam performance

predicting his/hers grades in an accurate manner, but also as a way to “make students aware of their learning behaviors and reflect these against the agreed goals”. For this, Hu *et al.* aim to employ “facial recognition, image and behavior recognition” in the system, to track and analyze students’ activities, such as their engagement levels and interactions within the classroom setting. This data can be then transformed into visual dashboards, provide both learners and educators with insights into different learning patterns, as well as help students self-reflect and adapt their strategies in order to achieve a better goal. Furthermore, teachers can utilize this data for immediate feedback, allowing them to intervene and assist students who can be either struggling or disinterested. On a critical level, while the research in this article can be a breakthrough in technology-powered education as a whole, the ever-growing concerns regarding data privacy can prove to be an obstacle in retrieving these kinds of data in a classroom setting.

Khosravi *et al.*’s article, *Intelligent Learning Analytics Dashboards: Automated Drill-Down Recommendations to Support Teacher Data Exploration* [80], also attempts to tackle the problem of choosing AI methods to help both students and educators in LADs in a unique way. As opposed to focusing on one specific AI method, the research team intended to, in their words, “contribute to the understanding of how human-in-the-loop AI methods (...) can be employed to develop successful LADs”, by analyzing multiple “state-of-the-art algorithms” from several different fields like online analytics processing, data mining and process mining. The approach, deemed Automated Insightful Drill-Down (AID) consists of “two algorithms that use drill-down trees and a scoring function” meant to analyze all possible drill-down strategies to recommend the most insightful ones which, for the authors, are sets of filtering rules able to identify a subpopulation of students that, on different metrics, deviate from the class mean. An example of a common use of a drill-down tree can be seen in Figure 3.5.

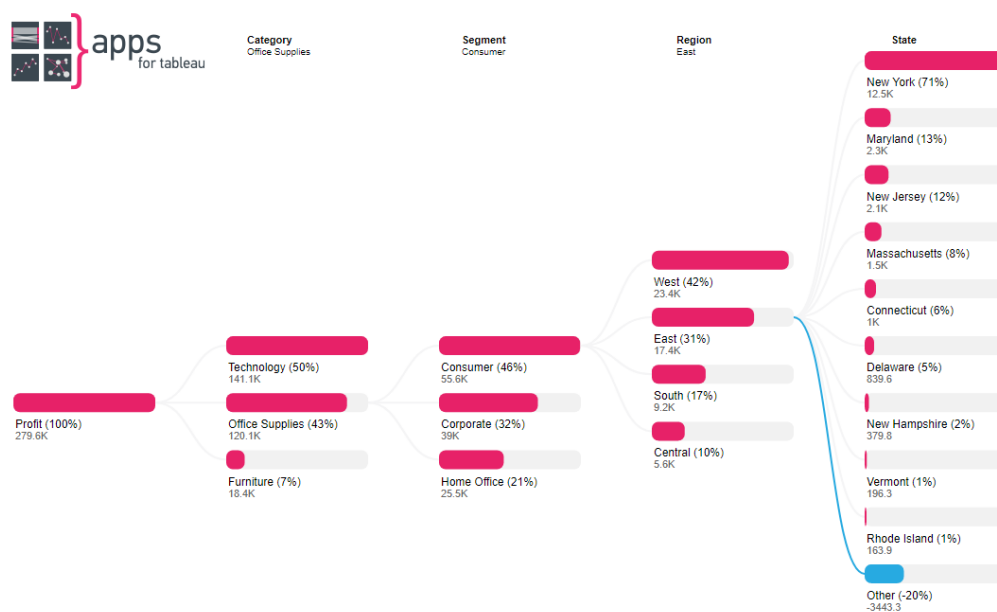


Figure 3.5: Example of a DrillDownTree by Tableau Exchange [1]

When it comes to the actual algorithms used to provide the wanted insights, the authors chose, to measure a student's performance-based variability, meaning the divergence between the performance distribution of a set of students and the rest of the class, the Kullback-Leibler Divergence function, not sustaining it on any particular research other than Kullback and Leibler's themselves in their 1951 article. [99]. Additionally, to compare the learning process of one set of students with the overall population, Khosravi *et al.* picked "Earth Movers' Stochastic Conformance Checking", a process mining technique that takes into account a student's performed learning activities, activity frequency and activity sequence. [100]

Duan *et al.*'s article [81], similarly to others in this category, to choose the most appropriate AI / ML techniques to employ in their prediction model, conducted a performance evaluation on a set of ML models that included logistic regression, k-nearest neighbors, naïve Bayes, decision tree and random forest regression on the features the researchers had previously extracted. To correctly extract the relevant features for the model to train on, the research team consulted with course instructors and, through intense collaboration with them, managed to construct a set containing both commonly used simple features in student success prediction, such as a student's click count on the course materials, tutorials or online discussions, as well as more complex features like the number of days before the deadline a student submitted a tutorial for the first time. While overcomplicating specific features could definitely lead to overfitting, as pointed out by Hawkins in the 2003 article *The Problem of Overfitting* [101], Duan *et al.*'s article makes a strong point in suggesting collaboration between the developer and instructors as the synergy between both parts can lead to the development of a not only more intuitive, but also more accurate LAD.

This evaluation, not alike of other articles in the set [75], took into account the model's accuracy, AUC [102] and F-measure [103], while also including precision and recall [104] as relevant metrics to deem the model useful and capable of performing on student data. While this article did not provide a table like Ramaswami *et al.*'s containing the specific results for each model, they concluded, from the evaluation process, that the random forest regression model performed the best being picked as the optimal one not just because of its performance, but also of its easy interpretability, which facilitated a discussion with the instructors.

Nguyen *et al.* [87], in a very direct, yet slightly contradictory way, state, based on their Literature Review, that there are several problems with creating a model to predict learning outcomes. The authors conclude that algorithms cannot accurately predict the learning outcomes of learners while also concluding that using only a single algorithm it is difficult to predict the learning outcomes of learners. Despite these remarks, the researchers tried to build a forecast learning outcomes model using some of the same features previous works have already built on. These features vary from the degree of interactivity, frequency of use, level of focus on the materials and even the times students actually interacted with the LMS, in this case, Moodle. As for the construction of the model itself, the research team decided to take a different approach from the other articles in this group, as they took the liberty of creating a system with three different modules: a Log Analysis Module, a Forecast Module and a Notification Module. As the second component is the only one of the set dealing with AI / ML, it will be explained in more detail.

The Forecast Module is, in Nguyen *et al.*'s words [87], “the core component of the predictive model” and utilizes ML techniques such as classifying, clustering and regression to analyze information, classify the students and predict their outcomes. To classify students as at-risk, or not at-risk, a regression model predicts the amount of interaction needed at the time to anticipate learning outcomes and uses that prediction to classify learners into specific classes. This classification phase is performed using the K-means algorithm, with the authors pointing out as the basis for this choice the Silhouette correlation coefficient result being vastly higher than that of other ML algorithms, namely Birch and Agglomerative Clustering. After the entire classification process, the newly labeled data is then used to create a forecasting model for future predictions to be made.

After experimenting with the model, the authors evaluated it by calculating the correlation coefficient in different time instances. Nguyen *et al.* concluded that, in the first weeks of the course, the prediction of attribute values is extremely inaccurate, gradually increasing in accuracy as the semester goes on. The researchers compiled the results of these calculations in several intervals, reaching remarkable Mean Squared Error (MSE) [105], Mean Absolute Error (MAE) [106] and R^2 rates [107] in the later stages of the semester, as can be seen in Table 3.9

Majumdar *et al.*'s article [79] focuses on the development of a module specifically designed to aid students with math quiz recommendations and explanations called Educational eXplainable AI Tool, or EXAIT, in short. This system attempted to help students complete their summer vacation assignments and built on the notion of Explainable AI, a concept gaining traction in the education domain. In an attempt to understand the effects of explainable recommendations on students and the self-regulated learning patterns exhibited by them while learning with explainable recommendations, the research team conducted a study on high school grade II mathematics students from six different classes. This study, focused on a summer vacation assignment, gave students an opportunity to see their visualized reading behaviors and presented them with recommended problems based on each student's prior reading behavior data in a dashboard. Each recommended problem would also have an explanation for its recommendation, the previously mentioned EX-AIT. To do so, the authors leveraged on Bayesian Knowledge Tracing (BKT), an algorithm already employed in many different intelligent tutoring systems and first developed by Corbett *et al.* [108]. Briefly, BKT is, as Pelánek puts it [109], a special case of a hidden Markov model. To understand this definition, it is crucial to first comprehend what a Markov and a Hidden Markov Model really are.

In short terms, a Markov model, named after the Russian mathematician Andrey Markov [110], is a stochastic model used to represent systems that evolve randomly under the Markov property. This property asserts that the probability of transitioning to a future state depends only on the present state and is independent of the sequence of states that preceded it. In logical terms, if the system transitions through states $A \rightarrow B \rightarrow C$, then the probability of transitioning to C depends only on the state B, not on A. As stated by Boyd and Lau in *An introduction to Markov modeling: Concepts and uses* [111], the Markov property can be both a blessing and a curse, since, on one hand, allows for system tractability that, due to overly complex computational operations would

Table 3.9: Nguyen *et al.*'s Results of calculating Correlation Coefficient in Predicted Interactive Activities over the Weeks

Weekspan	Predicted Interactive Activity	R^2	MSE	MAE
W3 - W7 Linear Regression	View	0.490	1.979	1.106
	Post	0.380	0.020	0.122
	Forum View	0.578	2.077	0.993
	Forum Post	0.935	0.0002	0.005
	Success Submission	0.496	0.0059	0.058
W3 - W15 Linear Regression	View	0.129	10.089	2.502
	Post	0.124	0.122	0.296
	Forum View	0.291	1.629	0.860
	Forum Post	-2.66	0.002	0.032
	Success Submission	0.144	0.031	0.161
W6 - W7 Linear Regression	View	0.977	0.088	0.224
	Post	0.939	0.0019	0.043
	Forum View	0.809	0.9395	0.269
	Forum Post	0.977	9.305	0.002
	Success Submission	0.939	0.0007	0.025
W6 - W15 Linear Regression	View	0.595	4.6945	1.686
	Post	0.290	0.0992	0.267
	Forum View	0.408	1.358	0.700
	Forum Post	-0.633	0.0010	0.027
	Success Submission	0.468	0.0193	0.129
W10 - W15 Linear Regression	View	0.876	1.428	0.875
	Post	0.756	0.034	0.153
	Forum View	0.637	0.833	0.550
	Forum Post	0.340	0.00043	0.0187
	Success Submission	0.652	0.012	0.0968
W13 - W15 Linear Regression	View	0.976	0.274	0.3374
	Post	0.927	0.010	0.0614
	Forum View	0.834	0.379	0.3517
	Forum Post	0.551	0.0002	0.0076
	Success Submission	0.912	0.0032	0.0462

otherwise be impossible to perform, while on the other hand, is not always consistent with the reality of real-world system behavior, since real systems tend to have their future behavior depend in more than their current state. Another characteristic of real-world systems lies in the visibility of every sequential state which, in many cases, is impossible, thus transforming a Markov chain into a Hidden Markov Model. As brilliantly explained by Sean R. Eddy in 1996 [112], a Hidden Markov Model is a model where the sequence of states is a Markov chain (the choice of the next state to occupy is dependent on the identity of the current state) yet the state sequence is not observed, therefore forcing an inference of the state sequence, rather than absolute knowledge of it.

To Pelánek [109], in BKT, skill is modeled as a binary variable, and learning is modeled as a transition from an unknown to a known state. To make predictions about a student's learning

process, the model uses probability data, such as the probability that a skill is initially learned, the probability of learning a skill in one step, the probability of an incorrect answer when the skill is learned (Slip) and the probability of a correct answer when a skill is unlearned (Guess), as well as domain data like the definition of the knowledge components.

Majumdar *et al.*, leveraging BKT parameters Guess and Slip calculated from the students' LMS logs, were able to generate explanations for the specific recommendations which, in their view, helped improve the transparency, persuasiveness and trustworthiness of the study. To evaluate EXAIT, the researchers separated the students in two groups, where the only differentiating variable lied in the possession of the AI recommendation module. In the *RecoCon* group, the one without the EXAIT module, only the quiz titles were displayed while in the *RecoExp* group both the quiz titles and the explanations for their recommendation were presented to the students. To analyze the entire system's usefulness and results, the research team also created a *RecoNone* group, where no student had interacted with the system, thus leaving no trace in the learning logs. Upon analyzing the performance results, also presented in Table 3.10, the authors concluded that while the difference between the three groups were significant, $F = 5.475$, $p = .005$, and the differences between *RecoExp* and *RecoNone* were also considerable, between *RecoCon* and *RecoNone* and between *RecoExp* and *RecoCon* were practically irrelevant.

Majumdar *et al.* also concluded that, after analyzing the behavior patterns, students in the *RecoExp* group showed higher proportions of active behavior, leading to infer that displaying explanations for these types of recommendation processes can lead to an increase in student motivation, despite of how they might influence the student's learning outcomes. While the researchers point out that the non-mandatory aspect of using the EXAIT module could have influenced the study, Majumdar *et al.* also point out that the providing of explanations of AI recommendations can benefit strongly from being placed within the aligned workflow of the learning task, as well as from being introduced in class before prompting independent use.

In conclusion, despite the concrete study results not being, in the research team's view, indicating of strong success in helping students and educators fulfill their expectations and improve their outcomes, this study offers very important arguments that can help in developing student-facing platforms, as well as in improving student motivation to adopt those same platforms, if available to them.

Table 3.10: Majumdar *et al.*'s Performance Scores

Group	N	Marginal Means	95% CI Lower	95% CI Upper	SE
RecoExp	60	63.083	57.81	68.356	2.676
RecoCon	66	59.045	54.018	64.073	2.551
RecoNone	96	52.198	48.029	56.367	2.115

Additional Statistics

$$F = 5.475, p = .005, \eta^2 = 0.048$$

When considering the other articles in the subdivision, an interesting trend regarding predictive analysis emerged. Despite not exactly in the same way, four of the remaining seven articles employ

a boosting algorithm to try to predict students' learning outcomes. Of these four, half of them use a variation of AdaBoost [82] [86] and the other half an algorithm deemed CatBoost [74] [75], popularized by Dorogush *et al.* [113], however, since the research team for the last two is the same and the articles content and contributions are extremely similar, only Ramaswami *et al.*'s work will be analyzed, since it is the latest one.

3.4.1.1 CatBoost

In layman's terms, CatBoost, or Categorical Boosting, is an open-source gradient boosting algorithm designed to handle categorical features efficiently in ML [114], taking advantage of dealing with the features during training, rather than preprocessing time. While Susnjak *et al.* [74] do not give a concrete explanation as to why CatBoost was the preferred algorithm, Ramaswami *et al.* [75] pointed out that, in a set with Random Forest [115], Naïve Bayes [116], Logistic Regression [117] and K-Nearest-Neighbors [118], when attempting to identify at-risk students, CatBoost outperformed all the other classifiers, as can be seen on Table 3.11.

Table 3.11: Ramaswami *et al.*'s Performance Scores of Classifiers with Standard Deviations

Classifiers	F-measure	Accuracy (%)	AUC
CatBoost	0.77 ± 0.02	78 ± 2.1	0.87 ± 0.02
k-nearest neighbours	0.71 ± 0.02	71 ± 2.4	0.72 ± 0.02
Naïve Bayes	0.68 ± 0.02	68 ± 2.3	0.71 ± 0.03
Logistic regression	0.67 ± 0.03	68 ± 3.0	0.73 ± 0.03
Random forest	0.67 ± 0.03	67 ± 2.4	0.74 ± 0.02

Upon analyzing the results, it becomes clear that, while not perfect, CatBoost can be a very strong candidate to consider when choosing an LAD prediction algorithm.

Ramaswami *et al.*, in an attempt to ensure the model's usability for students of different courses and backgrounds aimed to pick "course-agnostic and thus generic features" to build the classification models focusing mostly on students' average scores of prior courses, maximum scores achieved in prior courses, assignment scores and others. As a result, to ensure student features were not tightly coupled to his/her course's particularities, the researchers took it a step further by calculating the z-score (or standard score). This engineered feature relativized the student's score on a particular feature, by taking into account the score of their peers, thus capturing the degree of deviation from the class mean [119]

To train the model, the researchers chose a dataset comprising 4000 students and used an hold-out method while using data in a cumulative fashion to make predictions. This choice would allow for continuous model accuracy improvement as more data becomes available since a student's week 2 prediction would take into account his/hers week 1 data. The team also aimed to maintain the model's interpretability by using feature importance analysis, depicting the contribution of each variable towards the prediction making process. One particularly interesting choice by the research team was to display the model's accuracy to learners, as a measure to ensure confidence

in the reliability of the underlying model. This ensures students know that, while accurate, the model can fail, avoiding student disengagement due to false predictions of unsuccess.

3.4.1.2 AdaBoost

On the other hand, AdaBoost, or Adaptive Boosting, is a meta-algorithm formulated by Yoav Freund and Robert Schapire in the 1995's paper "*A decision-theoretic generalization of on-line learning and an application to boosting*" [120]. As explained by Schapire himself [121], AdaBoost "was the first practical boosting algorithm, and remains one of the most widely used and studied", and, in short, it works by combining different weak classifiers in order to form a stronger one by tweaking the training samples' weights until a good performance is achieved. [122].

Cechinel *et al.*'s explanation for picking AdaBoost as the deemed algorithm to predict students' success [82] ties to the already proved results the approach attained in previous experiments, specifically pointing out Buschetto *et al.*'s research in the area [123] as proof. Conversely, while Hellings *et al.*'s [86] explanation is a little more general mentioning previous cohorts, the authors sustain their choice in Iba *et al.*'s, Rojas *et al.*'s and Hu *et al.*'s research on the subject [124] [125] [126]. As opposed to Cechinel *et al.*, this article intends to use a combination of AdaBoost and Decision Stump, a binary classification ML algorithm that makes a prediction (or a decision) based on one single feature. [127] [128]

When it comes to picking the features to test the predictive model on, the article pair is heterogeneous in the approach.

For example, Cechinel *et al.* point out that students success, specially in distance learning courses, is highly correlated with their engagement inside LMS, citing Agudo-Peregrina *et al.*, Jayaprakash *et al.* and Smith *et al.*'s work as concrete proof of this thesis [129] [130] [131]. Knowing this, the authors point out that, since efficacy and efficiency of students has been "highly associated with different measures of their interactions inside the LMS", which in turn are correlated with their success, using the count of students interactions in the LMS could be a viable way to generate acceptable predictions, referencing Queiroga *et al.*'s work in both "*A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course,*" [132] and the "*Modelo de predição da evasão de estudantes em cursos técnicos a distância a partir da contagem de interações,*" [133], Detoni *et al.*'s work [134] and Buschetto *et al.*'s work once again [123] as researches that previously demonstrated this approach's viability.

Hellings *et al.*, on the other hand, base their prediction models on a mix of student background variables (like prior education level), and variables of online behavior (such as the number of practical assignments a student must complete or the average grade of those assignments), not particularly delving into the background for these choices or sustaining them in any previously conducted research.

3.4.2 Data Retrieval Based Categorization

As defined by Clow [135], LA is the application of Big Data techniques to improve learning, therefore, despite being self-explanatory, it is never enough to stress the importance of data in LA and, subsequently, LADs. By serving as the foundation for both the analysis and improvement of learning experiences, data holds an essential role in the field of LA, and without it, the potential for deriving insights and driving meaningful educational interventions is significantly diminished or even nullified. However, the utility of data in LA projects hinges critically on its quality and comprehensiveness, making Data Retrieval (DR) the first step in creating the foundation upon which LA efforts are built, as it shapes the scope, depth, and reliability of the insights that can be drawn. In typical LA projects, data can be collected from various sources such as LMS, assessments, or student interactions, enabling educators to identify trends, predict outcomes, and tailor interventions to individual learners' needs. Since the methods and sources of data collection are critical to the effectiveness of both LA and LADs, it is imperative to analyze the set of articles considering how they deal with data in this sense, aiming to uncover patterns, gaps or the best practices in data collection strategies for LA projects. After a deeper analysis, three categories were formed:

- Article defends DR from the LMS, or performs it
- Article defends DR from other sources, or performs it
- Article does not expand on DR processes

Around 57% of the articles either defended data collection from the LMS, or executed it during their research, which shows that relying on the LMS the students normally use can be useful when collecting student data. As for the articles present in group 3, while most of them discussed the development of an LAD, they did not present nor focus on ways to retrieve student data. In the interest of understanding in what other ways data can be retrieved in these types of projects, a slightly deeper analysis will be performed on the articles in group 2.

Apart from Hu *et al.*'s [77] unconventional idea of capturing student behavior from webcams and microphones, most of the other articles discussed one of two ways of collecting data: interviews with students [83] [136] or student responses on assignments and quizzes [84] [85]. It is, however, also important to notice that, while some of the articles present in set 2 did not defend or perform data collection operations on students' LMS, they did mention it either in their literature review or related work sections [77] [84], solidifying LMS-based retrieval as the dominant method for data collection when developing LA projects.

While interviews and self-reporting questionnaires can be less intrusive ways of gathering student information, it is undeniable that the level of detailed and objective data derived from those methods is significantly lower than from querying the platform where students usually spend most of their time studying on. As a result, since AI / ML algorithms for data analysis heavily benefit from precise and concrete data, both methods described can be ineffective ways of collecting data for LA projects. On the other hand, as most students' assignment and quiz responses are already

stored on LMS, whether open-source or self-developed, there is no need to dive deeper into the process of gathering data, as it is fairly analogous to the one performed in them.

3.4.3 Data Visualization Based Categorization

In a digital world where the use of dashboards has spread to the most various application domains, one thing remains to be unanimously claimed: Any dashboard's most valuable feature is the ability to allow users to analyze and understand data in the blink of an eye as, according to Ajibade and Adediran's article "*An Overview of Big Data Visualization Techniques in Data Mining*" [137], in order for any decision maker to make informed choices regarding enormous volumes of data, it becomes necessary to implement easily understandable presentations that allow the identification of either structures or patterns not as easily noticeable in simple data rows analysis.

Despite its importance, the choice of Data Visualization (DV) techniques to represent different information in LADs is not entirely agreed upon by the research community, or even strongly focused on. As a result, it becomes of crucial importance to analyze the results of the SLR in light of the different DV techniques either employed or defended by the authors and discuss them in detail.

After analyzing the SLR results and compiling their work and mentions of DV techniques, it became possible to divide the set of articles into two different categories, presented as follows:

- Article expands on DV Techniques used
- Article does not expand on DV Techniques used

It is, although, important to notice that, for many of the articles in the set, the names used by the authors to categorize the DV technique employed either deviate from each other while being the same technique ⁷, or just do not exist altogether, resulting in a complicated compilation process. To aid it and ensure coherence in the retrieval of the information, several steps were taken. Firstly, whenever possible, if an article did not explicitly mention in a table or in plain text the name of the DV technique, an analysis of the authors' proposed LAD screenshot or mock-up was performed, comparing the DV components of the finished work with the commonly known types. If an article did not provide any images of the LAD or descriptions of the DV techniques employed, it would be categorized in the second group. To create a cohesive list of DV techniques names to analyze, Severino Ribbecca's *The Data Visualization Catalogue* [138] was used, as it provides an extensive, yet intuitive list of DV techniques that aids in standardizing the names mentioned by different authors.

It becomes clear that the majority of the articles resultant from the SLR pay a great deal of attention to the way data is presented to the end-user, with most of the articles mentioning more

⁷An example of these discrepancies is the fact that some of the articles employ *Normal Distribution* or *Gaussian Distribution* curves with some characterizing them as Line Graphs and others labeling them by the curve designation. While mathematically correct, Gaussian or Normal Distributions are not names of DV techniques and, therefore, should be characterized as Line Graphs. This deviation is taken into account and these curves fall under the umbrella of the Line Graph concept.

than one DV technique employed in the making of the LAD. However, since just knowing which articles expand on it will not be fruitful to the development of an LAD from scratch, a deeper analysis is needed in the set of 17 articles that do mention DV techniques to understand what the prevalent trends, ideas and options are that aim to ease the user into the understanding of data. To do so, the main techniques pointed out by each of the 17 articles were retrieved and organized into Figure 3.6.

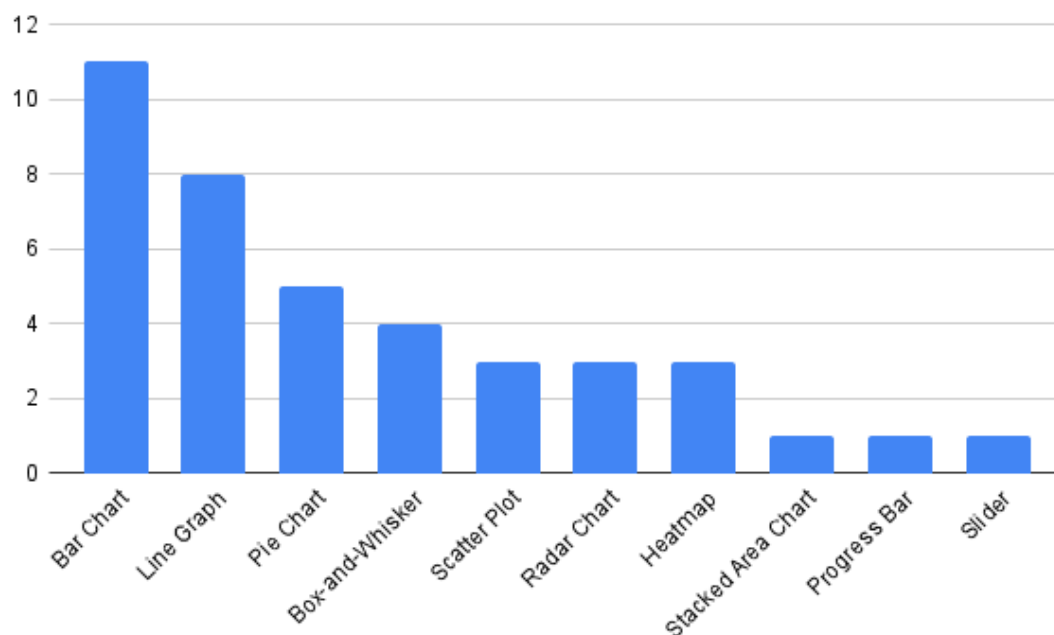


Figure 3.6: DV Techniques Across the SLR Results (Sorted by Publications)

After performing an analysis on the Figure it is possible to infer that, when it comes to the LADs resultant from the SLR, employing simple DV techniques proves to be the researchers' choices with techniques like Bar Charts or Line Graphs topping the ranking, with Pie Charts not far behind.

Notwithstanding the obvious trend of implementing uncomplicated DV techniques in the proposed LADs, it may still prove useful to understand what, if any, the articles' authors explanations for these choices are. Unfortunately, upon reanalyzing the research works, it is easy to point out that barely any argument is raised either in favor or against specific DV techniques or its employment in LADs, with the exception of Khelifi *et al.*'s works [72] [73]. This lack of reasoning from 15 of the 17 articles indicates that, although DV plays a major role in LADs fruitfulness and overall efficacy, its techniques can either be so self-evident that do not need an explanation to sustain their choosing, or not enough effort is placed upon the justification for its employment. Nevertheless, it is important to highlight that, while only Khelifi *et al.*'s articles provide a direct explanation for their choice of DV techniques, a considerable subset of the 17 articles do mention the importance of choosing the right technique to convey the message the data indicates in a simple, intuitive and

direct way. However, for the purpose of this analysis, as the objective of it was to determine the reasons behind the choices of specific techniques, these explanations were not deemed relevant.

As for Khelifi *et al.*'s articles' explanations, the authors pointed out that, while the choices were made with the LAD's target audience in mind, the specific graphs chosen to present the needed data had to be easy to interpret, explainable and user-friendly, to allow for non-data-savvy users the ability to understand and interpret data patterns without the need for background knowledge on the topic.

The choice for simpler DV techniques ensures intuitiveness in the LAD, prompting users to feel more confident when using it, thus skyrocketing its efficiency and value.

3.4.4 GUI-Based Categorization

The preference towards simplicity and conciseness is not something developers and researchers are unfamiliar with even outside the scope of LADs and AI as Edsger W. Dijkstra famously said, "Simplicity is a prerequisite for reliability". Despite the context that produced this renowned quote being quite blurry and inferred to be referring to his algorithm published in 1959 [139], in LADs, sometimes simplicity can also be the safest route to traverse.

As found out by Bach *et al.* [140] in the detailed *Dashboard Design Patterns* article, when it comes to the amount and type of information to be included in a dashboard, a frequent topic pointed out by users is Information Overload. In short, Information Overload is a phenomenon that, as described by Bawden *et al.* [141], "*occurs when information received becomes a hindrance rather than a help when the information is potentially useful*". Carefully constructing an interface that is neither too simple that makes the users miss out on important information nor too complex that confuses or exhausts them is a difficult process, that different developers tackle in different ways.

With this in mind, it becomes of major importance to analyze the articles through the lens of the GUI either implemented or defended by the authors, as knowing how the final work will interact with the user is crucial for its effectiveness.

As opposed to the previously performed analysis, reviewing an article, or in this case, an LAD, considering the way its GUI is developed is not a binary scan like in DR or in DV-oriented examinations, but a manifold process that, in detail, should take into account attributes such as the use of color, interactability or even actual performance capability of the final interface. However, as most of the articles did not place enough emphasis on the LAD's GUI final state or development process, not only does this detailed examination become less relevant, but it also becomes borderline impossible. Although, as explained, the information available in the articles is quite limited, there are still important issues to be addressed.

While not systematically researched due to the lack of information regarding color usage in the resultant LADs from the SLR, an interesting phenomenon to point out from this analysis deviates from the use of one specific DV technique, but an ongoing trend of employing different colors to convey concrete meanings and messages. In what can be described as a *Traffic Light System*, in a similar strategy to that of Course Signals [40], several of the articles defend the use

of color-coded visual indicators, such as green for positive or successful outcomes, yellow for caution or areas requiring attention, and red for critical issues or failures, thus the Traffic Light comparison [72] [73] [78] [86] [89] [136]. No other arguments were raised regarding color usage across the entire set apart from Susnjak *et al.*'s expressed attempt to minimize the use of color to avoid Information Overload (IO) and the use of a neutral pastel palate to reduce negative effects that can stem from a low data-to-ink ratio.

Another important aspect to highlight is that none of the 21 articles made any reference or implication regarding the LAD's GUI responsiveness towards mobile or other kinds of devices, allowing for the inference that either none of them was responsive, or that being responsive is not a major factor in its effectiveness.

As far as the user navigation process throughout the articles' LADs is concerned, a heterogeneous set of results was noticeable, however, due to the lack of emphasis placed on each LAD's GUI and somewhat forced decision to rely on each article's images and screenshots to make assumptions, it is important to approach the following conclusions with skepticism.

Table 3.12: GUI Navigation Process Divisions

Interface Navigation Design	Description
Navigation Menu	Interface where a user, to navigate, must press components (usually links) of a menu structurally organized
Self Contained	Interface that does not allow for a user's navigation since it contains every piece of information directly in the screen
Drill Down Operations	Interface where a user, to navigate, must explore deeper levels of content or categories (usually through filters) in an hierarchical structure

Since the articles do not provide specific strategy nomenclature for how the user is supposed to interact with the interface, the LADs were placed into three different groups⁸, listed in Table 3.12 along with their definitions.

The GUI Navigation Design distribution of the analyzed LADs can be seen in Figure 3.7.

⁸These concepts are not widely acknowledged by the research community nor the industry and are only defined to provide clarity and intuitiveness

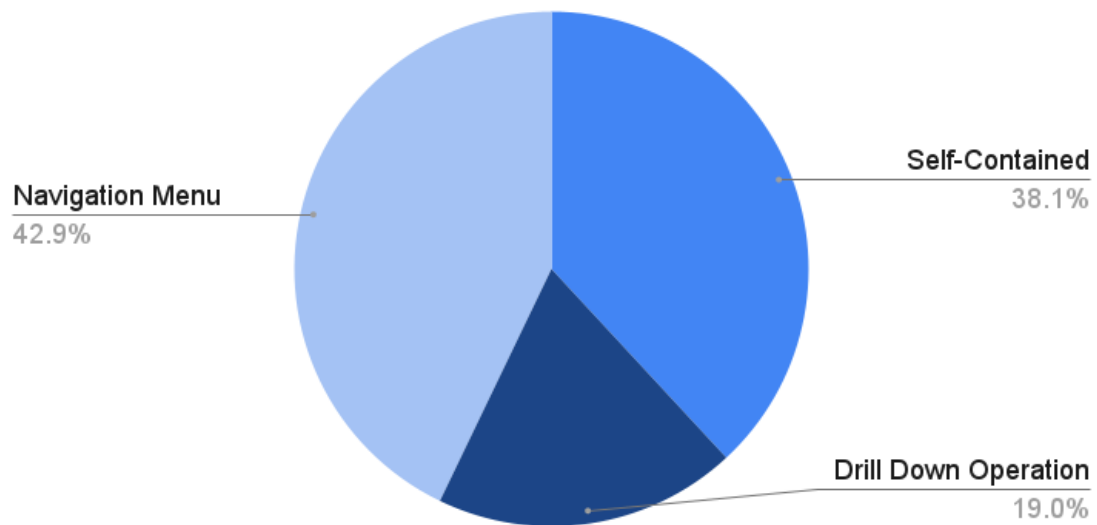


Figure 3.7: GUI Navigation Design Distribution

From the Pie Chart, it becomes easily concludable that, when it comes to the way a user should interact with an LAD, in the authors' view, there is not an uniquely correct way to develop its GUI, as the results are extremely dissonant with each other.

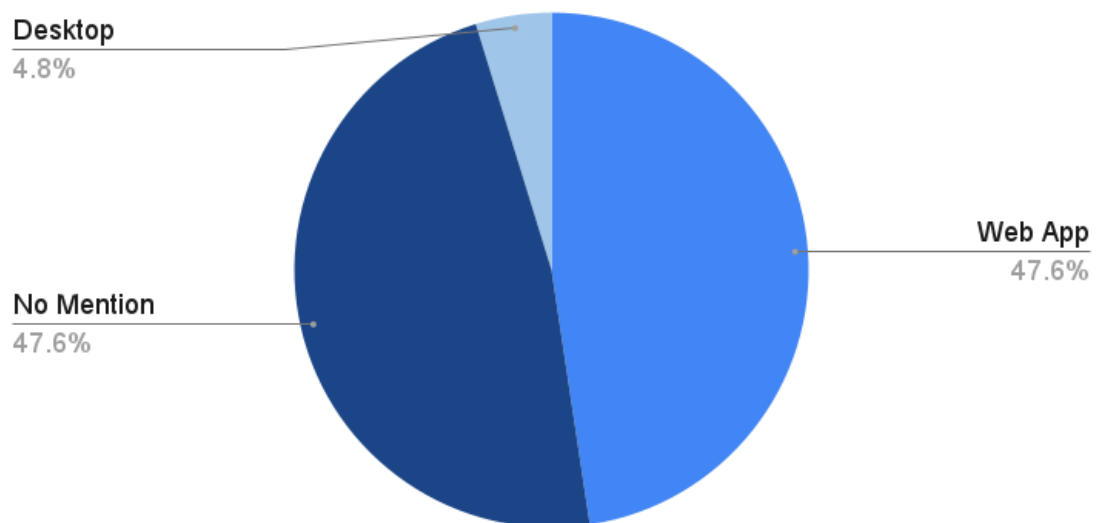


Figure 3.8: App Environment Distribution

When considering the LAD deployment environment, meaning whether its development was web or desktop-based, as can be seen on Figure 3.8, of the 11 articles that mentioned, or allowed for an inference of the deployment environment, 10 of them developed their GUI to be part of a web application, which, given the need for LMS integration, makes absolute sense. The remaining 10 articles provided no information regarding the deployment environment making it completely impossible to conclude with certainty whether the LAD was desktop or web based. Some articles

also provided information regarding the technology used to develop the LADs, which can be seen on Table 3.13 along with each article's deployment environment, if applicable.

Table 3.13: LADs Deployment Environment and Tech Stack

Article	Deployment Environment	Technologies Used for Development
[76]	Web App	(No Mention)
[72]	(No Mention)	(No Mention)
[77]	Web App	JS / PHP
[74]	Web App	Power BI / Python
[78]	(No Mention)	(No Mention)
[79]	Web App	(No Mention)
[80]	Web App	React JS
[75]	Web App	Power BI / Python
[73]	(No Mention)	(No Mention)
[81]	Web App	Tableau
[82]	Web App	HTML + CSS + JS + PHP + Mustache / PHP + Python
[83]	Web App	Power BI
[84]	Web App	JS + Node + Next + React / PostgreSQL
[85]	(No Mention)	(No Mention)
[86]	(No Mention)	(No Mention)
[87]	(No Mention)	(No Mention)
[88]	(No Mention)	(No Mention)
[89]	(No Mention)	(No Mention)
[90]	(No Mention)	(No Mention)
[92]	(No Mention)	(No Mention)
[136]	Desktop	MS Excel

3.5 Literature Findings

As expected, it is possible to conclude that the integration of AI into LADs can represent a transformative step in educational technology, by enabling both educators and students to derive actionable insights from extremely complex data. As pointed out, recent advancements in the research community underscore the multidimensional application of AI across categorization of students, predictions on their success, and even user interface design, shaping the evolution of these tools to align with pedagogical goals and technological advancements. This section aims to summarize the knowledge captured from the SLR Results Content Analysis, as well as define the current State of the Art for AI-Powered LAD Dashboards.

3.5.1 AI-Powered Data Analysis

When it comes to AI and ML, techniques that allow for predictive and prescriptive analytics have emerged as pivotal components in LADs, whether aiming to forecast student performance or provide targeted interventions.

To sum up, the articles, reviewed considering the way they approach AI / ML, collectively underscore the transformative potential of AI / ML in educational settings, particularly in enhancing LADs, predicting student outcomes, and providing personalized learning experiences. The research community demonstrates diverse strategies for using technology to empower both learners and educators, showing a trend of adopting advanced techniques —such as CatBoost, AdaBoost, and Bayesian Knowledge Tracing — suggesting a growing emphasis on balancing actual prediction accuracy with user trust and engagement, both increasingly crucial metrics when dealing with AI / ML. The set of challenges and limitations usually does not deviate from data privacy, model interpretability, and study sample size. While diverse in methodology and application, the importance of transparency, collaboration, and ethical considerations in developing AI-powered educational tools is an unanimous concern.

3.5.2 Data Retrieval

As for the data retrieval process, while leveraging sources as student responses on assignments or interviews can prove to be somewhat useful, they are nowhere near the research community's preferential decision on the source to gather data from. From the analysis, it becomes concludable that the integration of LMS as a foundational data source cannot be highlighted enough, as successfully incorporating vast logs of student interaction data produced by LMS such as frequency of LMS use, engagement with course materials, and submission timelines massively propels AI-augmented LADs efficacy and overall value.

3.5.3 Data Visualization Techniques

Visualization plays a crucial role in translating the complexity of AI-driven analytics into actionable insights. LADs incorporate graphical elements such as dashboards and progress indicators that empower educators to interpret data effectively without extensive technical expertise, enhancing user-friendliness and overall intuitiveness. Some notable approaches include visual traffic light indicators for student risk levels and dynamically generated graphs that present real-time insights into engagement and learning outcomes. When analyzing the specific DV techniques employed by LADs it is possible to see the preference towards simplicity in the form of Bar Charts, Line Graphs and Pie Charts, thus expanding on the intuitiveness and eliminating the requirement of tech or data-savviness to use an LAD.

3.5.4 GUI

As far as how interface design influences usability, accessibility, and data interpretability for educators and students, some conclusions can be taken. To emphasize user-centric design, modern LADs employ intuitive layouts with visual elements such as charts, graphs, and traffic-light indicators, simplifying the identification of trends and outliers in student performance. Interactively, features such as filtering and drill-down capabilities can also be usually employed to enable tailored analysis or to empower educators to focus on specific metrics or student groups. There is

Table 3.14: Critical Comparison of Existing LADs and their Limitations

Dimension	Traditional LADs	AI-Powered LADs	Critical Analysis
Data Processing	Descriptive analytics; aggregates basic usage metrics.	Uses ML models like CatBoost, AdaBoost, BKT.	Predictive power exists, but explainability and interpretability remain weak.
Personalization	Minimal or static; generic for all users.	Clustering and adaptive features attempted in some cases.	True personalization is rare and often shallow.
LMS Integration	Seamless but inflexible; tightly coupled to specific systems.	Often fragmented; requires data preprocessing and engineering.	Increased complexity reduces scalability in real-world deployments.
Learning Outcome Impact	Mostly anecdotal; unclear benefits.	Slight increase in engagement, limited effect on grades.	Lack of longitudinal studies and statistical rigor weakens validity.
Feedback Mechanism	Static reports, often delayed.	Attempts real-time or actionable insights.	Overwhelming or poorly contextualized feedback limits usability.
Explainability	Transparent due to simplicity.	Some use XAI (e.g., EX-AIT), but mostly opaque.	Complexity undermines trust; users struggle to interpret AI output.
Ethical Concerns	Low; uses minimal, anonymized data.	High; includes emotion/facial recognition, behavioral inference, bias in data collection	Raises serious privacy concerns and lacks established ethical frameworks.
Educator Involvement	Limited; often an afterthought.	Mixed; some co-design efforts.	Lack of pedagogical alignment results in low adoption and relevance.
Adoption Barriers	Low; simple and usable.	High; requires technical infrastructure and trust.	Complexity and opacity hinder institutional uptake.

also the constant fear of producing IO, forcing the constant limbo between “When is information enough?” and “When is information too much?”. As for the environment LADs are usually developed in, there is a preference towards Web-Based LADs, not explicitly explained, but most likely due to ease of use, seamless integration and remote collaboration.

3.5.5 Limitations and Gaps

In light of the findings from the SLR, it becomes imperative to critically examine the limitations and challenges associated with current LADs, particularly those augmented with AI capabilities. While the literature reveals several innovative contributions, it also surfaces a variety of persistent and systemic issues that continue to hinder the widespread effectiveness, adoption, and scalability of such tools.

Table 3.14 presents a synthesized comparison between Traditional and AI-Powered LADs, along key dimensions identified during the review process.

A primary challenge identified is the absence of standardized evaluation frameworks for LADs, particularly those incorporating AI. The literature reveals inconsistent metrics and validation approaches, which severely limit the comparability of results across studies and impede the formation of evidence-based best practices. This inconsistency extends to both system performance and educational impact assessments.

Furthermore, the increasing complexity and volume of educational data presents a significant obstacle for LAD scalability. Most dashboards struggle to effectively manage real-time data streams and integrate heterogeneous datasets across diverse educational contexts. Although some systems attempt to employ clustering or adaptive modeling techniques, true personalization remains underdeveloped and often superficial.

Another noteworthy issue concerns the explainability of AI-powered dashboards. Despite growing interest in XAI, many systems fail to integrate mechanisms that clearly communicate the reasoning behind AI-driven recommendations. This lack of transparency undermines user trust and limits adoption, particularly among educators and academic stakeholders unfamiliar with advanced data analytics.

In addition to explainability, usability challenges persist in the GUIs of many LADs. Dashboards frequently lack a balanced design that harmonizes usability with functional depth. Over-complicated visualizations or insufficient interactivity often impair the tool's effectiveness, particularly in high-pressure academic environments.

The reviewed studies also reveal a fragmented approach to system architecture and design philosophy. Many focus narrowly on specific features such as visualization tools or prediction models without articulating how these components should function cohesively within a holistic, user-centric system. This piecemeal development approach leads to reduced practical applicability and user engagement.

From an interoperability standpoint, LADs tend to be tightly coupled with specific LMS platforms (most commonly Moodle). While this facilitates immediate integration and data collection, it also introduces rigidity that limits cross-platform portability and institutional scalability.

Equally concerning is the limited involvement of educators in the design and implementation of LADs. Although a few studies highlight co-creation efforts, the vast majority treat educators as secondary users rather than primary stakeholders. This lack of pedagogical alignment often results in low adoption rates and questions the long-term relevance of the developed tools.

Ethical and regulatory issues further complicate the deployment of AI-powered LADs as the increasing use of sensitive data such as facial expressions, emotional indicators, and behavioral inferences raises legitimate concerns regarding privacy, consent, and potential biases in model predictions. Few studies offer clear frameworks or guidelines to address these challenges, leaving ethical gaps largely unaddressed.

Lastly, while some LADs report modest improvements in student engagement, there remains insufficient empirical evidence to confirm a direct causal relationship between LAD usage and enhanced academic performance. The scarcity of longitudinal studies, combined with limited sample diversity, significantly undermines the generalizability and impact of current findings.

Taken collectively, these challenges reveal a landscape marked by technological potential but constrained by fragmented design, limited standardization, ethical ambiguity, and a lack of pedagogical integration. These limitations substantiate the need for a new generation of LADs, particularly those that are AI-powered, designed with ethical safeguards, user trust, and educational relevance at their core.

3.5.6 Conceptual Framework

From the previously noted Literature Findings it was possible to design a Conceptual Framework that illustrates the integration of key components involved in an AI-powered LAD, represented in Figure 3.9

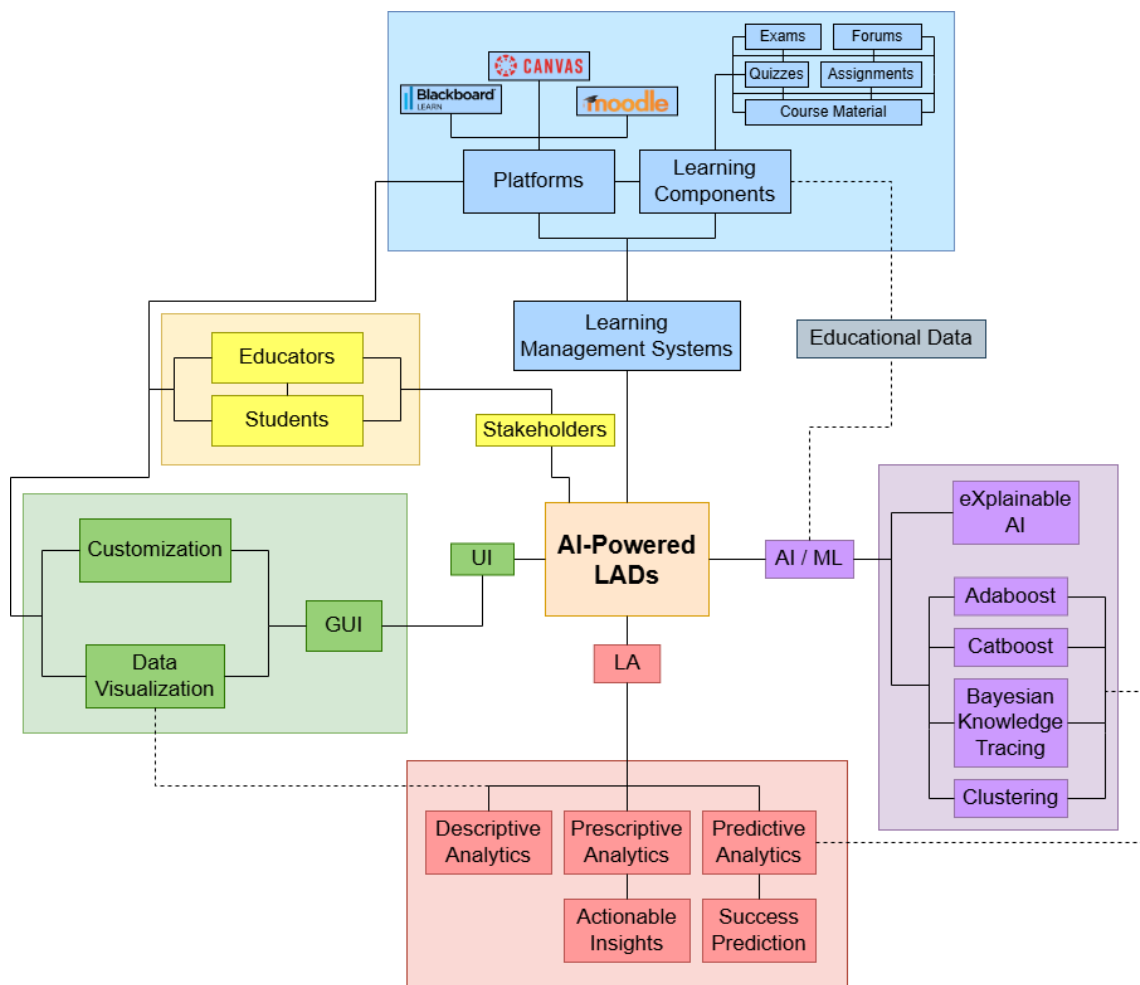


Figure 3.9: Conceptual Framework of an AI-powered LAD based on SLR Results and Findings

From the Conceptual Framework it is possible to understand that, although not extensively researched in the community, the topic of AI-Powered LADs is complex and results from the intersection of several different concepts. To better understand it, a deeper analysis becomes necessary.

First and foremost, despite every component playing a significant role in it, the most crucial part of an AI-Powered LAD and a traditional LAD is the employment of LA as a way to leverage knowledge and extract insights. This employment can usually be felt through the three different types of analytics previously described: Descriptive Analytics, Prescriptive Analytics, that have the ability to generate actionable insights and Predictive Analytics, that aid educators and students identify potential signs of success or failure in the early stages of the learning process. Although, in an AI-Powered LAD, the use of AI in LA can range from being minimal to being widespread, it is generally accepted that the biggest impact AI can have in LA lies in Predictive Analytics.

Through boosting algorithms such as Adaboost or Catboost or through regular AI/ML techniques like Bayesian Knowledge Tracing or Clustering, AI can improve the prediction process through a self-learning cycle constantly enhanced with the acquisition of new data. However, as opposed to using AI with little to no care towards transparency and interpretability, researchers have been highlighting another important field known as XAI. In short, XAI aims to preserve transparency during AI processes and attempts to minimize models' confusion and lack of understanding of how processing occurs. Since LADs tend to work with personal data and their predictions can produce undesirable effects in both students' and/or educators' motivation levels and, in extreme conditions, mental health, XAI's importance when designing any type of AI-Powered educational tool becomes exponentially larger and its use gradually becomes a standard requirement.

For these same AI processes to work, and to work correctly, data plays an impossibly overstatable role. This data is usually retrieved from learning or evaluation components such as quizzes, exams, assignments or even course materials made available by tutors or instructors and educators and is usually stored in LMS with platforms like Moodle, Canvas and Blackboard Learn standing at the top of the global LMS adoption rank. These systems have the ability to store enormous volumes of educational data, and, if used by the educational institution, play a crucial role in LADs' operations since retrieving their data and leveraging it correctly lay the foundation for an accurate AI prediction model operation and thus, a valuable tool.

Another important concept that tends to fly under the radar in studies and articles are an LAD's stakeholders. Despite the SLR results mainly focusing on one of the two represented, it is useful to look at both students and educators to understand the value an LAD can have on an educational setting. The importance of learners and instructors in an LAD not only applies to its placing on the center of the design process and understanding of needs, but also to their roles as main catalyzers of the data generation process since, by interacting with the LMS and with the learning environment in general, both students and teachers continuously generate data later fed to the AI models thus propelling the self-learning cycle previously described. Additionally, by being the main end-users of the tool, and the main enablers of its viability, not emphasizing the role of the stakeholders is a foolish mistake that can prove detrimental to the finished work.

Lastly, an important, but generally undervalued component of an LAD and any educational tool is the UI. Based on the SLR results and findings the most important part of an LAD's UI lies in the DV, as it conveys the message produced by LA that would otherwise be harder to interpret

or to derive conclusions from. DV can be expressed through simple techniques like graphs or charts or through more complex ways such as heatmaps or sliders, although most researchers tend to opt towards simpler ways of presenting data to enhance user-friendliness. However, despite not extensively covered by the articles, it is important to point out that, while providing intuitive and powerful ways to display data and information is crucial for an LAD's success, balancing its UI's usability with customizability might just be the difference between the tool's widespread adoption or lackthereof. As it also represents the main connection between the end-user and the system, designing an intuitive, interpretable and customizable UI can prove to be an important step towards presenting a usable, unique and valuable LAD.

The Conceptual Framework highlights that, while AI-Powered LADs hold immense promise and represent the current trend of AI adoption in the education field, their success lies in a delicate balance between technological advancements, stakeholder inclusion, ethical considerations, and user-centered design. Only by addressing these interconnected dimensions is it possible to realize the full potential of LADs in transforming educational landscapes and aiding students and educators fulfill their needs in a more direct, simple and efficient way.

Chapter 4

Proposed Solution

This chapter aims to describe the proposed approach to improve students and educators' learning outcomes through an AI-Powered LAD and define the methods to conduct the research process.

4.1 Approach Description

When considering the formulated Research Questions, to understand the exact impact an AI-Powered LAD might have in an education setting, specifically in students and educators, the main stakeholders of the LAD System, it is important to first outline its development process and main value proposition. To do so, the development methodology described in the following subsection places the end-users at the center of the development process, without direct involvement from them in it. This means, in short, that, while the LAD will not be developed in a co-design methodology, its main capabilities will be designed considering end-users needs, strengths and weaknesses. As a result, to effectively assess an AI-Powered LAD's effect in students and educators' outcomes, there are some guidelines to be considered that will lead the design process.

First and foremost, since the typical educational environment can be stress-inducing due to cognitive overload or assignments and deadlines, the finished LAD should not be stressing to interact and should have a soothing, simple UI design to avoid IO.

Secondly, to ensure the end-user can feel comfortable while using the LAD and its graphical appearance is not an obstacle to its usage, the ability to customize the aesthetic of the interface to one's liking should also be provided to the possible extent.

Thirdly, as ethical considerations and data protection policies are one of the main concerns of today's society and research community, the LAD should always maintain user privacy. Any data needed to be stored should follow the General Data Protection Regulation (GDPR) [142] policies and should be anonymized, whenever possible. To avoid internal class competition leading to a decrease in a student's motivation, students' individual learning data should never be directly disclosed to other students.

Fourthly, as today's higher education settings tend to have enormous amounts of students, the LAD's system must be scalable to avoid increased demands compromising performance to a

partial, or a full extent. This ensures that the system's ability to perform correctly is not a limiting factor to the system's adoption.

Last, but not least, every single piece of information displayed to the end-user via the interface should follow the principles of transparency and explainability. AI computations should comply with the values of XAI and should provide an interpretable explanation that does not rely on tech-savviness or AI knowledge to ensure its understanding. In addition, the system should not entirely rely on the end-user's digital proficiency to be usable and should be designed taking into account either technologically experienced or technologically challenged stakeholders.

With these guidelines clearly identified, it becomes necessary to point out what the final LAD's capabilities should be.

4.1.1 Solution Goals

Since most of the articles resultant from the SLR strongly focus on one group of stakeholders rather than on both, the proposed LAD will attempt to aid both educators and students in their daily operations. As the information available to the two groups of stakeholders is different, the division will be ensured through authentication processes. To understand the added value the LAD will provide to both stakeholders, it becomes necessary to divide the LADs capabilities in two groups.

Since success prediction is one of the main value propositions of predictive analytics in education, the need to include both the ability to predict potential at-risk students as well as an entire class grade average in a simple way is self-evident.

Another important ability to provide to educators is the generation of tailored advice messages directly sendable to specific students consisting of either support and encouragement words or specific steps to better achieve their goal. If presented this possibility, educators can speed up complex processes, enhance student motivation in a faster and more efficient way and, as a result, employ their time and efforts into other important tasks like class preparation or even daily life, increasing their own motivation as well.

Additionally, the final LAD should also enable educators to easily extract relevant pieces of information as well as actionable insights from educational data so they have a deeper understanding and knowledge of whether or not their strategies are being proven useful and fruitful.

Educators should also be able to analyze either a single specific student's or compare two different students' learning evolutions throughout the semester, both activities that, in grading processes, can prove to be important to ensure justice in evaluations. The educator should also have the opportunity to compare a single student to the class average, to understand possible discrepancies not easily interpretable otherwise.

When it comes to students, to help them understand how comfortable they should or should not be in a specific course, the ability to predict their own grade at any point of the course should be provided as well as their comparison with the class average. Students should also be able to check their progress at any time and, if deemed necessary, have a direct way of communicating with the educator to either ask for help or specific information.

A summary of the main activities the system should allow the stakeholders to perform is represented in the Venn Diagram in Figure 4.1

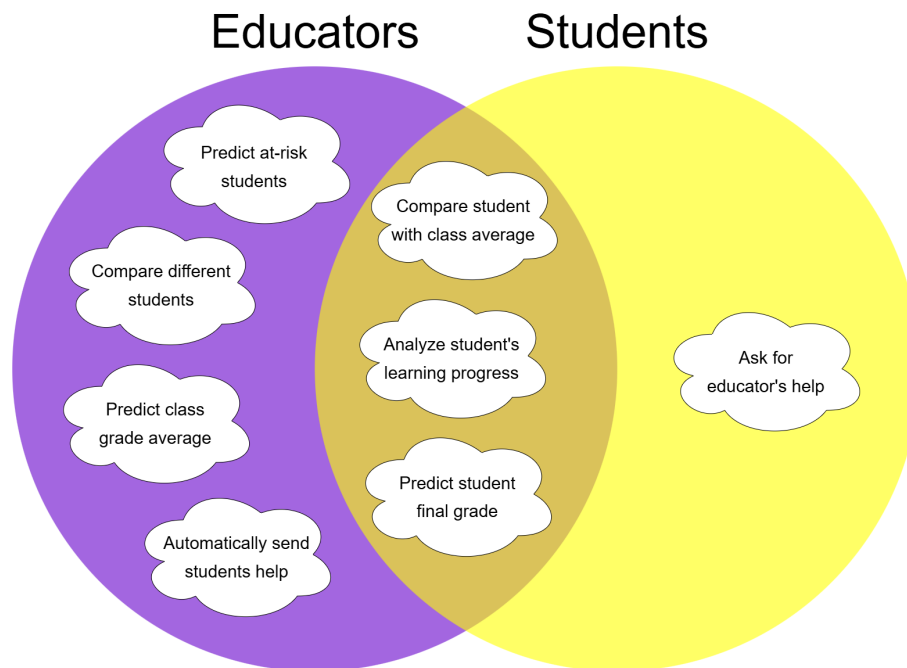


Figure 4.1: System's Main Capabilities

4.1.2 System Architecture

The architecture of the proposed solution represented in Figure 4.2 is structured in a modular and layered fashion, comprising three main domains: the Frontend, the Backend, and the Authentication layer, all supported by a centralized Database.

This design facilitates maintainability, scalability, and clear separation of concerns across the platform's core functionalities.

At the entry point of the system lies the Frontend, which hosts the UI. This interface serves as the primary interaction layer for both students and educators, allowing them to engage with predictive models, receive personalized messages, and visualize actionable insights and communicates directly with the backend to request data, trigger predictions, and fetch tailored guidance.

User credentials and access permissions are managed by a distinct Authentication module. This module verifies identities and assigns role-based access to users before redirecting them to the appropriate UI components. It integrates securely with the backend to enforce authorization protocols and prevent unauthorized data access.

The Backend acts as the core orchestrator of all computational and data processing workflows. Within it, the Data Retrieval and Processing component is responsible for fetching and pre-processing student event logs and course data from the Database, which utilizes a NoSQL architecture, specifically MongoDB, to accommodate flexible and nested educational data schemas.

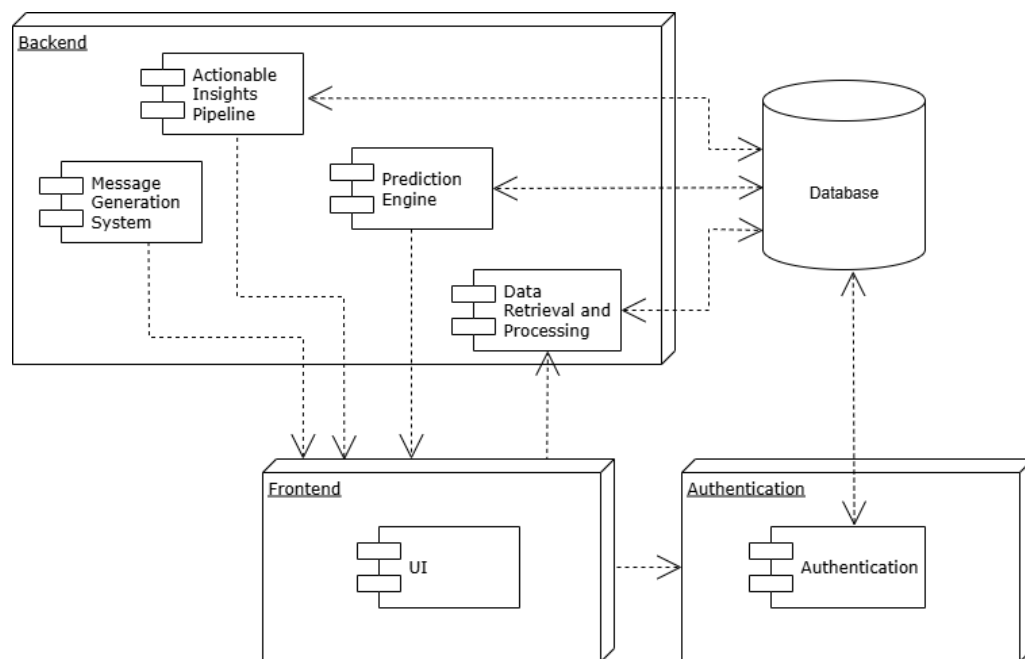


Figure 4.2: Proposed Solution's Architecture

Central to the system's intelligence capabilities is the Prediction Engine, which is invoked to perform both grade and motivation predictions. This component leverages historical and real-time data from the database to train and evaluate multiple machine learning models and incorporates self-reported feedback from students in the case of motivation modeling.

To complement predictive insights, the Actionable Insights Pipeline operates as an analytical subsystem, transforming raw data and prediction outputs into pedagogically relevant recommendations. These insights are contextually adapted for each educator or student and form the basis for the Message Generation System, which composes natural language summaries and personalized feedback, made available via the UI.

All components interact through well-defined APIs and message-passing mechanisms, ensuring a clean separation between data layers and service logic. The architectural choices made throughout the development of the solution prioritize transparency, modularity, and responsiveness to user interactions, aligning the platform with its overarching goal of enabling real-time, explainable learning analytics.

To clarify the interactions between the solution and its primary users, them being students and educators, two use case diagrams were constructed. As shown in Figure 4.3, students can, for instance, interact with the system by predicting their motivation or grade levels, viewing grade forecasts, checking their educational progress or setting academic goals.

Similarly, Figure 4.4 presents the use case diagram for educators, who are able to monitor student performance, identify at-risk learners, explore course-level analytics, and provide targeted feedback. These diagrams served as a foundational tool in guiding both the system's interface design and the usability evaluation tasks described in Chapter 6.

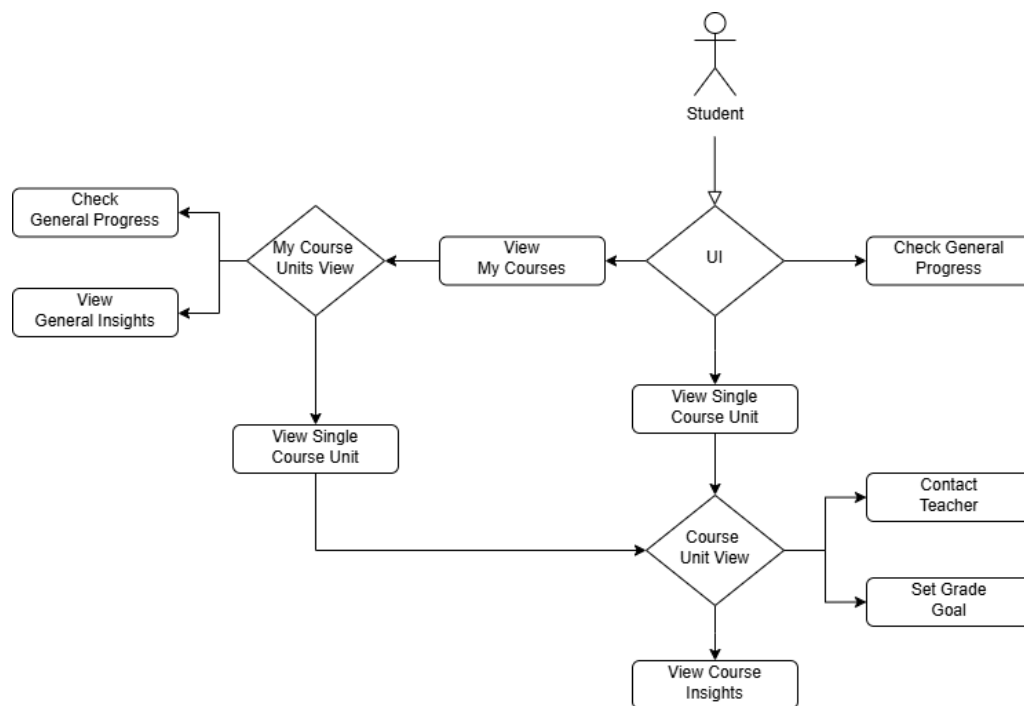


Figure 4.3: Student-Oriented Use Case Diagram

4.2 Methods

To ensure a holistic understanding of the LAD's impact on student and educator outcomes, the methods used to evaluate it must be carefully defined. The chosen methodology for the research design was the combination of both quantitative and qualitative methods, thus creating a mixed-methods approach. This choice appropriately addresses the research question as it combines precise measurements related, for example, with prediction accuracy with qualitative methods such as user experience and opinion, both extremely important topics in the context of an educational tool and enabling of a comprehensive evaluation of an LAD in a real-world educational context.

4.2.1 Research Design

The study conducted will fall under the umbrella of both explanatory and exploratory types, as it will both aim to evaluate the causal relationships between the usage of the AI-Powered LAD and student and educators' learning outcomes, as well as to investigate user behaviors and preferences associated with the actual use of the LAD. To allow for easier testing and reduce logistical complications, only higher education institutions from the district of Porto will be involved in the study and, to ensure data variety and reduce bias in such a small geographic frame, diverse courses' student population with different demographics will be selected to take part in the study. Additionally, to assess the LAD's adaptability and user-friendliness as well as to generalize its evaluation without tech-savviness induced bias, students with different levels of technological background will be selected.

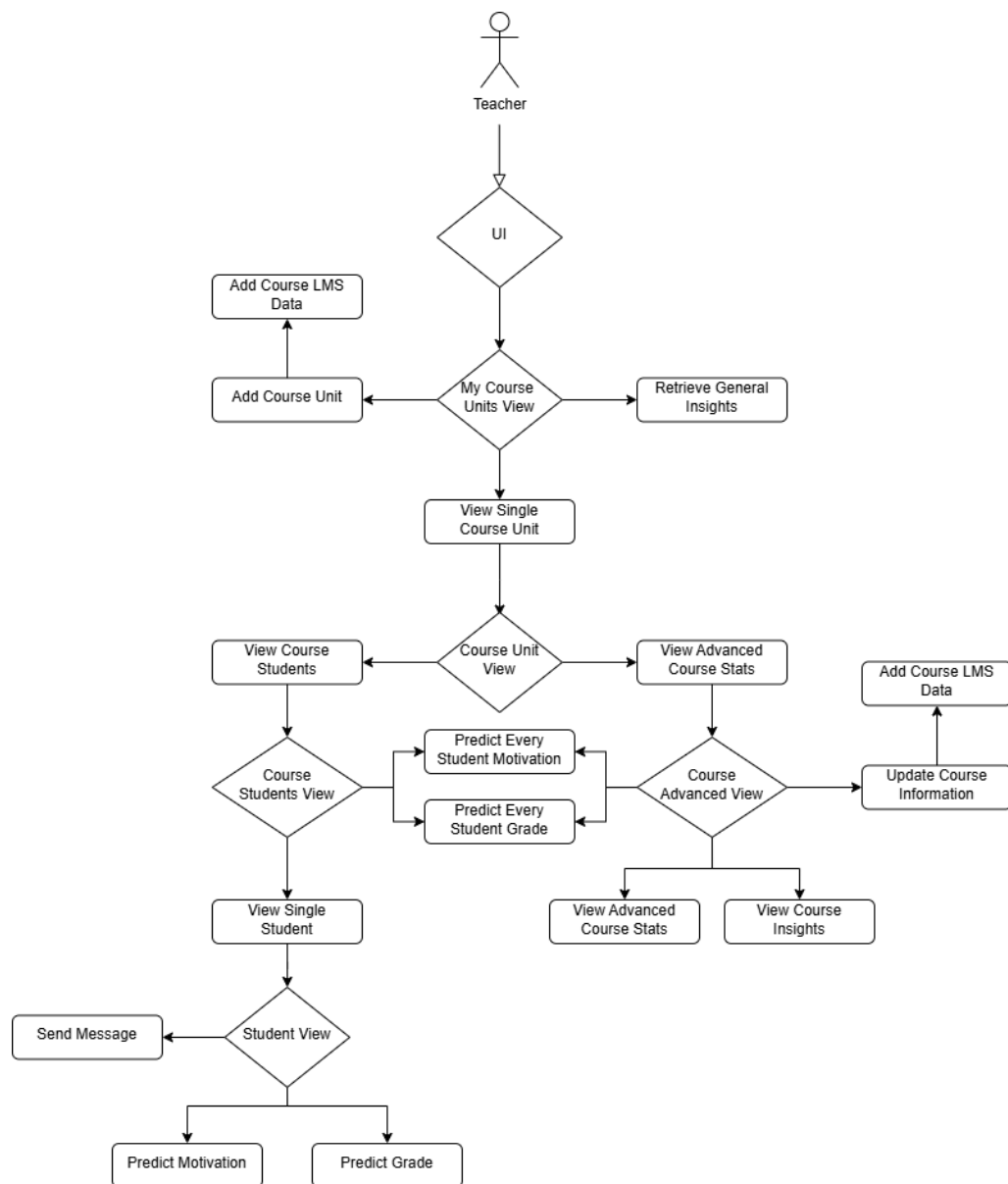


Figure 4.4: Educator-Oriented Use Case Diagram

In addition to the individuals analyzed in this study, around 70 students enrolled in higher education institutions and 10 educators responsible for teaching and assessing at those same institutions, data artifacts such as prediction logs and user feedback will also compose the study's subjects and be extensively examined. The sample size was defined taking into account its ability to ensure statistical power for quantitative analysis while remaining manageable for qualitative exploration and, to participate in the study, both students and educators must have a basic knowledge of the English language, must consent to the investigation and must lack any previous knowledge about the proposed solution.

Since the data intended to be collected will range from concrete and precise measurements to subjective and verbal feedback, both quantitative and qualitative tools will be used to aid the data

collection process. The quantitative instruments that will be used in this process are surveys such as John Brooke *et al.*'s System Usability Scale (SUS) [143], Single Ease Questions (SEQ) [144] and the LAD's analytics. In short, the SUS questionnaire is a standardized and widely used tool for measuring perceived usability, as it consists of ten statements that participants rate on a 5-point Likert scale ranging from 1 (Strongly disagree) to 5 (Strongly agree) with the statements alternating between positive and negative phrasing to reduce response bias. The SUS items are the following:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The SUS score is then calculated by first adjusting the responses: for odd-numbered items, 1 is subtracted from the user response; for even-numbered items, the user response is subtracted from 5. The adjusted values are then summed and multiplied by 2.5, resulting in a score ranging from 0 to 100, which is interpreted as a percentage of the ideal usability score. A SUS score above 68 is generally considered above average.

On the other hand, semi-structured interviews to understand user behaviors and opinions, as well as observational logs of the subjects' interactions will compose the qualitative instruments.

In ideal conditions, involving conducting interviews and / or focus groups mid-semester to understand user experiences and identify challenges as well as observing the differences between control and experimental groups regarding student outcomes would be the most precise evaluation procedure to conduct as it would allow for a more direct understanding of the tangible impacts the system would have in real-life scenarios and for a timeframe enabling of a continuous evolution analysis. However, since the available time period does not allow for such a strategy to prove effective or even achievable, the evaluation process will be divided into different procedures, aimed to test different activities in the system.

As Predictive Analytics is one of the main concepts associated with the developed system, assessing the LAD's prediction capabilities is of crucial importance. To do so, procedure A will work as follows:

- Provide real historical data from specific anonymous student at different points of the semester (one, two and three months into the semester)
- Perform the prediction on the system
- Log expected grade results
- Contrast with real data

On the other hand, assessing the system's usability for both students and educators is critical to understand how viable its application would be in real-world scenarios. Thus, procedure B will work as follows:

- Provide Informed Consent Form
- Administer pre-intervention survey
- Allow user to freely explore the system for 5 minutes
- Provide a set of tasks the user must accomplish
- Gather relevant information regarding task completion (number of clicks, time taken, help needed)
- Administer post-interview survey
- Perform semi-structured interview

To ensure reliability and consistency in the evaluation process, pilot tests will be conducted and the cross-referencing of the survey results with qualitative feedback and observational data will also be performed.

4.2.2 Data Collection

As the two procedures greatly vary in both aims and means to achieve them, the key variables and measurements to be directly retrieved from them depend on the evaluation procedure performed. For procedure A, the MSE, the MAE and the accuracy of the predicted grades will be analyzed, as understanding how close (or far) the students' predicted grades are to their actual grades is the most important factor towards being deemed useful. Notwithstanding, since procedure B aims to collect both subjective data such as user opinion and objective data like completion or incompleteness of specific tasks, the measurements to be retrieved range from perceived usefulness and usability, ability to complete the given tasks and perceived difficulty of completion (through SEQs).

Since the collected data can be expressed in both quantitative and qualitative forms, the techniques to analyze it must also be of these two types, with the quantitative techniques ranging from descriptive statistics to inferential analysis to identify causal links using IBM's Statistical Package for the Social Sciences [145] and the qualitative ones consisting of thematic analysis of the

semi-structured interview using the LiGRE tool [146] and categorization of observational notes. To perform the described analysis, the following step-by-step plan was created:

- Organize and preprocess data.
- Analyze quantitative data for trends and correlations.
- Perform thematic analysis of qualitative data to identify recurring patterns.
- Integrate findings for comprehensive interpretation.

Since the ethical considerations associated with research methodologies like the one presented are one of the research community's main concerns, especially since it involves data usage and gathering, some key principles need to be respected. As a result, not only will all data retrieved for or produced during the study be anonymized to prevent user identification, but its storage will be secure and encrypted, as well as permanently eliminated at the study's conclusion. To ensure user trust and clear understanding of data usage, all participants will be provided with an informed consent form disclosing both the study's GDPR compliant use of data as well as AI processes involved in the research to the participants.

4.2.3 User Insights and Design Motivation

Before beginning the design and implementation of the solution, to better understand current attitudes and behaviors regarding the use of LMS, two student surveys were conducted. The surveys, distributed in both Portuguese and English, gathered responses from undergraduate and graduate students across multiple programs with a clear goal to identify motivational triggers, study habits, LMS engagement patterns, and perceptions of LMS-related features such as notifications, deadlines, and peer interactions and the responses revealed several recurring themes that significantly shaped the design of the proposed solution.

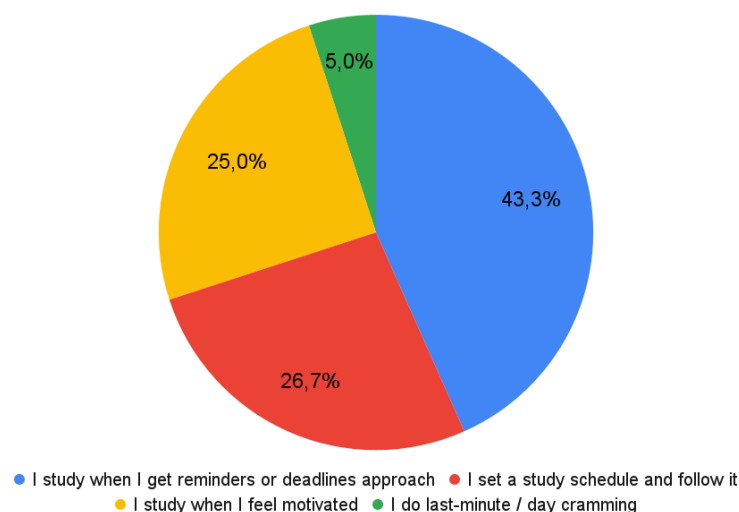


Figure 4.5: Student Motivation to Use LMSs - Response Distribution

The results, fully available in Appendix B, revealed that, while students commonly accessed LMS platforms such as Moodle or Canvas, their engagement was often driven by external reminders or deadlines rather than proactive study planning, as Figure 4.5 represents.

Additionally, many participants indicated that they procrastinated when materials were exclusively available online, felt that existing LMS tools lacked motivational support and only reviewed course content before exams, as shown in Figure 4.6.

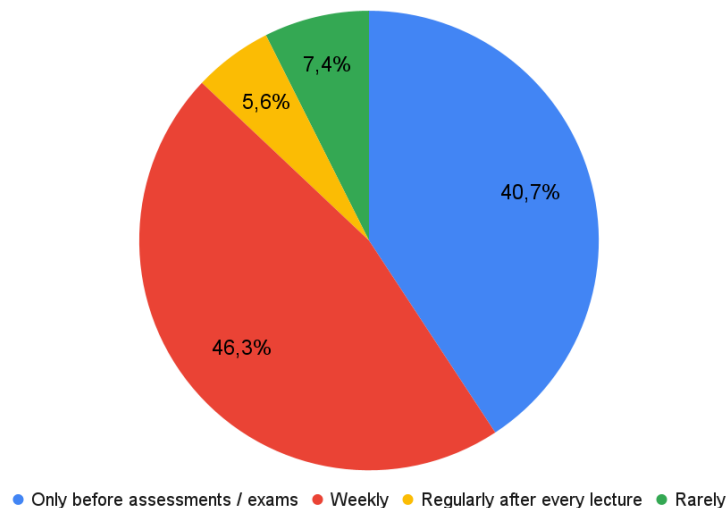


Figure 4.6: Student Regularity when Accessing LMS - Response Distribution

Interestingly, a majority expressed interest in features like personalized reminders, gamified elements, or more engaging UI designs, thus confirming the demand for a more interactive and intelligent layer atop traditional LMS systems.

When asked how they felt upon receiving overdue assignment notifications, most students reported anxiety or stress, yet continued working despite those feelings, which suggested that motivational features needed to be designed with emotional awareness not just as triggers for compliance, but as support mechanisms. As a result, a messaging module component was designed not only to alert but to encourage and guide the users without overwhelming them.

Interestingly, more than half of the students across both surveys stated they would find personalized study reminders helpful, with many noting that such features could help them stay on track. However, a cautious tone also emerged with some students preferring to test such features first, and others expressing a desire for autonomy. This duality led to the inclusion of opt-in controls and explainability mechanisms within the system, allowing users to choose their level of interaction and understand the logic behind the reminders or the predicted data.

On the social front, many students indicated that seeing peers engage in LMS discussions had limited motivational value, with a preference for independent learning being dominant. Consequently, while the proposed solution does not heavily emphasize social comparison to prevent

negative competition, features like progress overview and peer percentile positioning were designed to be informative rather than competitive, respecting the students' different motivational drivers.

These insights provided the foundation for several of the system's core modules, including the predictive models, intelligent messaging, and activity-driven feedback. The design choices throughout the system were directly informed by this early-stage feedback, ensuring that the solution was developed with real user behavior and expectations in mind.

4.2.4 Limitations

Despite being carefully constructed to avoid potential shortcomings, the evaluation methodology described has some limitations that can lead to potential issues. First of all, given the reduced evaluation period the study must adhere to, the capturing of long-term outcome effects resultant from the system's use might be difficult or borderline impossible. Nevertheless, to alleviate the potential effect of the shortness of the evaluation period, follow-up studies could be performed to assess the impact in a larger timeframe. Secondly, given the high geographical concentration of the study's participants and the possibility of sample bias, it might prove to be difficult to fully generalize about other institutions or populations with vastly different characteristics. However, to avoid it, all the participants will be selected from several different courses and will be randomly chosen to take part in the study.

Not only is this comprehensive evaluation protocol designed to ensure transparency, replicability, and validity, by employing a mixed-methods approach, defining clear variables, and adhering to ethical principles, the study aims to provide actionable insights into the effectiveness and usability of the AI-Powered LAD and its possibility of adoption in real-world education environments.

Chapter 5

Developed System

In order to correctly address and answer the four research questions formulated in the first chapter, an intuitive and complete system was developed designated Clarifying AI for Better Results in Academic Learning (*CABRAL*). This chapter aims to dissect the *CABRAL*'s capabilities and development phase and will be subdivided into five different sections, each attempting to explain a component in detail.

5.1 Technological Stack

In order to materialize the proposed functionalities of *CABRAL* into a coherent, modular, and scalable architecture, a diverse technological stack was employed across all subsystems, ranging from data processing to interface presentation. Each component was selected not only for its technical reliability and support ecosystem, but also for its compatibility with the project's research-driven and user-centered objectives.

5.1.1 Backend and Data Handling

The backend of *CABRAL* was developed using C# [147] with the .NET framework [148], enabling a structured and object-oriented approach to business logic, data orchestration, and analytical pipeline coordination. To properly accommodate the intricate and dynamic data structures inherent to the *CABRAL* system, particularly the flexible and evolving nature of educational logs and the interdependencies between various educational entities, a database system needed to be chosen. As an agile integration of diverse data formats originating from various LMS platforms without the rigid constraints imposed by traditional relational databases was necessary to ensure the system's functionalities, the schema-less nature of MongoDB [149] offered significant advantages as its inherent scalability also positions it as an ideal choice for a system designed to scale across different courses and a growing volume of learning analytics data, all while preserving data coherence and integrity.

5.1.2 Machine Learning and AI Components

All machine learning tasks were handled using Python [150], leveraging the scikit-learn framework [151] as the primary modeling toolkit, along with NumPy [152] for numerical operations and joblib [153] for model serialization and inference portability. A variety of models were tested and implemented during development, namely: Random Forest, CatBoost, Linear Regression, Dummy Regressor, and AdaBoost. To improve the system's transparency and interpretability, especially for non-technical stakeholders, SHAP (SHapley Additive exPlanations) was employed as an XAI framework, providing detailed insight into the decision-making rationale of the predictive models.

5.1.3 Frontend and Interface Design

The frontend of *CABRAL* was developed using React.js [154], with a codebase composed of both `.tsx` and `.js` components to support dynamic user interactions and real-time responsiveness. The styling layer was implemented in CSS [155], with the FontAwesome [156] icon set being the only UI design library employed in the final system. Visual and interaction design principles were heavily informed by Figma [157] mockups created during early prototyping stages, which prioritized usability under high cognitive load scenarios. No accessibility-focused frameworks were employed.

5.1.4 Deployment and Integration

The system was developed and tested in a local deployment environment. While API integration with LMS platforms such as Moodle or Canvas was not implemented in this iteration, the architecture remains open for such extensions in future work. Similarly, containerization technologies such as Docker [158] or orchestration tools like Kubernetes [159] were not employed. During development, Node.js [160] was used to run the application in a local development environment, as it provided a lightweight, efficient platform for serving the application and handling requests. Specifically, `npm run dev` was used to launch the system locally in a development mode, as it allows for faster iteration by running the application in a hot-reload environment. This command starts the development server, enabling real-time updates to the codebase without requiring manual restarts, which significantly improves the efficiency of debugging and testing during the development process.

5.2 Data Retrieval and Management

As managing and operating over data is crucial for any educational tool or system, a strong focus was placed on developing a fault-tolerant, scalable and capable pipeline that could extract information from LMS-retrieved logs submitted by educators into the system and organize it into a single database.

This component's work can be summarized in two phases: data cleaning and data processing.

The data cleaning phase works mostly to remove inconsistent or erroneous characters or data from the logs, with a strong emphasis on decoding HTML entities such as `'`. This phase ensures information quality and allows for a more reliable processing and storage.

The latter phase, on the other hand, aims to map educational events into entities like students, courses or interactions. Through a detailed processing of each log entry composed by both attribute extraction and existing record association, the system is able to represent students' actions, enrolments and learning progress.

Additionally, this dynamic structured modeling allows for a longitudinal analysis of a learner's progress, a more personalized monitoring from the educators and for an automated feedback mechanism that both reminds students of the necessary effort for success and empowers educators to help students in an otherwise impossible manner. Furthermore, this modeling also allows the system to scale across different courses while preserving the overall coherence of its learning analytics.

MongoDB's document-oriented model also aligns seamlessly with the system's need to represent complex, interconnected data and its ability to embed arrays and sub-documents within a single document allows for a holistic representation of interconnected data. This approach not only simplifies data retrieval and manipulation but also inherently supports the longitudinal analysis of learner progress and the personalized monitoring capabilities that *CABRAL* aims to deliver.

To allow for a more easily-understandable and simple data storage, four different classes were defined: *Student*, *Course*, *StudentCourseEvent* and *Prediction*. These classes are explained in the following sections:

5.2.1 *Prediction*

The *Prediction* class serves as a repository for the system's prognostications. Each instance encapsulates a calculated value, a reference to the course it pertains to, and the precise time the prediction was made. Crucially, it also tracks the accuracy of the forecast¹, categorizes its type, and associates it with a specific student id. Furthermore, it includes the given prediction shap_ values, offering insights into the factors influencing the prediction.

5.2.2 *Course*

The *Course* class defines the structure for course units within the system. Each course is identified by a distinct courseId and a descriptive courseName. It meticulously maintains lists of teachers and students associated with it, providing a comprehensive overview of its participants.

¹This "accuracy" attribute should not be confused with the commonly-perceived accuracy usually referenced in classification problems. The name was chosen to simplify the development process as the attribute aims to express either a confidence level in a certain prediction or a consistency value usually calculated through metrics such as Mean Hassanat Similarity Percentage

5.2.3 *Student*

The *Student* class aims to represent an individual learner within the educational framework. Identified by a unique `stdId`, each student is named, and their academic journey is meticulously logged. This includes grades and motivations, both captured as collections of *Prediction* objects, alongside their goals, which offer insights into their aspirations. Additionally, a list of courses provides a quick reference to their enrollments.

5.2.4 *StudentCourseEvent*

The *StudentCourseEvent* class acts as a detailed record of interactions between students and courses. Each event links a specific `courseId` with a `stdId`. A clear description elucidates the nature of the event, and the time it occurred is precisely recorded, ensuring a comprehensive historical log of student-course interactions.

5.3 Grade and Motivation Predictions

The core of *CABRAL* and perhaps the most crucial functionality to explain in detail in the entire system is the prediction of key components of academic success, namely grades and motivations.

This component's extreme importance in the composition of *CABRAL* and consequential emphasis placed in its development is twofold. Firstly, while not entirely representative of the true final grades and motivations of students, predictions derived from study interactions can work as reinforcement on their current behavior, whether it being positive or negative..

As a result, carefully balancing directness and abruptness becomes a complex task to take into account, as erroneously predicting a student as failing, even if they are not, can be detrimental to that student's future efforts, regardless of how fail-prone the component is disclaimed to be, while, on the other hand, mistakenly predicting a student as having an excessively high grade can lead to a reduction in the given student's effort to keep improving, as they might interpret the grade as already enough to meet their goal.

To understand in detail how the component works, it is crucial to scrutinize each of the sub-components, namely Grade Prediction and Motivation Prediction.

5.3.1 Grade Prediction

When developing a grade prediction pipeline such as the one represented in Figure 5.1 for an educational system or, more specifically, an LAD or AI-powered LAD, it is important to define early on how the model training and prediction process will work for different courses.

First and foremost, to train a grade prediction model for a specific course unit, an user must input information regarding student scores through the submission of an CSV file. To reduce the need for manual work, these files are expected to contain only two columns, the student's name and the final grade, information typically already stored by professors and administrators.

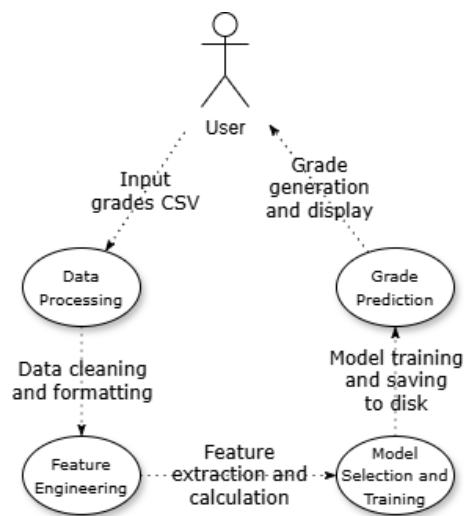


Figure 5.1: *CABRAL's* Grade Prediction Pipeline

Through this submission the system triggers the model training process and either creates a brand-new prediction model, or updates an already existing one.

At first glance, while generalizing the process to any course unit inputted in the system might feel like an easier and less computationally expensive alternative, this choice has severe drawbacks in the overall system's capability and adoption expectancy, due to several factors.

Firstly, since every course unit is different, namely its content and structure, a model trained on a generalized dataset might fail to capture the unique learning patterns and performance indicators specific to a particular discipline. Secondly, as students can interact with different course units in different ways, a course's LMS usage pattern can differ exponentially from another's, leading to student interaction imbalance, thus leading to erroneous predictions, especially when dealing with diverse course units. Thirdly, since each course can possess its own unique context, difficulty level, type of assignments, frequency of assessments, and interaction dynamics between students and instructors, ignoring course-specific features by adopting a generalized model would diminish the model's ability to make precise and relevant predictions. Lastly, a single model that accurately predicts grades across vastly different courses would be forced to find commonalities that might not be strongly predictive in all contexts, leading to suboptimal performance compared to a more specialized approach. As a result, an hybrid approach was chosen.

When a course unit is created in the system through user-submitted logs, two paths may follow. On one hand, if the course unit is an iteration of an already existing course, the grade prediction model used is of the already existing course, eliminating the need for additional model training. Otherwise, since there is no available data to train the model at the beginning of the course, the user who created the course unit is prompted to pick from one of the available models to temporarily attach to the course. Allowing the users to choose which model they want to temporarily make predictions is far more beneficial to the system's quality since educators and administrators generally have more knowledge on the specificities of the course units, empowering them to choose the prediction model of a course unit similar to the one in question, thus usually generating better

results.

The model training process was designed to be straightforward and understandable, as explainability was one of the main goals set for *CABRAL*.

Firstly, the component begins by ingesting and preparing student grade data, obtained from the user-submitted CSV file, involving a crucial cleaning phase where raw grade strings, often containing commas or percentage signs, are standardized into a numeric format, ensuring data integrity for subsequent analytical steps. Simultaneously, the system connects to the database to extract both comprehensive student profiles and detailed activity records for the specified course.

As the data sources do not offer a standardized format for the ingested data, harmonizing it is a pivotal step towards successfully predicting student metrics. To achieve this, student grade data, initially linked by student names, is seamlessly merged with the database's student records using unique student IDs, ensuring that activities are correctly associated with the students whose grades are being predicted. From the newly formed student activities, the system crafts meaningful features for the predictive models by, for example, extracting timestamps from event logs to derive insights into student engagement patterns, such as the hour of the day and day of the week they typically interact with course materials. These raw time-based features are then aggregated to create a concise summary of student behavior, such as the total count of activities and the average and standard deviation of their interaction times.

With the data meticulously prepared and features engineered, the system proceeds to the next phase by dynamically selecting a regression model based on a specified input, offering a choice among various robust algorithms including AdaBoost, Random Forest, Linear Regression, CatBoost, and even a Dummy Regressor for baseline comparison. Allowing the user to select the machine learning model rather than the system automatically selecting one significantly enhances the system's adaptability and utility for diverse educational contexts. Not only does this flexibility empower educators and researchers to choose algorithms that best suit specific course data characteristics and pedagogical goals but it allows them to optimize for either higher accuracy or greater interpretability as needed. Additionally, providing model choice fosters transparency and trust in the system's predictions, while also facilitating valuable experimentation and baseline comparisons for continuous improvement.

The chosen model is then trained on a carefully partitioned subset of the data, with a portion reserved for testing its performance on unseen information. Crucially, the training data is also preserved for future use, particularly for generating SHAP [161] explanations, which offer valuable insights into the factors driving each prediction.

Finally, the trained model undergoes a rigorous evaluation, assessing its predictive accuracy using a suite of standard metrics such as MAE, Root Mean Squared Error (RMSE)², and R-

2

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

squared. Additionally, a specialized metric, the Mean Hassanat Similarity Percentage (MHSP) [162]³ is calculated to provide a nuanced understanding of the model’s agreement with actual grades, especially for non-negative values.

Once the model is trained and persisted to disk, the system enables the users to predict their grades through simple clicks in the interface, as shown in Figure 5.2.

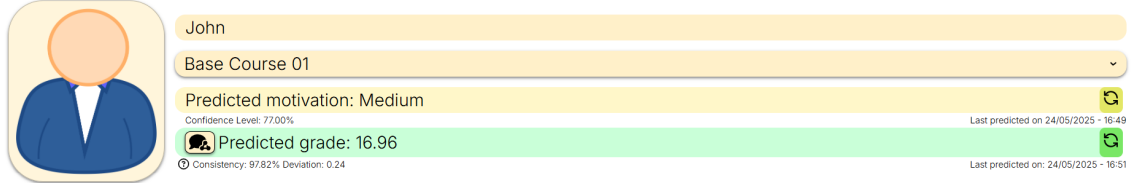


Figure 5.2: Prediction Buttons in *CABRAL*’s UI

5.3.2 Motivation Prediction

While having a sustained grade prediction can be extremely beneficial to a given student’s academic performance, it can, sometimes, be difficult to correctly assess students’ determination and effort at any given time in the semester. As a result, a motivation prediction component that aims to estimate students’ level of motivation by leveraging their LMS engagement data patterns was developed.

As building the student motivation prediction pipeline illustrated in Figure 5.3 is a fairly complex process, it is important to analyze its development in a deeper way.

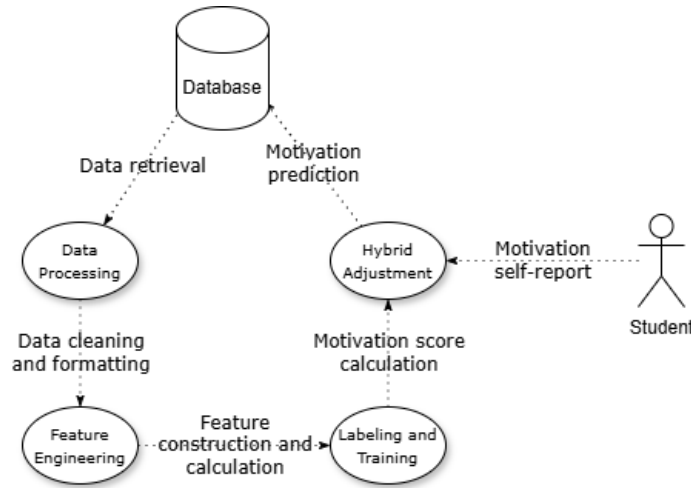


Figure 5.3: *CABRAL*’s Motivation Prediction Pipeline

First and foremost, the component connects to the system’s database to retrieve all relevant student activity logs and teacher information for a given course, while simultaneously filtering out

$$\text{HasS}(x, y) = \begin{cases} \frac{1 + \min(x_i, y_i)}{1 + \max(x_i, y_i)} & \text{if } \min(x_i, y_i) \geq 0, \\ \frac{1 + \min(x_i, y_i) + |\min(x_i, y_i)|}{1 + \max(x_i, y_i) + |\min(x_i, y_i)|} & \text{if } \min(x_i, y_i) < 0. \end{cases} \quad \text{MHSP} = \frac{\text{HasS}(x, y)}{n} \times 100\%$$

activities performed by teachers. This step ensures that the subsequent analysis focuses solely on student-generated data, which is essential for accurate motivational assessment.

Afterwards, to allow for a standardized understanding of students' behavior, raw activity descriptions are transformed into categorized action types, such as `quiz_interaction`, `resource_view`, or `forum_post`, and temporal features like hour and dayofweek are extracted from timestamps, providing insights into when students are most active. These actions are then aggregated per student, creating a rich profile of their engagement, including counts of different activity types and summaries of their activity timing.

As not all activities hold equal weight in indicating motivation, the system then normalizes these activity counts using a Min-Max Scaler [163] and applies a heuristic weighting scheme, represented in Table 5.1.

Table 5.1: Activity Weights

Activity	Weight
<code>resource_view</code>	10.0
<code>quiz_interaction</code>	2.0
<code>forum_post</code>	15.0
<code>forum_view</code>	7.5
<code>assignment_view</code>	2.0
<code>course_view</code>	20.0
<code>unique_active_days</code>	20.0

While appearing arbitrary at first, these values were carefully pondered during the development phase and reflect the different importance of different actions when trying to assess a student's motivation.

For example, activities representing posts on a forum receive a very high weight because they signify active contribution, critical thinking, and a willingness to engage with peers and course material beyond passive consumption. Similarly, activities representing viewing a course and its resources are intended to capture consistent presence and overall engagement with the learning environment.

While seemingly basic actions, frequent and regular viewing of the course page or content as well as posting a message on a forum indicate that a student is consistently checking in, staying connected, and maintaining a presence, which is a strong foundational indicator of sustained motivation, even if not always accompanied by deeper interactions. The `unique_active_days` value also receives the highest weight as this metric directly measures consistency and persistence over time. A student who interacts with the course on many different days demonstrates sustained effort and discipline, which are paramount aspects of motivation, rather than sporadic bursts of activity.

Viewing a forum, while less impactful than posting, still signifies a student's passive engagement with community discussions, indicating they are following conversations and absorbing information from their peers, which is a step beyond merely existing in the course.

On the other hand, actions like interacting with a quiz and viewing an assignment are given lower weights because they represent more fundamental, often prerequisite, steps in the learning process. Merely viewing a quiz or an assignment does not inherently denote high motivation as it could simply be a check-in or a necessary preliminary action before deeper engagement or completion. Their low weight reflects that these are necessary but not sufficient indicators of strong motivational drive.

These weighted activities are then synthesized into a single motivation score, then manually labeled into discrete categories ('Low', 'Medium', 'High') based on predefined thresholds, serving as the target variable for the classification model.

The choice for discrete categories stems from the fact that, as opposed to student grades, motivation values are harder to quantify and, when quantified, harder to interpret and act upon. Displaying single words to convey the student's motivation level simplifies the decision making process and ensures a consistent message.

Finally, with the features meticulously prepared and the motivation labels established, the system proceeds to the model training phase, employing a Random Forest Classifier to learn the complex relationships between student activity patterns and their motivation labels. The complete set of features used to train the motivation prediction model is represented in Table 5.2.

Table 5.2: Feature Set used in Motivation Prediction Model

Feature Name	Description
resource_view	Normalized count of resource views
quiz_interaction	Normalized count of quiz views or attempts
forum_post	Normalized count of forum posts
forum_view	Normalized count of forum views
assignment_view	Normalized count of assignment-related views
course_view	Normalized count of general course views
unique_active_days	Normalized number of distinct days active in course
hour_mean	Mean hour of day when student was active
hour_std	Standard deviation of active hours
dayofweek_mean	Mean day of week for activity (0=Monday to 6=Sunday)
dayofweek_std	Standard deviation of active days

Having trained and persisted a model to disk, the system is then able to predict a given student's motivation level at any point in the semester. Nevertheless, this solution is far from perfect, as, similarly to grade predictions, facing an incorrect motivation prediction can prove to be detrimental to a student's actual desire to keep working and therefore affect their educational outcomes. To solve this problem, an hybrid approach that integrates the traditional machine learning model with a Q-learning reinforcement learning [164] component was developed.

As a student's motivation is a fairly subjective value and can only be confirmed or denied by the student, incorporating their own self-assessment provides the most reliable ground truth available for this kind of prediction task, as it allows the system to validate and adjust its outputs based on the individual's perspective, thus ensuring the model aligns more closely with the student's actual



Figure 5.4: Emojis employed in *CABRAL*

experience. As a result, periodically and before being able to retrieve the system's prediction for their motivation, students are prompted to answer a single question in the system: How was your motivation this week? This question has only three possible answers, all represented by the same emojis used to convey their motivation in other pages of *CABRAL*: sad, neutral and happy, represented in Figure 5.4.

This answer is then submitted into the system and triggers a Q-learning update on the Q-table created during the model training phase. It records the current student's state, the model's prediction, and the actual reported motivation into a separate database for feedback logging. Based on whether or not the prediction matched the reported category, a reward is calculated, and the Q-table is updated using the Q-learning algorithm, allowing the system to iteratively improve its predictions over time by adapting to real-world feedback, making the motivation assessment more precise and responsive. Over time, if the Random Forest model's prediction confidence falls below a set threshold, the system consults a dynamic Q-table that learns optimal actions (motivation predictions) based on past experiences and environmental states. This Q-table's values, normalized and combined with the Random Forest's probabilities using a lambda weighting factor, yield a final, refined motivation prediction.

5.4 Actionable Insights

As translating raw data and predictive analytics into meaningful educational interventions can prove to be a difficult task to accomplish, not displaying actionable insights to both educators and students would turn the sophisticated data processing and machine learning processes into merely academic exercises, reducing their inherent value. As a result, an actionable insights pipeline with the ability to generate personalized data-driven feedback for both types of *CABRAL* users based on activity logs and predicted grades was developed, represented in Figure 5.5.

To address both student and educator needs, the pipeline operates at two distinct levels: one focused on individual student learning and the other on course-wide trends, each representing a module of the overall component.

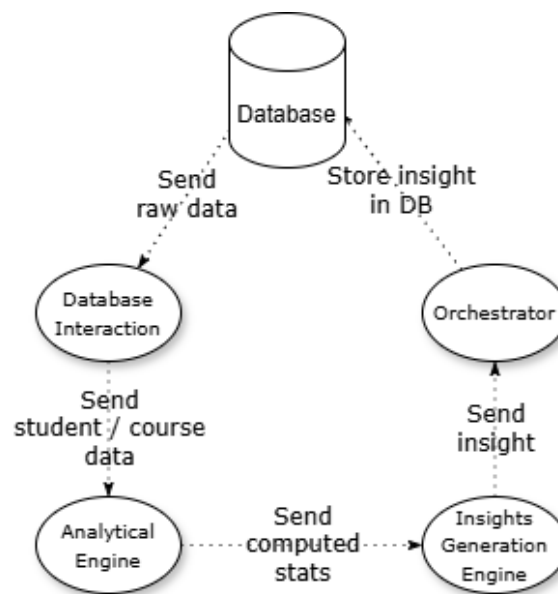


Figure 5.5: Actionable Insights Pipeline

While fulfilling different user needs, these two levels share the same initial steps, making it important to understand how the common execution takes place before straying into specifically addressing one user type or the other. To do so, it is important to analyze each module on its own, to then observe the process as a whole.

5.4.1 Database Interaction

The first module of the component serves as the system's robust data abstraction layer, managing all interactions with the MongoDB database through an handler class, such as establishing a reliable connection to the database and its "InsightLogs" collection.

To precisely query student data, a few methods are developed inside this module, namely `get_last_student_data_by_stdId_and_courseId`, `get_other_course_data_by_courseId` and `get_all_student_data_by_stdId`.

The first method, `get_last_student_data_by_stdId_and_courseId`, fetches the most recent activity record for a single student within a given course, `get_other_course_data_by_courseId` retrieves the latest activities for all other students in a specific course (excluding the target student), enabling peer-based comparisons and `get_all_student_data_by_stdId` gathers a comprehensive history of a student's engagement across all courses, forming the basis for broader insights. This meticulous data retrieval ensures that the subsequent analytical stages operate on the most relevant and timely information.

5.4.2 Analytical Engine

Building upon the data provided by the previous module, the analytical engine is responsible for contextualizing student performance and activity, defining clear grade bands and activity labels,

thus establishing a standardized vocabulary for performance levels and activity labels. The vocabulary for performance levels is represented in Table 5.3.

Table 5.3: Grade Band Labels

Grade Band	Label
0-5	very low scorers
5-10	low scorers
10-15	average performers
15-20	top performers

On the other hand, the vocabulary for activity labels is represented in Table 5.4.

Table 5.4: Activity Labels

Activity Code	Label
quiz_interaction	quizzes
forum_view	the discussion forums
resource_view	course materials
assignment_view	assignments
course_view	the course page
choicegroup_view	group choices
other	other activities

While defining specific grade bands and activity labels might look like a detail in the overall system, these two labels are fundamental for *CABRAL* to generate meaningful insights, as these vocabularies create a common language for interpreting complex data. On one hand, the grade bands dictionary helps transform continuous numerical grades into intuitive qualitative categories such as “top performers”, while the activity labels one converts technical activity logs into user-friendly terms (e.g., “quizzes”), ensuring that feedback is both clear and is able to resonate with users. This standardization is absolutely crucial for contextualized statistical analysis and benchmarking, since it allows students’ performance and activities to be accurately compared against relevant peer groups. Moreover, these categories directly facilitate human-readable insight generation by enabling the system to articulate complex data points in natural language, making the feedback actionable and maximizing its impact on student engagement and academic outcomes.

Apart from helper functions like `get_grade_band` and `compute_percentile` that enable the system to pinpoint a student’s standing precisely, the primary method the module employs is the `compute_statistics` one, which processes student activity records to generate aggregated metrics for each grade band. For every activity type defined in Table 5.4, the method calculates the average engagement and compiles the distribution of values within each performance band, providing crucial comparative context needed for generating meaningful insights. When executed, this method allows the system to determine if a student’s activity level in a particular area is high or low relative to others within their predicted grade bracket, helping both educators identify at-risk students, and students compare themselves with the rest of the class while maintaining privacy compliance.

5.4.3 Insight Generation Engine

With the necessary data retrieved and its statistics computed, the insight generation module takes center stage as the intelligence layer responsible for crafting human-readable, actionable feedback.

Despite having helper functions designed to retrieve intermediate data needed for further processing, the two main methods that compose the core of the module are the `generate_insights_for_student` and the `generate_insights_for_course`. As they accomplish different tasks in the component, it is important to analyze them individually.

The first method, `generate_insights_for_student`, focuses on providing personalized, granular feedback to individual students to foster self-regulation and targeted improvement by taking an individual student's profile (including their predicted grade and activity counts) and comparing their activities against the averages and distributions of their grade band. After the comparison, the method identifies the student's primary focus areas, i.e. where the student is the most active, the student's overall engagement, it detects whether or not the student is overly focused on one activity type or is instead engaging broadly across key areas, and pinpoints if the student's activity in a specific area is in the top percentiles or notably lower than their peers in the same grade band. With messages such as "You're in the top 10% for using quizzes!" or "Your activity in forums is lower than most students at your level. Consider using it more." Being displayed in the screen, not only are the students more knowledgeable of their effort when compared to their classmates, but they are also encouraged to change their studying patterns or to place more effort in specific areas. This method can both be used to analyze a student's profile in a single course or in a comprehensive, aggregated view across all their courses, providing different actionable feedback depending on the case.

The second method, `generate_insights_for_course` shifts the focus from singular learner profiles to the collective dynamics of an entire course cohort, empowering educators with a comprehensive understanding of overall performance trends and engagement patterns and providing a crucial macroscopic view of an entire classroom, much needed in, for example, cumulative analysis.

By operating with the aggregated course data, meaning all the student records for the specified course and the statistical averages and distributions segmented by grade bands, this method produces a multi-faceted set of insights to be displayed to educators, such as the grade performance breakdown, the overall activity engagement comparison, the engagement by grade band and the behavioral balance within the cohort. While simple insights like the course's grade performance breakdown may be easily calculated without the need for a system like *CABRAL*, other feedback like overall activity engagement comparison and engagement by grade band directly correlate the students' LMS usage to their predicted outcome, helping educators easily discern which learning resources or interaction modalities may be resonating widely with which students and which may be neglected, thus guiding tutors' decisions on resource management or educational strategies. Presenting these aggregated, course-level insights not only equips educators with a holistic perspective, but also enables them to make data-informed decisions regarding curriculum adjust-

ments, pedagogical strategies, and targeted support initiatives that benefit the entire student body, thereby enhancing the overall efficacy of the learning environment.

5.4.4 Orchestrator

The orchestrator module serves as the central director and entry point for the entire insight generation process, bringing together the functionalities of the database interaction, analytical engine, and insight generation modules. Based on the task needed to fulfill, this module can either conduct the execution to retrieve course-specific insights, or general insights, both for students or educators. If both a `studentId` and `courseId` are provided, the `get_insights_for_student_on_course` method is invoked, focused on analyzing a student's behavior within a specific course by retrieving their latest data and comparing it against the activities of other students in that same course. This pathway allows for highly contextualized feedback, addressing specific engagement patterns relevant to that particular unit of study. Otherwise, if only a `studentId` is provided, the `get_general_insights_for_student` function takes over. Here, the system fetches all of the student's most recent activity records across all their enrolled courses. It then aggregates these activities and computes an overall average grade to create a holistic student profile. Insights generated in this mode are broader, reflecting the student's general learning habits and overall academic journey, deliberately filtering out course-specific nuances to offer a generalized view of their engagement.

Regardless of the execution path, this module directs the flow by calling upon the database module to fetch the necessary data, then asking the analytical engine component to compute the comparative statistics, and finally the insight generation module to transform these statistical comparisons into natural language insights. The ultimate output is a JSON array of these generated insights, designed for seamless integration into *CABRAL*'s user interface, empowering both students and educators with immediate, personalized, and actionable feedback derived directly from their learning behaviors.

5.5 User Interface

As is the case with any user-facing system, but particularly with a LAD, developing a consistent, appealing, and user-friendly interface is a crucial step in achieving the intended goals. While this aspect of development is often underestimated, it undeniably plays a significant role in determining whether users will engage meaningfully with the system, being the main connection between the user and *CABRAL*. In the context of a LAD, however, the interface must do more than just look polished, as it needs to present educational data in a clear, intuitive, and actionable manner. This includes not only choosing the right visualizations, but also organizing content hierarchically and ensuring the design supports, rather than hinders, the user's ability to interpret and act upon the information. Additionally, the interface must be designed to minimize IO by guiding attention to what matters most, providing appropriate levels of detail, and enabling smooth navigation through the dashboard. Achieving this balance is far from simple, since creating an effective UI requires

careful consideration of user needs, iterative design, and user testing to ensure the final product not only functions well but also genuinely enhances the learning experience.

To effectively guide the development process, two primary design goals were established for the final system UI. Firstly, the interface should feature a soothing and minimalist aesthetic that promotes focus and reduces distractions, by employing a clean layout, calming color schemes, and intuitive visual hierarchy to help users engage with the content effortlessly. Secondly, the system should be fully accessible to users regardless of their level of technical expertise, ensuring that interacting with the dashboard does not require any prior experience with complex software or data interpretation. The UI must therefore be intuitive, with clear navigation, familiar interaction patterns, and helpful guidance or tooltips where necessary. Together, these goals ensure that the system remains inclusive and approachable while still maintaining functional depth and visual clarity, thus fostering a positive user experience for a diverse audience of educators, students, or administrators.

To support the goal of achieving a soothing and simple design, particular attention was devoted to the visual language of the interface. Recognizing that educational environments can often be stress-inducing for both students and educators, deliberate aesthetic choices were made to help counteract this pressure and promote a sense of calm. One of the most significant decisions was the selection of a soft, neutral color palette, intentionally avoiding overly bright, saturated, or high-contrast tones that could distract or visually fatigue users, instead opting for a pale, warm-yellow beige as the dominant background color, seen on Figure 5.6.

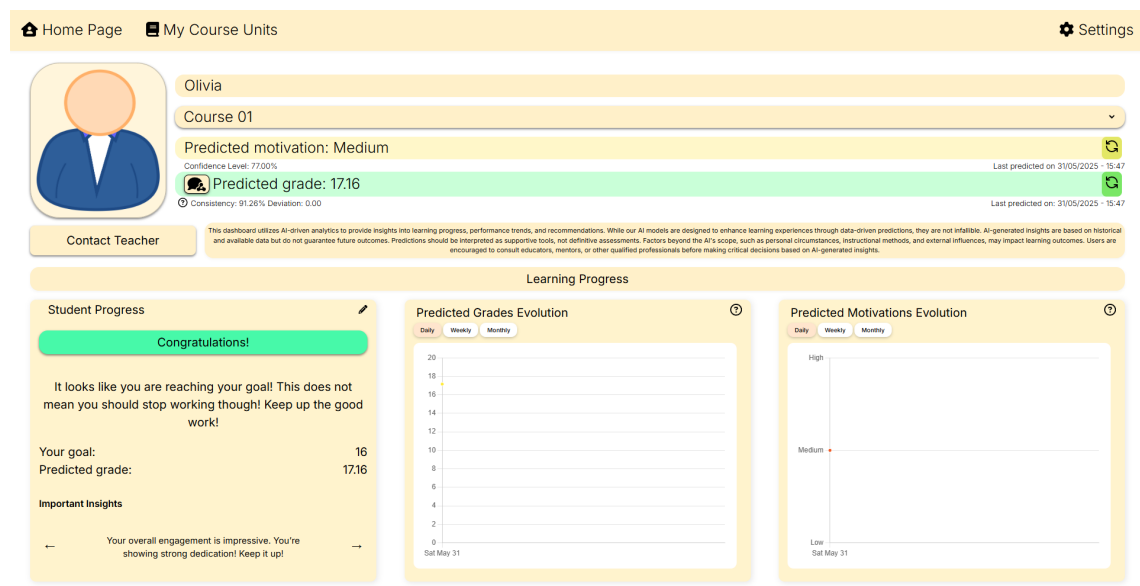
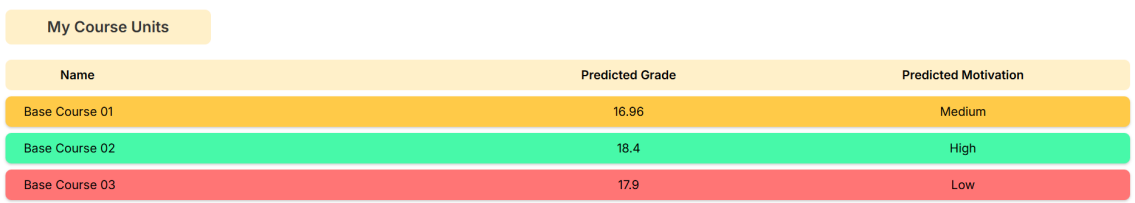


Figure 5.6: CABRAL's main page UI

This hue was carefully chosen not only for its gentle, unobtrusive appearance, but also for its functional benefits as it serves as a subtle canvas that allows more vibrant visual elements such as charts, alerts, or progress indicators to stand out clearly without overwhelming the viewer.

Additionally, its warmth and low visual intensity help to reduce eye strain, making the interface more comfortable for prolonged use.

To further enhance clarity and provide intuitive visual cues, especially when conveying key information such as student motivation levels or academic performance, a straightforward traffic light color-coding system was implemented. This familiar visual convention using red, yellow, and green indicators allows users to quickly grasp the significance of certain data at a glance, reducing the cognitive effort required to interpret numerical or textual feedback. An example of this system can be seen in Figure 5.7, where these indicators are strategically integrated into the dashboard to communicate critical insights without the need for extensive explanation.



My Course Units		
Name	Predicted Grade	Predicted Motivation
Base Course 01	16.96	Medium
Base Course 02	18.4	High
Base Course 03	17.9	Low

Figure 5.7: Traffic Light System employed to convey Student Predictions

Conversely, the overall layout of the interface was guided by a minimalist design philosophy, with a strong emphasis on clarity, focus, and functional simplicity. At any given time, only the most essential and contextually relevant information is displayed, allowing users to concentrate on their immediate tasks without being overwhelmed by superfluous content. This intentional reduction of on-screen elements helps minimize distractions and supports a more goal-oriented user experience. To achieve this, unnecessary or decorative visual components were consciously eliminated, creating a clean and unobtrusive environment that prioritizes usability and avoids IO. As a result, information is grouped logically and presented in a clear hierarchy, ensuring that related elements are both easily identifiable and accessible which improves readability but also enhances the system’s overall navigability, allowing users to locate and interpret data quickly and efficiently.

To allow for a more widespread user adoption the system was designed to adapt seamlessly to a wide variety of screen sizes and device types, from large desktop monitors to tablets and smartphones, ensuring that users can interact with the dashboard through their preferred platform without sacrificing clarity, functionality, or performance.

Another key element in supporting legibility and reducing cognitive load throughout the interface was typography as thoughtful typographic choices help establish a clear and intuitive visual hierarchy, allowing users to process information more easily and navigate the system with minimal effort. By employing generous spacing between elements, consistent alignment, and deliberate use of whitespace the UI and its components not only contribute to a sense of visual balance and calm but, most importantly, reinforce the overall design goal of simplicity.

The primary typeface used across the system is Inter by Rasmus Andersson, a font specifically designed for digital interfaces. The choice behind Inter stems from its exceptional legibility,

clarity, and adaptability across various screen sizes and resolutions as well as its clean, modern aesthetic that complement the minimalistic visual language of the dashboard, while still allowing for effective typographic differentiation given its extensive range of weights and styles. To create a visible hierarchy of information, varying font weights were employed strategically highlighting headers, labels, and key data points so that users can quickly scan and identify the most important elements at a glance without needing a deep inspection.

On the other hand, the overall layout of the interface was crafted with a minimalist philosophy in mind, having only the most relevant information presented at any given time, reducing distractions and helping users focus on their immediate goals. Unnecessary visual elements were stripped away to prevent clutter, and thoughtful grouping of related components further streamlined the experience. Importantly, the interface was designed to be fully responsive, ensuring seamless usability across a range of devices so users can interact with the system through whichever platform they prefer without compromising functionality or clarity.

To ensure that users' varying levels of digital or technological proficiency would not hinder their overall experience with the platform, a strong emphasis was placed on intuitive design and the use of straightforward, accessible interfaces as the main goal of *CABRAL* was to be easily navigated and understood by all users regardless of their familiarity with digital tools or data interpretation. In keeping with this objective, when presenting educational data such as individual student grade distributions or the overall progress of a class, the choice to implement clear and easily interpretable data visualization methods was deliberately made. This meant that, rather than relying on complex or highly technical chart types, simpler and more universally recognizable formats, such as donut charts and radar charts were employed as these visualizations have an undeniable ability to convey key patterns and insights at a glance, without requiring prior experience with analytical tools. This approach not only supports quick comprehension but also helps reduce cognitive load, allowing users to focus on the meaning behind the data rather than the mechanics of the visualization itself. Examples of these visualization strategies in use can be seen in Figure 5.8, where the design facilitates immediate interpretation while maintaining visual clarity and consistency with the overall user interface.

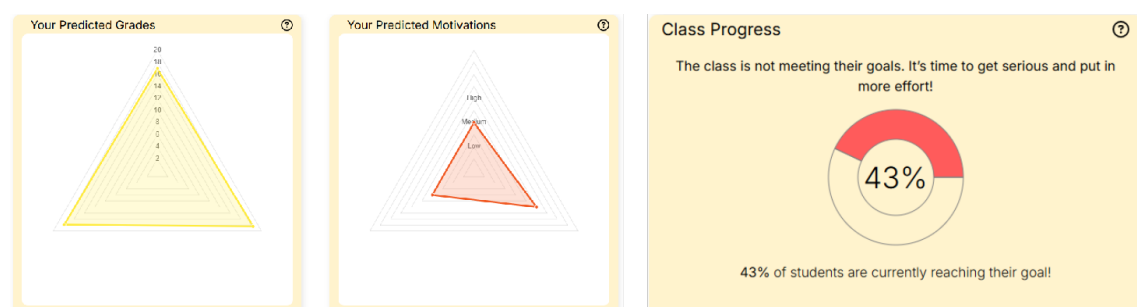


Figure 5.8: Examples of *CABRAL*'s Visualization Techniques

Given that a key feature of *CABRAL* is its emphasis on XAI, as well as broader principles of trust and transparency, specific interface elements were incorporated to support user understanding

of both the data presented and the underlying processes at work. To this end, contextual help buttons were strategically placed throughout the interface. These buttons, visually represented by a circular icon containing a question mark, serve as intuitive and easily recognizable entry points for users seeking additional clarification or deeper insight and have a twofold purpose: demystifying the operations of AI-driven components by explaining how certain predictions or insights are generated, and enhancing the interpretability of complex data visualizations and metrics presented on screen. When activated, each button opens a context-specific modal window that provides supplementary information tailored to the component in question it being explanations of key terms, detailed breakdowns of how specific values were calculated, or step-by-step walkthroughs of the predictive models' training and inference processes.

By making this information readily accessible and situating it directly within the user's flow, the system supports informed decision-making while fostering a sense of control and transparency, especially important in the context of educational environments, where both students and educators must feel confident in the validity and fairness of automated insights. An example of this functionality is illustrated in Figure 5.9, which demonstrates how a student can access detailed explanations of why a particular course was flagged as at risk.

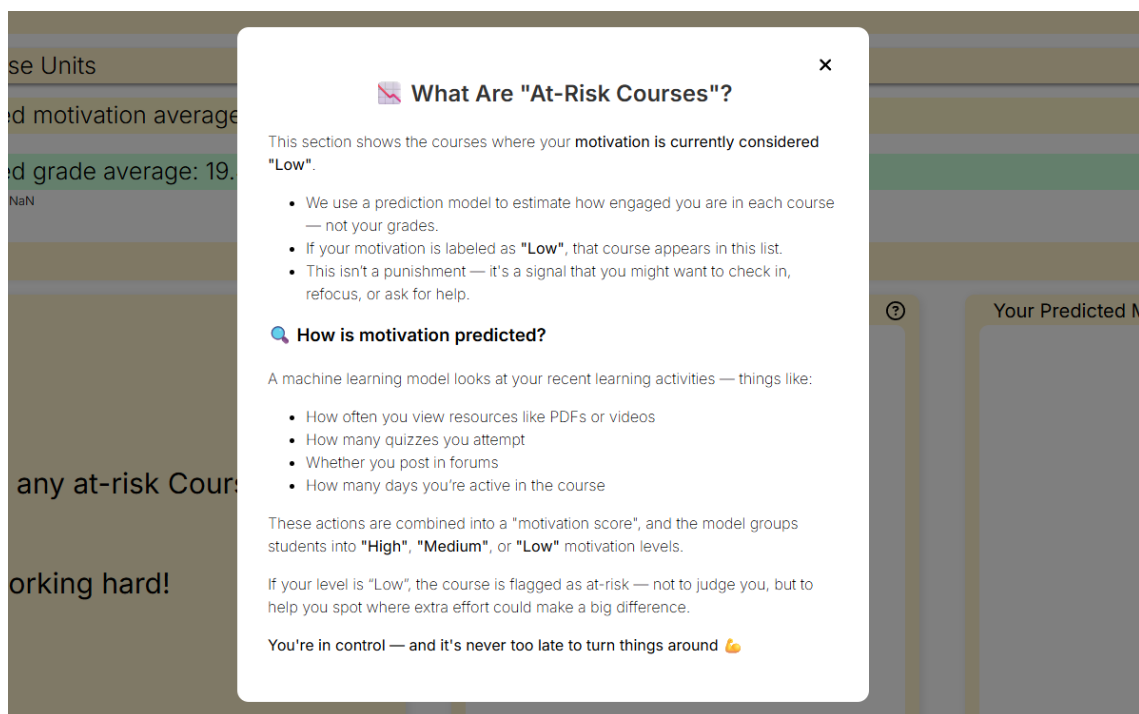


Figure 5.9: At-Risk Courses Help Modal

5.6 Messaging Generation System

Although the digitization of the average educational environment allows for a more autonomous and self-regulated learning experience, timely and personalized support from educators still remains a significant factor in promoting students' academic success. As a result a dedicated digital

tool aimed at facilitating meaningful communication between students and professors was developed that, integrated directly into *CABRAL*, aims to lower the barrier for students seeking assistance and for educators seeking to provide it by automatically generating customized message drafts based on their academic data and current context within a course. Rather than requiring students to initiate contact from scratch, an often intimidating or uncertain process, this component provides them with thoughtfully pre-written message templates that they can review, edit, and send with ease to their professors. Furthermore, not only are these drafts tailored to specific situations, such as requesting help after a drop in performance, inquiring about missed deadlines, or seeking clarification on course material, but also leverage available student data to populate the messages with relevant details such as recent grades and predicted motivations, making the communication more specific and actionable for instructors.

The primary purpose of this tool is to reduce the time, hesitation, and emotional effort that students often experience when crafting academic messages, especially in situations where they may feel uncertain, overwhelmed, or intimidated. By providing a structured and contextually appropriate message draft, the messaging generator ensures that communication remains clear, respectful, and aligned with academic norms, while still allowing room for personal expression and customization.

On the other hand, the tool is designed to support not only student-to-educator communication, but also educator-to-student outreach, as instructors can use similar mechanisms to initiate proactive messages based on student performance indicators, offering guidance, encouragement, or interventions when necessary. An example of this feature is illustrated in Figure 5.10, where users can see how a message draft is generated in response to specific academic triggers or conditions.

Home Page My Course Units Settings

John
Base Course 01

Grade Prediction
16.96
You are reaching your goal!

Motivation Level
Medium
It looks like you are a bit motivated!

Message Settings
Tone: Formal
Length: Short
Topic: I want to improve my motivation
Generate Message

This dashboard utilizes AI-driven analytics to provide insights into learning progress, performance trends, and recommendations. While our AI models are designed to enhance learning experiences through data-driven predictions, they are not infallible. AI-generated insights are based on historical and available data but do not guarantee future outcomes. Predictions should be interpreted as supportive tools, not definitive assessments. Factors beyond the AI's scope, such as personal circumstances, instructional methods, and external influences, may impact learning outcomes. Users are encouraged to consult educators, mentors, or other qualified professionals before making critical decisions based on AI-generated insights.

Your Message

Dear Professor,
I am writing to discuss my current progress in your course. With a predicted grade of 16.96, I am concerned that my motivation levels, which are currently medium, may be impacting my performance. I would greatly appreciate any guidance or advice you can offer to help me improve my motivation and achieve my academic goals.
Sincerely,
John

Message Stats
This message will be sent to
Adam
your professor for the course Base Course 01.
This message has 59 words!
As this message was generated with the help of an AI model, it would be wise to double-check your words!
Are you sure you are comfortable with your message?
Send Message

Figure 5.10: Message Generation System

Within this component, students and educators are given the ability to customize key aspects

of the generated message to better suit their needs and communication preferences namely the desired tone, message length and general topic or intent of the message. This allows the tool not only to adjust the tone and linguistic style of the message, but also to shape its content in a more meaningful way as a student predicted to be struggling may receive a message draft that emphasizes empathy and encouragement, while a high-performing student may be guided to ask more targeted questions or seek advanced feedback. Similarly, when an educator is composing a message to a student, the system can suggest language that balances support and constructive feedback based on the student's academic context.

Once the necessary information is collected, a large language model is leveraged to generate a fully-formed, human-readable message tailored to the user's context. This is accomplished through a Python-based implementation that programmatically constructs a detailed prompt and submits it to the Groq API [165], which interfaces with the *llama-3.3-70b-versatile* language model.

The choice behind Groq API stems from its speed, reliability, and ability to support high-quality language models. Additionally, the API supports advanced open-source models like LLaMA 3 [166], ensuring the generated content is coherent, respectful, and context-aware, making it a strong fit for educational environments where communication must be both fast and thoughtful. Groq also offers a straightforward developer experience and cost-effective usage, making it suitable for scaling the feature across many users while maintaining privacy and efficiency.

Depending on whether the message is being sent by a student or a teacher, the script dynamically assembles a different version of the prompt, ensuring the generated content is appropriately framed, respectful, and situationally relevant. The prompt used to generate the drafts is the following:

"You are a helpful assistant designed to generate a well-crafted message from a professor to a student. The purpose of this message is to provide meaningful, respectful communication tailored to the student's academic situation and the professor's intent. The professor has the following preferences and context: 1. Tone: tone - The professor prefers a tone tone in this communication. 2. Message Length: length - The professor wants the message to be length in length. 3. Message Topic: The professor is addressing the following topic with the student: "topic". This might include grade updates, encouragement, feedback, clarification, or guidance on coursework. 4. Student's Academic Context: Based on current understanding: - Predicted Grade: grade - Observed Motivation Level: motivation 5. Student Name: student - The message is intended for student. Your task is to generate a thoughtful and professional message from the professor to the student. The message should be clear, kind, and aligned with the professor's tone and intent, while showing consideration for the student's performance and motivation. Do not include any instructions or comments outside the message itself. Just generate the message the professor would send. Important: The generated message should not include any accented characters or special symbols. Please ensure the text is plain, UTF-8-safe, and uses only standard ASCII characters."

The prompts explicitly instruct the model to adhere to specific stylistic and technical constraints, such as returning the output using only standard ASCII characters and avoiding including any instructions, metadata, or formatting beyond the message itself. These constraints ensure the final result is a clean, ready-to-edit draft that can be directly presented to users without the need for additional filtering or sanitization.

Once the message is generated, it is immediately displayed within the interface for the user to review and retain full control over, either accepting the message as is, making minor adjustments, or completely rewriting it to better suit their needs or personal communication style. To further support this process, the interface includes a real-time word counter and clear visual indicators that help users assess the length, tone, and potential clarity of the message, particularly useful features in educational contexts where effective communication often depends on conciseness and appropriateness. If the message sender is a student, the interface also prompts them to select the specific educator to whom the message will be addressed, ensuring that the communication is correctly routed and contextually relevant.

The entire subsystem is built with transparency and user empowerment at its core, including a clear AI disclaimer that informs users the message they are viewing was generated with the assistance of artificial intelligence. This notice not only encourages users to carefully review and, if necessary, modify the message before sending, but also ensures they remain in full control of their communication. By providing this guidance, the system fosters trust and promotes responsible use of AI-generated content, ultimately complying with the previously stated XAI principles and with a positive use of AI.

Chapter 6

Evaluation and Discussion

In order to thoroughly evaluate a system like *CABRAL*, it is essential to define and consider the two core dimensions that contribute to its overall effectiveness and value and evaluate each in a personalized way.

As a system's success depends not only on its functionality, but also on its ability to support meaningful, low-friction interactions, the first dimension to analyze in *CABRAL* pertains to the user experience. As any user-facing platform, especially in an educational context, must be accessible, engaging, and intuitive to a broad and diverse user base, assessing how easily and effectively users can interact with the system to achieve their goals, taking into account usability, clarity, emotional response, and overall satisfaction becomes a necessary task to accomplish in order to understand the possibility of widespread adoption of a system like this one.

The second dimension, on the other hand, focuses on the system's core functionality and predictive power. In an LAD like the one developed, where much of its value stems from intelligent, data-driven insights, it becomes essential to examine the accuracy, robustness, and consistency of the underlying models, namely through evaluating how well the system can predict academic outcomes, personalize feedback, and support timely interventions. Understanding this technical dimension is vital not only to ensure that the system delivers meaningful and trustworthy results, but also to assess its practical viability and long-term integration within real educational contexts.

To address both these dimensions in a structured manner, an evaluation protocol was designed consisting of two distinct yet complementary phases: a usability evaluation, focusing on the interface and user interaction, and a predictive evaluation, aimed at assessing the accuracy and trustworthiness of the system's underlying models and recommendations. This chapter aims to address this evaluation protocol through a detailed presentation of the evaluation methods, results, and their implications.

6.0.1 Usability evaluation

To better understand the users' experiences, emotions and difficulties derived from using *CABRAL*, two usability studies, one for students and one for educators, were designed and conducted consisting of two separate but sequential phases: an experimental phase and a feedback-driven phase,

both intended to offer a comprehensive view of the platform's effectiveness from both behavioral and reflective standpoints.

In the first students' phase, participants access a *CABRAL* prototype and interact with it by assuming the role of a fictional student named Alex. In this simulated context, they are guided through four simple tasks in the system, designed to illustrate *CABRAL*'s main functionalities, listed below:

Task 1. Predict your motivation in Course 02

Task 2. Send a message to your professor of Course 01 to improve your motivation

Task 3. Predict your grade in Course 02

Task 4. Set a goal grade of 15 for Course 01

On the other hand, for the first educators' phase, participants are asked to complete five simple tasks, also designed to portray the system's main functionalities, listed below:

Task 1. Predict Gray's motivation in Course 01

Task 2. Check the advanced information for Course 01

Task 3. Add a new course unit with the default logs

Task 4. Send a message to Gray to help him improve his motivation

Task 5. Predict every student's grade in Course 01

Following the interaction with the prototype, participants move on to the second phase, in which they complete a short questionnaire to provide feedback on their experience and basic information to contextualize the responses such as age, degree currently trying to achieve and course for students and courses usually taught and previous experience with LADs for educators. In this questionnaire participants are asked to answer SEQs for each task performed in the prototype, to name the difficulties in each task, to fill in the SUS questionnaire and one simple question: could a prototype like this help you in your academic life?

This feedback is not only essential to evaluate the system's intuitiveness and clarity, but also, and most importantly, its potential to support real educational needs.

Although the initial plan involved conducting the evaluation in person, a number of practical constraints and possible limitations derived from a more individualized procedure posed serious questions regarding the executability of such an approach within the available timeframe and resources.

First and foremost, because the evaluation period coincided with the month of June, a notoriously demanding time within academic calendars, particularly for students and educators engaged in final assessments and grading activities, a substantial risk of low availability and participation emerged. Either due to exam-related fatigue, enormous academic workloads, or even a transitional

mindset anticipating the upcoming academic break, as students and educators might feel overwhelmed in the time period allocated for this procedure, the likelihood of a lower-than-desired user sample size or response rates constituted a threat to the diversity, representativity and overall reliability of the collected feedback.

Secondly, as the initially planned evaluation procedure would require flexible access to adequately equipped campus facilities along with precise scheduling coordination among all parties involved, correctly organizing a large in-person evaluation procedure would undoubtedly introduce unnecessary complexity and risk significant delays to the research process, potentially compromising both its pace and consistency.

Furthermore, while performing this research in-person would allow for a clearer understanding of the actual user experience with the system given the constant observation of the user behavior, the influence of external pressures or social desirability biases often associated with supervised sessions could prove to impact participants' natural responses, potentially skewing the authenticity of the feedback collected.

To solve these problems, a remote unsupervised procedure was designed, as conducting the study in a remote setting allowed participants to interact with *CABRAL* under conditions that closely resemble their everyday academic contexts thus generating more candid and ecologically valid feedback. Moreover, the widespread normalization and adoption of remote tools for learning, communication, and research further validated the feasibility and effectiveness of this format, since the remote and autonomous structure of the usability study not only aligned with participants' digital habits but also enhanced the inclusiveness, flexibility, and overall reach of the evaluation process. Despite seeming simpler and less demanding to coordinate, planning and executing a remote, unsupervised evaluation in a clear and efficient manner nonetheless introduces a new challenge that must be carefully addressed: how will users interact with *CABRAL*?

Although deploying *CABRAL* as a fully hosted system on a server might seem like the most simple solution, doing so introduces a range of practical challenges, especially hard in the context of a dissertation. From needing ongoing server maintenance and associated costs, to ensuring reliable uptime and security, and handling technical support for diverse user environments, integrating a dynamic, interactive prototype like *CABRAL* required a robust backend infrastructure and continuous updates and interaction logging techniques, which can be resource-intensive, heavy to implement and budget wrecking.

As a result, a flexible, cost-effective alternative that could simplify distribution and user access while maintaining the important data gathering functionalities would need to be sought. Various options were explored, including custom web hosting services, dedicated usability testing platforms, and open-source tools, but these often proved either too expensive, technically complex to implement within the project timeline, or lacked sufficient support for the level of interactivity required by the prototype.

To accommodate these needs without incurring the technical and financial overhead of full deployment, a decision was made to develop a high-fidelity interactive prototype using Figma [157], represented in Fig 6.1.

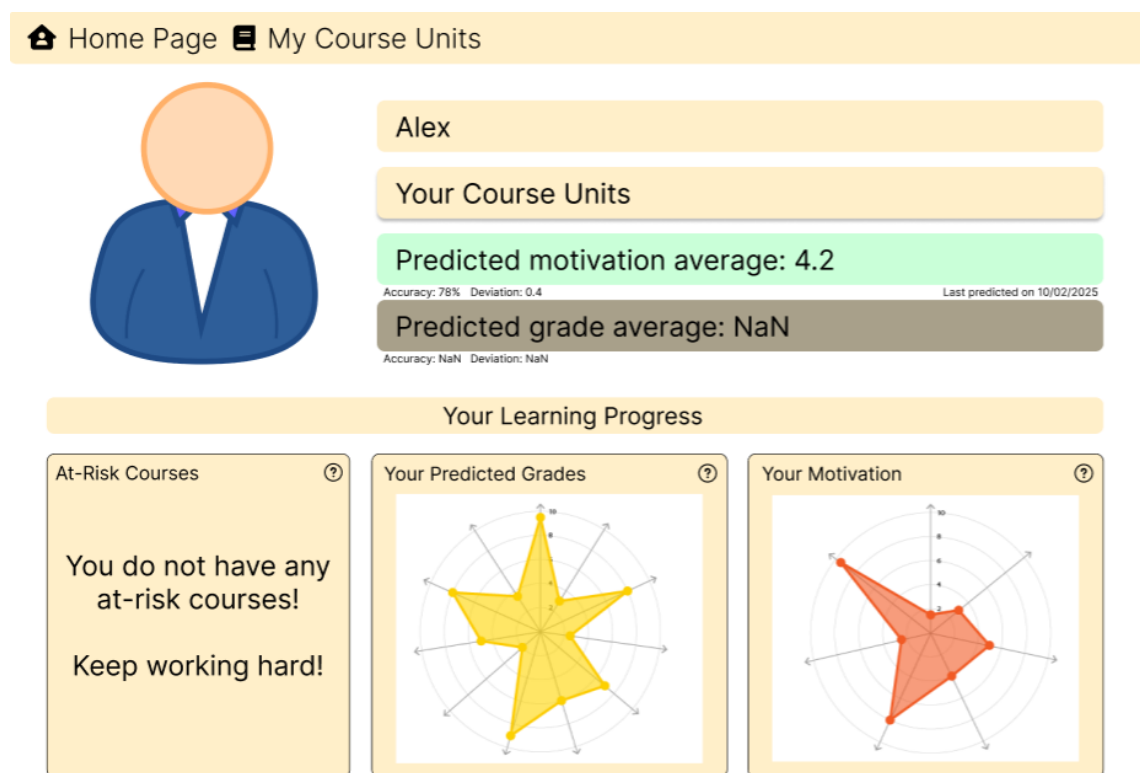


Figure 6.1: *CABRAL*'s Home Student Page in the Developed Figma Prototype

This design environment allowed for rapid, collaborative prototyping that could accurately represent *CABRAL*'s core functionalities, including clickable navigation, simulated predictions, goal-setting flows, and communication prompts. Crucially, Figma enabled the construction of a realistic, visually polished version of the system without requiring backend development, which proved essential for maintaining the project's agility and scope within the dissertation's constraints.

Ultimately, the Maze [167] platform was selected as the most suitable tool for deploying this prototype, offering an ideal balance between ease of use, interactivity, data collection capabilities, and affordability, while allowing smooth integration to create a cohesive and practical testing environment for *CABRAL*'s usability evaluation. An example of an user's interaction with the Maze questionnaire is represented in Figure 6.2.

Despite being a fitting solution, using Maze to conduct the entire process turned out to be impractical, as its free tier imposed strict limitations on the number of blocks per test, making it unfeasible to embed the complete evaluation journey within a single session. To overcome this constraint, a Google Forms [168] survey was introduced as a complementary second phase, allowing participants to submit detailed feedback about their experience with the system in a structured, asynchronous manner, in a way that preserved continuity and ensured all relevant insights could be collected without breaching technical constraints.

To recruit participants for the usability evaluation procedure, a brief and informative invitation message was crafted and distributed to students and educators of multiple Portuguese universities,

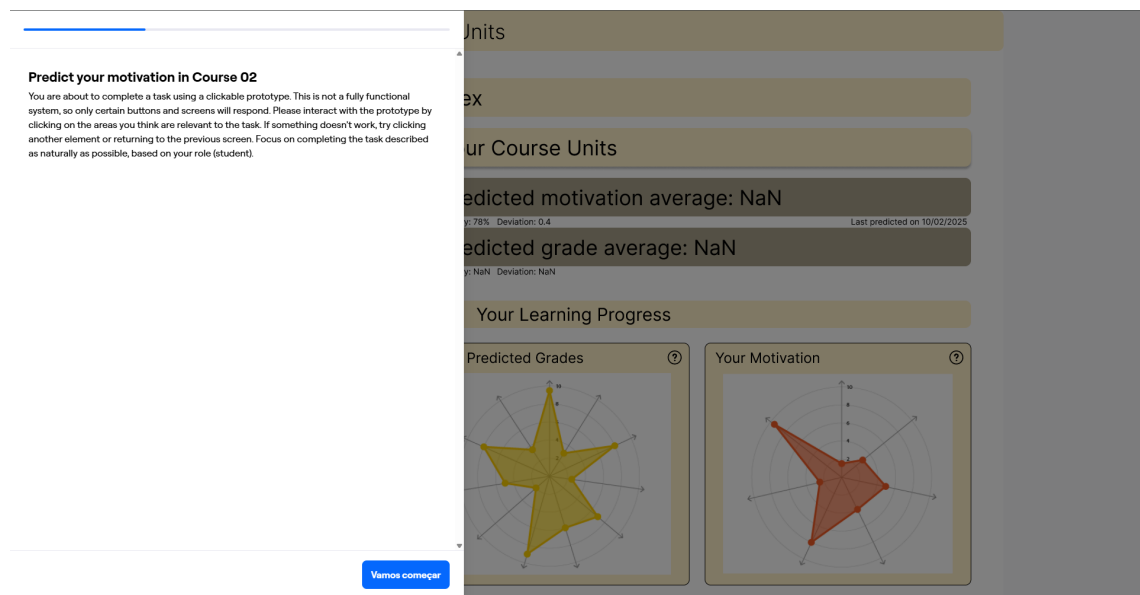


Figure 6.2: Example of User Interaction with the Maze Questionnaire

not previously in contact with *CABRAL* during its development. The goal of the message was to clearly explain the purpose of the study, outline the steps involved, and encourage voluntary participation while maintaining transparency and compliance with ethical standards.

6.0.2 Predictive evaluation

While evaluating a system's usability and user experience is a complex and detailed task to accomplish, understanding how to assess *CABRAL*'s innate capabilities is not a simple process for multiple reasons. Firstly, as some of the system's functionalities such as organizing and displaying educational data are relatively binary in nature, evaluating *CABRAL* in light of such features becomes a more clear-cut process, as the system either performs these tasks effectively or does not. Secondly, since the message generation component outlined in Subsection 5.6 heavily relies on generative AI, examining this component's performance is particularly challenging. Given the non-deterministic nature of model outputs, responses may vary across different sessions, exhibit ambiguity or even reflect subtle biases, making it extremely difficult, if not infeasible, to conduct an evaluation procedure to systematically analyze this component's performance. Thirdly, since students' motivation is a fluctuating and an difficult-to-quantify variable, determining whether or not the developed AI models accurately estimate a student's motivation level at a given time in the semester can be an herculean task, as motivation is as inherently subjective, as it is challenging to measure. Additionally, the novelty of *CABRAL*'s architecture and the limited availability of established benchmarks for similar systems further complicates any possible comprehensive assessment that could be performed on the system as a whole.

Despite these challenges, it is still possible to reliably evaluate specific components of *CABRAL*, such as the grade prediction module. Because grade prediction deals with objective, quantifiable

outcomes that can be directly compared with historical records, a reliable ground truth for the module’s validation is obtainable, enabling the use of well-established machine learning evaluation metrics and frameworks such as including accuracy-like measures and error rates, thus facilitating systematic and reproducible assessment. Furthermore, unlike more subjective or non-deterministic components, the grade prediction models produce consistent outputs given the same inputs, making their performance easier to measure and compare. Additionally, the clear success criteria for predictions and the ethical handling of anonymized academic data further support a rigorous and meaningful evaluation process.

As a result, during *CABRAL*’s development phase, a set of machine learning models was trained and evaluated using academic data sourced from three distinct course units from different semesters: Project Management Laboratory (LGP) [169] and Cyberphysical Systems and Internet of Things (SCI) [170] from the Master’s in Informatics and Computing Engineering syllabus, at the Faculty of Engineering of the University of Porto (FEUP) [171] and Programming (PROG) [172] from the Bachelor’s in Electrical and Computer Engineering syllabus at FEUP.

This data was obtained through Moodle activity reports provided by the respective educators and was integrated into the system via the previously described data extraction process. Crucially, the entire workflow adhered strictly to the GDPR, with all student records anonymized prior to storage in the system’s database and processed solely for analytical purposes under a privacy-by-design framework. To further safeguard user privacy, comprehensive data cleaning and preprocessing steps were undertaken to ensure the removal of any personally identifiable information, with the dataset used for training and evaluation being ethically sourced, clearly limited in purpose, and handled with a strong emphasis on data minimization. These privacy-conscious measures not only ensured legal and ethical compliance but also contributed to fostering user trust, also an essential factor for the potential deployment of the system in real-world educational settings.

To provide a larger contextual understanding of the datasets used for training and evaluating the predictive models, Table 6.1 summarizes key information about the three course units analyzed in this study, such as the total number of enrolled students who met attendance requirements, the average final grade (after rounding to two decimals), and the number of students who failed each course.

Table 6.1: Summary of Course Data

Course	# of Students	Average Grade [0-20]	Failing Students
LGP	302	16.62	1
PROG	277	11.53	60
SCI	28	16.96	0

While these courses aim to reflect the majority of typical course experiences in a college student’s academic life, it is important to highlight that different courses demonstrate varying degrees of Moodle engagement due to several factors, like the educators’ motivation to use the LMS, the availability of online resources for each course, or specific teaching strategies leading some courses to rely heavily on it for assessments and interactions, and others to only doing it

minimally. Undeniably, this discrepancy affects both the volume and the richness of the available training data, and thus, forces predictive accuracy to be interpreted in the context of course-specific digital activity, instead of in a generalizable way.

To ensure consistency, all student grades not previously rounded by the educators were rounded to the nearest integer on a 0–20 scale, with students who failed due to attendance violations being excluded entirely, as no meaningful grade could be inferred from their learning progress.

In order to explore how different modeling strategies would perform on the available course data, two distinct types of models were trained: simple and smart models.

On one hand, the simple models were designed using a minimalist feature set derived solely from the temporal characteristics of student activity, namely the average hour of access, the distribution of weekdays in which interactions occurred, and the raw count of logged events. This allowed the evaluation of baseline predictive performance without relying on complex behavior interpretation. The complete set of features used to train the simple models is represented in Table 6.2

Table 6.2: Feature Set used in "Simple" Grade Prediction Models

Feature Name	Description
description_count	Total number of activity log entries (i.e., interactions)
hour_mean	Mean hour of day when student was active (0–23)
hour_std	Standard deviation of hour of activity
dayofweek_mean	Mean day of the week for activity (0=Mon to 6=Sun)
dayofweek_std	Standard deviation of day of week activity

In contrast, the smart models incorporated more nuanced, semantically rich features extracted from log data by categorizing student actions into pedagogically meaningful event types like quiz interaction, forum posting or assignment views. These features were then normalized by student engagement span and supplemented with behavioral diversity metrics like the activity ratio and unique active days, resulting in a more context-aware dataset. The complete set of features used to train the smart models is represented in Table 6.3.

Both types of models were trained using a consistent set of algorithms: AdaBoost, Random Forest, Linear Regression and CatBoost. The selection of predictive models in this study was guided by both empirical relevance and findings from the preceding SLR. Specifically, AdaBoost, Random Forest, Linear Regression, and CatBoost were chosen as they represent some of the most frequently cited and benchmarked algorithms in educational data mining literature and they offer a balance between interpretability, robustness to noise, and the ability to capture non-linear relationships within student behavior and performance data. By employing both linear and ensemble-based models, the system can assess predictive stability across varying data complexities and feature interactions. These algorithms were then evaluated using the standard error metrics MAE, RMSE, R^2 along with the MHSP, to offer a multi-dimensional comparison of model efficacy and assess model performance from a different perspective. These metrics were chosen based on their widespread use and relevance in educational data mining and predictive modeling as MAE,

Table 6.3: Feature Set used in "Smart" Grade Prediction Models

Feature Name	Description
assignment_view_per_day	Views of assignments per active day
choicegroup_view_per_day	Interactions with choice groups per active day
course_view_per_day	Course page views per active day
forum_post_per_day	Forum posts per active day
forum_view_per_day	Forum views per active day
log_report_per_day	Accesses to log reports per active day
other_per_day	Actions not categorized elsewhere per active day
quiz_interaction_per_day	Quiz views or attempts per active day
quiz_response_per_day	Submitted or updated quiz responses per active day
resource_view_per_day	Resource (PDF, file, etc.) views per active day
avg_hour	Average hour of activity (0–23 scale)
std_hour	Standard deviation of activity hour
avg_dayofweek	Average day of week of activity (0=Mon to 6=Sun)
std_dayofweek	Standard deviation of day of week of activity
unique_active_days	Number of unique days the student was active
activity_ratio	Ratio of unique active days to total possible span

RMSE, and R^2 are among the most commonly cited error metrics in the literature, providing a reliable means to evaluate the accuracy and generalization of models. MAE is frequently used for its simplicity and clear interpretation, offering a direct understanding of average prediction errors while RMSE, on the other hand, is often preferred for its sensitivity to larger errors, which is useful for identifying models that may struggle with outliers or extreme values. Lastly, R^2 is widely used to assess how well the model explains the variance in the target variable, offering a measure of explanatory power that is fundamental in regression analysis.

Unlike traditional error-based measures, MHSP evaluates the relative similarity between predicted and actual grades, offering an interpretable percentage that resembles the traditional classification metric “accuracy” and reflects the closeness of predictions regardless of scale or direction. Contrarily to MAE or RMSE, which penalize raw numerical differences, the MHSP metric emphasizes relative agreement, which is particularly meaningful in educational contexts. For instance, if a model predicts 65 for a student who scored 70 and 85 for one who scored 90, both predictions would receive high similarity scores, even though the absolute errors are nonzero, a behavior desirable in formative settings, where predicting a student’s general performance category is often more valuable than exact point estimates. In short, a MHSP score closer to 100% indicates that the model’s predictions closely approximate the true grades on a proportional basis while scores below 80% may suggest that the model, while perhaps minimizing absolute errors, fails to preserve the relational structure of the grades, an aspect that may be crucial for tasks such as ranking students or identifying outliers.

In addition to MHSP, traditional regression metrics such as MAE, RMSE, and the coefficient of determination R^2 were employed to provide a well-rounded evaluation framework grounded in widely accepted statistical conventions. MAE offers a straightforward interpretation by averaging

the absolute differences between predicted and actual values, thus providing a clear measure of the typical prediction error regardless of direction. RMSE, on the other hand, builds upon MAE by squaring the individual errors before averaging and then taking the square root, thereby penalizing larger deviations more heavily. This sensitivity to outliers makes RMSE valuable in scenarios where extreme mispredictions could carry disproportionate consequences, such as misclassifying high-performing or at-risk students.

Finally, the coefficient of determination R^2 was included to contextualize the model's explanatory power, indicating the proportion of variance in the observed data that is captured by the predictions. An R^2 score approaching 1 suggests that the model captures most of the underlying data patterns, while lower scores signal poor alignment with the target variable. Unlike error metrics, R^2 facilitates a relative assessment, offering insights into how well the model performs compared to a naive baseline such as the mean. When considered collectively, these metrics not only benchmark the model's predictive precision but also support nuanced interpretations across practical, statistical, and pedagogical dimensions.

6.1 Evaluation Results

After following the dual-phase evaluation protocol previously outlined, it is important to analyze the results obtained from the two central dimensions of system assessment: user experience and predictive performance. By organizing the findings in alignment with the system's core goals, facilitating intuitive, emotionally resonant interactions while delivering reliable and intelligent academic predictions, a holistic view of *CABRAL*'s practical effectiveness in an educational environment can be provided, as the usability results reflect the extent to which the system succeeded in creating an accessible and supportive learning analytics experience, while the predictive results offer a technical validation of the system's capacity to model student performance accurately and meaningfully. Through this two-pronged analysis, we can better understand not only the system's immediate impact on users and its predictive accuracy, but also the broader implications for its adoption, scalability, and integration into authentic educational environments.

6.1.1 Usability Results

To definitively assess whether *CABRAL* is a usable system, it is essential to deeply analyze the results obtained from the user evaluation, both through objective measurements such as average task duration, misclick rates and success rates, perceived metrics like SEQ and SUS, and engineered ratios that can shed light on complex patterns otherwise unnoticeable. As a result, this subsection aims to examine three different axes of this procedure:

1. Performance, that focuses all objective quantitative metric related to how users interact with the system
2. Perception, that covers subjective or self-reported measures

3. Patterning, that includes engineered or cross-analyzed metrics that might reveal deeper insights between performance and perception

Only after scrutinizing all three axes is it possible to reach a well-supported conclusion about *CABRAL*'s usability, and, by extension, its viability in a real-world educational environment.

6.1.1.1 Performance

Before attempting to understand how users' performance varied with each task fulfilled, it is crucial to know the sample size in the analyzed and how it varied throughout the study's two parts. Since there was a light drop-off in both user groups in the experimental part of the procedure the sample size considered for the study and represented in Table 6.4 pertains to the number of users that attempted to accomplish the first task, regardless of their success, rather than the number of users that finished the procedure as a whole or even each part. In the feedback-driven phase, on the other hand, there was no drop-off, as every user that initiated this part carried it through until the end.

Table 6.4: Sample Size Distribution across *CABRAL*'s Evaluation Procedure Phases

Phase	Students	Educators
Experimental	75	14
Feedback-driven	25	7

Given the glaring discrepancy between the two phases' sample sizes, it is important to explore the possible reasons behind the substantial drop-off, particularly in the students user group. First and foremost, users were only redirected to the feedback-driven upon completing the experimental phase, meaning that any participant who abandoned the study during the initial phase would be unable to both access and complete the feedback-driven form, thereby affecting the second part's sample size. Additionally, as directly reported by some of the participants to the research team, Maze-related technical issues occasionally occurred, such as slow response times or even screen freezing. In these cases, refreshing the page would trigger the message "You have already completed this study!", effectively locking participants out and preventing them from finishing the phase. This issue was particularly compounded by the decision of allowing only one submission per device, implemented to prevent response duplication and unwanted bias in the results. As this friction in the system is unpredictable and unpreventable, it possibly compromised several users from carrying out the study completely. Furthermore, since most participants were addressed through institutional emails and had no prior relationship with the researchers conducting the study, the absence of incentive or reward may have hindered their motivation leading them to abandon the study after the first part, thus causing the drop-off. Other possible causes may stem from personal time constraints from participants leading to indefinite postponement or plain abandonment due to academic or personal obligations or user fatigue due to the possible mentally-demanding nature of completing the different tasks in the initial phase of the procedure.

Table 6.5: Progression of Student and Educator Responses throughout *CABRAL*'s Usability Evaluation Experimental Phase

Task #	# Educator Responses	# Student Responses
1	14	75
2	14	62
3	13	61
4	13	59
5	13	N/A (There is no Task 5 for Students)

Upon analyzing the drop-off between the two different phases, it is also necessary to analyze the inner drop-off within each individual part of the experiment conducted.

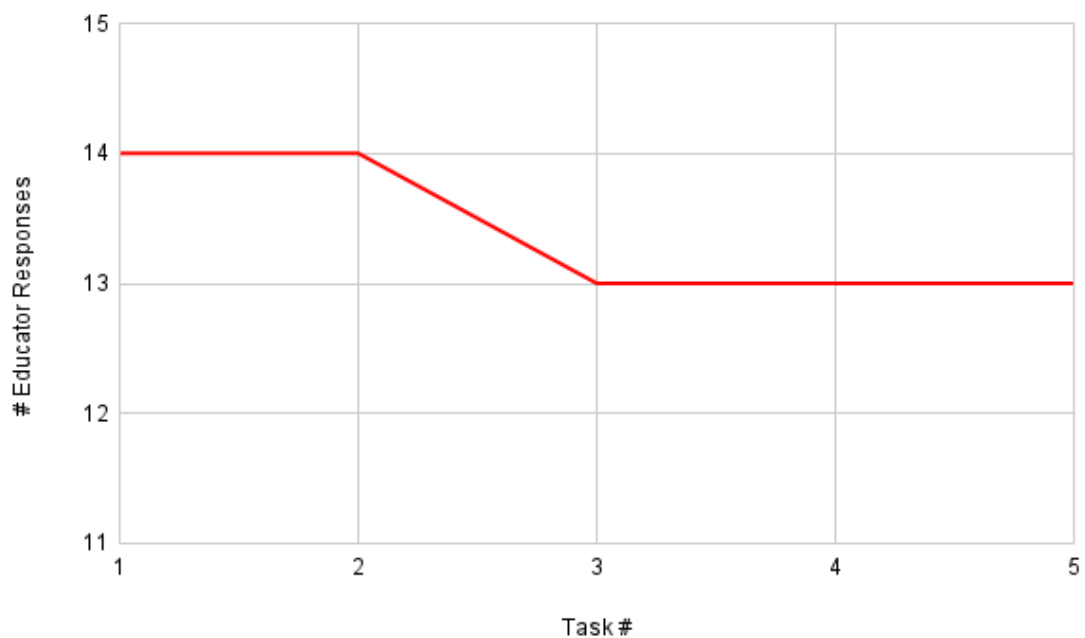


Figure 6.3: Progression of Educator Responses throughout *CABRAL*'s Usability Evaluation's Experimental Phase

While every participant who began the feedback-driven phase completed it, the experimental phase, as represented in Figure 6.3, Figure 6.4 and in Table 6.5, indeed showed an inner drop-off in both participant groups, although undeniably larger in the student group.

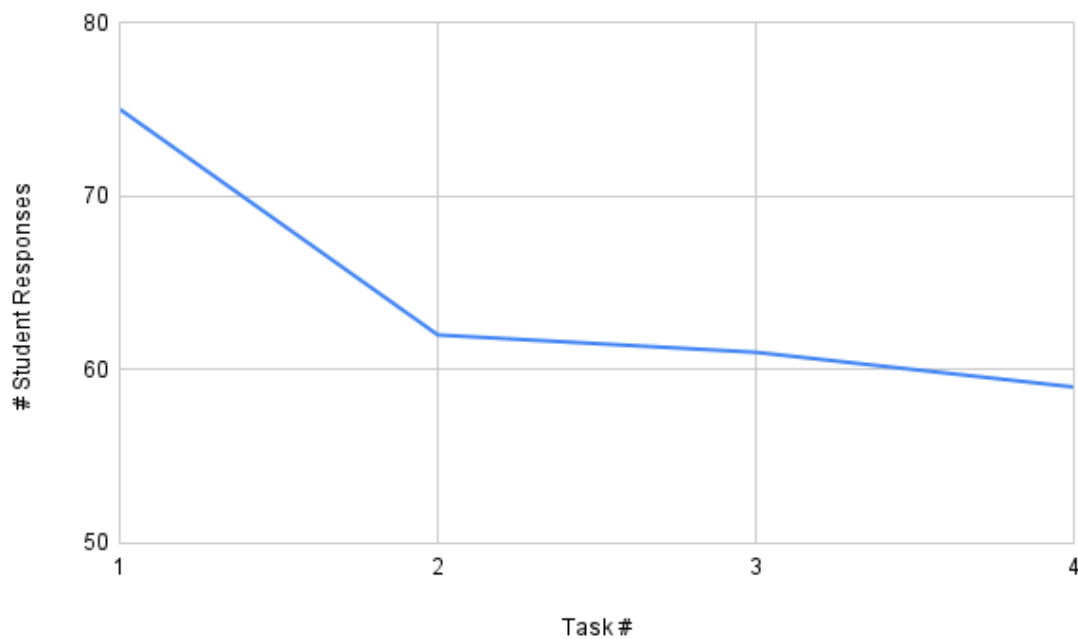


Figure 6.4: Progression of Student Responses throughout *CABRAL*'s Usability Evaluation Experimental Phase

Although it is difficult to determine the exact reasons behind the considerable inner drop-off in the experimental phase of the study, it is plausible that some of the same factors that led to abandonment between phases may still apply in the context of a single phase. For instance, lack of motivation and reward as well as time constraints and academic or personal obligations may have also contributed to early drop-offs, especially in the first task, when participants may have realized the time and effort required to complete the entire set.

Having established the participant sample size for each phase and the number of participants who completed each task in the experimental part, it is now crucial to deeply examine the actual users' performance and success when using *CABRAL*. To this end, three main metrics will be scrutinized:

1. Success rate, that symbolizes the percentage of users that completed the task successfully
2. Misclick rate, that represents the percentage of clicks outside the correct path of task execution
3. Average duration, meaning the average amount of time users took to complete, or give up on a task

Firstly, by analyzing the absolute success rate values throughout the entire experimental phase represented in Table 6.6, several conclusions can be drawn.

Table 6.6: Educator and Student Success Rates throughout each Task

Task #	Educator Success Rate	Student Success Rate
1	78.6	98.7
2	92.9	93.5
3	92.3	98.4
4	84.6	100
5	100	N/A (There is no Task 5 for Students)

Given that, for every task, both student and educator success rates are consistently high, it becomes clear that the participants were generally able to complete the proposed tasks in the high-fidelity prototype consistently, strongly suggesting that users would most likely be able to complete those same tasks within the actual *CABRAL* system. As demonstrated in Fig 6.5, students' success rates remain remarkably stable and high throughout, suggesting a consistently intuitive interaction with the prototype while educators' performance, while similarly strong overall, exhibit a modest upward trend, particularly from Task 1 to Task 2.

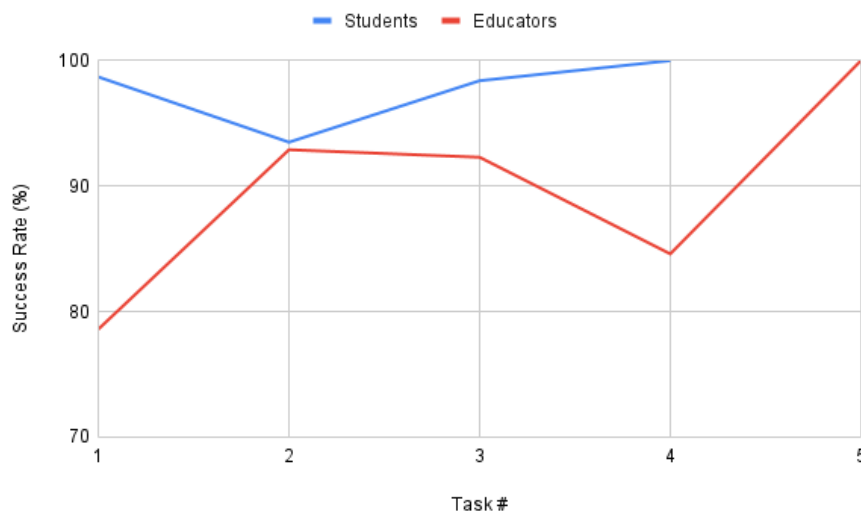


Figure 6.5: Success Rate Progression

However, as the tasks were not randomized for each user, it is difficult to determine whether upward trend in success rates, especially among educators, is due to increasing user familiarity with the interface or to variations in the inherent difficulty of the tasks. As a result, examining the success rate progression alone is not sufficient to isolate the cause behind performance improvement throughout the experiment. Interestingly, even though the two different task sets were designed to have the same level of difficulty to accomplish, students consistently achieved slightly higher success rates for each task, possibly indicating either greater ease with digital tools, fewer hesitations in exploratory interaction or even more intuitive engagement with the prototype.

Furthermore, the stable success rate across tasks for both groups also suggest a consistent and accessible interaction design, with little evidence of escalating difficulty or increasing task

complexity further validated by the lack of abrupt fluctuations or sharp declines across either group as represented in Fig 6.5.

Before scrutinizing the misclick rates and drawing definitive conclusions from the value distributions, it is important to consider that a key limitation of the prototype used in the study stems from it being configured to only allow clicks on predefined execution paths, a constraint clearly explained to users. As a result, not all misclick rates indicate errors or attempts to interact with inactive elements in *CABRAL* and some even might be considered expected, as supported by the hotspot visualization in Fig 6.6.

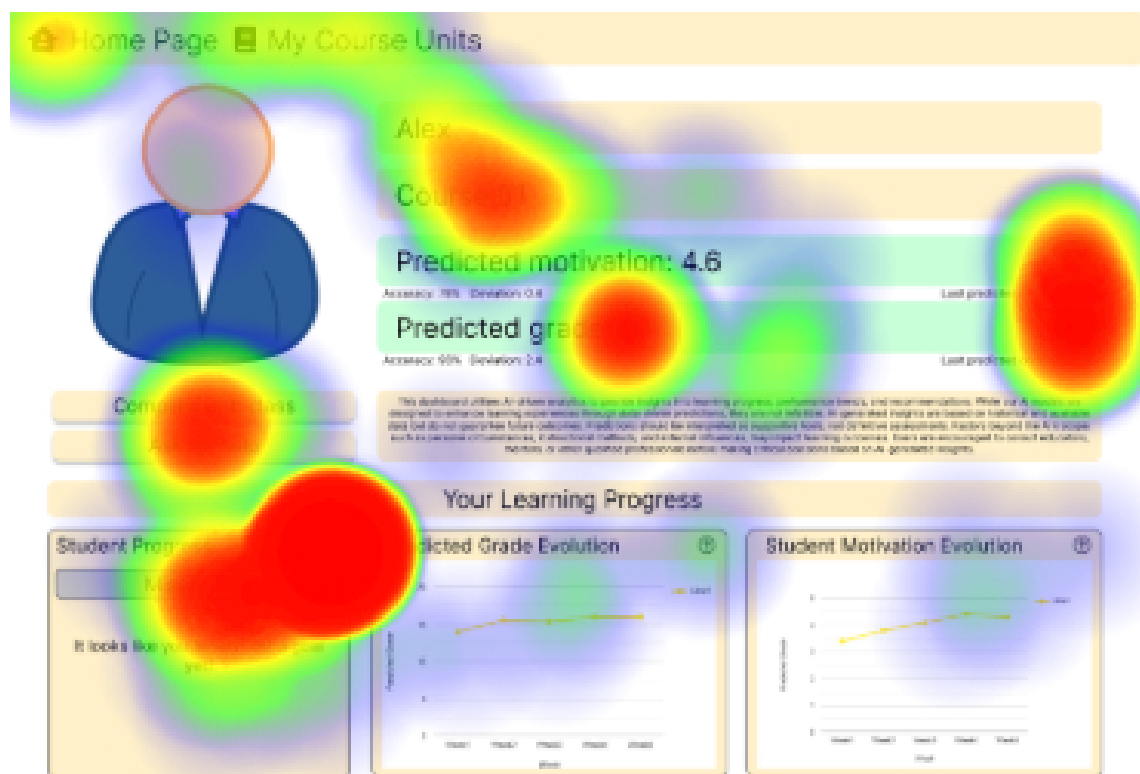


Figure 6.6: Hotspot Visualization of User Clicks on a *CABRAL* Screen

In this specific case, users frequently clicked on areas which, in the fully functional *CABRAL* system, would redirect them to other valid pages or provide alternate routes toward successfully completing the task at hand.

Nonetheless, as these values may yield valuable insights regarding *CABRAL*'s usability, a deeper analysis on the values represented in Table 6.7 remains both relevant and informative.

Table 6.7: Educator and Student Misclick Rates throughout each Task

Task #	Educator Misclick Rate	Student Misclick Rate
1	73.7	78.6
2	42.3	69
3	61	49.9
4	56.7	52.9
5	22	N/A

In contrast to the consistently high success rates observed across all tasks, as seen in Figure 6.7 the misclick rate data reveals not only a larger fluctuation throughout different tasks, but also a higher nuance in user interaction.

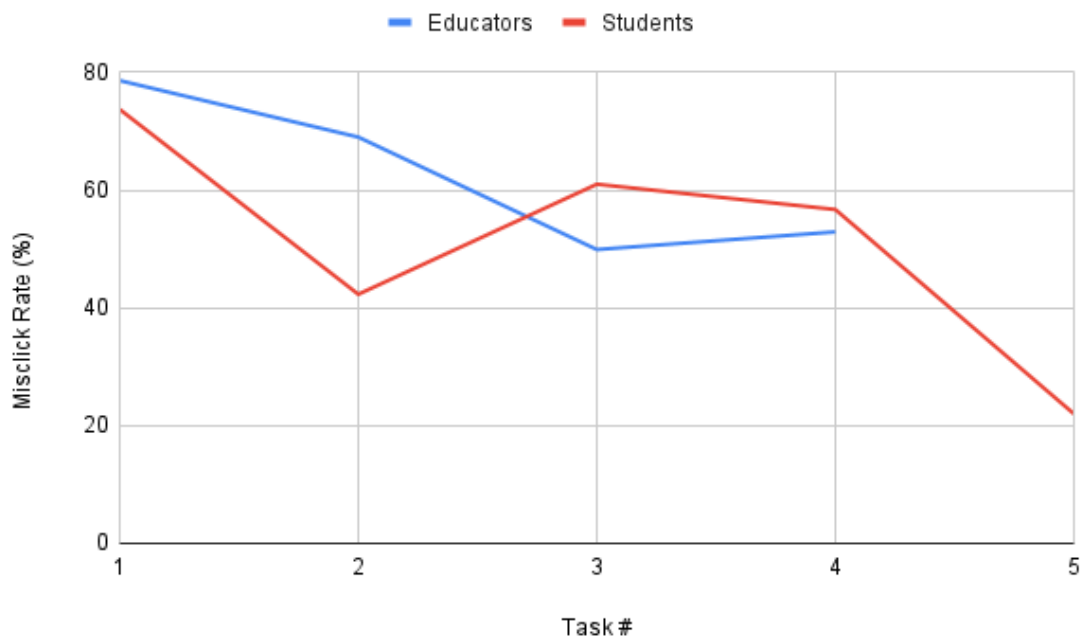


Figure 6.7: Misclick Rate Progression

As expected, Task 1 represents a challenge for participants given the demonstrably high rates in both groups, potentially indicating a lack of immediate clarity in the interface or uncertainty derived from first navigating the prototype. The sharp contrast between the high success rate values and the high misclick rates could suggest that, while users eventually found their way to reach the desired task outcomes, the path was not always straightforward or even intuitive, especially at the outset. As the tasks progressed, however, both educator misclick rates, reaching a value of 22% in the final task, and student rates, in a more erratic manner, trended downward as the study progressed, further implying a possible learning curve or an increased confidence in navigation, amplified by a repetitive use of the interface. These tendencies strongly indicate that, while the system remained usable, certain design elements may have prompted hesitation or

exploratory clicking, particularly during earlier interactions, highlighting areas where affordances or visual hierarchy could be refined to better guide user behavior.

To correctly interpret the average task duration values expressed in Table 6.8, it is important to consider several factors.

Table 6.8: Educator and Student Average Duration for each Task

Task #	Educator Average Duration	Student Average Duration
1	126.9	50.6
2	16.5	31.7
3	29.6	11.4
4	63	29.4
5	11.9	N/A

While tracking the progression of absolute average duration values across tasks is useful and may reflect improvements or challenges in user performance, raw time data alone can be misleading. As to each task requires users to traverse through an unique number of screens, a potential structural bias relative to the number of necessary screens to complete a task might be introduced, since tasks that require more transitions naturally take longer to complete. To mitigate this and better assess efficiency, a derived ratio was calculated between the average task duration and the number of distinct screens needed to complete each task. This adjusted metric, visualized in Figure 6.8, offers a more equitable basis for comparing user performance across tasks with differing structural demands and, upon analysis, it reveals a clearer and more meaningful understanding of user efficiency.

Unsurprisingly, Task 1 shows the highest values for both educators and students, 31.7 and 16.87 respectively, and, after the initial peak, both groups exhibit a marked reduction in the ratio values, suggesting a rapid improvement in efficiency and system fluency further supporting the hypothesis that the initial contact with the system prompted hesitation or unfamiliarity. Notably, while educators consistently display higher ratio values than students across all tasks, this difference is most pronounced in the first task and then narrows progressively, implying either a steeper learning curve for the educator group or, due to a probable lower experience with digital tools, a more cautious deliberate interaction style. Overall, the decreasing trend across both groups reinforces the notion of an intuitive and learnable interaction design, an essential quality when evaluating a tool targeting users with diverse technological expertise levels.

On the other hand, while average duration values provide valuable insights and are reflective of user performance, they often fail to capture the full picture either due to the influence of extreme outliers or limited sample sizes. To address this, dispersion graphs for each of the tasks and participant groups were constructed and are represented in Appendix A. To properly understand the graphs in each figure and the variously colored data points, it is important to clarify the definitions of unfinished tasks as well as direct and indirect task completions, concepts that may initially seem contradictory given that the prototype's flow was designed to allow only direct completion paths.

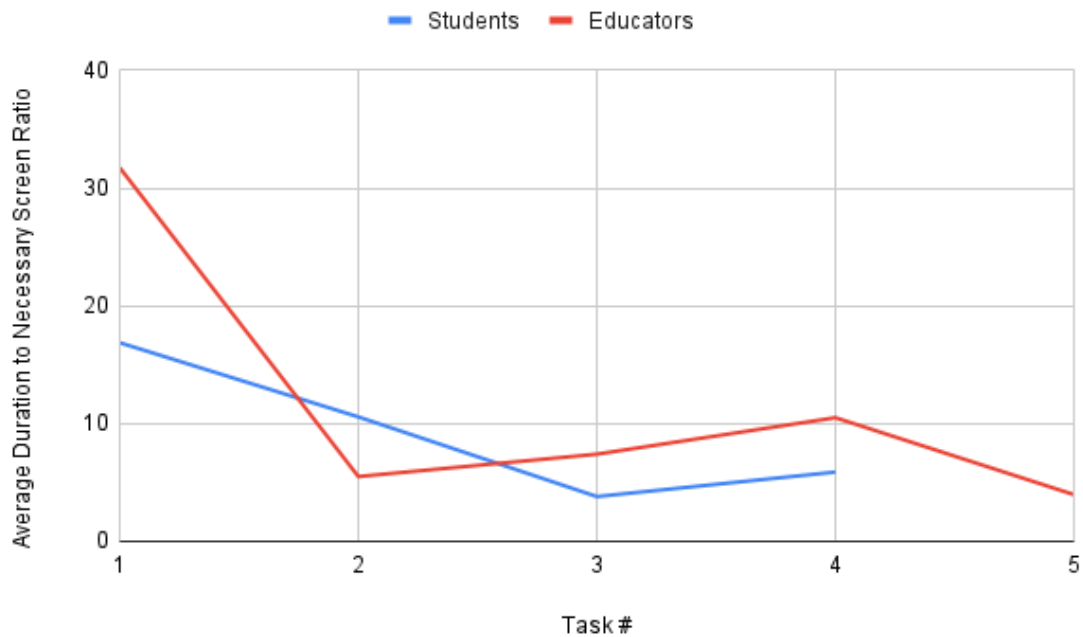


Figure 6.8: Average Duration to Necessary Screen Ratio Progression

Simply put, a task is considered directly completed when a participant follows the expected execution path without any backtracking, traversing each necessary screen only once. Conversely, an indirect completion occurs if the participant visits one or more screens multiple times, indicating backtracking. Tasks are labeled unfinished when participants abandon the task before reaching the final screen but still proceed to the next task in the sequence.

By analyzing the student task durations, it becomes clear that Task 1 exhibits the greatest variability in completion times, an expected outcome given it represents users' initial exposure to the prototype and that, beyond isolated outliers, completion times for Tasks 2, 3, and 4 appear relatively stable, reinforcing the reliability of the collected data and its relevance to real-world student behavior. Interestingly, the unfinished and indirect completion times in Task 2 closely resemble those of direct completions, indicating perhaps that possible backtracking in the execution path may not necessarily reflect user confusion or uncertainty, but could instead be attributed to factors such as unintentional navigation or even user fatigue. As in both Task 1, Task 3 and Task 4 the number of unfinished and indirect completions is either zero or one, these instances are too few to allow definitive conclusions.

On the other hand, the educator task durations reveal a less consistent distribution across all tasks, likely influenced by the varying digital among the surveyed users and the limited sample size. Notably striking outliers include Task 1's durations of 529.33s and 285.78s, more than four and two times the average for the task, respectively, and Task 4's 300.9s, nearly five times the average duration for the task. Interestingly, all these extended durations correspond to direct completions of the task suggesting that prolonged times do not necessarily stem from navigation errors

or task abandonment. Although these values are potentially indicative of moments of uncertainty or limited prototype intuitiveness, Task 1's extreme outlier and its near four-minute separation to the closest iteration points to the influence of external interruptions or user disengagement and should, therefore, be interpreted with caution rather than as a definitive reflection of the system's usability.

Overall, the performance analysis reveals that participants were able to successfully complete tasks with consistently high success rates and a clear trend of improved efficiency over time, as evidenced by decreasing misclick rates and reduced average durations. The initially higher misclick and duration values in early tasks suggest a brief learning curve and some initial uncertainty, particularly among educators, likely due to varying digital proficiency and, despite isolated outliers and technical constraints, the data collectively indicate that the prototype supports an intuitive and learnable interaction experience for users. While no formal benchmark exists, success rates above 90% combined with decreasing misclick trends are generally considered indicative of effective interaction design in prototypes, making users' performance with *CABRAL* highly promising, as a whole.

6.1.1.2 Perception

Even though users' performance metrics are extremely powerful in conveying a system's usability, understanding users' subjective experiences is equally important, especially for a system like *CABRAL*, intendedly designed to be functional and usable in highly stressful educational contexts and appealing to a wide range of users. As a result, the analysis now turns to, once again, three key dimensions:

1. SEQ values, that capture users' perception of task difficulty
2. SUS scores, that reflect the overall system usability
3. General user feedback left by users and other comments that provide qualitative insights into their interaction with the system.

Firstly, to correctly interpret the SEQ values represented in Table 6.9, it is important to comprehend what SEQs actually measure.

Table 6.9: Average SEQ Values for each Task

Task #	Educators	Students
1	2.43	2.16
2	1.57	1.76
3	1.43	1.4
4	1.57	1.76
5	1.43	N/A

In short, SEQ assesses users' perceived difficulty upon completing an individual task by asking them "How difficult did you find Task x?" and retrieving a value on a scale from one to five, where

higher values reflect higher difficulty and vice-versa. Having clarified this definition, it is now possible to analyze this metric's progression.

As noted in the Performance subsection, the fixed order of task execution across all participants makes it difficult to definitively determine whether the steadily decreasing SEQ values represented in Figure 6.9 result from growing familiarity with the prototype or from inherent differences in task complexity.

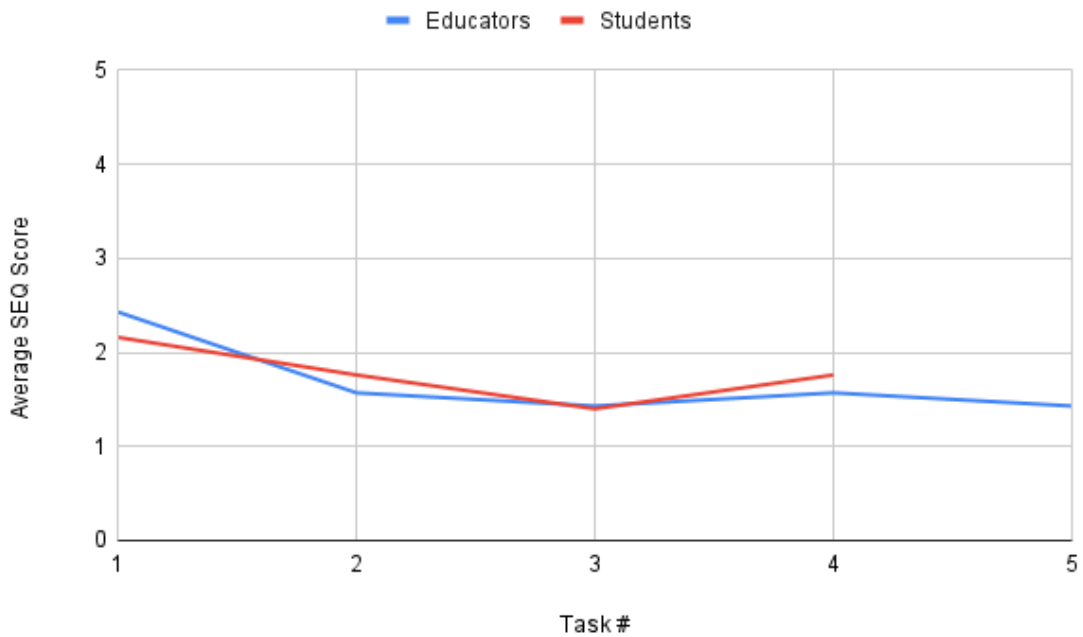


Figure 6.9: Average SEQ Value Progression for each Task

Nevertheless, given that the overall values remain consistently low across both user groups, it can be reasonably inferred that users generally found the tasks easy to complete, an interpretation that aligns with the previously discussed high success rates and reinforces the conclusion that users experienced minimal friction while interacting with the system.

A closer look at the SEQ scores also reveals a slight difference between user groups, with educators generally reporting marginally higher perceived difficulty than students, particularly in the earlier tasks, which further supports the previously postulated theory that differences in baseline digital proficiency or prior exposure to similar systems may influence the smoothness of initial interactions.

The SUS scores ranging from 0 to 100 represented in Figure 6.10, on the other hand, aim to judge the holistic usability of the system, rather than the difficulty of individual tasks and are indicative of how usable the participants found *CABRAL* as a system.

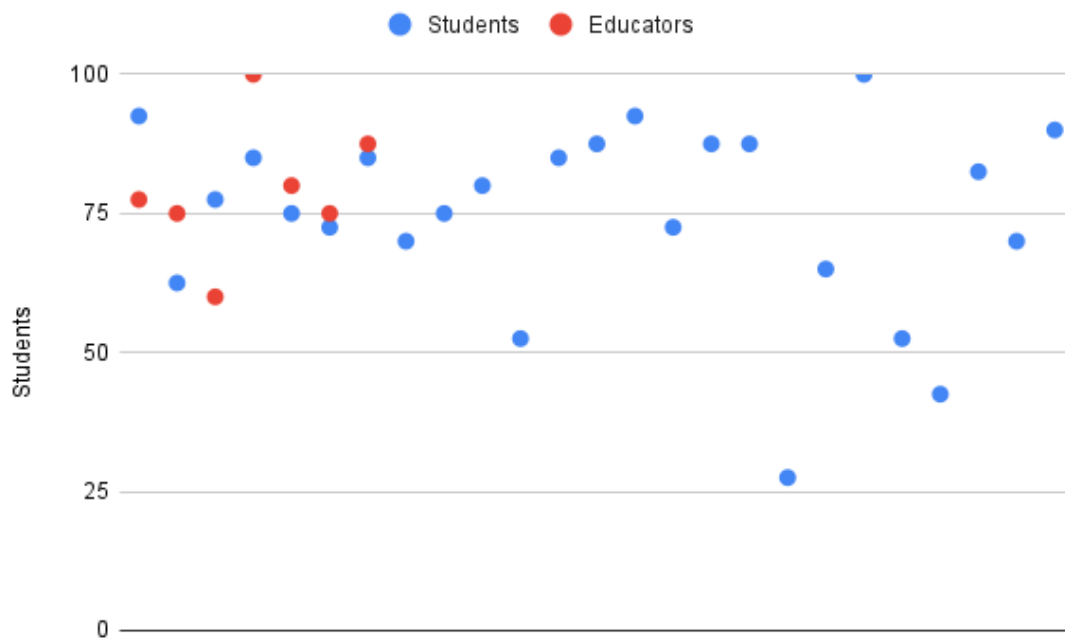


Figure 6.10: SUS Scores Distribution

As seen in the table and figure, both educators and students generally awarded *CABRAL* high SUS scores, with most ratings being above 70, a commonly interpreted threshold as representative of a good usable system, and average scores of 79.29 for educators and 74.8 for students, placing the system roughly in the 85th and 70th percentiles respectively, further highlighting *CABRAL*'s strong perceived usability.

In addition to the generally high SUS scores, it is worth noting some variability within the ratings represented in Figure 6.11 and Figure 6.12.

By analyzing the histograms in the figures it becomes clear that, while most participants rated *CABRAL* above the threshold of 70, indicating good usability, a few outliers, such as an educator score as low as 27.5, suggest individual differences in user experience or isolated difficulties that an experimental part analysis could help clarify. However, due to privacy constraints and the lack of personal data collection, it is impossible to cross-reference task performance with SUS scores to explain these outliers. Surprisingly and seemingly contradicting the performance results discussed, educators provided slightly higher average SUS scores than students, indicating a better perceived usability in the former group.

While no concrete conclusion can be reached regarding the causes for this discrepancy, strong possible factors could be differing expectations or familiarity with similar systems that, in comparison, could influence the users' perception.

Nevertheless, it is important to consider limitations such as the limited number of responses, as well as the unbalanced distribution between user groups which may influence the generalizability of these results.

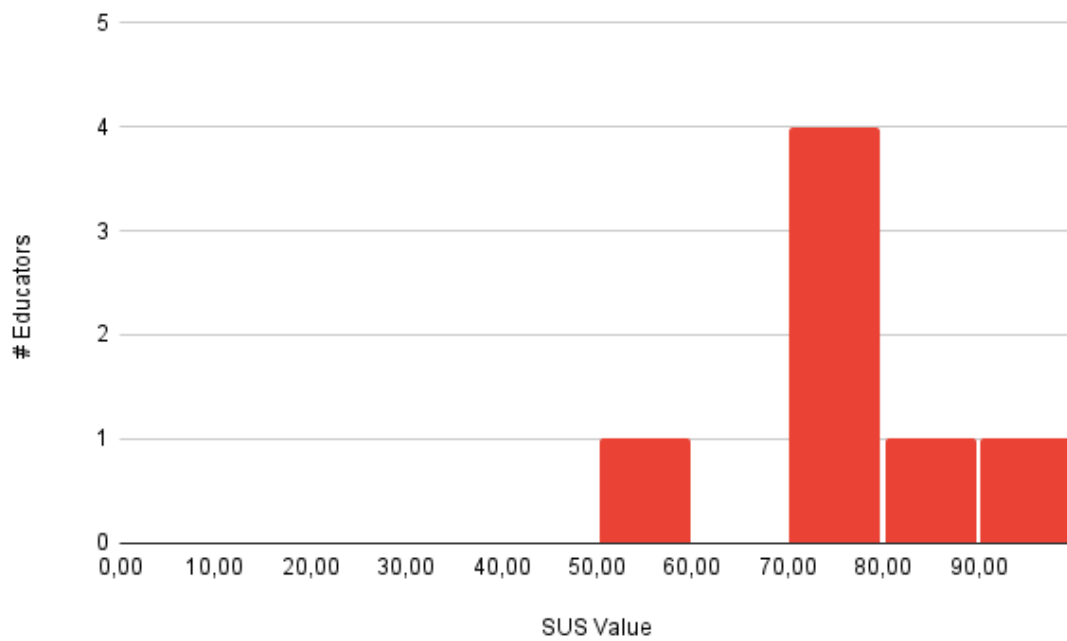


Figure 6.11: Educator SUS Values Histogram

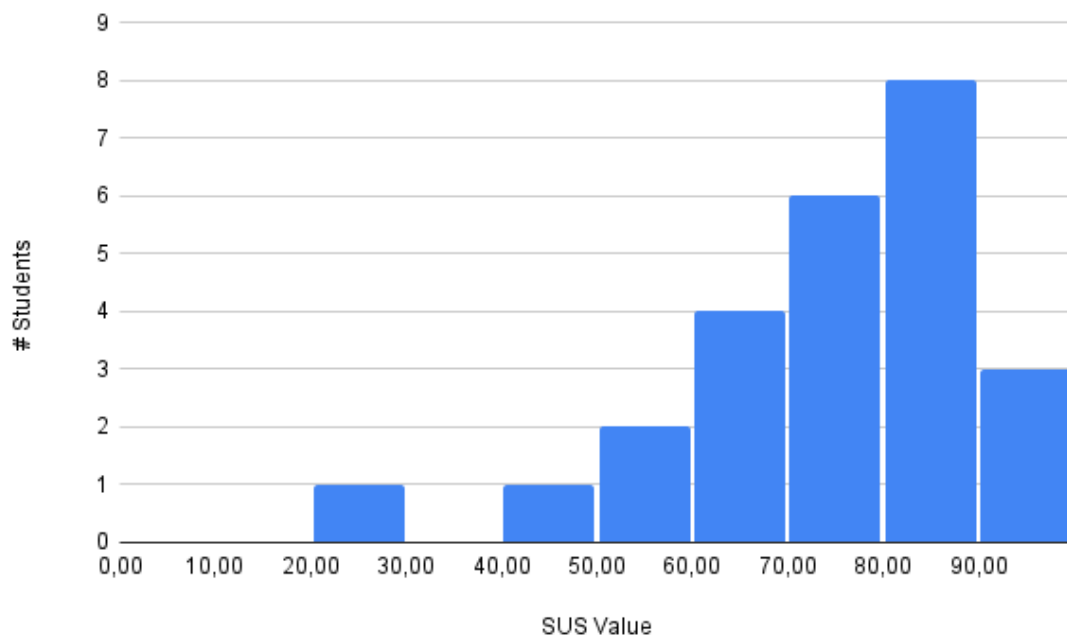


Figure 6.12: Student SUS Values Histogram

When it comes to participants' qualitative remarks about *CABRAL*, it is important to both analyze task-related feedback and general impressions in order to form a broad understanding of users' opinions on the system.

Although educators task-related comments were sparse and limited mostly to the first task, as student feedback is more distributed it is still useful to analyze general participant opinions transcribed from the feedback questionnaires. Even though the entire questionnaire was administered in English, some users chose to respond in Portuguese due to their geographic distribution and nationality. As such, to preserve uniformity and interpretability, all remarks were directly translated to English and only edited for clarity and grammatical purposes.

Regarding the first task, one general concern that appeared among participants was not directly related to its inherent difficulty or even accomplishment of the task, but instead, as previously postulated, to the fact it represented the initial point of contact between users and the system. Some feedback includes: *"Being the first task, I wasn't sure how to proceed, so it took some time to understand the goal of the entire usability test, but once I figured that it was only to click on the supposedly correct buttons, then it all went smoothly from there. You should have a first 'demo' task where you demonstrate a 'successful' task being completed."*; *"I had to navigate forwards and backwards through the interface to find out how to perform the task. Perhaps the fact of being the first interaction with the system made the task harder than it would be afterwards. Anyway, it was not very intuitive."*; *"Not difficult but it being the first task and there being a plot about motivations in the first page my first instinct was to click it before trying to go to a page of the course"*. *"It took me a while to understand what needed to be done."* *"I did not find where to click first"*.

These remarks emphasize a lack of onboarding or guidance at the outset and suggest that a short demo task could have helped users better understand the expected interaction.

Another interesting trend to take into consideration stems from users' perceived clickability of different elements, as two students point out: *"I didn't know what 'predicting your motivation' meant in practice. The predict button was darker so I thought it didn't work."* and *"After selecting the course I thought I needed to click on the pencil or somewhere in the first box below. I thought there were maybe some boxes to check and that would generate the predicted motivation that appeared in the long box at the top. Only then I clicked in the long box and it worked."*. These perceptions, however, point more to the limitations of the prototype than to the full system, where such elements are interactive and responsive and thus should not be interpreted as indicative of the system's usability, but instead of the actual prototype. Another student pointed out their opinion on the interface, saying: *"I didn't really like the interface, I found it a bit cluttered and difficult to understand where things are."*. Although this feedback symbolizes a single opinion, it reflects *CABRAL*'s core design priority of providing a clean and simple interface, making it worth future consideration.

Task 2's student feedback, on the other hand, shifted towards a specific problem in the task, rather than to an holistic view of the prototype or system. As the task pertained to the message generation component of *CABRAL*, some participants expressed several issues with the task ex-

ecution, indicating both confusion towards the need of AI for generating messages and, in some cases, even discomfort towards its usage: *“I just found a bit weird that the message was automatically generated.”*; *“It is not suggested that you need to use AI to help generate the message.”*; *“Did not realize at first that I first needed to generate the message”*; *“I didn’t quite understand the concept of being AI generated, I think it should be something more human”*. These remarks not only highlight shortcomings in task explanation but also suggest a broader concern: how to balance automation with user autonomy and comfort. Other feedback pointed to more concrete interface issues, such as the location of buttons or wording such as: *“I didn’t know where to click first, my gaze had to search the page and it wasn’t oriented.”*, *“The generate message button was not in the same place at the message screen”*; *“A lot of random clicks in the beginning trying to find the right button. I simply didn’t associate “Send a message to your professor” with “Ask for help”. After that it was really simple”*. Interestingly, while some of this feedback reinforced existent design decisions, such as substituting the “Ask for help” button for a “Contact Teacher” button, a choice made previously to the feedback reception. Another opinion reinforced the notion that perceived ease of use improved with familiarity, as one user wrote: *“Now I was more familiar with the system and just needed to find where I would send the message.”*.

Though Task 3 received no relevant comments, Task 4 elicited multiple remarks pointing to the same usability concern: difficulty identifying the input field for goal setting. Comments include: *“I was trying to write the number, and also I didn’t get the place to do so in the beginning, it was very small”*, *“It took me a while to find the place to set the goal”*, *“Struggled in finding the place to set up the goal, could be more visible”*. These beliefs not only point to the need to revise both the size and positioning of this UI element to improve visibility and user flow.

General feedback, on the other hand, reflected by the responses to the question “Could a prototype like this help you or your students in your / their academic life?”, mostly provided a comprehensive view of *CABRAL* and pose a great opportunity to understand the overall strengths and weaknesses of the system as well as its perceived value.

While student opinions regarding *CABRAL*’s potential slightly varied, among educators responses were consistently positive, arguing that a system like the one evaluated shows great promise in helping either their students or their daily work-lives, although acknowledging areas for improvement: *“Definitely. From an educator perspective this is quite useful. Keep improving it and expanding it with more metrics.”*; *“I would have to see it working on a real setting, but it looks like an idea that has some promise.”*; *“It could be useful for me to follow the students’ progress more closely/frequently. Not being a psychologist, I cannot really tell whether it would be effective in helping the students. Probably, it would.”*; *“Possibly yes, but the UI needs much work. This prototype seems rather initial and to be more convincing, it would need much more attention to design. Regardless, if the features are working, the conceptual idea is promising.”*. One educator, reinforcing one of the core principles of the *CABRAL* development phase, underlined the need for transparency in the predicted data: *“I think it could help, although transparency on how the “predicted” data was generated would be critical”*. As neither students nor educators were shown any of the XAI-based features that compose *CABRAL*, an essential principal guiding the devel-

opment of the system, having users mentioning transparency unsolicited is undoubtedly positive and underscores the significance of explainability, as it proves that in the age of technology and AI-based solutions, interpretability and knowledge is crucial when using a system.

Students, on the other hand, mostly believed *CABRAL* had the potential to improve their academic lives, albeit with some students mentioning UI improvements: *“Will be very useful, but it still needs more UI improvements”*; *“Of course. It would help me to keep track the my current and target status.”*; *“I think it would be very helpful”*; *“Yes, I believe it could be useful to track my academic performance”*; *“Yes! I have some difficulty deciding which subject is my priority and organising my studies in general.”*; *“I don’t know, I think it’s something interesting but it needs to be more intuitive and dynamic. It’s something I look for daily, so I would use it.”*. Notably, a distress voiced by some students aligned with one of the initial concerns that guided *CABRAL*’s predictive features’ development phase, namely how wrongful motivation or grade prediction could perhaps hinder actual student motivation: *“I think it could work in favor of it, but at the same time I think that if my predicted motivation for the course was low, I would start the course somewhat discouraged”*; *“Possibly. Didn’t quite understand how the predictions work.”*; *“I believe so, but it might give me a little anxiety.”*. Similarly to XAI, these reactions reaffirm the importance of designing AI systems that are not only effective but also psychologically considerate, an approach that was a central pillar of *CABRAL*’s development.

In summary, the qualitative feedback provided by the surveyed users offered valuable insights that complement the quantitative findings, shedding light on user perceptions, expectations, and emotional responses. While several usability issues emerged, particularly related to initial orientation, element visibility, and terminology, these were often tied to the limitations of the prototype rather than the underlying system, thus not being entirely reflected of *CABRAL*’s innate nature. Participants, especially students, demonstrated increasing ease with use over time, suggesting that familiarity reduces friction and both user groups acknowledged *CABRAL*’s potential to positively impact academic experiences, with educators particularly highlighting its promise for student monitoring and support. Importantly, recurring themes such as transparency, intuitive design, and the psychological impact of predictive features underline the need for thoughtful, user-centered development in educational AI systems.

6.1.1.3 Patterning

To complement the isolated analysis of both performance and perception metrics, exploring potential patterns and correlations between the two dimensions, may help achieve the goal of uncovering deeper insights into the user experience with *CABRAL*. While performance data such as task duration, misclick rate, and success rate offer an objective lens into how users interacted with the prototype, perception-based measures such as SEQ and SUS scores provide a subjective understanding of how users interpreted and felt about these interactions and, by examining how these two domains intersect, it becomes possible to identify areas of alignment, contradiction, or group-specific trends that may not be immediately visible in separate analyses, ultimately offering a more integrated and nuanced view of *CABRAL*’s usability and overall effectiveness.

However, while this analysis is useful and potentially revealing, some limitations that might compromise the results are still worthy of consideration and need the proper acknowledgement. Most notably, as previously discussed, the inability to cross-reference individual user responses given the lack of personal data captured in the questionnaires prevents the formulation of a detailed breakdown regarding how specific users' performance aligned their perceptions of the system. Additionally, the difference in sample sizes from the experimental to the feedback-driven phases might also compromise the interpretability of the uncovered patterns resultant from this analysis as, especially in cases where averages reflect the input of over fifty users in one phase versus fewer than thirty in another.

In addition to these structural constraints, several contextual limitations may also affect the reliability of the uncovered patterns. Firstly, as the analysis was conducted on an interactive prototype rather than the final, fully functional system, certain mismatches between performance and perception, such as high misclick rates coexisting with low SEQ scores, may be artifacts of unresponsive or visually static interface elements that would not persist in the complete version of *CABRAL* and thus should be interpreted carefully. Secondly, while SEQ and SUS are validated instruments, their nature undoubtedly remains subjective and susceptible to external factors unrelated to usability, such as user expectations, mood, or prior familiarity with similar tools, which may weaken their alignment with performance metrics. Lastly, given the absence of real stakes in the usability testing scenario, participants' performance behavior and subsequent perceptions may not fully mirror how they would interact with *CABRAL* in a real educational setting, which introduces uncertainty when interpreting performance-perception alignment.

With these restrictions properly acknowledged, it becomes necessary to define the axes of analysis that will guide this subsection:

1. The relationship between task completion success and perceived task difficulty
2. The alignment between misclick rates and perceived task difficulty
3. The connection between average task duration and perceived task difficulty
4. The link between number of necessary screens and perceived task difficulty
5. The connection between participant remarks and absolute performance metrics

As such, the following analysis explores each of the proposed axes in turn, seeking to understand the interplay between objective performance and subjective perception measures.

To best interpret the potential correlation between participants' success and their perceived difficulty for each task, it is necessary to refer to the scatter chart represented in Figure 6.13.

As the graph reveals, while it is true that the task with the lowest success rate was rated as more difficult, it is impossible to definitively argue that users' success directly influenced their responses to the administered SEQs since the data points across the chart are mostly clustered and offer little variability. This clustering, combined with the generally high success rates across tasks and the limited number of distinct task cases, hinders the ability to draw a strong or generalizable

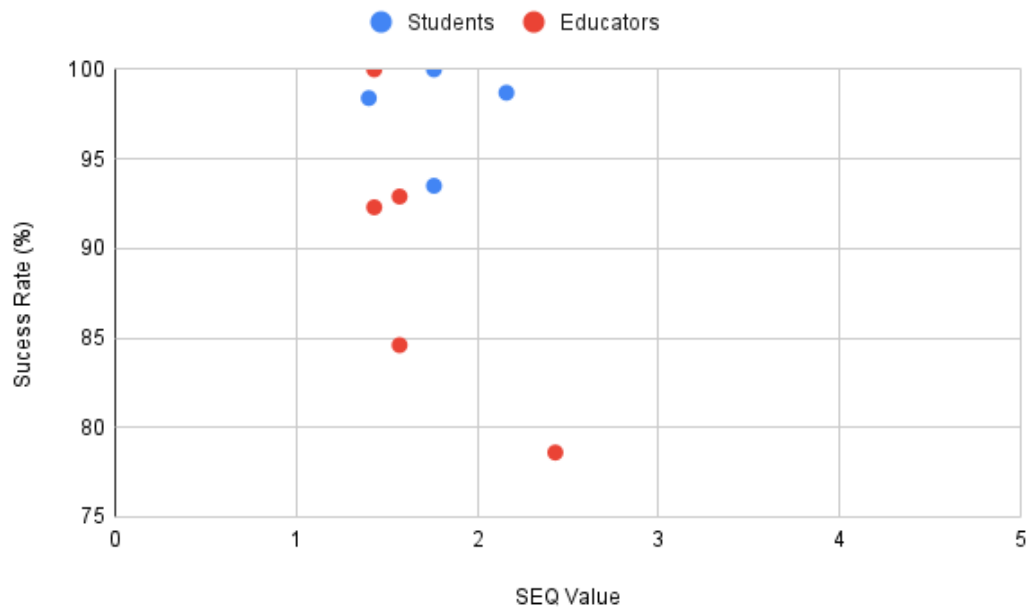


Figure 6.13: Success Rate vs. SEQ

conclusion meaning that, while users found the hardest task the one where they had the least success, this does not translate to the emergence of a robust trend.

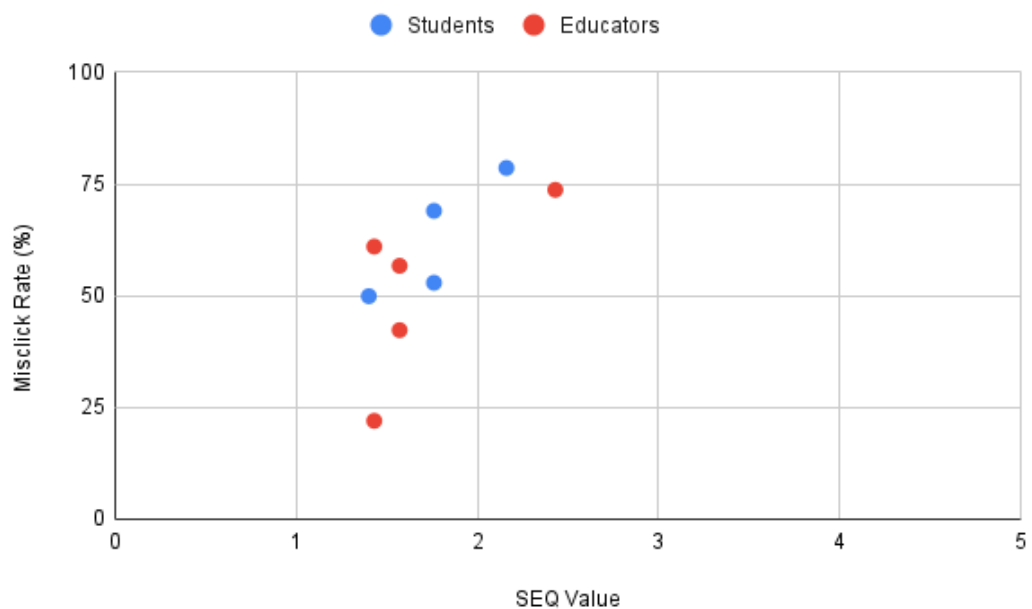


Figure 6.14: Misclick Rate vs. SEQ

On the other hand, the relationship between misclick rates and perceived difficulty suggests a much stronger and clear pattern, as Figure 6.14 supports.

Even without an in-depth reading, the visible upward trend in the graph reveals a near direct proportionality between SEQ values and misclick rates, and reinforces the idea that usability issues reflected in error-prone interactions are directly registered in participants' subjective experience, further supporting the idea that perceived difficulty and error frequency are closely intertwined in usability evaluations.

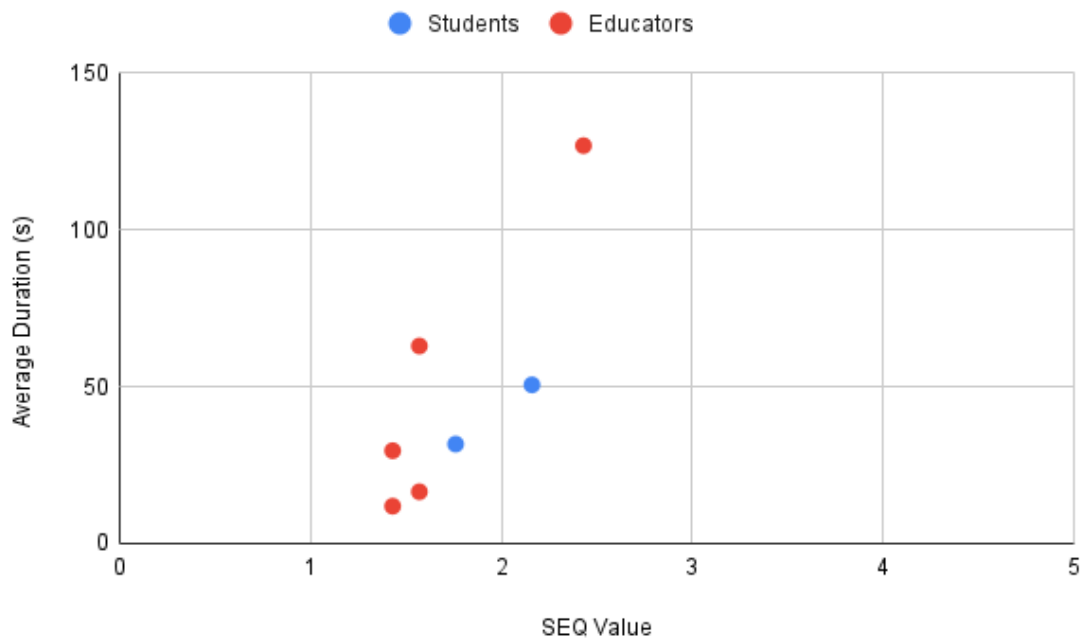


Figure 6.15: Average Task Duration vs. SEQ

Similarly, the results represented in Figure 6.15 also point towards a positive relationship between average task duration and perceived complexity. As longer durations tended to correspond to higher SEQ values, the results indicate that extended interaction time may have been perceived as a sign of complexity or inefficiency, suggesting that performance time played a role in shaping users' perceptions of usability.

Conversely, the results depicted in Figure 6.16 are not indicative of larger patterns. As values narrowly ranged between three and six screens per task, it becomes extremely difficult to conclude whether the task's execution path had any impact on its perception by users.

Lastly, before analyzing the possible relationships between participants' remarks and absolute performance metrics, it is important to clarify the analytical approach. As responses to the "Could a prototype like this help you or your students in your / their academic life?" primarily reflect perceptions on the entire system rather than feedback on specific tasks, only task-related feedback derived from responses to the questions "How difficult did you find task x?" are considered in this analysis, as they more accurately reflect users' perceptions of individual tasks.

Given that these responses often focused on concrete usability concerns or specific interface issues, one possible strategy for uncovering patterns is to correlate the number of comments per

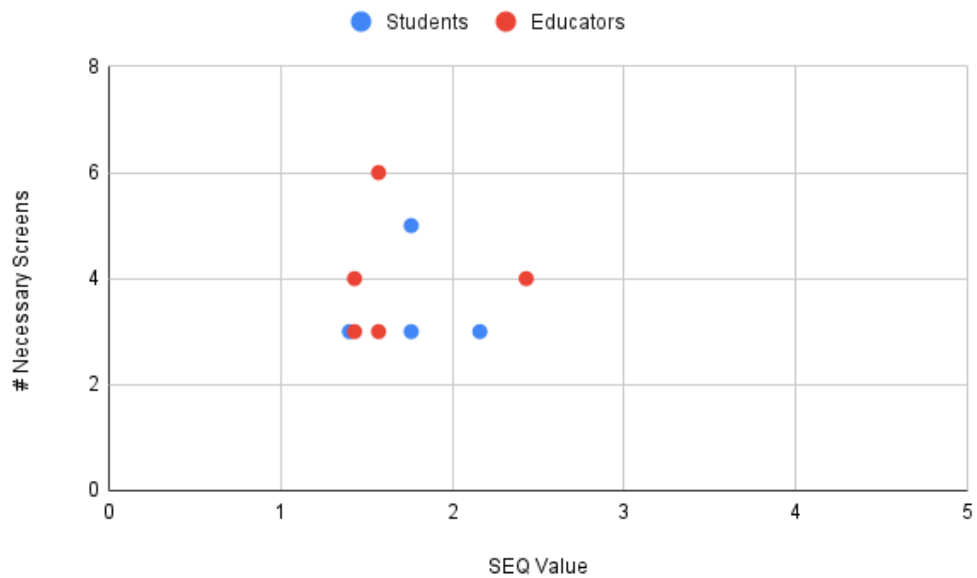


Figure 6.16: Necessary Screens vs. SEQ

task with the three core performance indicators: success rate, misclick rate, and average task duration. In this context, a greater volume of feedback can be interpreted as a proxy for user friction or confusion, enabling the identification of potential links between participant concerns and objective task difficulty.

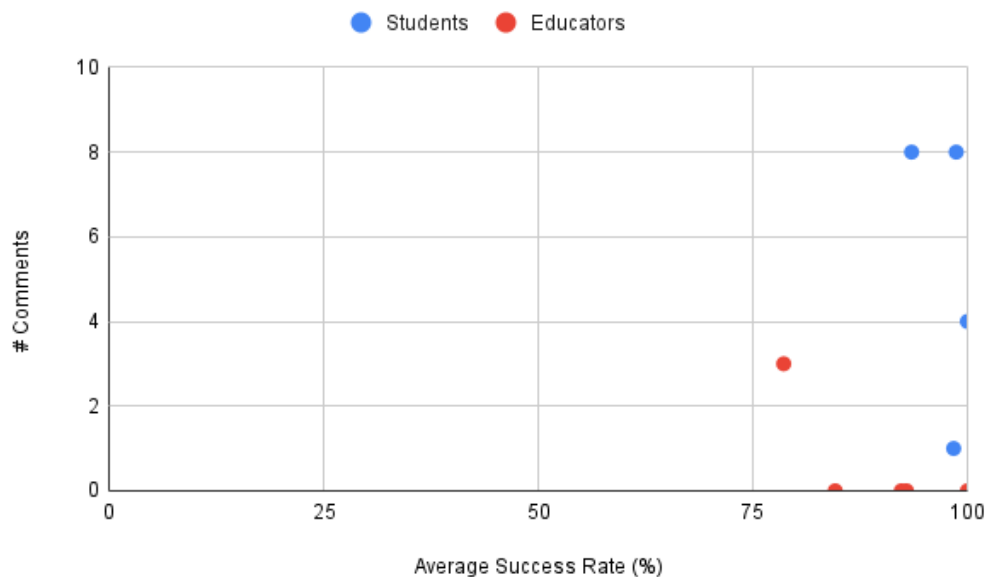


Figure 6.17: # Comments vs. Success Rate

First and foremost, after examining the chart in Figure 6.17, a surprising trend are revealed.

Contrary to the assumption that tasks with lower success rates would elicit more participant concerns, the graph suggests a slight upward trend, indicating that tasks with higher success rates may have actually received more comments. While this inverse relationship is intriguing, the narrow range and clustering of success rates across tasks limit the interpretability of the data and prevent any firm conclusions about whether success directly influenced participants' inclination to express difficulties.

By contrast, the graph in Figure 6.18 appears to show a more interpretable and consistent pattern.

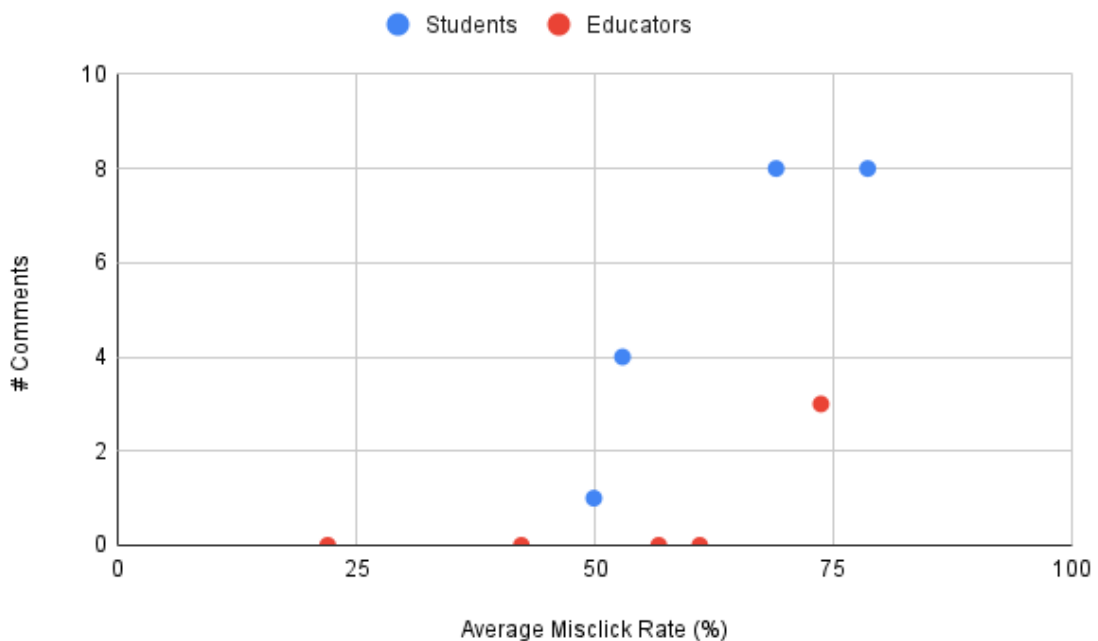


Figure 6.18: # Comments vs. Misclick Rate

Unlike success rate, the average misclick rate seems to exhibit a linear correlation with the number of participant remarks, supporting the notion that increased error-prone interactions prompted more user concerns and reinforcing the idea that misclicks serve as a visible trigger for user frustration or confusion, often encouraging feedback.

Last, but not least, the data represented in Figure 6.19, though somewhat skewed by the outlier of Task 1's significantly longer completion time, still hints at a broader trend.

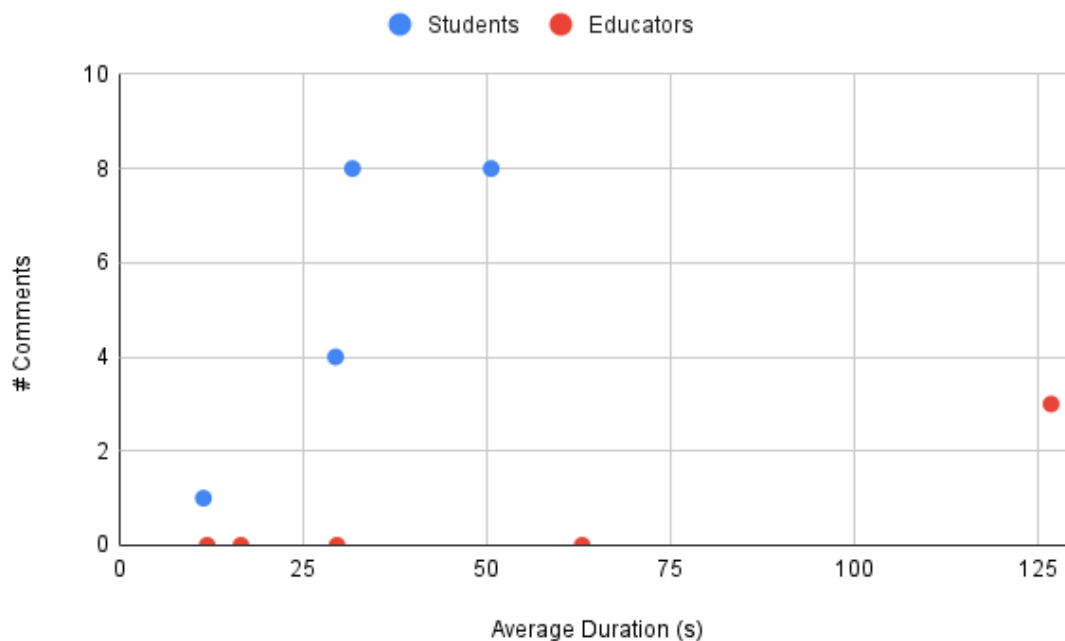


Figure 6.19: # Comments vs. Average Task Duration

When isolating the outlier, it is possible to observe that tasks with longer average durations tended to generate more comments, suggesting that time-on-task may have influenced participants' perception of difficulty and encouraged them to voice concerns.

Taken together, the analysis of cross-metric patterns between objective performance indicators and subjective user perceptions reveals a nuanced but insightful portrait of *CABRAL*'s usability. While some axes, such as success rate, did not present clear or generalizable correlations with perceived task difficulty or user feedback, others offered more compelling evidence of alignment. Notably, the misclick rate and task duration metrics emerged as meaningful predictors of perceived complexity, both correlating with higher SEQ values and a greater number of user remarks. These findings suggest that users are particularly attuned to interaction inefficiencies and interface errors when forming subjective judgments about task difficulty and, although constrained by several methodological and contextual limitations, the patterning analysis contributes valuable perspective by bridging performance and perception data, reinforcing the importance of interpreting usability holistically rather than through isolated metrics.

6.1.2 Predictive Results

To allow for a clearer understanding of how each model's performance is influenced by the unique data and characteristics of individual courses, it is important to analyze the distribution of evaluation metrics on a per-course basis. By breaking down results in this manner, we can observe patterns of variability that emerge across different learning contexts, providing insight into the robustness, generalizability, and contextual sensitivity of each predictive approach. This analysis

not only enables a more nuanced interpretation of the system’s strengths and limitations, but ultimately supports more informed decisions regarding model selection and system deployment in diverse educational scenarios.

To facilitate a clearer interpretation of model performance across the different predictive setups, each course’s evaluation data was compiled into a figure containing four grouped bar chart subplots, with each subplot representing a distinct performance metric. In each bar chart, every group of bars corresponds to a specific model, and within each group, individual bars represent the performance of different model types (simple and smart). This visualization strategy allows for a more accessible and direct comparison between models and configurations, highlighting the impact of feature enrichment on the prediction quality across multiple evaluation criteria.

6.1.2.1 LGP Model Performance Review

As the course with the biggest cohort of all three and with a fairly normal grade disparity, the LGP course evaluation results play a huge role in reflecting how each model performs in largely populated course units, but also in showing the main advantages and disadvantages of employing more nuanced AI models as opposed to simpler ones in these types of educational environments.

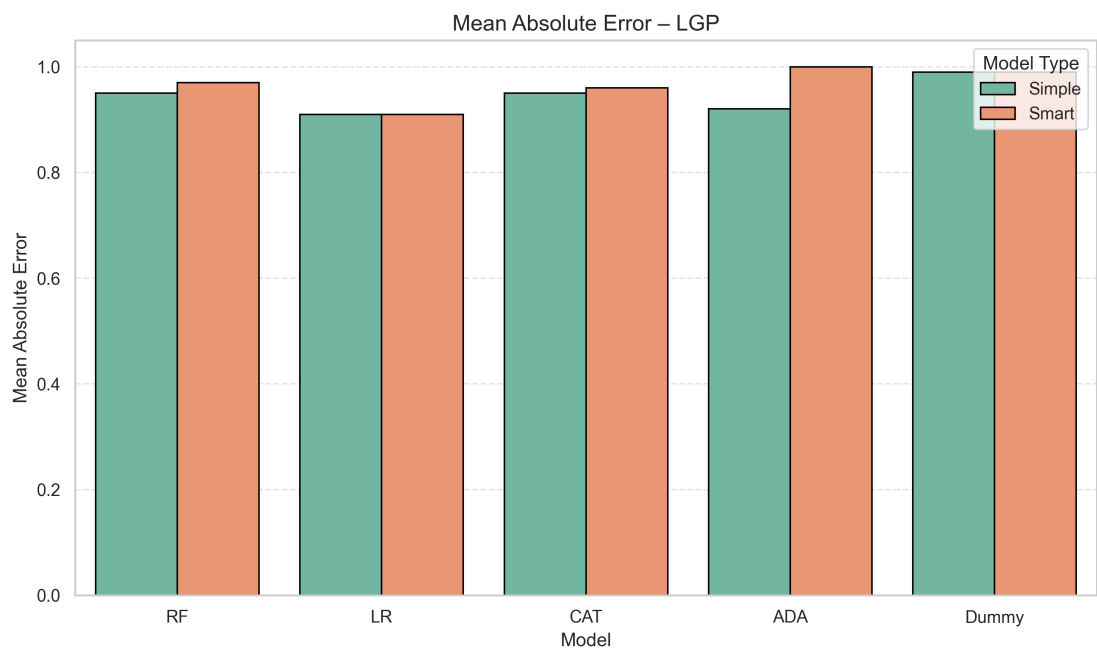


Figure 6.20: LGP - MAE Distribution

When analyzing the MAE values across the different models represented in Figure 6.20, one conclusion becomes easy to reach, the smart models achieve, in general, either the same or higher error rates than their simpler counterparts.

Although the sample size is relatively low to emphatically assume the reasons behind this disparity, it is possible to speculate on some of the possible causes that might have influenced this gap.

A possible explanation stems from the fact that, as smart models rely on enriched features derived from student interaction logs, if these new features are not sufficiently informative or weakly correlated with actual student grades, they can prove to be counterproductive and dilute the model's ability to learn the true signal, thus leading to higher error rates. Additionally, as smart models have a greater number of features and lead to an increased model complexity, they are more prone to overfitting, modeling idiosyncrasies in the training set rather than general learning patterns. It is also important to note that, while the Random Forest, Linear Regression and CatBoost models confirm the theory of simplicity outperforming complexity, this disproportion is particularly glaring when it comes to the AdaBoost model, as the more advanced model fails a substantial amount more than its more basic version and is the only one outperformed by the dummy regressor.

As for the RMSE subplot illustrated in Figure 6.21, the results vary slightly more.

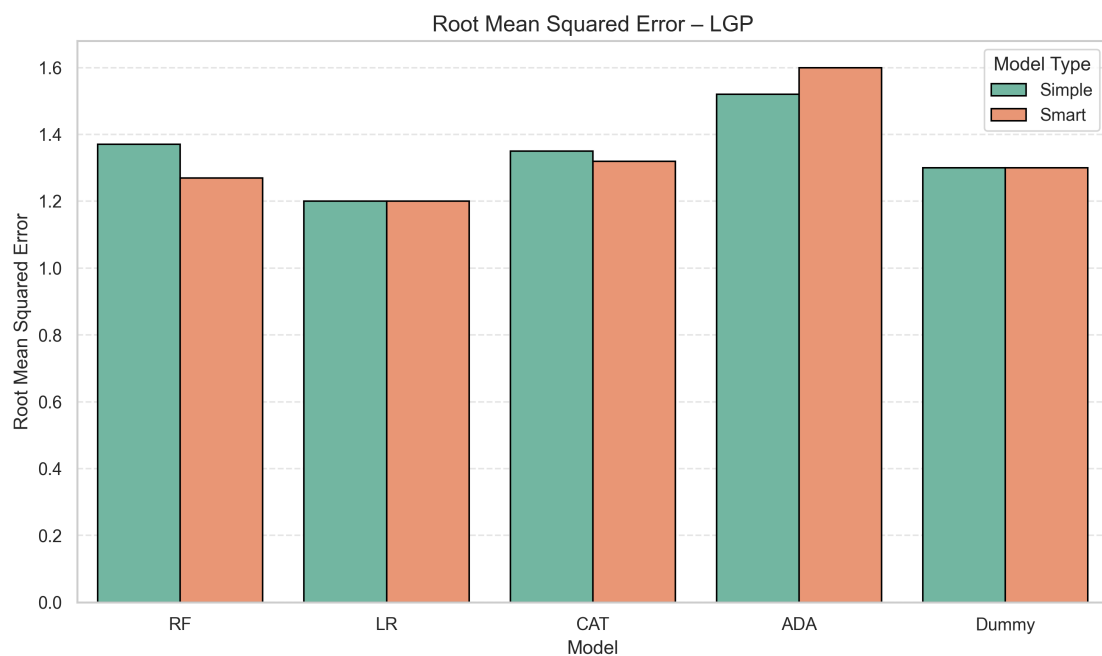
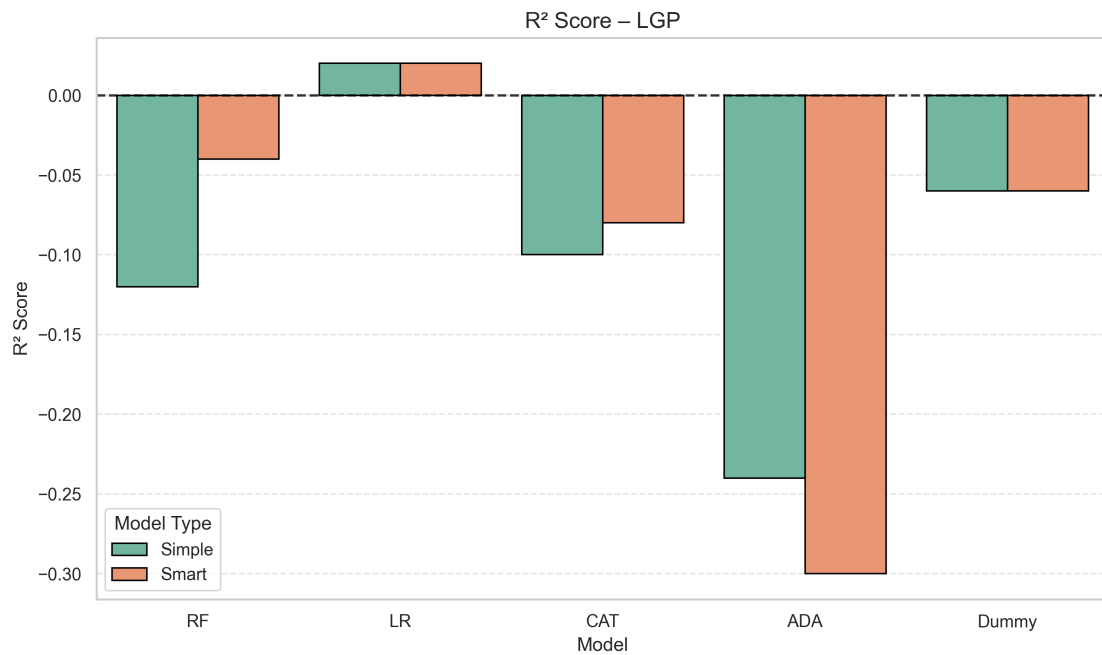


Figure 6.21: LGP - RMSE Distribution

While models like the smart Random Forest, and the smart CatBoost overperform their simpler versions, the AdaBoost algorithm distances itself from the others as it is the only of all four that, once again, heavily reduces its performance when employed in a more refined model environment. When compared to the dummy regressor, all models with the exception of both Linear Regressions and smart Random Forest achieve higher RMSE values, possibly explained by the metric's sensitivity to outliers and the models failure in addressing them.

The R^2 scores represented in Figure 6.22, on the other hand, show unpromising results in nearly all the models assessed.

Figure 6.22: LGP - R^2 Distribution

With the exception of Linear Regression, where both model types achieve similar low positive results, all the evaluated models show a negative coefficient of determination value, thus enabling the conclusion that these models fail to adequately explain the variance in the data and perform worse than a simple mean-based prediction. Nevertheless, although Linear Regression are the only models achieving positive R^2 values, these scores are also extremely low when compared to the theoretical range of 0 to 1, prohibiting this outlier from being immediately interpreted as a fitting choice and forcing it to be analyzed not in context of its counterparts, but of its absolute values.

Last, but not least, the MHSP distribution seen in Figure 6.23 shows that, although slightly variant from model to model, all predictive setups achieve values of over 94.5%, with both Linear Regression models achieving values of over 95%, indicating that all models' predictions strongly resemble the actual grades on a proportional basis.

Curiously, as opposed to all others, only the Linear Regression models show an advantage in employing refined computed features as opposed to LMS-retrieved ones, with simple AdaBoost, once again, severely benefitting from a lower number of simple features over a larger amount of more nuanced variables.

As a whole, it is possible to conclude that, although there are some disparities between models, all predictors achieved low MAE and RMSE values and extremely high MHSP scores and that the smart Linear Regression came up as the top performer with MAE=0.91, RMSE=1.20, R^2 =0.02 and MHSP=95.05%. These values indicate that, although predicting student grades is an immensely complex task to accomplish, there is an unexpected strong linear relationship between engagement metrics and final grades, thus strengthening the theory that an higher LMS engagement leads to an improved learning outcome. Furthermore, the reduced performance gap across the different

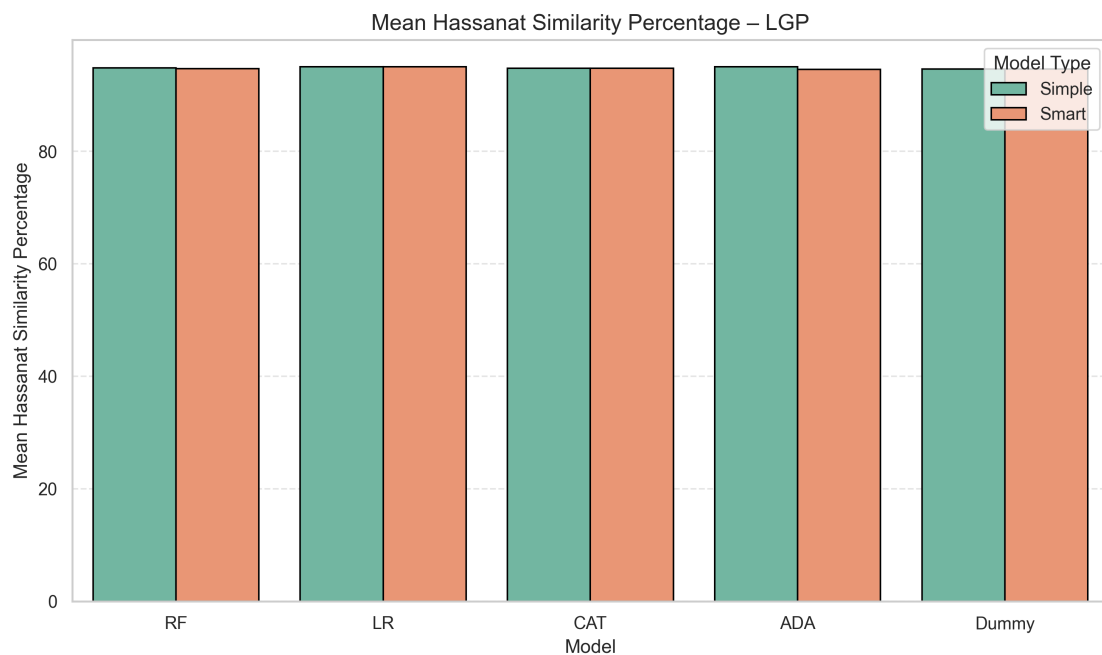


Figure 6.23: LGP - MHSP Distribution

models also suggest that LMS-derived features can provide a sufficient signal for a reliable grade prediction process, especially in high-engagement contexts.

6.1.2.2 PROG Model Performance Review

To better understand the PROG results this subsection aims to dissect, it is important to understand the unique characteristics this course presents.

Firstly, since the course presents a very wide grade disparity between the enrolled students with over 20% of them entirely failing the syllabus and a lower average grade when compared to other courses, it is more than likely that the high error rates in both the R^2 and RMSE metrics stem from reduced predictive model stability. This disparity not only introduces a high degree of variance and heteroscedasticity into the data but, more importantly, makes it difficult for models to generalize effectively. Furthermore, such imbalanced distributions of grade data points often result in skewed data where outliers disproportionately affect model training and evaluation, especially in models sensitive to variance such as regression algorithms.

By analyzing in detail the MAE subplot in Figure 6.24, one conclusion becomes easily drawn, apart from the simple AdaBoost and the smart Linear Regression, all models outperformed the dummy regressor, meaning they provide better predictive accuracy than a naïve baseline and that they are indeed learning some useful patterns in the student engagement data.

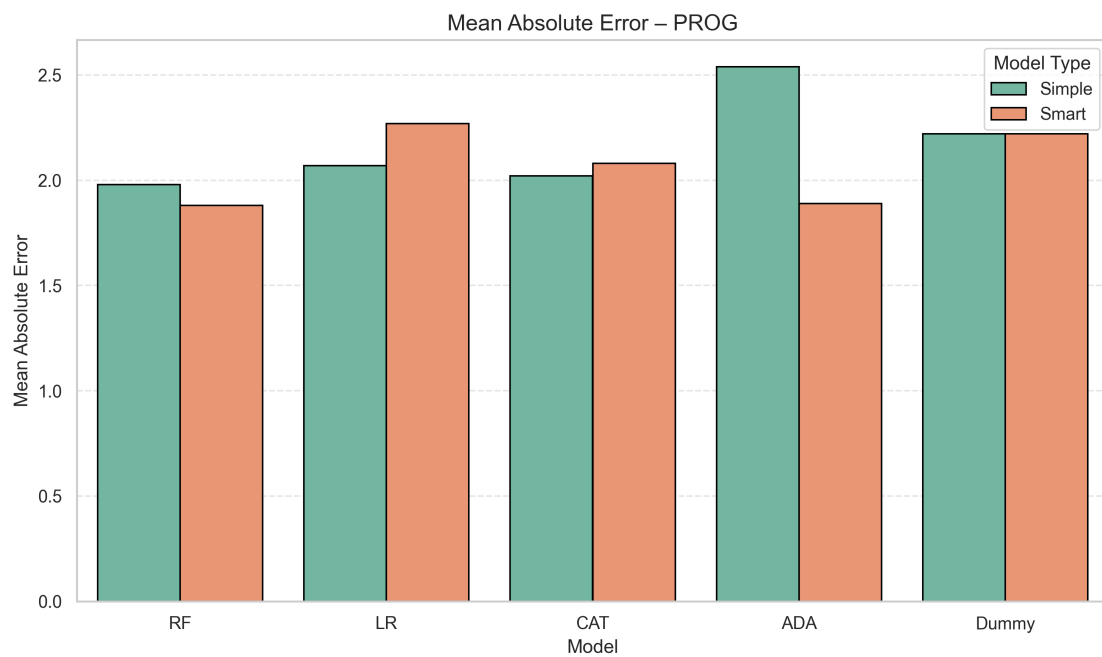


Figure 6.24: PROG - MAE Distribution

Additionally, as in half the smart model outperforms the simple and in the other half the simple one achieves better results, it is impossible to conclude whether or not employing semantically rich features in model training actually enhances the chances of lower MAE rates.

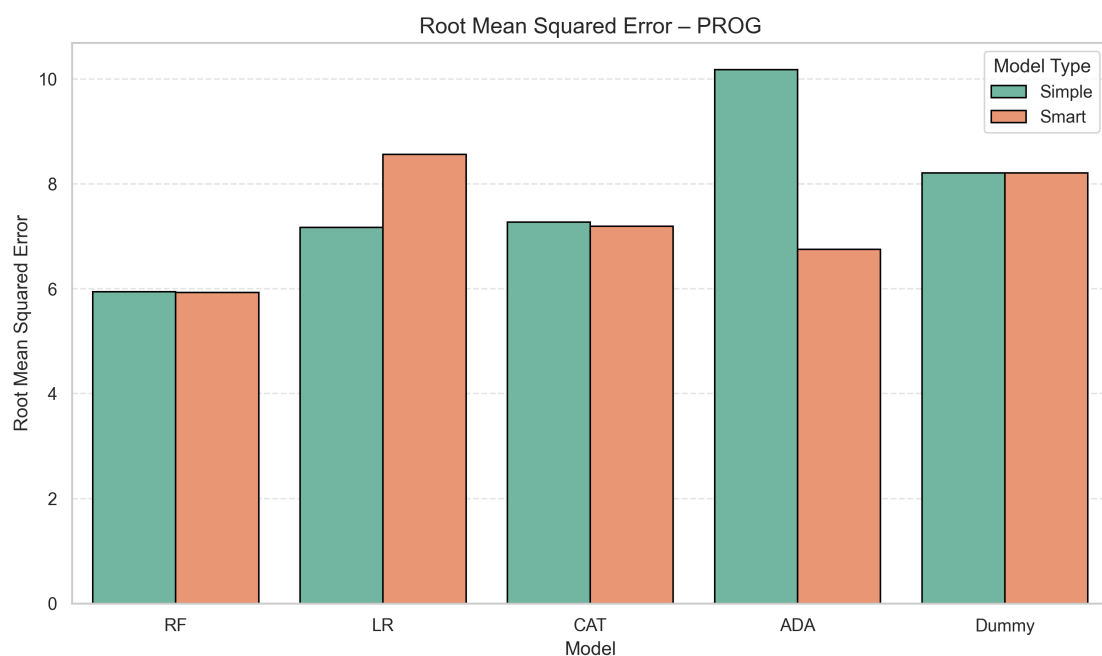


Figure 6.25: PROG - RMSE Distribution

The RMSE subplot represented in Figure 6.25, on one hand, partly sustains similar conclusions

to the MAE one, as, once again, only the simple AdaBoost and the smart Linear Regression achieve a higher error rate than the dummy regressor.

It is, however, also important to notice the smaller disparity between model types in this metric as opposed to MAE, as both the Random Forest and CatBoost model groups present a very small value discrepancy, leading to the inference that, in courses with an imbalanced distribution, both simple and smart models are similarly affected by outliers.



Figure 6.26: PROG - R^2 Distribution

Surprisingly, even with the higher-than-expected error rates in both previous subplots, the distribution of R^2 scores shown in Figure 6.26 reveals more encouraging outcomes, with most models managing to attain positive values suggesting that, to a certain extent, the models are capturing meaningful variance within the data.

However, as was the case with the previous metrics, due to the performance imbalance inside each model group, it becomes substantially difficult to conclude whether or not smart or simple models are the best choice in this type of course since the results vary notably across model families. While in both boosting-based approaches, the refined smart versions outperform their simpler counterparts, in the case of Linear Regression, the simple model achieves better results. Random Forest models, on the other hand, yield nearly identical R^2 values, further reinforcing the idea that no single model type or data strategy consistently dominates in this learning context.

In spite of the R^2 improvement over the previous course, the MHSP distribution represented in Figure 6.27 makes it easier to conclude that, when compared to LGP, every model struggles to generate predictions that closely resemble individual students' actual results, given the models considerably lower results.

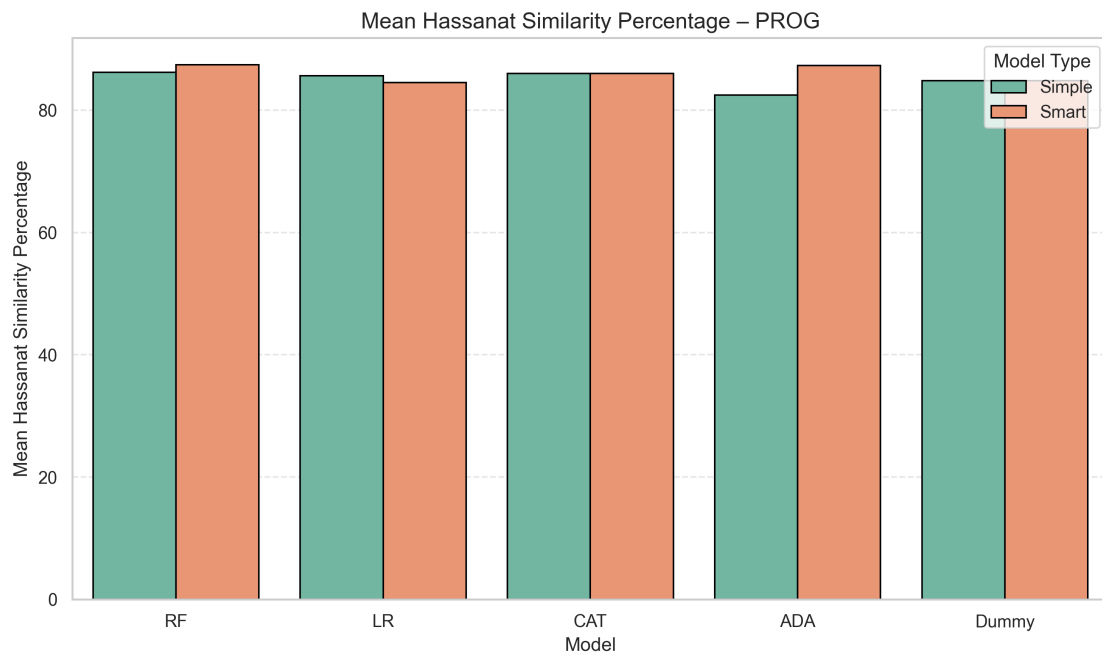


Figure 6.27: PROG - MHSP Distribution

This reduction can be easily explained by the aforementioned grade variety that makes it undeniably more difficult to make informed predictions and the high percentage of failing students, potentially skewing the engagement metrics data and compromising the performance of each model. Nevertheless, while significantly lower than in LGP, every model attained a MHSP value of over 80%, with both the smart versions of Random Forest and AdaBoost achieve values of above 87%. As a result, the models' potential of handling noise and class imbalanced should not be disregarded and should, instead, be considered within the context of the educational environment assessed.

In a macroscopic view of the entire PROG metrics plot, and when comparing it with LGP, a course with a similar number of students, interesting trends emerge. Firstly, the notably lower performance of both Linear Regression models when compared to the previous course could suggest that, in underperforming cohorts, the relationship between student behavior and performance is non-linear, noisy or weakly defined. This conclusion is further supported by the comparison with more flexible models such as the AdaBoost and the Random Forest, which consistently outperform the linear models in these contexts as these ensemble methods, by design, are capable of capturing more intricate patterns and nonlinear dependencies within the feature space, something that Linear Regression, due to its simplicity and reliance on strong assumptions of linearity, fails to achieve.

Secondly, as opposed to LGP, where AdaBoost suffers from complex and numerous model features, in PROG the complete opposite happens, as the smart version severely outperforms the simpler counterpart in every analyzed metric. This reversal might be explained by the fundamentally different data characteristics of the PROG course when compared to LGP, namely its higher variability and wider grade disparity. In this context, the enriched features used by the smart Ad-

aBoost model may provide valuable additional information that helps capture subtle behavioral patterns and interactions, which simpler features fail to represent adequately. As a consequence, the increased complexity and feature richness allow the smart model to better handle noise and outliers, improving its predictive accuracy while in LGP, where the data distribution is more balanced and linear relationships stronger, the additional complexity may lead to overfitting or signal dilution, explaining why simpler models perform better there.

Lastly, the apparent contradiction derived from PROG exhibiting higher error rates than LGP, but better R^2 values can also be fairly explained by understanding what the latter metric actually measures. As R^2 , unlike simpler error metrics like MAE or RMSE that reflect absolute prediction error, quantifies the proportion of variance in the target variable that is explained by the model, when assessing a more diverse course with a wider range of student learning outcomes like PROG, the increased variance in the target variable, meaning the student grades, provides more room for the model to capture some of that dispersion. As a result, even if absolute error values remain high, the predictive models can still achieve relatively strong R^2 values by successfully modeling part of that spread. In contrast, LGP's more homogeneous grade distribution limits the variance that can be explained, making it harder for models to achieve high R^2 scores, even when absolute errors are relatively low.

In summary, the evaluation of predictive models results for the PROG course can reveal key insights into how model complexity and feature richness interact with highly variable learning environments. First and foremost, the undeniable success of ensemble models, such as the smart AdaBoost configuration and the Random Forest models, demonstrates the value of nuanced, interaction-based features in capturing the diverse behaviors and outcomes of students in underperforming or imbalanced cohorts. Additionally, while in fairly homogeneous courses simpler models thrive due to more uniform grade distributions, PROG benefits from flexible, non-linear models capable of handling noise and heterogeneity in the data, in spite of the actual methods used. Despite higher absolute error rates, the relatively strong R^2 values indicate that these models are indeed able to explain meaningful variance, supporting their use in contexts where student performance is more unpredictable. The top performing configuration across all evaluated setups was the smart Random Forest model, with MAE=1.88, RMSE=5.93 and $R^2=0.23$, MHSP=87.46%. These results reinforce the notion that, in courses with a disproportionately distributed grade spectrum, committee-based approaches are particularly well-suited for predicting student grades, as they offer greater resilience to noise and variability in learner behavior. Ultimately, the PROG results underscore the importance of aligning modeling strategies with the unique structural characteristics of each course to maximize predictive reliability and relevance.

6.1.2.3 SCI Model Performance Review

The SCI results, however, must be analyzed in a different light.

As a cohort of only 28 students, an high average grade and no failing students, this course shows little room for comparison with the two previous analyzed courses since neither its size nor its grade dispersion is even remotely comparable to either LGP or PROG.

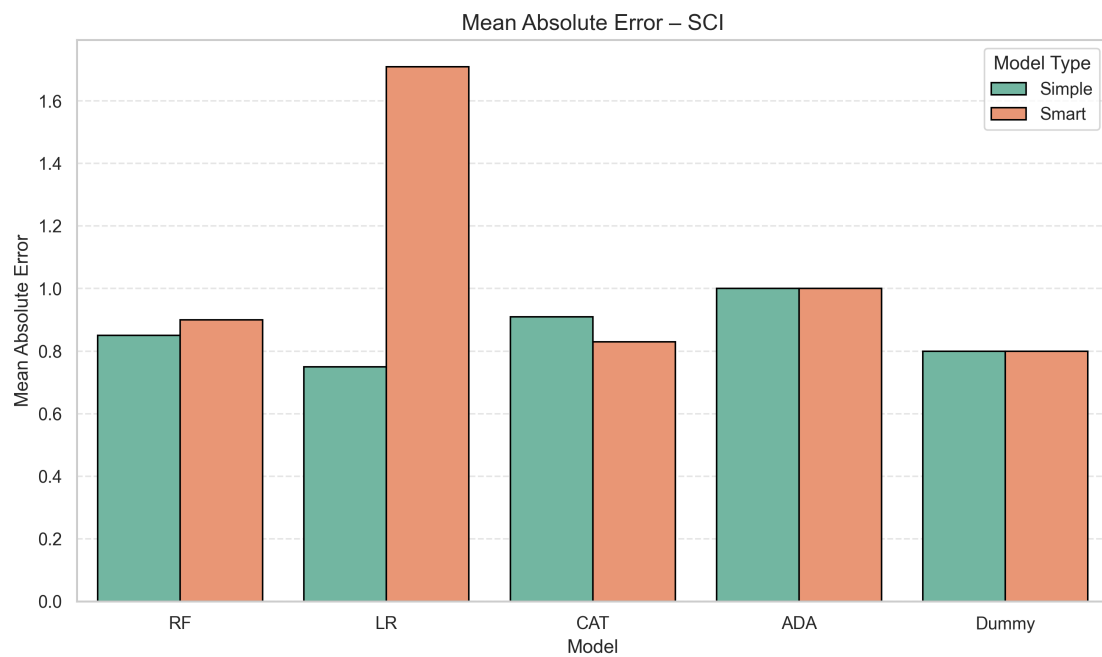


Figure 6.28: SCI - MAE Distribution

After a deeper analysis on the MAE and RMSE subplot in Figure 6.28 and Figure 6.29, respectively, two notions become evident.

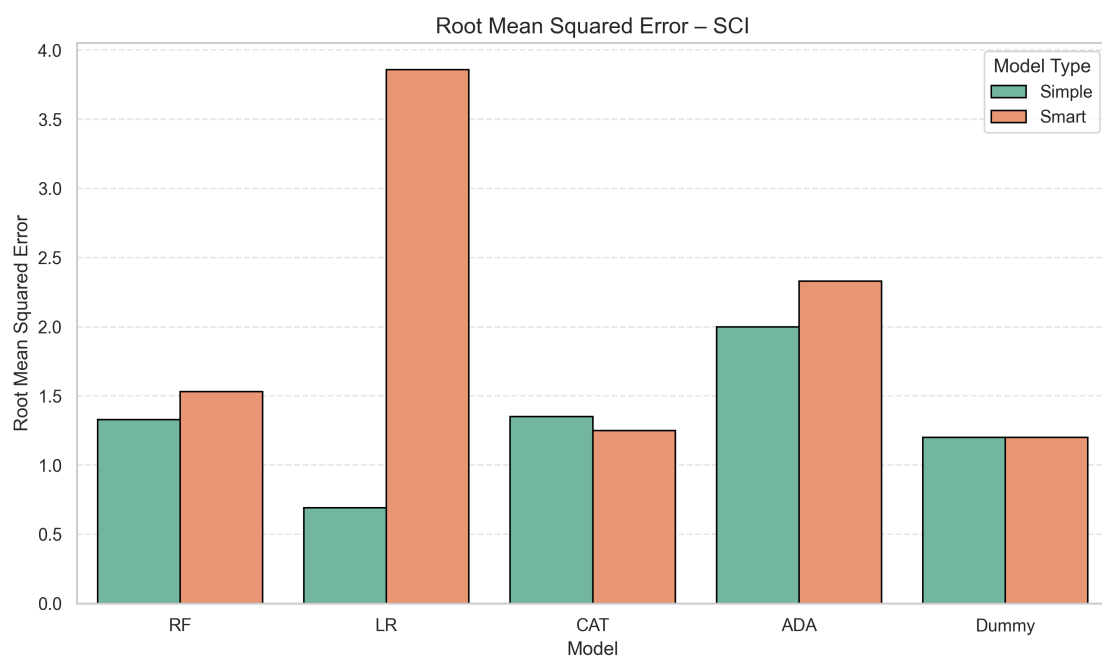
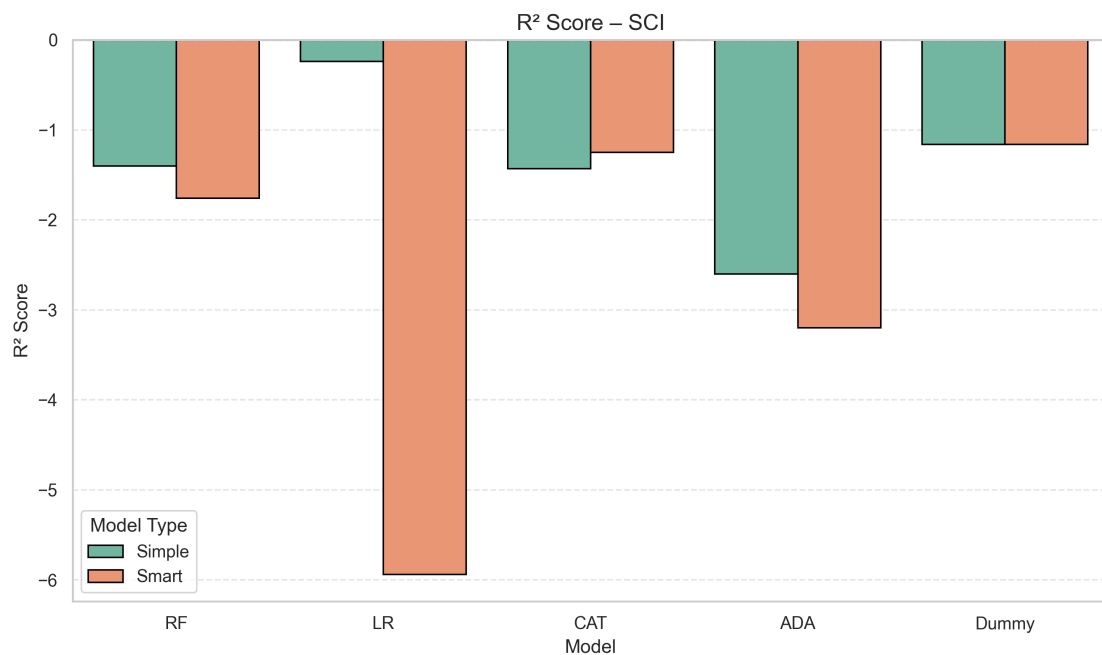


Figure 6.29: SCI - RMSE Distribution

Firstly, it is impossible to infer what model type is prone to less error as the results across all

Figure 6.30: SCI - R^2 Distribution

model sets are dissonant. Secondly, and more importantly, with the exception of the simple Linear Regression, all models attain higher error rates than the dummy regressor, which shows that these models fail to capture meaningful patterns in the data and perform worse than a naive baseline that simply predicts the average outcome. These results suggest only poor generalization capabilities, possibly due to overfitting, lack of informative features, or low variance in the target variable that reduces the learning opportunity for more complex models. The exception observed in the simple Linear Regression model, which managed to outperform the dummy regressor across all three courses, warrants closer examination. Unlike more complex models, as Linear Regression relies on a direct linear relationship between the features and the target variable, it is less prone to overfitting in small or noisy datasets, an advantageous simplicity when the underlying structure of the data is not overly complex or when the feature set contains weak or moderately predictive variables. In the context of this analysis, it's likely that Linear Regression benefitted from the relatively small feature space and the linear tendencies present in the data, allowing it to capture basic trends without being misled by noise.

The consistently negative R^2 scores illustrated in Figure 6.30, on the other hand, imply that all models have a limited explanatory power when facing a course with a low number of students and fairly homogeneous final grades, sustaining the notion that generalizing from a limited and uniform dataset is a challenging task, often leading to poor model performance regardless of complexity.

Additionally, the MHSP value distribution represented in Figure 6.31 shows a partly similar, conclusion, as, once again, with the exception of the simple Linear Regression, all models attain lower values than the dummy regressor.



Figure 6.31: SCI - MHSP Distribution

In spite of this comparison, these scores are surprisingly positive, as all models, except smart Linear Regression, achieve results of over 94.5%, showing a strong alignment between predicted grades and actual student grades with the best performing model setup in the course being simple Linear Regression with MAE=0.75, RMSE=0.69, $R^2=-0.24$ and MHSP=95.77%.

Overall, despite being trained on a small dataset with little to no final grade imbalance, while difficult to understand whether or not more refined models are better options rather than simpler ones, the SCI model performance results show surprising promises, especially when it comes to the MHSP metric, further illustrating that, even in homogeneous and well-performing classes, predictive models can still produce sustained plausible predictions, even with limited generalizability.

However, as understanding how ML models perform in different educational environments requires more than just raw performance metrics, it becomes central to analyze the key patterns and insights derived from comparing the predictive outcomes across the three distinct course types.

6.1.3 Results Interpretation

6.1.3.1 Model and Feature Complexity vs. Data Structure

Firstly, even with a relatively limited volume of evaluation data, it becomes undeniably clear that each model's performance greatly depends on the structure of the dataset, and that there is a consistent relationship between data structure and model complexity. As a result, a critical axis of analysis lies in the effect of feature enrichment, made clear by the distinction between smart models that utilize engineered behavior-based features and simple models that solely rely on raw LMS-derived variables.

In the LGP course, which features a large and relatively uniform distribution of student grades, simpler models such as basic Linear Regression consistently outperform their more complex, feature-rich counterparts, proving that the stable, high-engagement context allows linear models to effectively capture the primary signal in the data without being misled by noise or unnecessary feature complexity. In this course context, added model sophistication appears to introduce overfitting rather than improved generalization, particularly evident in the poor performance of advanced smart models like AdaBoost and Random Forest. Most notably, in the case of AdaBoost, the enriched feature-set not only fails in comparison to the simpler counterpart, but also to the baseline dummy regressor, suggesting that in well-structured, low-noise datasets, additional complexity may dilute meaningful patterns leading to either overfitting or misalignment between features and target outcomes. In the specific case of LGP, the predictive strength of the simple raw LMS-derived engagement data seems to already be sufficient to make an informed prediction, proving that further enrichment not always translates into a better model performance.

In contrast, the second course, characterized by high grade variability, a significant failure rate, and substantial noise in student engagement patterns demonstrates the value of model flexibility. Having this particular dataset, PROG benefits from more complex and flexible models like smart ensemble approaches such as Random Forest and AdaBoost, given their unique ability to capture nonlinear patterns and manage variability across disperse student outcomes. As the smart versions of the predictive models almost always outperformed their simpler configurations, it is inferable that, in this specific educational context, the engineered behavioral features may be able to better reflect subtle engagement dynamics that correlate more closely with student learning outcomes than raw LMS metrics alone, suggesting that enriching feature sets within flexible models may be a productive strategy when dealing with outlier-heavy datasets and irregular data patterns.

Meanwhile, in the SCI course setting, simple models once again prevail, as complexity tends to exacerbate instability. Due to the course's limited cohort size and extremely narrow grade variance, the value of feature enrichment significantly diminishes as most smart models not only fail to improve upon their simple versions, but perform substantially worse, most likely due to overfitting on a small, stable dataset. This performance drop leads to the conclusion that the simplicity of learning environment favors models that do not rely heavily on nuanced or high-dimensional input spaces, suggesting that, as a whole, feature enrichment is only necessary when the data demands it.

In conclusion, these trends reinforce a core principle: model complexity should mirror data complexity. When the available educational data is structured, stable, and low in variance, simpler predictive setups are not only sufficient, but actually superior to more complex configurations, whereas in more fragmented or imbalanced learning contexts, greater model flexibility and expressive power become crucial to achieving meaningful and sustained predictions.

6.1.3.2 Metric Interpretability

Another interesting analysis that should be performed over the evaluation results lies in each metric's interpretability and actual contextual relevance in shaping meaningful conclusions about

model performance. While absolute error metrics such as MAE and RMSE offer a direct understanding of how close predicted results are from actual student grades, their effectiveness in conveying meaning heavily depends on the underlying dataset characteristics. For example, as in the case of LGP grades are relatively consistent and errors are smaller, both metrics give a reliable sense of model precision, whereas in PROG, the presence of extreme outliers and significant grade disparity inflates RMSE disproportionately, making it less informative when used alone and reflective of a noisy, less robust signal than MAE.

The R^2 score, however, introduces a different kind of complexity as unlike MAE and RMSE, which measure the scale of errors, this metric measures the proportion of variance explained by the model. This effectively means that, in courses that exhibit PROG-like high variability in student performance, models can, in spite of high error rates, achieve moderately positive R^2 values, since there is more variance to explain. On the other hand, in more uniform courses such as LGP and SCI, the opposite happens, as even small absolute error rates may result in low or even negative R^2 scores, making the metric appear less favorable. These results further support the conclusion that, alike other metrics, the coefficient of determination should be interpreted within the context of the dataset's grade distribution and engagement imbalance, and not used in isolation to assess model quality.

Lastly, the MHSP offers a complementary perspective to the other measures, since it captures the proportional alignment between predicted and actual grades. Furthermore, its consistently high values across all models and courses suggest that the different models are often correctly capturing the relative positioning of student outcomes, even if other metrics signal poor performance, or if the absolute predicted values are off. These characteristics make MHSP especially valuable in educational contexts where rank order and student outcome proportionality may be more pedagogically relevant than raw point accuracy in formative assessments, student progression tracking, or identifying relative learning gains across diverse cohorts.

6.1.3.3 Model-Specific Behavior Trends

As opposed to broader discussions of model complexity and metric interpretability, closely analyzing individual model behaviors across different courses or, in other words, each model's personality when facing different educational data structures, may help reveal distinct and consistent performance patterns that would otherwise go unnoticed.

To this end, a table was constructed for each model group, summarizing the behavior of both simple and smart versions across the three course datasets, along with key takeaways highlighting the observed performance patterns. The Linear Regression results, represented in Table 6.10, are remarkably stable and consistent, especially in datasets with either low variance or reduced sample sizes.

In the particular cases of LGP and SCI, the simple version of the model achieves surprisingly strong results across nearly all metrics, outperforming many of its more complex counterparts. However, in the PROG course, defined by an irregular and imbalanced dataset, the struggle of Linear Regression becomes evident, as the smart variant notably underperforms with a worsened

Table 6.10: Linear Regression Performance across Datasets

Course	Version	MAE	R ²	MHSP	Observed Behavior and Take-away
LGP	Simple	0.91	0.02	95.03	Performs best; stable due to consistent grades. Simplicity suffices in structured data.
LGP	Smart	0.91	0.02	95.05	No gain over simple; enrichment has no impact.
PROG	Simple	2.07	0.07	85.61	Performs poorly; struggles with noise and imbalance.
PROG	Smart	2.27	-0.11	84.52	Smart version underperforms; sensitive to noise.
SCI	Simple	0.75	-0.24	95.77	Performs very well; low variance favors linearity.
SCI	Smart	1.71	-5.94	91.26	Degrades severely; overfits small cohort.

MAE, an heavily inflated RMSE and a negative R^2 . This substantial drop leads to the conclusion that, in noisy and highly variable environments, the model's assumptions break down, and adding engineered complex features only amplifies this weakness. Altogether, while Linear Regression proves to be highly effective in clean, structured datasets, it is less suited for data with non-linearities or noise, making it less reliable when predicting student learning outcomes in complex educational environments.

Random Forest, on the other hand, displays a strong adaptability to dataset complexity and variability, as shown in Table 6.11.

Table 6.11: Random Forest Performance across Datasets

Course	Version	MAE	R ²	MHSP	Observed Behavior and Take-away
LGP	Simple	0.95	-0.12	94.80	Underperforms linear models; complexity unnecessary in uniform data.
LGP	Smart	0.97	-0.04	94.70	Slightly worse; no benefit from enrichment.
PROG	Simple	1.98	0.23	86.21	Decent performance; benefits from ensemble learning.
PROG	Smart	1.88	0.23	87.46	Best results here; handles complexity and noise well.
SCI	Simple	0.85	-1.40	95.32	Poor generalization; overfits small data.
SCI	Smart	0.90	-1.76	95.06	Slightly worse; instability increases.

Specifically in the PROG environment, the Random Forest configurations show a particularly strong performance, with the smart setup achieving the best overall MAE, a competitive RMSE

and the highest R^2 among all assessed models, further demonstrating its capacity to manage noise and imbalanced data in an effective manner. The LGP and SCI datasets, however, provide an environment where the Random Forest models fails to thrive. While the LGP results are decent, with slightly higher error rates than Linear Regression, its complexity does not translate into significant gains, highlighting a case where the data's simplicity does not warrant an overly-complex model. Alternatively, in SCI, the smart version actually worsens across all metrics when compared to the simpler counterpart, suggesting the possibility of overfitting in the small, homogenous sample. Overall, Random Forest's capabilities shine in challenging, unstructured datasets but may be unnecessarily complex or even slightly counterproductive when dealing with regular datasets.

Conversely, the results represented in Table 6.12 prove that AdaBoost configurations are extremely sensitive to context, with performance swinging dramatically depending on the dataset.

Table 6.12: AdaBoost Performance across Datasets

Course	Version	MAE	R^2	MHSP	Observed Behavior and Take-away
LGP	Simple	0.92	-0.24	95.00	Fragile; underperforms despite consistent data.
LGP	Smart	1.00	-0.30	94.55	Degrades significantly; overfitting likely.
PROG	Simple	2.54	-0.32	82.47	Very weak; smart version needed for complexity.
PROG	Smart	1.89	0.12	87.32	Strong improvement; flexibility handles outliers well.
SCI	Simple	1.00	-2.60	94.64	Performs worst; instability in small data.
SCI	Smart	1.00	-3.20	94.69	No gain from enrichment; remains fragile.

In PROG, for example, the smart version of the model delivers substantial improvements over its simpler version cutting MAE by over 25% and achieving the second-highest MHSP in the dataset, proving AdaBoost's strength in aggregating weak learners to model complex patterns, particularly in irregular and disperse datasets. On the other hand, in both LGP and SCI, the results are particularly disappointing. In the case of LGP, the simple version of the model attains worse values in both RMSE and R^2 than a dummy regressor, pointing to either overfitting or misalignment with the data's structure. Additionally, in SCI both variants deliver some of the worst results in the entire model set, with both highly negative R^2 scores and an inflated RMSE, indicating that iterative reweighting and boosting mechanisms may be too reactive to minor variations in already stable datasets and better applied in complex, messy environments.

CatBoost, unlike other models, presents an uniquely nuanced picture, with performance closely tied to the quality and relevance of the enriched features applied in the dataset.

As the results in Table 6.13 portray, both models versions perform quite similarly across different datasets, exhibiting relatively stable values in both the MAE and MHSP metrics.

Table 6.13: CatBoost Performance across Datasets

Course	Version	MAE	R ²	MHSP	Observed Behavior and Take-away
LGP	Simple	0.95	-0.10	94.78	No edge over simpler models; complexity unnecessary. Very marginal change; feature enrichment not helpful.
LGP	Smart	0.96	-0.08	94.75	
PROG	Simple	2.02	0.06	86.03	Moderate performance; potential underutilization. Nearly identical; feature quality likely insufficient.
PROG	Smart	2.08	0.07	86.03	
SCI	Simple	0.91	-1.43	94.98	Unstable; struggles with small, uniform data. Minor improvement; still underperforms.
SCI	Smart	0.83	-1.25	95.45	

In LGP, while the smart counterpart improves slightly on RMSE, it offers marginal gains overall, further suggesting that the added feature complexity neither helps, nor harms the model's performance in a significant way. This indifference towards engineered features is likely due to CatBoost's inherent algorithm capacity to handle both categorical and numerical variables well, even with absolutely no feature enrichment. Moreover, in the case of PROG, CatBoost once again benefits modestly from the smart variant, particularly in R^2 , improving from 0.06 to 0.07, reinforcing the notion that feature complexity in this case offers marginal performance improvement, if any. This trend is further supported by the SCI dataset, where the smart configuration slightly improves in both the RMSE and MHSP metrics, possibly hinting that carefully engineered features can offer slight resilience in small sample environments. Taken together, the findings indicate that CatBoost's performance is closely tied to dataset characteristics and the quality of feature enrichment, resulting in stable but unremarkable outcomes that lack the volatility observed in more reactive models like AdaBoost.

6.1.3.4 Educational Strategy Implications

As highlighted in the previous subsections, different models and model types thrive under distinct educational conditions, solidifying the idea that the structure and characteristics of educational datasets demand tailored predictive modeling approaches that are closely aligned with the surrounding pedagogical context. Consequently, to make better informed decisions about which predictive setup to employ in a given learning environment, it becomes essential to closely examine the strengths and limitations of model behavior across each course since, by examining how different models perform within each course's unique structure, it becomes possible to identify not just which models work best, but why they succeed or fail in specific contexts, allowing for more pedagogically aligned decisions, and ultimately ensuring that modeling efforts support meaningful and context-aware educational interventions.

In the case of SCI, a highly homogeneous environment characterized by consistently high and tightly clustered student performance, simple models such as Linear Regression often outperform their more complex counterparts. In these scenarios, raw error metrics such as RMSE might lose interpretability given the uniformity of the dataset suggesting that alternative evaluation methods like MHSP may offer greater pedagogical insight, as they reflect relative performance and proportional accuracy.

Conversely, noisy, imbalanced distributions and wider student performance gaps in courses like PROG strongly benefit from the flexibility of smart ensemble methods. As committee-based models like Random Forest and AdaBoost have a predisposition to capture more complex, non-linear relationships, these configurations are able to highlight both under and over-performers in a more effective way, making these setups particularly suitable for context with either irregular engagement patterns or outlier-driven grade behavior, in contrast to simpler models. In these cases, combining both R^2 and MHSP helps provide a fuller picture of model utility, particularly helping educators differentiate between at-risk students and successful ones.

Finally, balanced courses like LGP lie in between both ends of the spectrum, often performing well with both simple models and modestly enriched ones, with both MAE and MHSP values offering a reliable interpretive power.

6.1.4 General Findings

Overall, both the Usability and Predictive Evaluation procedures enabled a comprehensive assessment of *CABRAL* as a complex educational system, highlighting its strengths, limitations, and areas for refinement.

To begin, the focus will be placed on the usability-related findings, which provide a clear picture of how users navigated, interpreted, and responded to the system during the evaluation.

Firstly, the consistently above 90% success rates across tasks for both participant groups not only indicate that the system is highly usable, but also that task completion is intuitive and. While educators showed a slightly lower initial success score, there is a clear improvement trend throughout the study, likely due to growing familiarity with the system.

Furthermore, as early tasks exhibited higher misclick rates and longer durations, particularly Task 1, *CABRAL* presents a brief learning curve, further supported by the subsequent tasks' decreasing values in both measures, suggesting that users' adapted quickly and that *CABRAL* became more efficient with use.

Additionally, the low SEQ values throughout the system are indicative of a low perceived task difficulty and the slight discrepancy between students and educators' reports may reflect higher digital fluency in the former group.

The average SUS scores are also extremely promising, as they place *CABRAL* in the 70th and the 85th percentile for students and educators, respectively. The fact that most participants rated the system above the usability benchmark of 70 indicates positive perceptions of system-wide usability, further solidifying *CABRAL*'s value as a user-oriented tool.

On the other hand, the strong correlations found between misclick rates and SEQ scores as well as average task duration and SEQ implies that participants' subjective difficulty ratings were likely driven by visible interaction issues such as errors or prolonged task executions, proving that overly complex interfaces can be cause of confusion and overall frustration with a system.

Qualitative feedback reported by users further reinforced these observations, as users highlighted initial orientation issues with the procedures, including uncertainty about the system's goals or clickable areas, suggesting the need for a tutorial or a demo task. UI suggestions pointed towards clearer button labeling, improved field visibility and reduced visual clutter, with some psychological concerns being raised by students regarding automated motivation predictions, emphasizing the need for emotionally sensitive design choices.

Despite early UI confusion and comments about unclear instructions or button placement, *CABRAL* demonstrated rapid user adaptation, supporting the idea that the system is learnable, scalable, efficient in interaction design and suitable for deployment with modest refinement.

Following this, attention turns to the predictive evaluation findings, which shed light on the accuracy and contextual behavior of the system's AI-driven grade predictions.

While most measures varied across different courses, MHSP scores were consistently high, often exceeding 94.5%, indicating a strong proportional alignment between predicted and actual student grades, even when traditional error metrics such as RMSE and MAE.

Moreover, in stable, low-variance environments, simpler models tended to outperform more complex ones, as added feature richness sometimes hurt performance. Conversely, in chaotic, highly variable datasets, feature enrichment and ensemble models were essential for accurate predictions.

Some models, like AdaBoost proved to be very sensitive to data context, performing well in noisy environments but poorly in simpler ones, likely due to overfitting. On the other hand, CatBoost delivered stable yet unremarkable results with feature enrichment having little to no impact on its performance, probably due to the model's native handling of categorical variables. Lastly, Random Forest models were mostly effective in noisy, unstructured data but unnecessary in more balanced environments, with smart configurations shining in complex contexts, but degrading in simpler ones.

Overall, the predictive evaluation demonstrated that, while feature enrichment can be helpful in imbalanced educational settings, the principle that model complexity should match data complexity should be treated as a guiding requirement when selecting predictive strategies for educational applications.

6.1.5 Research Questions Reflections

As this research reaches its conclusion, it becomes important to return to the core questions that sparked its inception in Chapter 1, as they served as the foundation for guiding the development, implementation, and evaluation of the *CABRAL* system.

- **RQ1: Can an AI-Powered LAD help students and educators improve their learning outcomes?**
- **RQ2: What are the main capabilities of existing AI-Powered LADs?**
- **RQ3: How can advanced AI techniques, such as predictive and prescriptive analytics, improve the accuracy and utility of insights provided by LADs?**

After an extensive SLR, system design and development, and empirical testing through usability and predictive evaluations, it is now possible to reflect on what was learned and how each question has been addressed in order to provide clarity on the contributions and implications of the work conducted.

RQ1 asked whether an AI-powered LAD could actually help students and educators improve their learning outcomes and, from the evidence gathered through both quantitative metrics and qualitative feedback, the answer is a resounding yes, despite important nuances. *CABRAL*, the system developed in this work, was received positively by participants and, while students noted its usefulness in organizing their academic progress and identifying areas for improvement, educators appreciated the ability to monitor class-wide trends and detect struggling learners early on. Usability testing confirmed that the system was intuitive to use, with high task success rates and favorable SUS scores. At the same time, student feedback raised important emotional considerations, particularly regarding the psychological effects of motivation predictions. These concerns suggest that while AI can be a powerful ally in education, it must be implemented with sensitivity to human experience. Nevertheless, *CABRAL* demonstrated that, when designed thoughtfully, AI can play a transformative role in improving educational outcomes.

On the other hand, RQ2 explored the main capabilities of existing AI-powered LADs. To better address this question it is important to analyze the conducted SLR, which revealed a growing ecosystem of tools that combine data visualization, performance tracking, and predictive analytics. While the review identified clear limitations, it uncovered many systems that are currently capable of integrating with LMS and offer real-time insights into student engagement, motivation, and achievement. However, many dashboards lacked standardization in interface design and failed to emphasize explainability or emotional impact, two aspects that became central to *CABRAL*'s development. By addressing these gaps, the system presented in this thesis pushes the boundaries of what such dashboards can offer, especially in terms of user-centered design and emotional intelligence.

The final research question, RQ3, focused on how advanced AI techniques, particularly predictive and prescriptive analytics, could enhance the accuracy and utility of LAD-generated insights, a question answered through extensive model evaluation across three distinct academic environments. From the analysis, a nuanced understanding emerged: not all models perform equally well in every context. For example, in highly structured, predictable settings, simple models like Linear Regression were surprisingly effective, while on the other hand, in noisier, more unbalanced courses, ensemble methods such as Random Forest and AdaBoost proved invaluable. The

thesis also utilized the MHSP metric, a novel way of evaluating predictions that better reflects educational needs by prioritizing ranking consistency and outcome categories over raw numeric precision. These findings underscore that AI models must be matched thoughtfully to the data and educational context they serve and that when they are, they can produce insights that are not only accurate, but genuinely actionable.

In reflecting on these three questions, it is clear that the work accomplished through this thesis has not only addressed them, but done so in a way that opens the door for further exploration. *CABRAL* stands as a working example of how AI, analytics, and thoughtful design can intersect to support better teaching, deeper learning, and a more responsive educational system.

6.2 Ethical Considerations in *CABRAL*

Given that predictive analytics form a core component of the *CABRAL* system, it is essential to critically examine the ethical implications of their use. While such technologies undoubtedly offer enhanced support for educators and students, overlooking concerns related to fairness, privacy, and accountability could turn a promising educational tool into a counterproductive system—one that inadvertently harms the classroom environment rather than improving it.

1. **Bias and Fairness:** As predictive models rely on historical data, they may reflect existing inequalities or systemic biases that, if not carefully mitigated, can lead to inaccurate or unfair predictions particularly affecting underrepresented or marginalized student groups. For instance, a model that disproportionately flags students from certain demographics as at-risk could reinforce negative stereotypes rather than provide equitable support. As a result, ensuring diverse training data and continuously auditing models for disparate impact are critical steps toward fairness.
2. **Privacy and Consent:** Since student data is inherently sensitive, collecting, storing, and analyzing this data, especially for predictive purposes, requires stringent privacy protections and ethical data governance. Consequently, both students and educators should be clearly informed about how their data is used, and institutions must ensure that consent is meaningful and revocable with anonymization techniques and secure data handling protocols being crucial to preserve the privacy and safety of all stakeholders involved.
3. **Transparency and Explainability:** The use of black-box algorithms, where predictions are made without clear explanation, can erode trust in the system. For LADs to be ethically viable, they must incorporate XAI approaches that allow stakeholders, and particularly educators and students, to understand the rationale behind predictions. This is especially important when such predictions could influence interventions or decisions about student support or are directly displayed to students, as they can have an undeniable impact on their motivation, if not handled carefully.

4. **Accountability:** Ultimately, predictive analytics tools should augment human decision-making, not replace it. Educators should retain final responsibility for how insights are used, and institutions must establish clear accountability structures for model performance and decision outcomes. Without this, there is a risk of overreliance on automated suggestions, potentially undermining professional judgment and promoting a less than desirable educational environment that can lead to misinformed interventions, decreased educator autonomy, and erosion of trust between students and instructors. These considerations highlight the need for ongoing evaluation not only of model performance but also of ethical impact. As LADs continue to evolve, integrating ethical design principles and engaging users in co-design processes will be crucial to ensuring that these tools support equitable, transparent, and responsible educational outcomes.

Chapter 7

Conclusions and Future Work

The research described in this document set out with a clear and ambitious goal: to investigate whether AI-powered LADs could meaningfully enhance student motivation and support educators in fostering more effective learning experiences. To that end, the study introduced *CABRAL*: a context-aware, explainable LAD designed around students' real needs and usage habits, informed by survey feedback, predictive analytics, and human-centered design principles.

The journey began with a review of existing LADs, which revealed a field rich in predictive capability but often lacking in emotional design, explainability, and personalized intervention where *CABRAL* could enter as a response to these gaps. The system was built as an interactive, web-based dashboard that combined motivational prediction, performance insights, and prescriptive feedback in a user-centered interface. Alongside technical development, two surveys were conducted to better understand how students interact with LMS platforms, what motivates or hinders their engagement, and which features might make learning tools more impactful. These insights became the foundation upon which *CABRAL*'s core features such as personalized reminders, behavioral tracking, and simplified progress visualization were built.

A rigorous evaluation followed, involving both predictive model benchmarking and usability testing where results demonstrated that both students and educators found *CABRAL* intuitive and supportive. Task success rates were consistently high, SUS scores surpassed standard benchmarks, particularly among educators and users appreciated the ability to track progress, anticipate challenges, and receive timely, context-aware support. Although some concerns were raised, especially regarding emotional reactions to low motivational predictions, these were regarded as constructive feedback for future iterations rather than flaws in the concept itself, as many of the issues raised were already being addressed through existing features.

Conversely, the predictive evaluation added further depth as a range of machine learning models from simple linear regression to ensemble methods like AdaBoost and Random Forest were tested across three course environments. Results confirmed that model complexity must align with data complexity: simple models performed well in stable environments, while advanced models were better suited to noisy, imbalanced academic contexts. The usage of MHSP as a performance metric offered a more education-relevant view of prediction quality, showing how models could

align proportionally with student outcomes even when traditional metrics appeared inconclusive.

Taken together, the findings of this research affirm that AI-powered dashboards like *CABRAL* can indeed enhance educational experiences, not only by delivering accurate predictions, but by doing so in ways that are explainable, empathetic, and aligned with user needs as while existing LADs may offer powerful algorithms, without thoughtful interaction design, their impact is limited. *CABRAL* demonstrated that blending advanced analytics with user-centered principles can yield a tool that is both functional and meaningful.

Over the course of the project, two significant contributions to the academic community emerged in the form of research publications.

The first [173] is a SLR (currently under preparation), which synthesizes existing work on AI-powered learning analytics dashboards and identifies design gaps in current solutions.

The second is a publication focused on *CABRAL*, the dashboard developed in this thesis, which presents its architecture, predictive performance, and educational relevance. This work introduces a real-time AI-powered LAD capable of predicting both academic performance and motivational engagement using behavioral data extracted from LMS platforms and was submitted to ICELIE 2025 [174], under the name “*An AI-Powered Learning Analytics Dashboard for Real-Time Student Performance and Motivation Monitoring*”.

These publications not only extend the impact of this research beyond the scope of this thesis but also contribute to the broader discourse on the ethical and effective application of AI in education.

7.1 Limitations

Despite the encouraging results and positive reception of the proposed system, it is important to acknowledge the limitations that shaped the scope and interpretation of this research as every design and evaluation decision, while grounded in sound methodology, was influenced by practical constraints related to time, access, and technical feasibility. These limitations do not undermine the validity of the findings but rather help contextualize them within the boundaries of a prototype-driven, academically framed investigation. Addressing these shortcomings in future iterations will be essential to ensure that the system evolves into a more robust, generalizable, and ethically responsible educational tool.

1. **Limited Sample Size and Diversity:** The user surveys and usability tests were conducted with a relatively small sample of students and educators. This limits the generalizability of the findings, particularly regarding variations in LMS usage habits, academic cultures, or motivational factors across institutions, disciplines, or countries. A more diverse participant pool could reveal additional needs or behaviors not captured in this study.

2. **Short-Term Evaluation:** The evaluations were conducted over a short time frame without long-term engagement. As a result, the impact of *CABRAL* on sustained motivation, academic performance, or behavioral change could not be fully assessed. Longitudinal studies are needed to evaluate long-term effects.
3. **Controlled vs. Real-World Contexts:** Testing occurred in semi-controlled environments with defined tasks, rather than being embedded in live courses. This may have influenced user behavior, limiting the ecological validity of the findings. Real-world deployment may reveal different usage patterns influenced by course difficulty, instructor behavior, or peer interaction.
4. **Dependence on LMS Data Quality:** *CABRAL*'s predictive features depend on the quality and consistency of LMS data. In contexts where data is sparse or poorly structured, prediction accuracy may degrade. Institutions with limited digital infrastructure might struggle to derive value from the system.
5. **Model Transparency vs. User Trust:** Despite including explainability features, a gap remains between algorithmic transparency and users' intuitive understanding. Some participants expressed discomfort with certain predictions, suggesting a need to better align interpretability with user expectations and emotional responses.
6. **Ethical and Psychological Considerations:** Predicting motivation and performance raises ethical concerns. Even with supportive intentions, predictions may lead to stress, reduced self-efficacy, or over-reliance on algorithmic outputs. These risks must be carefully managed, especially at scale.

7.2 Future Work

While *CABRAL* demonstrates significant potential, there are several avenues that future research could explore to extend and refine this work. One natural progression would be to deploy the system in real classroom environments over an extended academic term, allowing for a more accurate evaluation of *CABRAL*'s long-term impact on student motivation, engagement, and academic performance, and providing insights that are difficult to capture in short-term testing scenarios.

Another important direction involves expanding the system's adaptive capabilities. Allowing students to personalize how they receive insights, for example by selecting the level of guidance or motivational messaging, could help balance support with autonomy, accommodating different learner preferences. This would also build on feedback from participants who expressed varying comfort levels with automated interventions.

Further development could also involve further integrating emotional and mental health considerations into the system's design. As some users reported anxiety when presented with certain predictive features, future versions of *CABRAL* might include emotionally intelligent feedback

mechanisms, opt-in affective analytics, or softer nudges that promote motivation without increasing stress.

Additionally, scaling the predictive models to work across different institutions, course types, or learning platforms would improve the generalizability of the system. This would involve training and validating the models on more diverse datasets, potentially uncovering new patterns or behaviors that were not visible within the scope of this study.

Finally, further investigation into how instructors respond to and use the insights provided by the dashboard could add an important dimension. While *CABRAL* was designed for both students and educators, more focused research on how it influences teaching strategies, communication, and interventions could help fine-tune the system to better support both ends of the learning relationship.

Bibliography

- [1] Tableau Exchange. <https://exchange.tableau.com/pt-br/products/117>. Accessed: 2025-01-06.
- [2] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.
- [3] António Inácio Correia Nogueira, Carlos Manuel Santos Assunção Rodrigues, and Joaquim de Sousa Morais Ferreira. *Formar hoje, educar amanhã*. G.C Gráfica de Coimbra, Lda., 9 1990.
- [4] Eduaide.ai. <https://www.eduaide.ai/>. Accessed: 2025-06-28.
- [5] Slidesgo. <https://slidesgo.com/>. Accessed: 2025-06-28.
- [6] Chatgpt. <https://chatgpt.com/>. Accessed: 2025-06-28.
- [7] Namsoo Shin, Shawn Y. Stevens, and Joseph Krajcik. Designing assessments to track student progress. In *Proceedings of the 9th International Conference of the Learning Sciences - Volume 1*, ICLS '10, page 580–587. International Society of the Learning Sciences, 2010.
- [8] Francesca Maridina Malloci, Paola Mejia, Agathe Merceron, Anna Monreale, and Daniela Rotelli. 1st international workshop on user-centered practices of knowledge discovery in educational data. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, page 534–537, New York, NY, USA, 2024. Association for Computing Machinery.
- [9] Iuliana Dobre. Learning management systems for higher education - an overview of available options for higher education organizations. *Procedia - Social and Behavioral Sciences*, 180:313–320, 5 2015.
- [10] What is a learning management system (lms)? <https://www.techtarget.com/searchcio/definition/learning-management-system>. Accessed: 2025-06-28.

- [11] Moodle. <https://moodle.org/>. Accessed: 2025-06-28.
- [12] Canvas. <https://www.instructure.com/canvas/>. Accessed: 2025-06-28.
- [13] Niklas Meißner, Nadine Koch, Sandro Speth, Uwe Breitenbücher, and Steffen Becker. Unveiling hurdles in software engineering education: The role of learning management systems. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*, ICSE-SEET '24, page 242–252, New York, NY, USA, 2024. Association for Computing Machinery.
- [14] Anthology. Blackboard. <https://www.anthology.com/products/teaching-and-learning/learning-effectiveness/blackboard>. Accessed: 2025-06-28.
- [15] Smithsonian Institution. Get the latest news from the smithsonian. https://www.si.edu/object/pressey-teaching-machine%3Anmah_1367149. Accessed: 2025-01-06.
- [16] Stephen Petrina. Sidney pressey and the automation of education, 1924-1934. *Technology and Culture*, 45(2):305–330, 2004.
- [17] Nor Aziah Alias and Ahmad Marzuki Zainuddin. Innovation for better teaching and learning: Adopting the learning management system. *Malaysian online journal of instructional technology*, 2(2):27–40, 2005.
- [18] Nurul Nadirah Mohd Kasim and Fariza Khalid. Choosing the right learning management system (lms) for the higher education institution context: A systematic review. *International Journal of Emerging Technologies in Learning*, 11(6), 2016.
- [19] Beth Davis, Colleen Carmean, and Ellen D Wagner. The evolution of the lms: From management to learning. *Santa Rosa, CA: e-Learning Guild*, 24, 2009.
- [20] David A Falvo and Ben F Johnson. The use of learning management systems in the united states. *TechTrends: Linking Research & Practice to Improve Learning*, 51(2), 2007.
- [21] EdTech. Edutechnica. <https://edutechnica.com/2017/03/12/lms-data-spring-2017-updates/>, Mar 2017. Accessed: 2025-01-06.
- [22] Imed Bouchrika. <https://research.com/education/lms-statistics>, Oct 2024. Accessed: 2025-01-06.
- [23] Society for Learning Analytics Research. <https://www.solaresearch.org/about/what-is-learning-analytics/>, Aug 2024. Accessed: 2025-01-06.
- [24] Dan Ye. The history and development of learning analytics in learning, design, & technology field. *TechTrends*, 66(4):607–615, 2022.

- [25] Edward L Thorndike. Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2(4):i, 1898.
- [26] John B Watson. Psychology as the behaviorist views it. *Psychological review*, 20(2):158, 1913.
- [27] International Educational Data Mining Society. <https://educationaldatamining.org/>. Accessed: 2025-01-06.
- [28] Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan. A systematic review on educational data mining. *Ieee Access*, 5:15991–16005, 2017.
- [29] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, 40(6):601–618, 2010.
- [30] Arun P Sanjeev and Jan M Żytkow. Discovering enrollment knowledge in university databases. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 246–251, 1995.
- [31] Cristobal Romero and Sebastian Ventura. *Data mining in e-learning*, volume 4. WIT press, 2006.
- [32] Cristobal Romero and Sebastian Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.
- [33] George Siemens and Ryan SJ d Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254, 2012.
- [34] Ashish K Sharma, Durgesh M Sharma, Neha Purohit, Saroja Kumar Rout, and Sangita A Sharma. Analytics techniques: descriptive analytics, predictive analytics, and prescriptive analytics. *Decision intelligence analytics and the implementation of strategic business management*, pages 1–14, 2022.
- [35] Wannisa Matcha, Nora’ayu Ahmad Uzir, Dragan Gašević, and Abelardo Pardo. A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, 13(2):226–245, 2020.
- [36] Joris Klerkx, Katrien Verbert, and Erik Duval. Learning analytics dashboards. *Handbook of learning analytics*, 1:143–150, 2017.
- [37] Katrien Verbert, Xavier Ochoa, Robin De Croon, Raphael A Dourado, and Tinne De Laet. Learning analytics dashboards: The past, the present and the future. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 35–40, 2020.

- [38] Steve Tally. <https://www.purdue.edu/uns/x/2009b/090827ArnoldSignals.html>, Sep 2009. Accessed: 2025-01-06.
- [39] Matthew D Pistilli and Kimberly E Arnold. Purdue signals: Mining real-time academic data to enhance student success. *About campus*, 15(3):22–24, 2010.
- [40] Purdue IDC. Course signals explanation. <https://www.youtube.com/watch?v=KxYewO9iAgw>. Accessed: 2025-01-06.
- [41] Kimberly E Arnold and Matthew D Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 267–270, 2012.
- [42] Jeongwon Lee and Dongho Kim. From awareness to empowerment: self-determination theory-informed learning analytics dashboards to enhance student engagement in asynchronous online courses. *Journal of Computing in Higher Education*, pages 1–41, 2024.
- [43] UNESCO. Higher education: figures at a glance. Electronic, 2024. Programme and meeting document [212123]. Accessed: 2025-01-06.
- [44] UNESCO. Covid-19 education response. <https://www.unesco.org/en/covid-19/education-response>, n.d. Accessed: 2025-01-06.
- [45] Chih-Tsung Hung, Shou-En Wu, Yi-Hsien Chen, Chen-Yeu Soong, Chien-Ping Chiang, and Wei-Ming Wang. The evaluation of synchronous and asynchronous online learning: student experience, learning outcomes, and cognitive load. *BMC Medical Education*, 24(1):326, 2024.
- [46] Veda C. Storey, Roman Lukyanenko, Wolfgang Maass, and Jeffrey Parsons. Explainable ai. *Commun. ACM*, 65(4):27–29, March 2022.
- [47] David Gunning and David Aha. Darpa’s explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.
- [48] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.
- [49] Nada Hashmi and Anjali S Bal. Generative ai in higher education and beyond. *Business Horizons*, 2024.
- [50] Dall-e. <https://openai.com/index/dall-e/>. Accessed: 2025-01-17.
- [51] Turnitin. <https://www.turnitin.com/>. Accessed: 2025-01-17.
- [52] Jiahong Su and Weipeng Yang. Unlocking the power of chatgpt: A framework for applying generative ai in education. *ECNU Review of Education*, 6(3):355–366, 2023.

- [53] Yeonjeong Park and Il-Hyun Jo. Development of the learning analytics dashboard to support students' learning performance. *Journal of Universal Computer Science*, 21:110 – 133, 2015. Cited by: 176.
- [54] Wannisa Matcha, Noraayu Ahmad Uzir, Dragan Gasevic, and Abelardo Pardo. A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective, 4 2020.
- [55] Yaëlle Chaudy and Thomas M Connolly. *Integrating assessment, feedback, and learning analytics in educational games: Literature review and design of an assessment engine*, volume 4-4. 2021. Cited by: 0.
- [56] Natercia Valle, Pavlo Antonenko, Kara Dawson, and Anne Corinne Huggins-Manley. Staying on target: A systematic literature review on learner-facing learning analytics dashboards. *British Journal of Educational Technology*, 52:1724 – 1748, 2021. Cited by: 43.
- [57] Kimberly Williamson and Rene Kizilcec. A review of learning analytics dashboard research in higher education: Implications for justice, equity, diversity, and inclusion. In *ACM International Conference Proceeding Series*, pages 260–270. Association for Computing Machinery, 3 2022.
- [58] Sdenka Zobeida Salas-Pilco, Kejiang Xiao, and Xinyun Hu. Artificial intelligence and learning analytics in teacher education: A systematic review, 8 2022.
- [59] Carlos Alberto Peláez, Andrés Solano, Emma Adriana De La Rosa, Jesús Alfonso López, Jorge Andrick Parra, Johann Alexis Ospina, Gabriel Mauricio Ramírez, and Fernando Moreira. Learning analytics in the design of interactive multimedia experiences for elementary education: A systematic review. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 9 2023.
- [60] Vamsi Krishna Nadimpalli, Florian Hauser, Dominik Bittner, Lisa Grabinger, Susanne Staufer, and Jürgen Mottok. Systematic literature review for the use of ai based techniques in adaptive learning management systems. In *ACM International Conference Proceeding Series*, pages 83–92. Association for Computing Machinery, 6 2023.
- [61] Riordan Alfredo, Vanessa Echeverria, Yueqiao Jin, Lixiang Yan, Zachari Swiecki, Dragan Gašević, and Roberto Martinez-Maldonado. Human-centred learning analytics and ai in education: A systematic literature review, 6 2024. Accessed: 2025-01-06.
- [62] Rogers Kaliisa, Kamila Misiejuk, Sonsoles López-Pernas, Mohammad Khalil, and Mohammed Saqr. Have learning analytics dashboards lived up to the hype? a systematic review of impact on students' achievement, motivation, participation and attitude. In *ACM International Conference Proceeding Series*, pages 295–304. Association for Computing Machinery, 3 2024. Focuses on the impact of Learning Analytic Dashboards on students' learning outcomes.

- [63] Manuel Valle Torre, Catharine Oertel, and Marcus Specht. The sequence matters in learning - a systematic literature review. In *ACM International Conference Proceeding Series*, pages 263–272. Association for Computing Machinery, 3 2024.
- [64] Lucas Paulsen and Euan Lindsay. Learning analytics dashboards are increasingly becoming about learning and not just analytics - a systematic review. *Education and Information Technologies*, 29:14279–14308, 8 2024.
- [65] Fan Ouyang and Liyin Zhang. Ai-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review, 8 2024.
- [66] Lixiang Yan, Roberto Martinez-Maldonado, and Dragan Gasevic. Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In *ACM International Conference Proceeding Series*, pages 101–111. Association for Computing Machinery, 3 2024.
- [67] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.
- [68] Elsevier. Scopus database. <https://www.scopus.com>, 2025. Accessed: 2025-06-28.
- [69] Association for Computing Machinery (ACM). Acm digital library. <https://dl.acm.org>, 2025. Accessed: 2025-06-28.
- [70] IEEE. Ieee xplora digital library. <https://ieeexplore.ieee.org>, 2025. Accessed: 2025-06-28.
- [71] Institute of Education Sciences (IES). Eric (education resources information center). <https://eric.ed.gov>, 2025. Accessed: 2025-06-28.
- [72] Tesnim Khelifi, Bénédicte Le Grand, and Nourhène Ben Rabah. Explainable learning analytics dashboard: Enhancing understanding of insights derived from educational data. *International Journal of Computers and their Applications*, 31:83 – 94, 2024. Cited by: 0.
- [73] Tesnim Khelifi, Nourhène Ben Rabah, Bénédicte Le Grand, and Ibtissem Daoudi. Ex-lad: an explainable learning analytics dashboard in higher education. In *EPiC Series in Computing*, volume 97, page 38 – 51, 2024. Cited by: 1; All Open Access, Bronze Open Access.

- [74] Teo Susnjak, Gomathy Suganya Ramaswami, and Anuradha Mathrani. Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19, 2022. Cited by: 69; All Open Access, Gold Open Access, Green Open Access.
- [75] Gomathy Ramaswami, Teo Susnjak, and Anuradha Mathrani. Effectiveness of a learning analytics dashboard for increasing student engagement levels. *Journal of Learning Analytics*, 10:115 – 134, 2023. Cited by: 3; All Open Access, Gold Open Access.
- [76] Damien S Fleur, Wouter van den Bos, and Bert Bredeweg. Learning analytics dashboard for motivation and performance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12149 LNCS:411 – 419, 2020. Cited by: 14.
- [77] Yung-Hsiang Hu. Effects and acceptance of precision education in an ai-supported smart learning environment. *Education and Information Technologies*, 27:2013 – 2037, 2022. Cited by: 22.
- [78] Nuo Cheng, Wei Zhao, Xiaoqing Xu, Hongxia Liu, and Jinhong Tao. The influence of learning analytics dashboard information design on cognitive load and performance. *Education and Information Technologies*, 2024. Cited by: 0.
- [79] Rwitajit Majumdar, Kyosuke Takami, and Hiroaki Ogata. Learning with explainable ai-recommendations at school: Extracting patterns of self-directed learning from learning logs. In *Proceedings - 2023 IEEE International Conference on Advanced Learning Technologies, ICAALT 2023*, page 245 – 249, 2023. Cited by: 1.
- [80] Hassan Khosravi, Shiva Shabaninejad, Aneesha Bakharia, Shazia Sadiq, Marta Indulska, and Dragan Gašević. Intelligent learning analytics dashboards: Automated drill-down recommendations to support teacher data exploration. *Journal of Learning Analytics*, 8:133 – 154, 2021. Cited by: 27; All Open Access, Gold Open Access.
- [81] Xiaojing Duan, Chaoli Wang, and Guieswende Rouamba. Designing a learning analytics dashboard to provide students with actionable feedback and evaluating its impacts. In *International Conference on Computer Supported Education, CSEDU - Proceedings*, volume 2, page 117 – 127, 2022. Cited by: 13; All Open Access, Hybrid Gold Open Access.
- [82] Cristian Cechinel, Mateus De Freitas Dos Santos, Caio Barrozo, Jesiel Emerim Schardosim, Eduardo De Vila, Vinicius Ramos, Tiago Primo, Roberto Munoz, and Emanuel Marques Queiroga. A learning analytics dashboard for moodle: Implementing machine learning techniques to early detect students at risk of failure. In *Proceedings - 2021 16th Latin American Conference on Learning Technologies, LACLO 2021*, page 130 – 136, 2021. Cited by: 2.

- [83] Riina Kleimola, Laura Hirsto, and Heli Ruokamo. Promoting higher education students' self-regulated learning through learning analytics: A qualitative study. *Education and Information Technologies*, 2024. Cited by: 0; All Open Access, Hybrid Gold Open Access.
- [84] Iris Groher, Michael Vierhauser, and Erik Hartl. A learning analytics dashboard for improved learning outcomes and diversity in programming classes. In *International Conference on Computer Supported Education, CSEDU - Proceedings*, volume 2, page 618 – 625, 2024. Cited by: 0.
- [85] Songran Liu, Kousuke Mouri, and Hiroaki Ogata. Learning analytics data flow and visualizing for ubiquitous learning logs in lms and learning analytics dashboard. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12203 LNCS:548 – 557, 2020. Cited by: 0.
- [86] Jan Hellings and Carla Haelermans. The effect of providing learning analytics on student behaviour and performance in programming: a randomised controlled experiment. *Higher Education*, 83, 2022. Cited by: 22; All Open Access, Hybrid Gold Open Access.
- [87] Viet Anh Nguyen, Quang Bach Nguyen, and Vuong Thinh Nguyen. A model to forecast learning outcomes for students in blended learning courses based on learning analytics. In *ACM International Conference Proceeding Series*, page 35 – 41, 2018. Cited by: 17.
- [88] Sotiris Kotsiantis, Nikolaos Tselios, Andromahi Filippidi, and Vassilis Komis. Using learning analytics to identify successful learners in a blended learning course. *International Journal of Technology Enhanced Learning*, 5:133 – 150, 2013. Cited by: 42.
- [89] Alyssa Friend Wise and Yeonji Jung. Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, 6:53 – 69, 2019. Cited by: 118; All Open Access, Gold Open Access.
- [90] Dabbebi Ines, Gilliot Jean-Marie, and Iksal Sebastien. User centered approach for learning analytics dashboard generation. In *CSEDU 2019 - Proceedings of the 11th International Conference on Computer Supported Education*, volume 2, page 260 – 267, 2019. Cited by: 10; All Open Access, Green Open Access, Hybrid Gold Open Access.
- [91] Saketh Kodumuru, Brendan Lucas, Vivek Sabanwar, Sachin Patil, Deepa Avudiappan, Parth Parikh, and Kavi Arya. Towards developing a learning analytics dashboard for a massive online robotics competition. In *2022 IEEE Global Engineering Education Conference (EDUCON)*, pages 1020–1025, 3 2022.
- [92] Iryna Sapsai, Yeimy Paola Valencia Usme, and Joerg Abke. Learning analytics dashboard for educators: Proposed project to design with pedagogical background. In *Proceedings of the 5th European Conference on Software Engineering Education*, pages 38–47. Association for Computing Machinery, 2023.

- [93] Francisco García-Peñalvo and Andrea Vázquez-Ingelmo. What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai. 2023.
- [94] Muhammad Yasir Mustafa, Ahmed Tlili, Georgios Lampropoulos, Ronghuai Huang, Petar Jandrić, Jialu Zhao, Soheil Salha, Lin Xu, Santosh Panda, Sonsoles López-Pernas, et al. A systematic review of literature reviews on artificial intelligence in education (aied): a roadmap to a future research agenda. *Smart Learning Environments*, 11(1):1–33, 2024.
- [95] Wordcloud python package. <https://pypi.org/project/wordcloud/>. Accessed: 2025-01-19.
- [96] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- [97] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
- [98] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.
- [99] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [100] Sander JJ Leemans, Anja F Syring, and Wil MP van der Aalst. Earth movers’ stochastic conformance checking. In *Business Process Management Forum: BPM Forum 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*, pages 127–143. Springer, 2019.
- [101] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.
- [102] Area under the curve. <https://www.cuemath.com/calculus/area-under-the-curve/>. Accessed: 2025-01-11.
- [103] Ethan Zhang and Yi Zhang. *F-Measure*, pages 1147–1147. Springer US, Boston, MA, 2009.
- [104] Will Koehrsen. Precision and recall in classification models. <https://builtin.com/data-science/precision-and-recall>. Accessed: 2025-01-11.
- [105] Ken Stewart. Mean squared error. <https://www.britannica.com/science/mean-squared-error>, Dec 2024. Accessed: 2025-01-11.
- [106] M Waqar Ahmed. Understanding mean absolute error (mae) in regression: A practical guide. <https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e8> Aug 2023. Accessed: 2025-01-11.

- [107] Sahin Ahmed. ML series 5: Understanding r-squared in regression analysis. <https://medium.com/@sahin.samia/understanding-r-squared-in-regression-analysis-2d8246a63dbb>, Feb 2024. Accessed: 2025-01-11.
- [108] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4:253–278, 1994.
- [109] Radek Pelánek. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User modeling and user-adapted interaction*, 27:313–350, 2017.
- [110] Rony Chow. Andrey markov: A mathematician on a mission. <https://www.historyofdatascience.com/andrey-markov-a-mathematician-on-a-mission/>, Dec 2021. Accessed: 2025-01-06.
- [111] Mark A Boyd and Sonie Lau. An introduction to markov modeling: Concepts and uses. In *Annual Reliability and Maintainability Symposium*, 1998.
- [112] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.
- [113] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.
- [114] Oluseye Jeremiah. Catboost in machine learning: A detailed guide. <https://www.datacamp.com/tutorial/catboost>, Sep 2024. Accessed: 2025-01-06.
- [115] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [116] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29:103–130, 1997.
- [117] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Wiley New York, 2000.
- [118] Klaus Hechenbichler and Klaus Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. 2004.
- [119] Scott Nevil. Z-score: Meaning and formula. <https://www.investopedia.com/terms/z/zscore.asp>. Accessed: 2025-01-06.
- [120] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

- [121] Robert E Schapire. Explaining adaboost. In *Empirical inference: festschrift in honor of vladimir N. Vapnik*, pages 37–52. Springer, 2013.
- [122] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28, 2016.
- [123] Luiz Antonio Buschetto Macarini, Cristian Cechinel, Matheus Francisco Batista Machado, Vinicius Faria Culmant Ramos, and Roberto Munoz. Predicting students success in blended learning—evaluating different interactions inside learning management systems. *Applied Sciences*, 9(24):5523, 2019.
- [124] Wayne Iba and Pat Langley. Induction of one-level decision trees. In *Machine learning proceedings 1992*, pages 233–240. Elsevier, 1992.
- [125] Raúl Rojas et al. Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep*, 1(1):1–6, 2009.
- [126] Ya-Han Hu, Chia-Lun Lo, and Sheng-Pao Shih. Developing early warning systems to predict students online learning performance. *Computers in Human Behavior*, 36:469–478, 2014.
- [127] Devin Schumacher. Decision stump. <https://serp.ai/decision-stump/>, Jul 2023. Accessed: 2025-01-06.
- [128] Ray Hsu. Decision stump. <https://medium.com/geekculture/decision-stump-b8e93c1f54d7>, Mar 2023. Accessed: 2025-01-06.
- [129] Ángel F Agudo-Peregrina, Santiago Iglesias-Pradas, Miguel Ángel Conde-González, and Ángel Hernández-García. Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning. *Computers in human behavior*, 31:542–550, 2014.
- [130] Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.
- [131] Vernon C Smith, Adam Lange, and Daniel R Huston. Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of asynchronous learning networks*, 16(3):51–61, 2012.
- [132] Emanuel Marques Queiroga, João Ladislau Lopes, Kristofer Kappel, Marilton Aguiar, Ricardo Matsumura Araújo, Roberto Munoz, Rodolfo Villarroel, and Cristian Cechinel. A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences*, 10(11):3998, 2020.

- [133] Emanuel Marques Queiroga, João Ladislau Barbará Lopes, Ricardo Matsumura de Araújo, and Cristian Cechinel. Modelo de predição da evasão de estudantes em cursos técnicos a distância a partir da contagem de interações. *Revista Thema*, 15(2):425–438, 2018.
- [134] Douglas Detoni, Cristian Cechinel, Ricardo Araujo Matsumura, and Daniela Francisco Brauner. Learning to identify at-risk students in distance education using interaction counts. *Revista de Informática Teórica e Aplicada*, 23(2):124–140, 2016.
- [135] Doug Clow. An overview of learning analytics. *Teaching in Higher Education*, 18(6):683–695, 2013.
- [136] Sarah Alcock, Bart Rienties, Maria Aristeidou, and Soraya Kouadri Mostéfaoui. How do visualizations and automated personalized feedback engage professional learners in a learning analytics dashboard? In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 316–325. Association for Computing Machinery, 2024.
- [137] Samuel Soma Ajibade and Anthonia Adediran. An overview of big data visualization techniques in data mining. *International Journal of Computer Science and Information Technology Research*, 4(3):105–113, 2016.
- [138] Severino Ribecca. The data visualisation catalogue. <https://datavizcatalogue.com/>. Accessed: 2025-01-06.
- [139] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, Dec 1959.
- [140] Benjamin Bach, Euan Freeman, Alfie Abdul-Rahman, Cagatay Turkay, Saiful Khan, Yulei Fan, and Min Chen. Dashboard design patterns. *IEEE Transactions on Visualization and Computer Graphics*, 29:342–352, 1 2023.
- [141] David Bawden, Clive Holtham, and Nigel Courtney. Perspectives on information overload. In *Aslib proceedings*, volume 51, pages 249–255. MCB UP Ltd, 1999.
- [142] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016. Accessed: 2025-01-19.
- [143] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.
- [144] Jeff Sauro and Joseph S. Dumas. Comparison of three one-question, post-task usability questionnaires. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1599–1608, 2009.

- [145] Ibm spss statistics. <https://www.ibm.com/products/spss-statistics>, Apr 2024. Accessed: 2025-01-19.
- [146] Ligre software. <https://ligresoftware.com/>, Oct 2023. Accessed: 2025-01-19.
- [147] Microsoft. C# programming language. <https://learn.microsoft.com/en-us/dotnet/csharp/>, 2024. Accessed: June 2025.
- [148] Microsoft. .net: The free, open-source, cross-platform framework for building modern apps. <https://dotnet.microsoft.com/>, 2024. Version 7.0.
- [149] Inc. MongoDB. Mongoddb. <https://www.mongodb.com>, 2025. Accessed: 2025-06-28.
- [150] Python Software Foundation. Python programming language. <https://www.python.org/>, 2024. Version 3.12.3.
- [151] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [152] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020.
- [153] Gaël Varoquaux, Olivier Grisel, Vlad Niculae, and Alexandre Gramfort. Joblib: running python functions as pipeline jobs. <https://joblib.readthedocs.io/>, 2024. Version 1.4.2.
- [154] Inc. Meta Platforms. React: A javascript library for building user interfaces. <https://reactjs.org/>, 2024. Version 18.2.0.
- [155] Bert Bos, Håkon Wium Lie, Chris Lilley, and Ian Hickson. Cascading style sheets level 2 revision 1 (css 2.1) specification. <https://www.w3.org/TR/CSS2/>, 2011. W3C Recommendation, World Wide Web Consortium (W3C).
- [156] Inc. Fonticons. Font awesome: The iconic font and css toolkit. <https://fontawesome.com/>, 2024. Version 6.0.0.
- [157] Inc. Figma. Figma: Collaborative interface design tool. <https://www.figma.com>, 2025. Accessed: 2025-06-28.

- [158] Inc. Docker. Docker: The open-source platform for automating the deployment, scaling, and management of containerized applications. <https://www.docker.com/>, 2024. Version 23.0.0.
- [159] The Kubernetes Authors. Kubernetes: Open-source system for automating deployment, scaling, and management of containerized applications. <https://kubernetes.io/>, 2024. Version 1.26.0.
- [160] OpenJS Foundation. Node.js: Javascript runtime built on chrome’s v8 javascript engine. <https://nodejs.org/>, 2024. Version 18.16.0.
- [161] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 4765–4774. Curran Associates, Inc., 2017.
- [162] Ahmad B. Hassanat, Mohammad Khaled Alqaralleh, Ahmad S. Tarawneh, Khalid Almomhamadi, Maha Alamri, Abdulkareem Alzahrani, Ghada A. Altarawneh, and Rania Alhalaseh. A novel outlier-robust accuracy measure for machine learning regression using a non-convex distance metric. *Mathematics*, 12, 11 2024.
- [163] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>, 2011. Accessed: 2025-06-28.
- [164] Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, and Jong Wook Kim. Q-learning algorithms: A comprehensive classification and applications. *IEEE access*, 7:133653–133667, 2019.
- [165] Inc. Groq. Groq api. <https://groq.com>, 2024. Accessed: 2025-06-28.
- [166] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [167] Inc. Maze. Maze: The usability testing and research platform. <https://www.maze.co>, 2025. Accessed: 2025-06-28.
- [168] Google LLC. Google forms. <https://forms.google.com>, 2025. Accessed: 2025-06-28.

- [169] Faculdade de Engenharia da Universidade do Porto. Laboratório de gestão de projetos. https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_id=370247, 2024. Accessed: 2025-06-28.
- [170] Faculdade de Engenharia da Universidade do Porto. Sistemas ciberfísicos e internet das coisas. https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_id=540796, 2025. Accessed: 2025-06-28.
- [171] Faculdade de Engenharia da Universidade do Porto. Faculdade de engenharia da universidade do porto (feup). <https://www.fe.up.pt>, 2025. Accessed: 2025-06-28.
- [172] Faculdade de Engenharia da Universidade do Porto. Programação. https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_id=540304, 2024. Accessed: 2025-06-28.
- [173] Luís Cabral, Rui Pinto, and Gil Gonçalves. Ai-powered learning analytics dashboards: A systematic review of applications, techniques, and research gaps. June 2025.
- [174] Ieee international conference on e-learning in industrial electronics (icelie 2025). <https://icelie.org>, October 2025. Madrid, Spain.

Appendix A

Usability Evaluation Task Durations

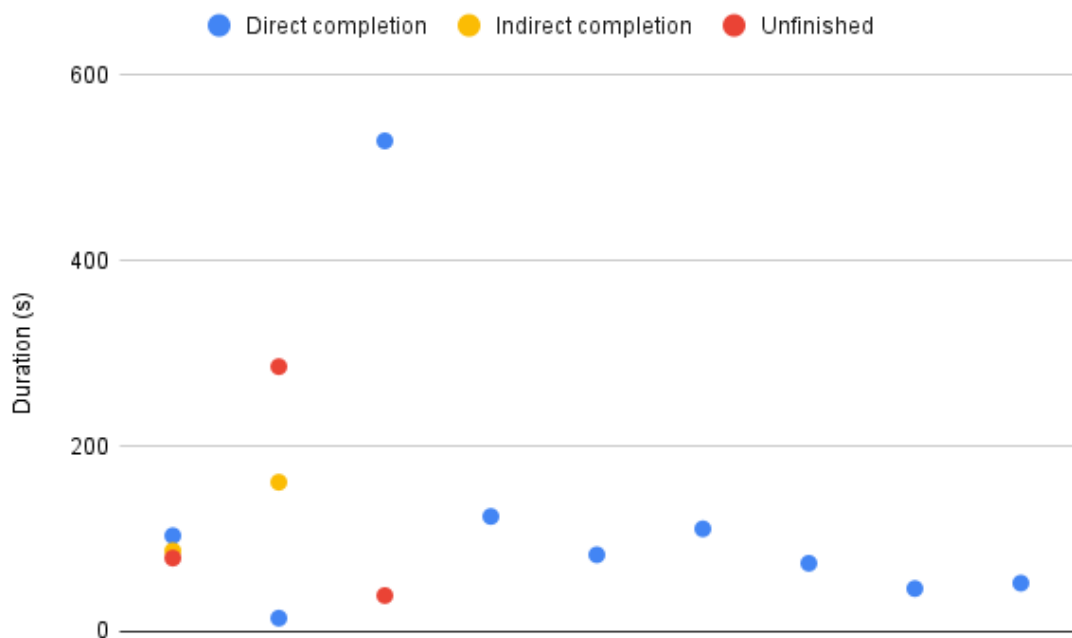


Figure A.1: Task 1 - Educator Durations

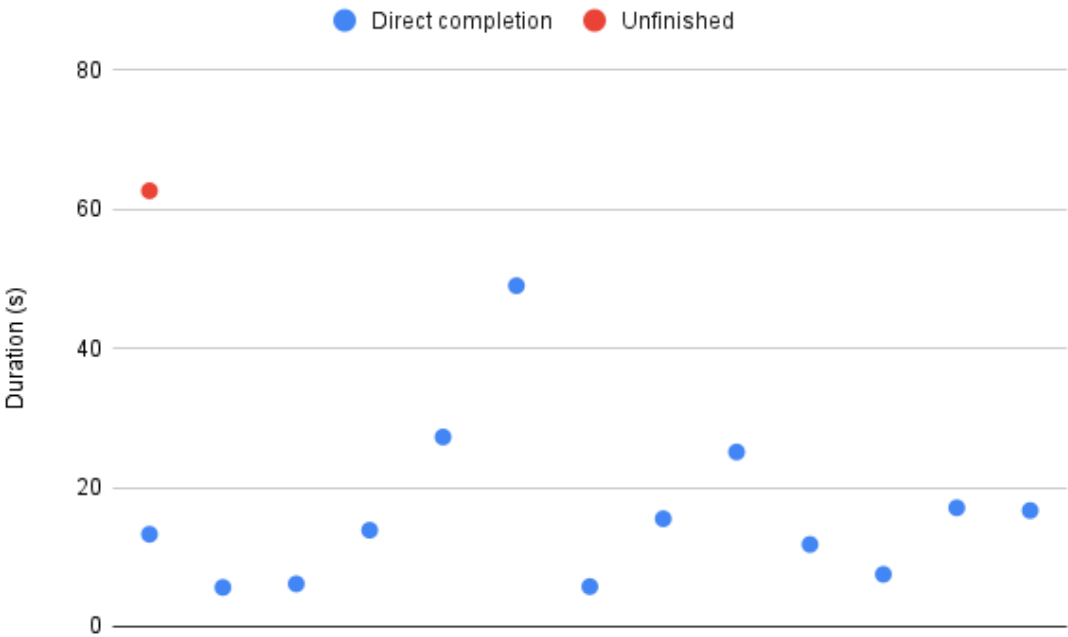


Figure A.2: Task 2 - Educator Durations

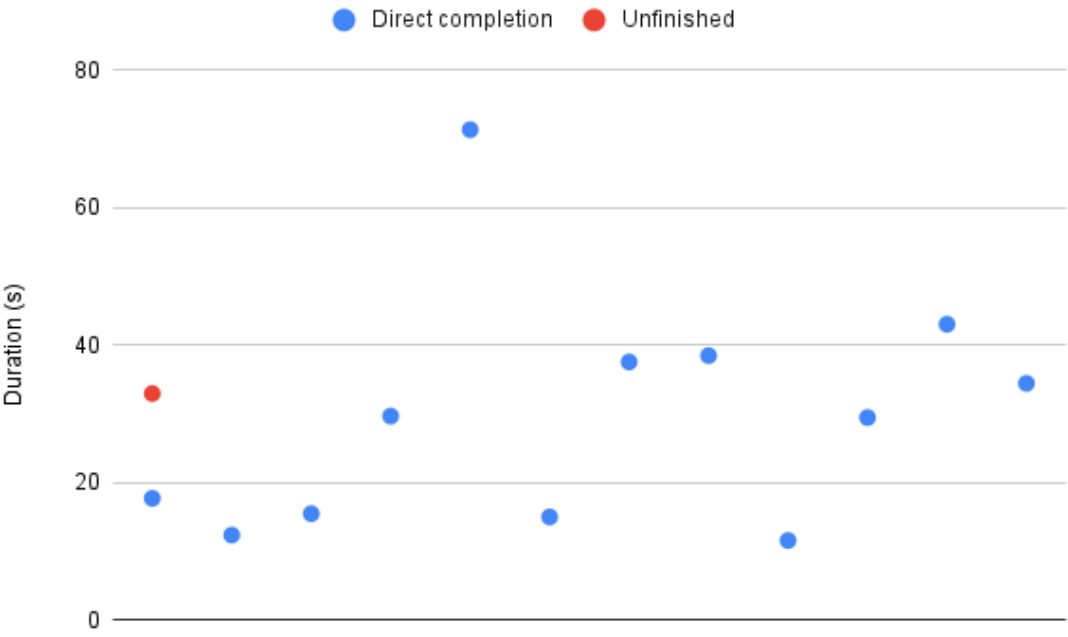


Figure A.3: Task 3 - Educator Durations

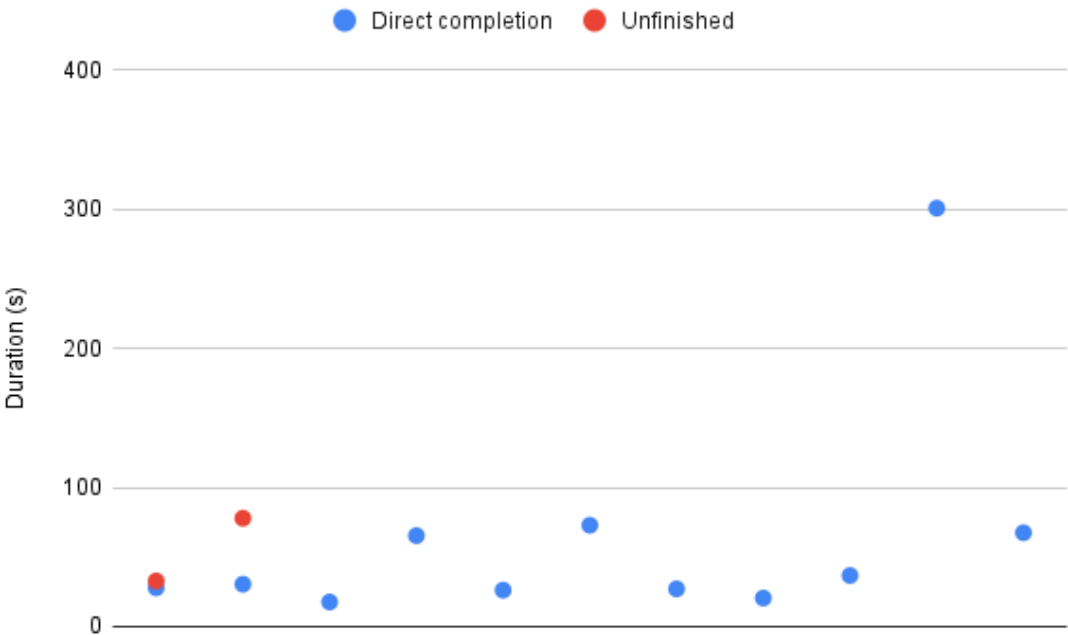


Figure A.4: Task 4 - Educator Durations

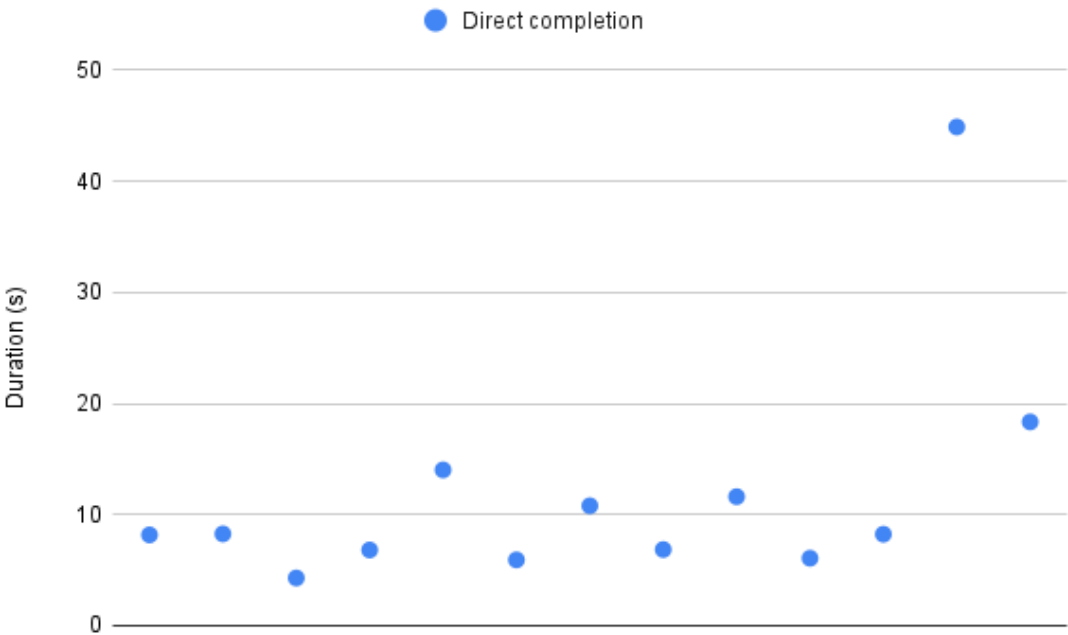


Figure A.5: Task 5 - Educator Durations

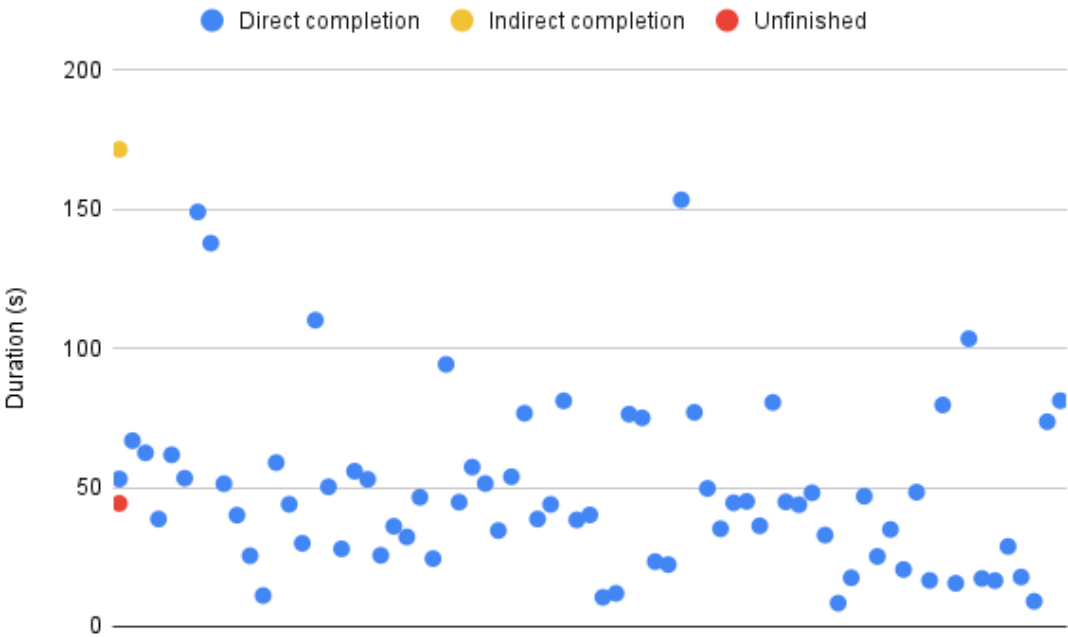


Figure A.6: Task 1 - Student Durations

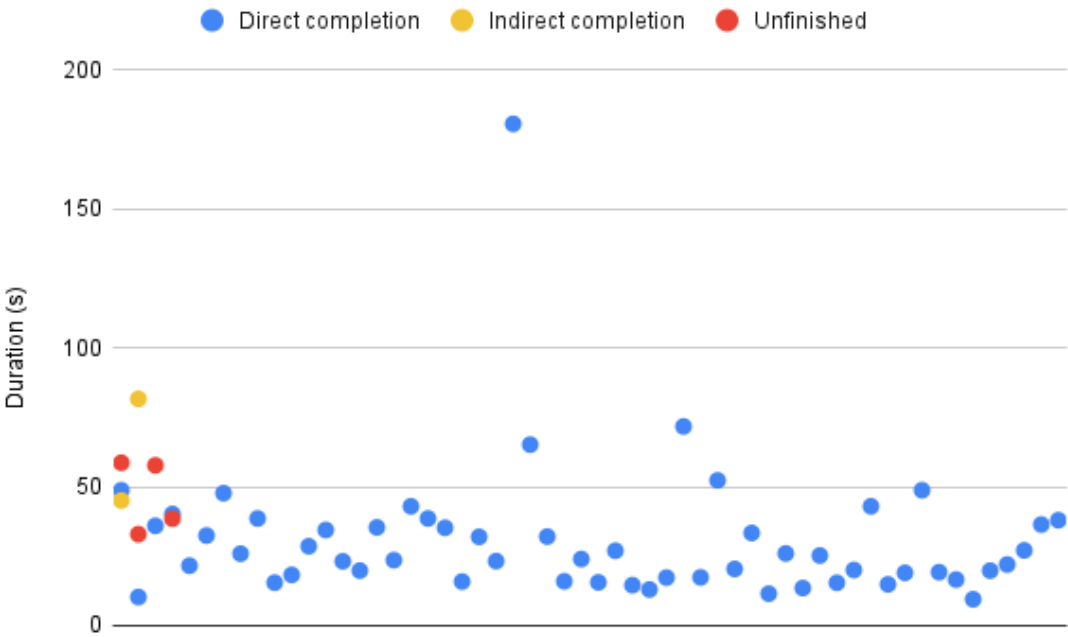


Figure A.7: Task 2 - Student Durations

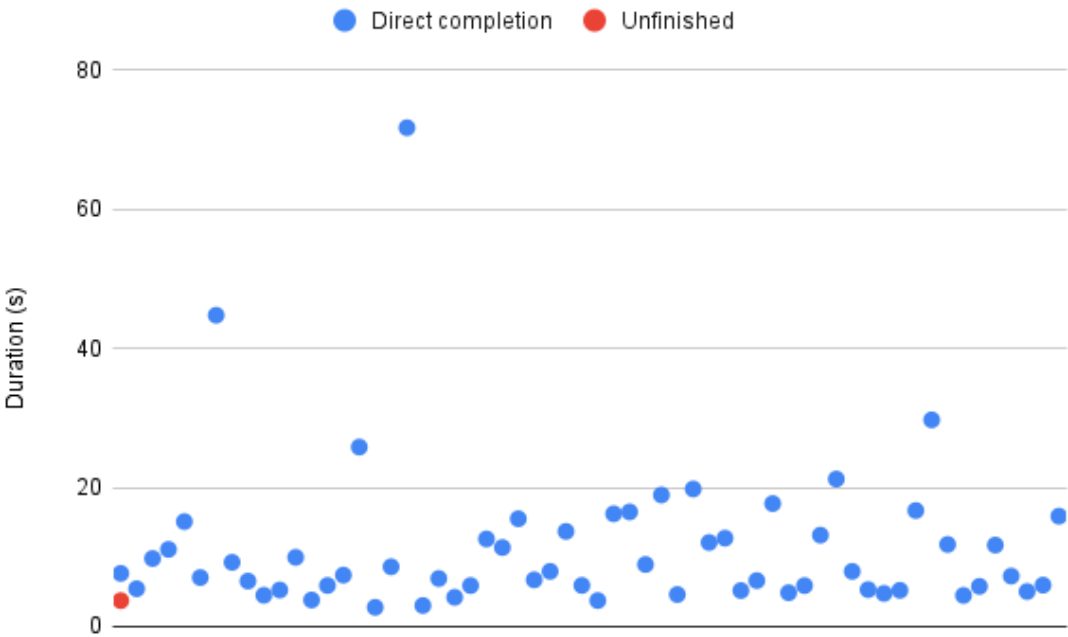


Figure A.8: Task 3 - Student Durations

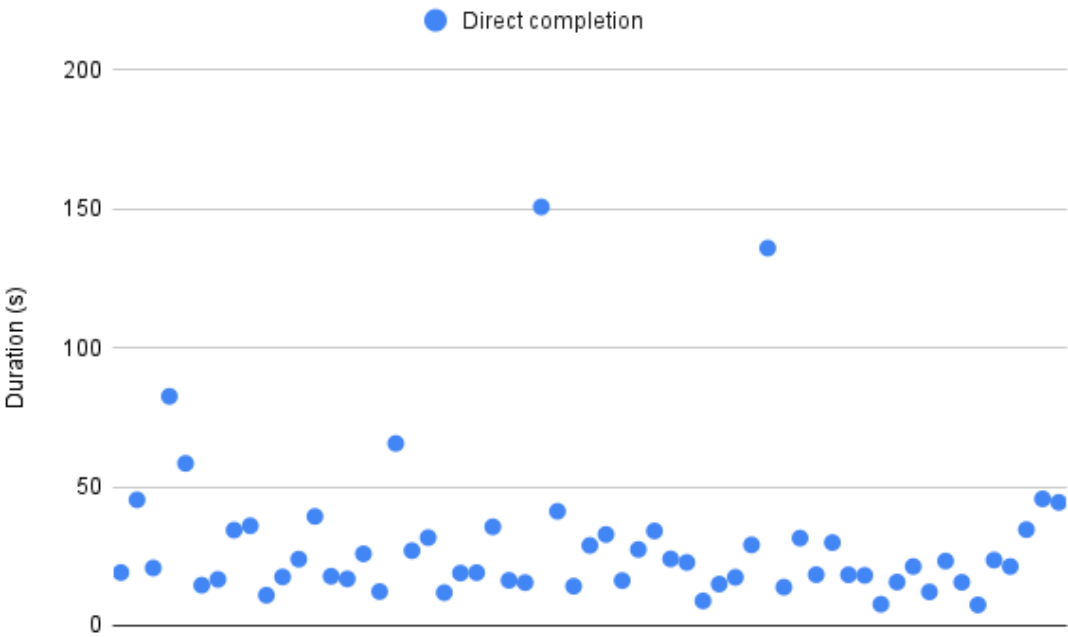


Figure A.9: Task 4 - Student Durations

Appendix B

Preliminary Survey Responses

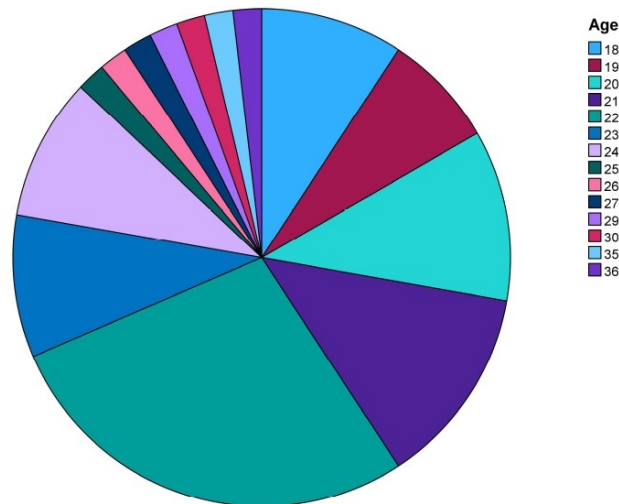


Figure B.1: Preliminary Survey - Q1 Response Distribution

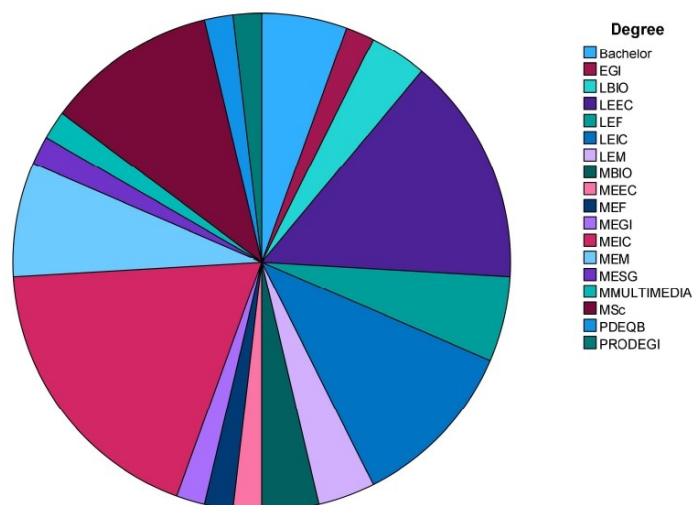


Figure B.2: Preliminary Survey - Q2 Response Distribution

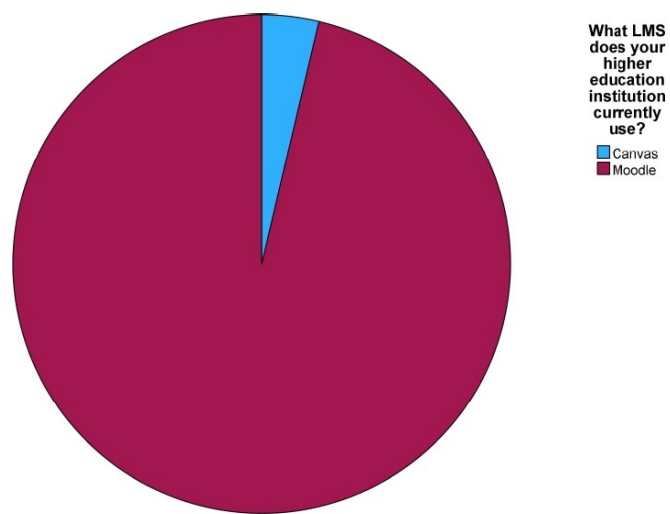


Figure B.3: Preliminary Survey - Q3 Response Distribution

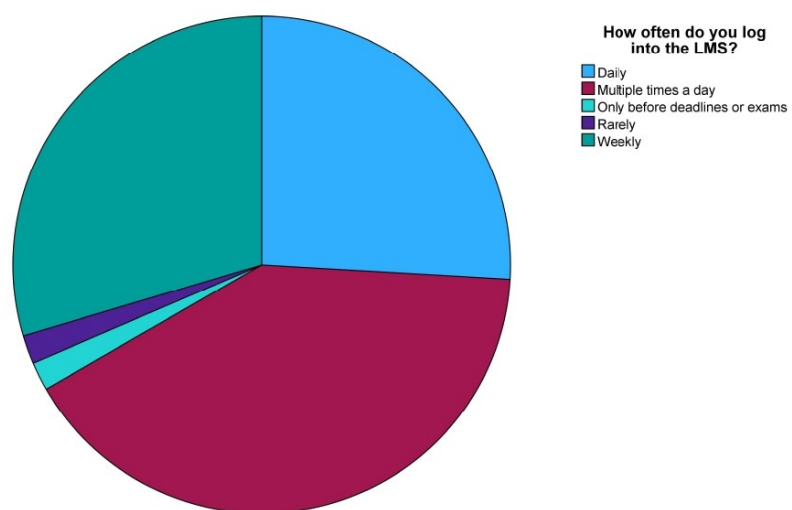


Figure B.4: Preliminary Survey - Q4 Response Distribution

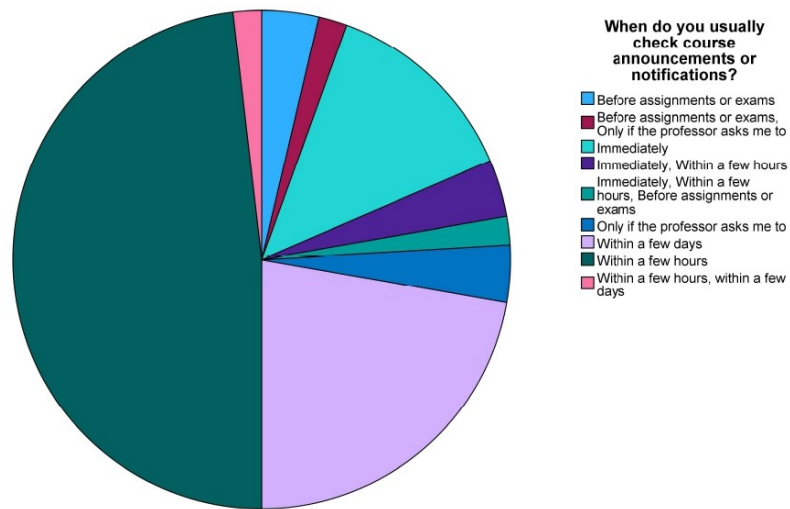


Figure B.5: Preliminary Survey - Q5 Response Distribution

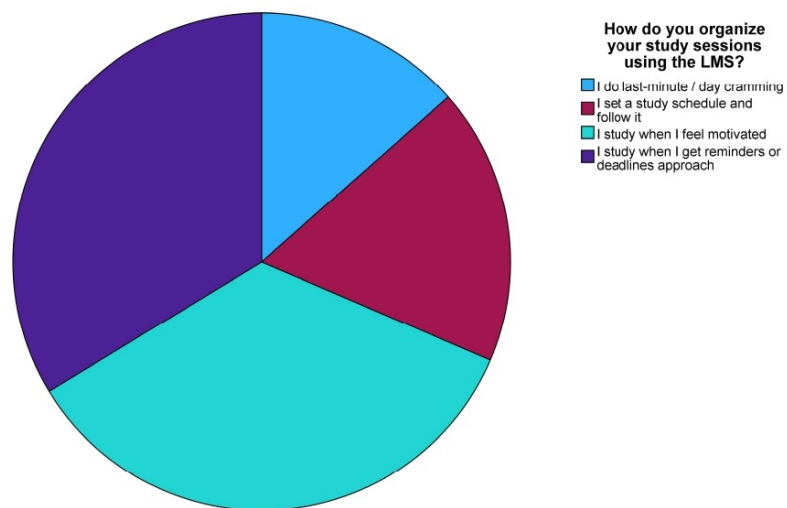


Figure B.6: Preliminary Survey - Q6 Response Distribution

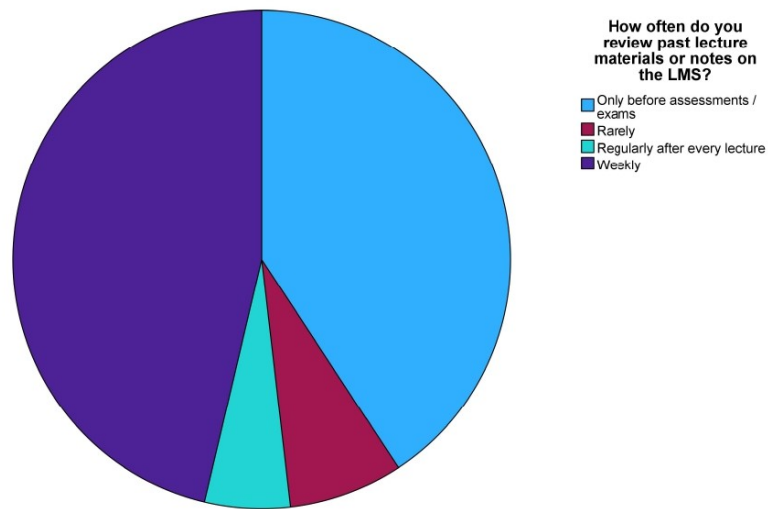


Figure B.7: Preliminary Survey - Q7 Response Distribution

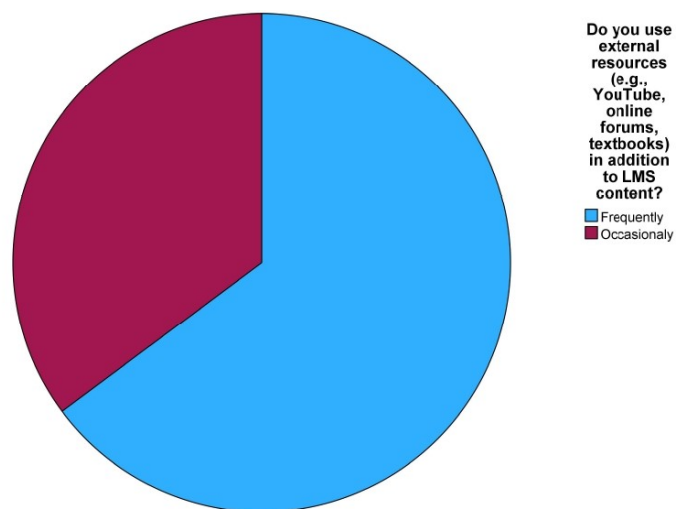


Figure B.8: Preliminary Survey - Q8 Response Distribution

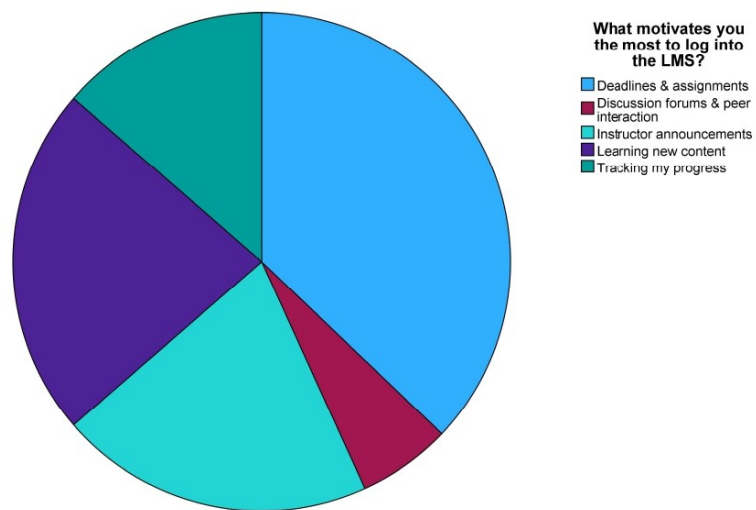


Figure B.9: Preliminary Survey - Q9 Response Distribution

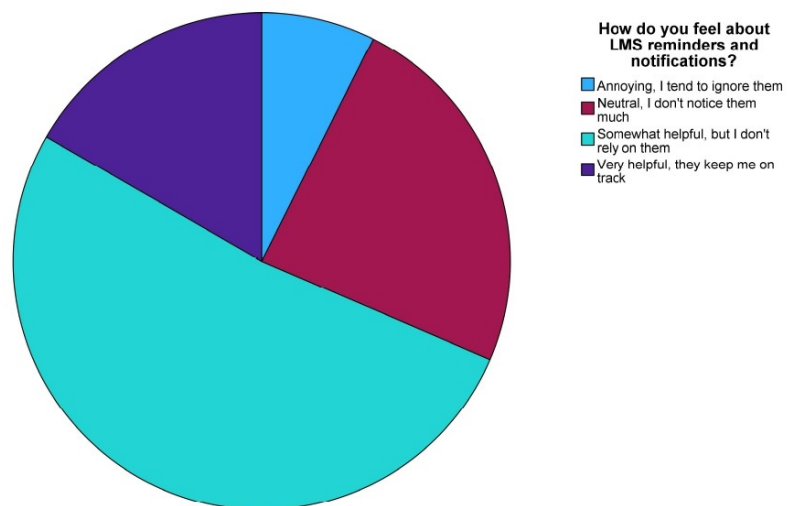


Figure B.10: Preliminary Survey - Q10 Response Distribution

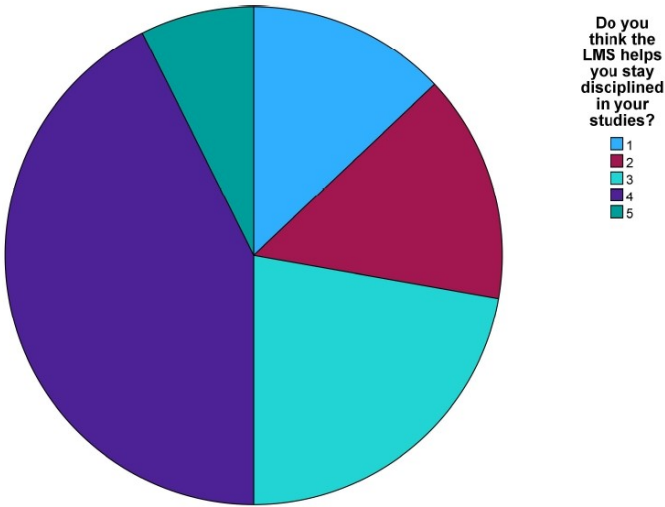


Figure B.11: Preliminary Survey - Q11 Response Distribution

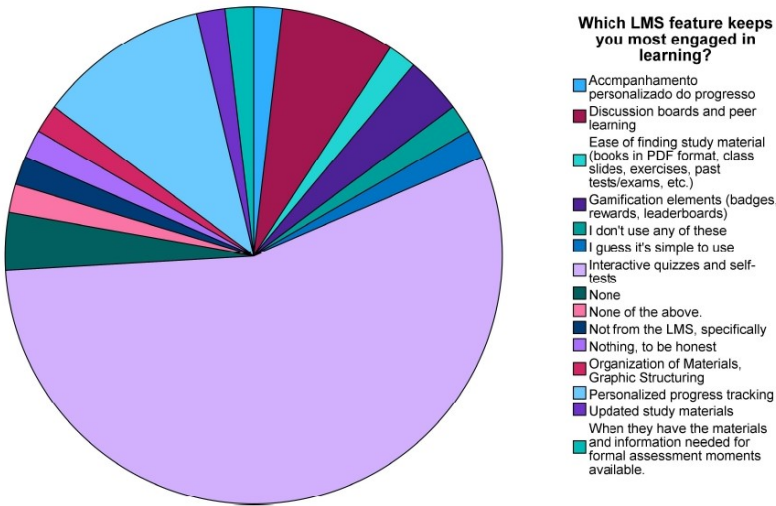


Figure B.12: Preliminary Survey - Q12 Response Distribution

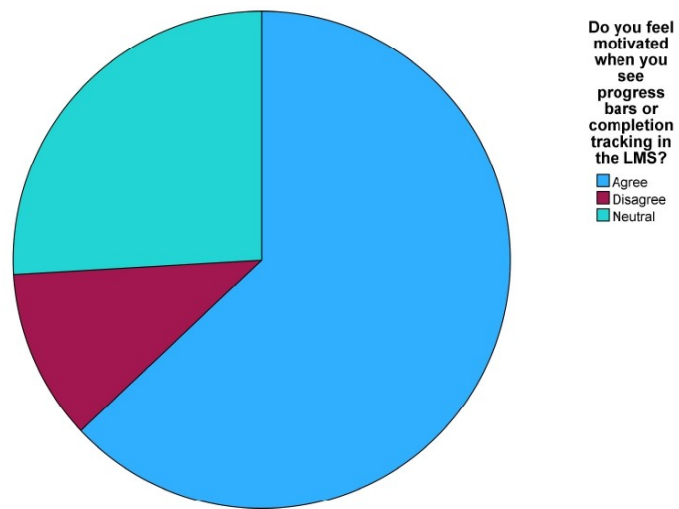


Figure B.13: Preliminary Survey - Q13 Response Distribution

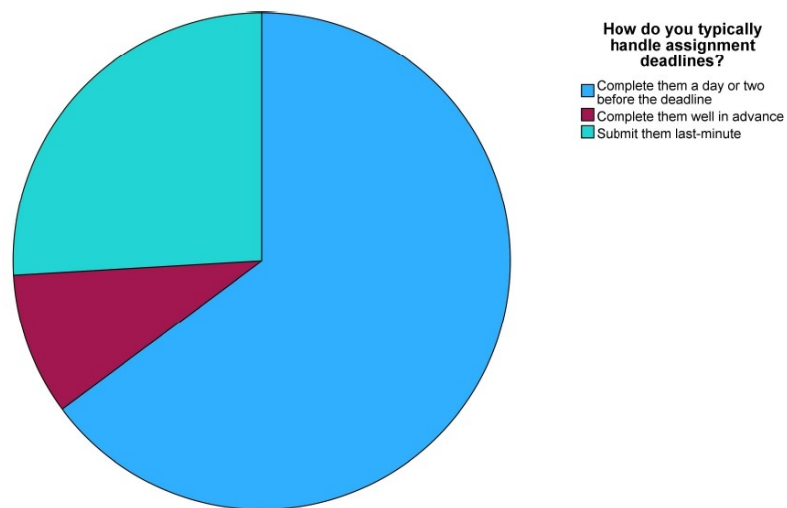


Figure B.14: Preliminary Survey - Q14 Response Distribution

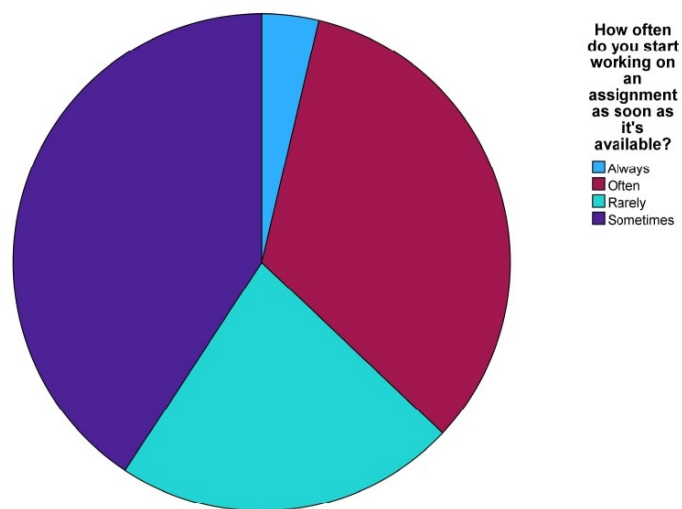


Figure B.15: Preliminary Survey - Q15 Response Distribution

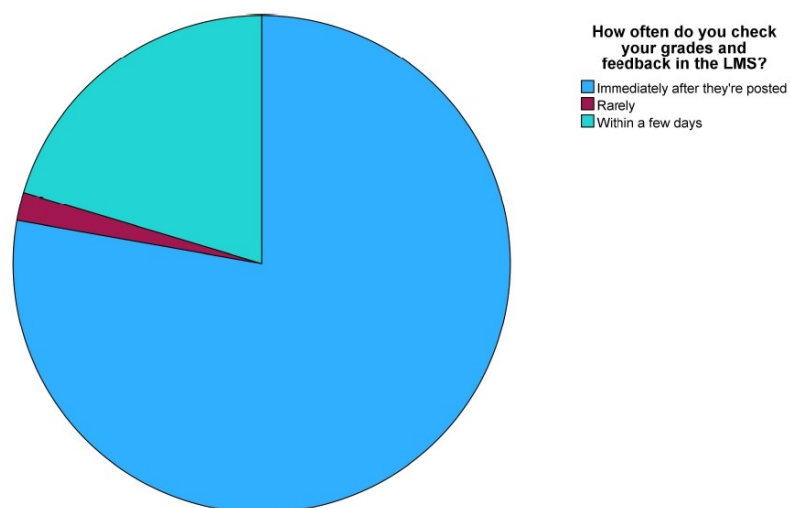


Figure B.16: Preliminary Survey - Q16 Response Distribution

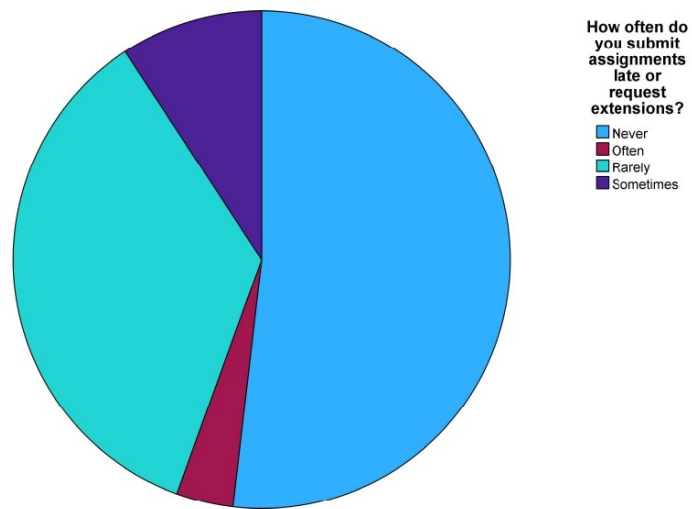


Figure B.17: Preliminary Survey - Q17 Response Distribution

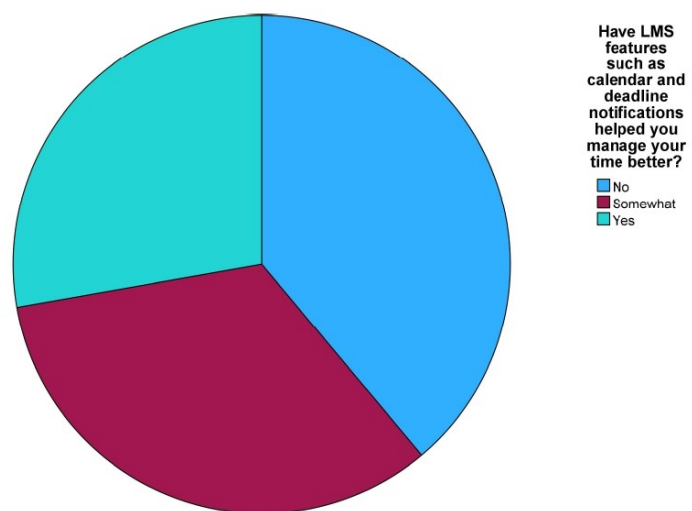


Figure B.18: Preliminary Survey - Q18 Response Distribution

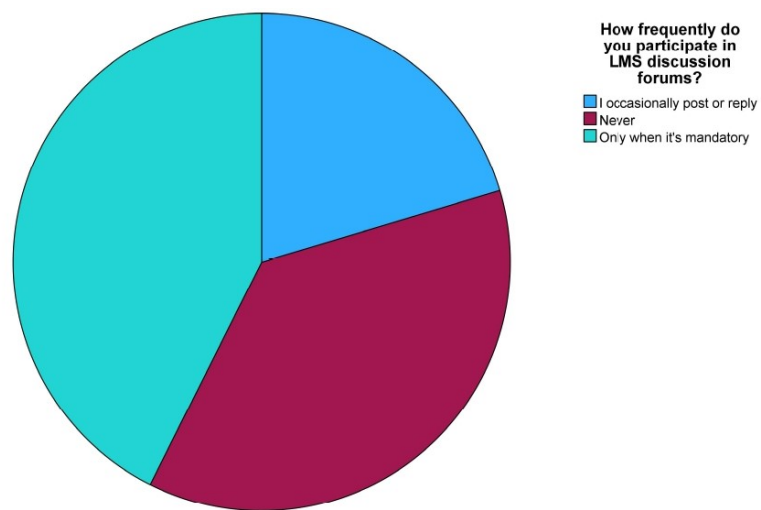


Figure B.19: Preliminary Survey - Q19 Response Distribution

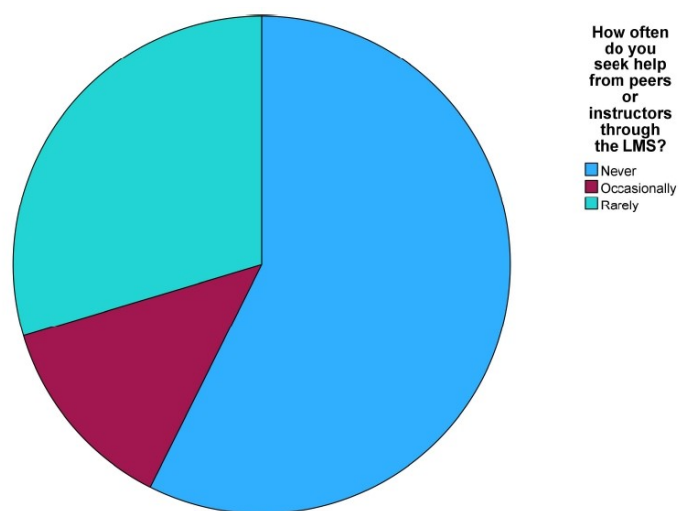


Figure B.20: Preliminary Survey - Q20 Response Distribution

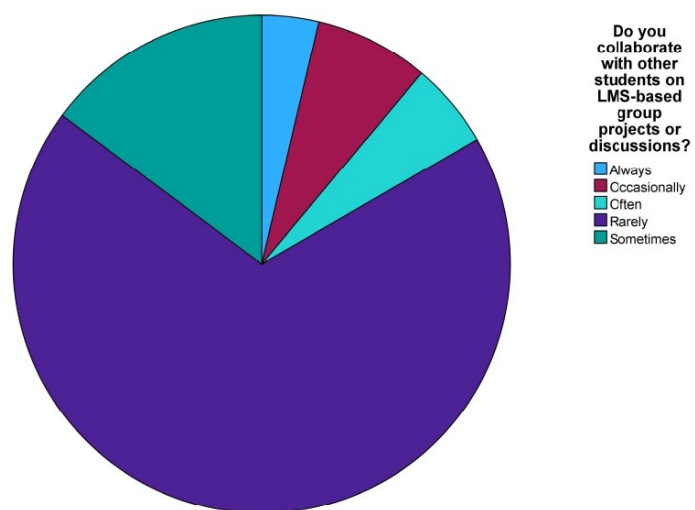


Figure B.21: Preliminary Survey - Q21 Response Distribution

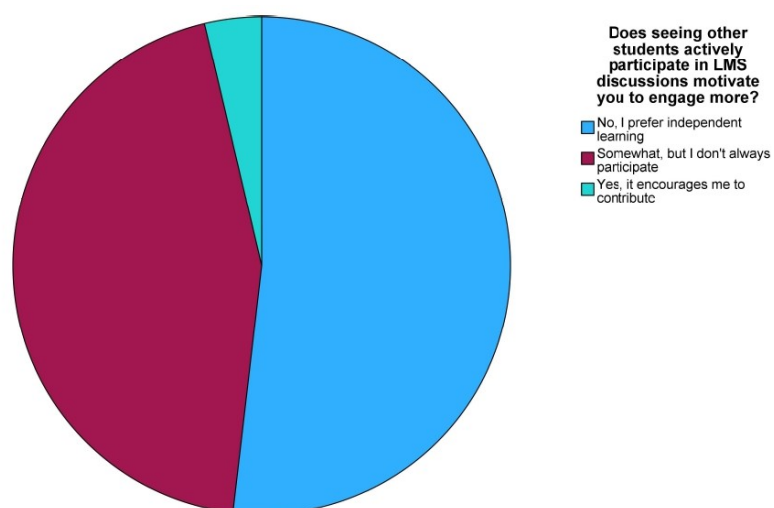


Figure B.22: Preliminary Survey - Q22 Response Distribution

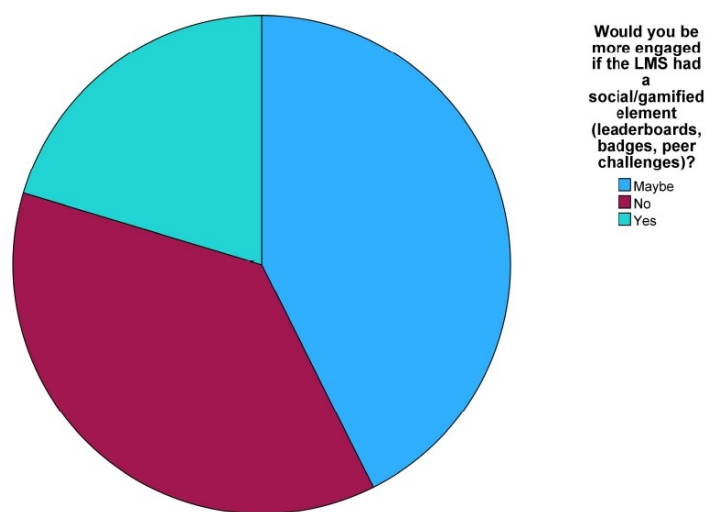


Figure B.23: Preliminary Survey - Q23 Response Distribution

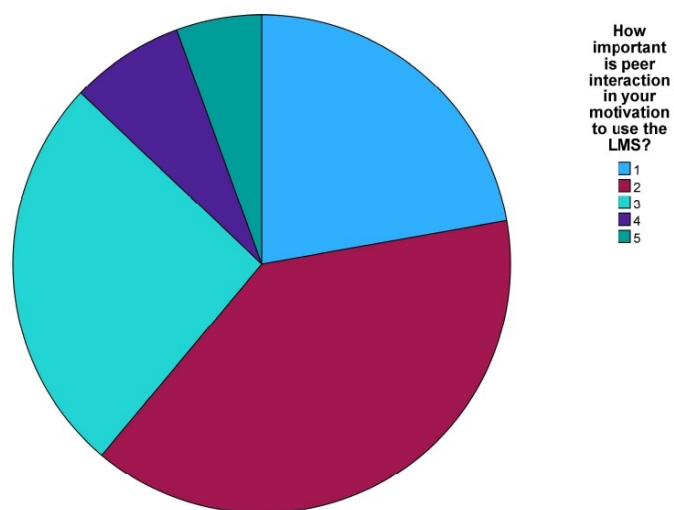


Figure B.24: Preliminary Survey - Q24 Response Distribution

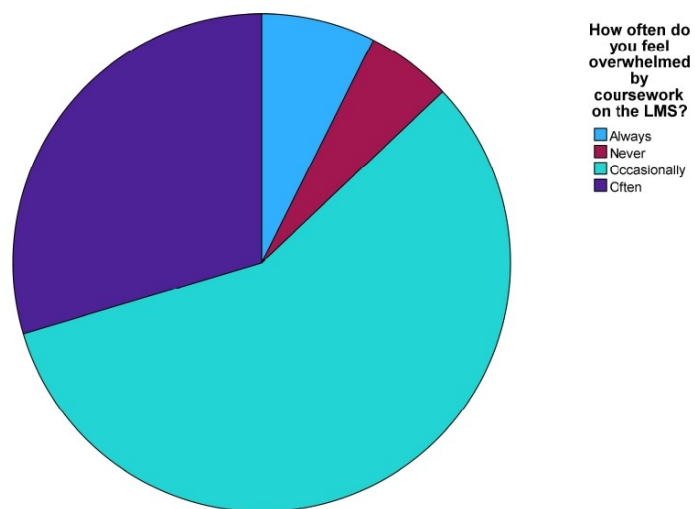


Figure B.25: Preliminary Survey - Q25 Response Distribution

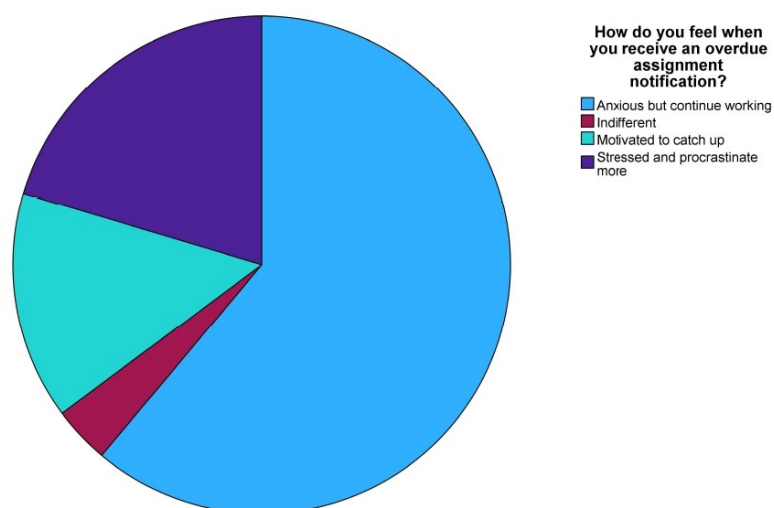


Figure B.26: Preliminary Survey - Q26 Response Distribution

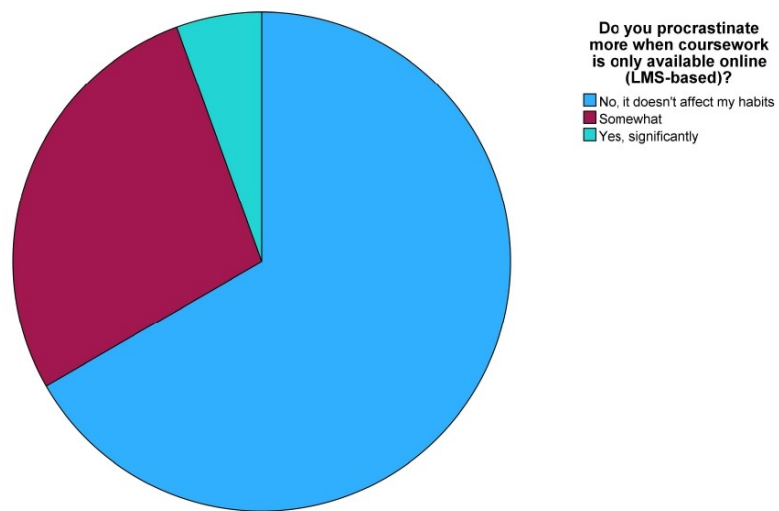


Figure B.27: Preliminary Survey - Q27 Response Distribution

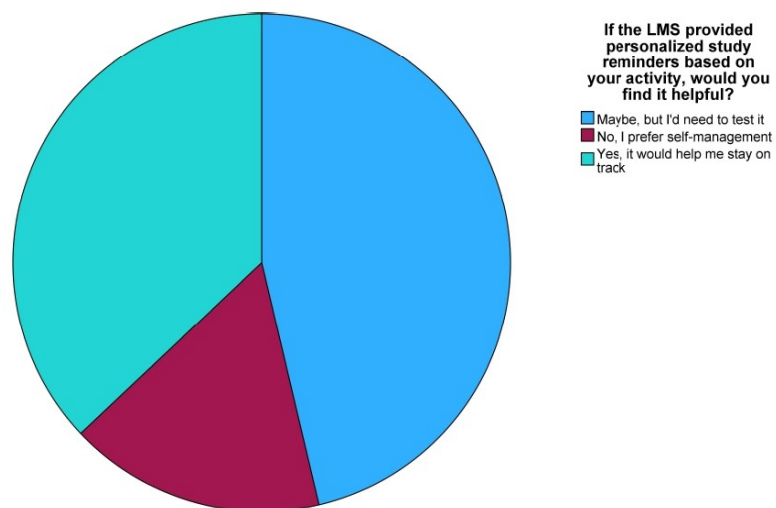


Figure B.28: Preliminary Survey - Q28 Response Distribution

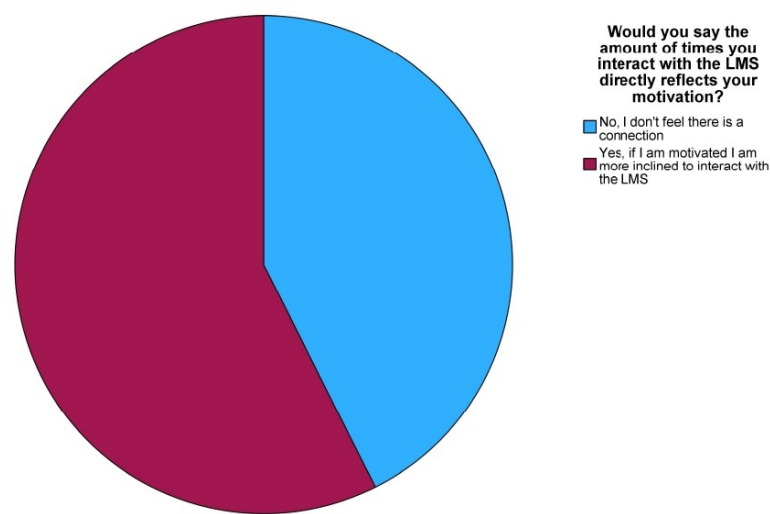


Figure B.29: Preliminary Survey - Q29 Response Distribution

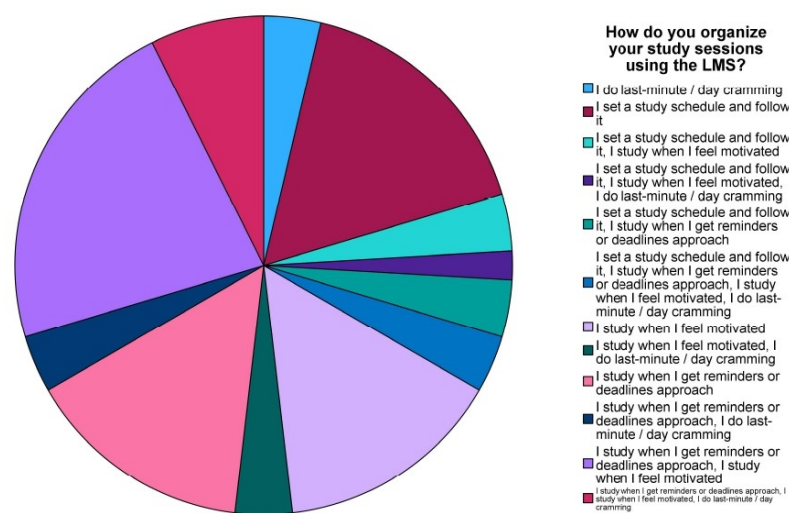


Figure B.30: Preliminary Survey - Q30 Response Distribution

Table B.1: Preliminary Survey - Q31 Response Distribution (Part A)

Would you say LMS plays a huge role in your study sessions and why?
Yes, because it helps having all resources from every class in the same place
A bit. Even though I only use it to transfer the files, or sometimes to solve some Moodle Model Exams, it is more organized than the Documents on SIGARRA, especially when by week allowing to know what and in which order I should see the documents
Yes, it is usually well organised, allowing me to easily catch up on all the topics, contributing to the efficiency of my study sessions, while also being a good way of keeping in touch with professors to clarify any questions.
Yes, it's where the teachers make the course content available, necessary for exams and projects.
Yes, because it helps us organize our information sources all in one place.
They are important to keep everything accessible and organised
Not really, it's just an alternative to a place where the things from the classes are posted
It plays a role in getting professors' notifications and access to the course content. Back in the pandemic, all the recordings of classes were available, which was extremely useful. But other than that, no more impact is made. It is just a communication channel and an official repository for the class's lecture notes and assignment submissions.
LMS if well used by the professors can be time saving tools that gather everything you need for your study sessions.
yes
Yes. Being a working student I could rarely attend the physical classes otherwise.
It helps alot since it's an easy and organised way to check the classes' materials. I struggle to manage with classes that only use sigarra's contents tab.
Comfortable way of storing/finding info about the subjects

Table B.2: Preliminary Survey - Q31 Response Distribution (Part B)

A maior parte das unidades curriculares não tiram o proveito que deviam da plataforma, portanto normalmente o conteúdo está na página de Conteúdos no Sigarra em vez de estar no Moodle. Mas quando os professores usam a plataforma, aí sim tem um grande papel. Adoro principalmente como fazem as questões dentro do moodle.
Sim, porque consigo ter maior parte dos materiais de estudo no msm lugar e organizados
Yes since most of the time it is where teachers put the material for students to study for their curricular unit.
Yes, because it helps centralize the course materials in an organized and structured way.
Yes
Having lesson material easily available helps a lot
sim, é onde estão todos os materiais fornecidos pelos docentes, onde acompanho a matéria e a minha principal fonte de estudo
Yes. Having materials on Moodle helps me keep track of what I have already studied and what I still need to study. Also, if professors organize the content properly, studying becomes easier.
Probably? But I couldn't name a direct role it plays in my study beyond providing me materials.
Yes, it does. The LMS is a great tool because it gathers all the study resources in one place, making them easily accessible for review. Announcements on Moodle also appear in student emails, which is very convenient. More specifically, it influences how I plan my study sessions—for example, by reviewing slides first and then practicing with exercises.
Y, otherwise the course's materials would be scattered around, making it harder for us students to access it.
yes because it gives us all the material needed to study
It helps remember what as lectured in classes and usually its a great source of resumes and exercises
Yes, in order to study according to the study plan of each curricular unit (in order to stay within the scope of it).
more or less
It's important to find the class materials but sometimes can find better resources on those contents elsewhere(depending how well structured the course is)
Yes, because it contains the class materials.
It's only used if professors upload class summaries there; that's the only functionality it has for me. Otherwise, I don't miss Moodle if the professor doesn't upload class content.
No.
The LMS plays an important role in my studies depending on how the course page is structured by the professor. If the professor takes advantage of the platform's features, students can access not only class materials but also books, useful links, articles, videos, etc.
Yes, because it's where we access the content made available by the professors.
Yes, it's where professors upload all the study materials.

Table B.3: Preliminary Survey - Q31 Response Distribution (Part C)

Depending on the course, yes, it plays a very important role in my study, as it helps me understand which parts of the material are most important to study, what type of exercises are usually used in exams, and also allows me to review content from theoretical classes and prepare in advance for practical or lab sessions.
Yes.
Yes, because the most important study materials are found there.
Sharing materials between students and professors.
Yes, because it provides the selected content, and sometimes professors upload additional exercises there.
Yes, for convenience and more available exercises.
Yes, largely due to the organization it enables.
Yes, it gathers all the course information in one place in an easy-to-use way.
Only in the sense that it has the content uploaded there. Other than that, it doesn't help at all with studying.
Yes, because I have direct access to class and course materials from the professor teaching a given subject.
Yes, as all study materials are available there in a more visually appealing and organized manner.
Yes, because it's through them that I can better prepare for assessments.
Yes, as I can always check and even save useful documents to my computer/phone for my studies. Unfortunately, not all professors upload/organize information on Moodle the same way, but still, it's great to have a single place where I can find my study material for all the courses I'm enrolled in.
It offers timely, comprehensive, and concise information on needed materials as well as course announcements.
Yes, slides are a solid foundation.
Because it contains the slides/exercises from class, and while I usually prefer studying new concepts through external resources, I use Moodle to ensure that what I learn from those resources aligns with the course's approach.
Yes, it allows me to access exercises and content provided directly by professors.
At the moment no, because I don't need any information from Moodle for my thesis, except for the deadlines and presentation dates. But I do think it's important if I have a course for which I need access to slides/material/theoretical content to study.
Yes, because the content is there.

Table B.4: Preliminary Survey - Q32 Response Distribution (Part A)

What is your primary challenge in staying motivated while using the LMS?
Distributing the effort through all assignments/classes
Some Moodle are not well organized making everything harder
Procrastination
My study motivation is not related to the LMS itself
I get easily distracted with other things on my phone, laptop, etc, when logging in the LMS.
When it doesn't test my knowledge, I feel that at the end of the study session I spent time reviewing slides for nothing
Bad organization of contents
Most of the time, it is just a list of content, like PDF attachments, that I need to read to get ready for an exam or project. It can be un motivating to see that list of materials.
The contents themselves
Most of the time the activitives and deadlines doesnt reflect the reality
No challenge. It just helps to be updated with the university without any specific communication.
My study habits aren't consistent enough to make full use of the LMS resources at times.
.
Again, my teachers usually don't use Moodle. If they did staying motivated would be a lot easier, since I tend to do any Moodle specific assignments immediatelly.
Falta de concentração
The LMS layout not being very user-friendly.
None in particular I think.
Idk
overwhelming
Para mim é apenas uma fonte de informação, como se fosse uma drive. acho que esta questão não se aplica
First, the will to study. Second, I'm more of a pen-and-paper person, and since I don't have a tablet, studying on my PC screen bores me more quickly than usual.
It's the content itself, not the system.
Sometimes, the material of the discipline gets crammed in there without much organization, making it unnecessarily difficult to find specific information. Sometimes, information about deadlines is actually hidden, and one must hear about them through colleagues.
When is a group project, sometimes only a student needs to deliver the project. So, the members of the group that did not have to submit it, receive a warning saying that the deadline is missed, so an anxious feeling sets before realizing that the work has actually been delivered on time.
i don't know
I don't think I really have one.
deadlines :D

Table B.5: Preliminary Survey - Q32 Response Distribution (Part B)

If I can do it anytime, why do it now? Just leave it for later
I don't use it as a source of motivation.
The GUI of my university's LMS (Moodle). I feel disappointed with the tools for tracking course progress; I wish it provided more data about my learning progress.
The fact that I have to study
Once again, it depends on how the instructors structure things, which is not consistent.
I don't know
Sometimes it's not very convenient when I want to search for things or classmates on the platform, and it would be useful to have exam dates synchronized and appearing on Moodle.
Not getting distracted by deadlines or notifications from other courses
Not feeling overwhelmed by the amount of information provided upfront.
Online distractions
Avoiding stress when seeing the large amount of content
Time it takes me to complete studying each topic
The practicality of the available content, and I think forums are interesting, but since few people engage (when it's optional), I end up not feeling very motivated
The fact that professors/courses don't have a standardized format. Each professor uses a different structure—some with PDFs, others with external links, others with nothing at all... it's a mess.
None
.
I don't identify any. I don't even know where to find the forum or social interaction space on the LMS, but in-person discussion forums would certainly be interesting.
Studying in a way that allows me to get the best grade I can, even if it's not necessarily the highest possible
Getting good grades
Being able to keep my studies balanced across all subjects, since only a small fraction of them post frequently on the LMS
N/A
.
The absurd amount of files per course, but the FEUP Moodle overall doesn't help students at all—it's simply a kind of file drive.
Avoiding opening other more appealing sites
I only see the LMS as a platform for submitting assignments and checking materials or information needed for assessments. I don't think there's an incentive to use it for anything else, but I also don't know if there needs to be.
Dealing with having too much to do

Appendix C

Usability Survey Responses

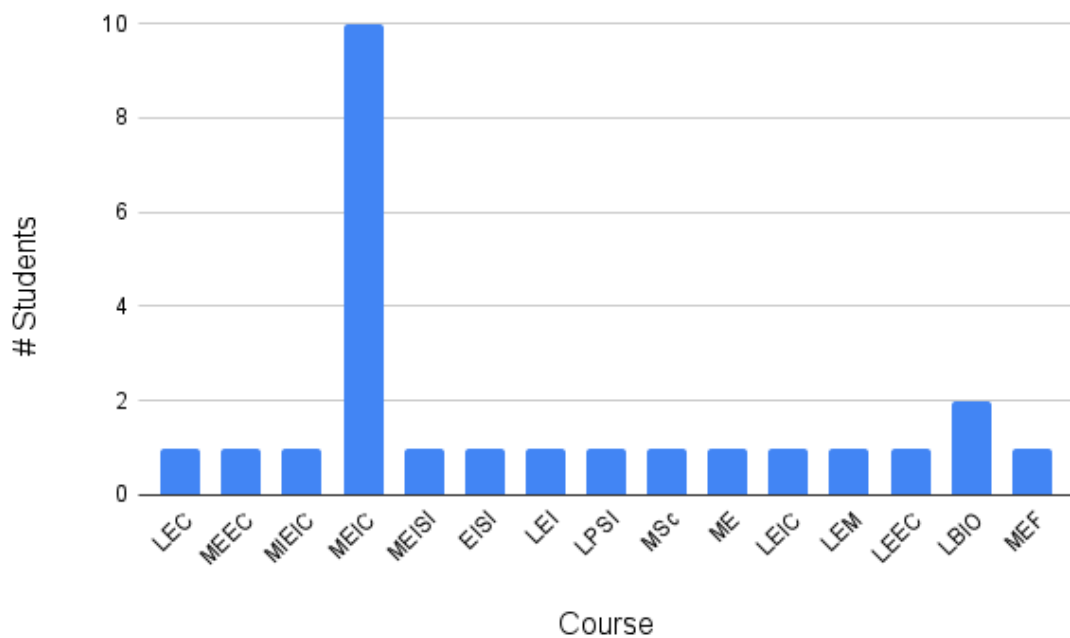


Figure C.1: Usability Questionnaire - Course - Student Response Distribution

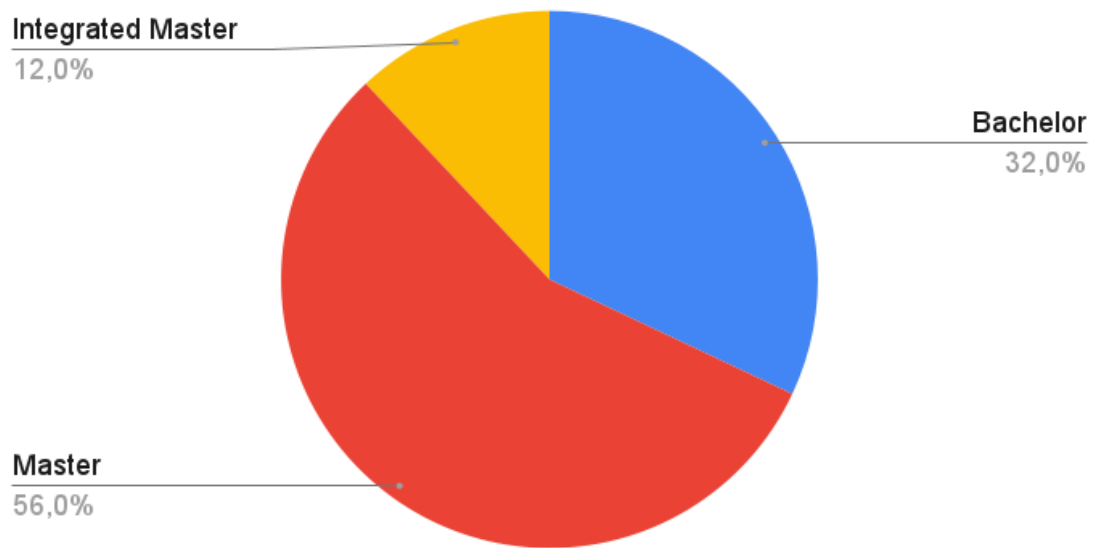


Figure C.2: Usability Questionnaire - Degree - Student Response Distribution

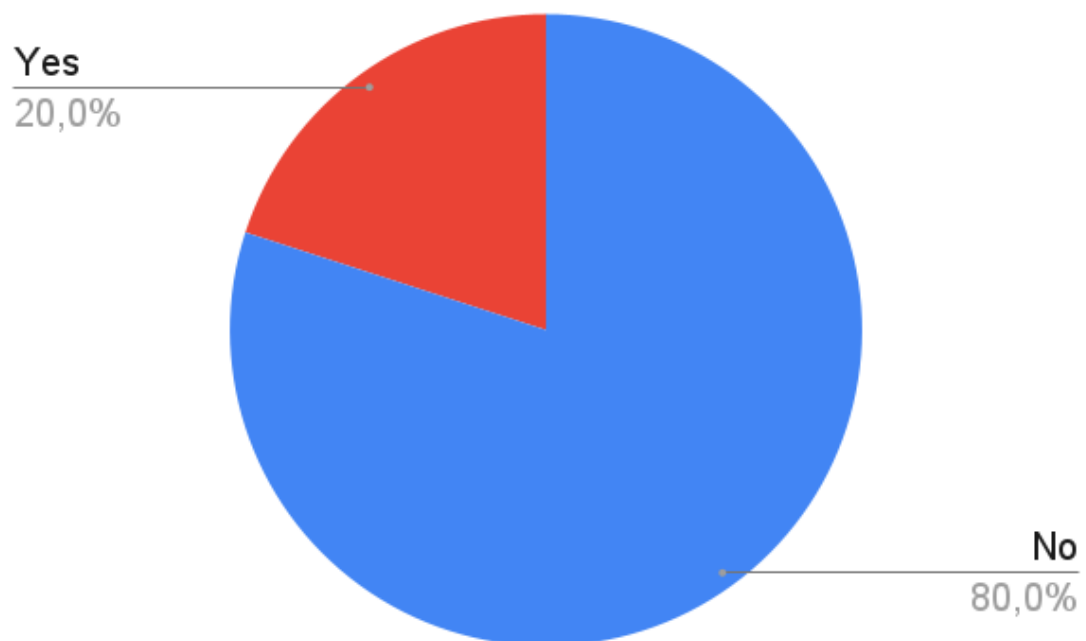


Figure C.3: Usability Questionnaire - "Are you a mobility student?" - Student Response Distribution

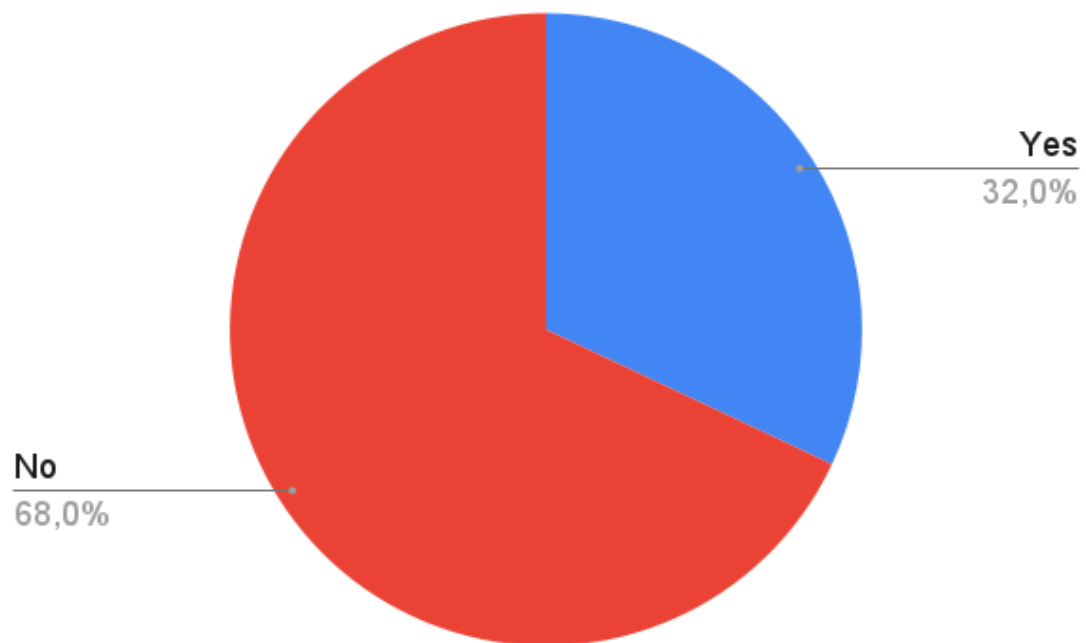


Figure C.4: Usability Questionnaire - "Are / Were you enrolled in any Lifelong learning course?" - Student Response Distribution

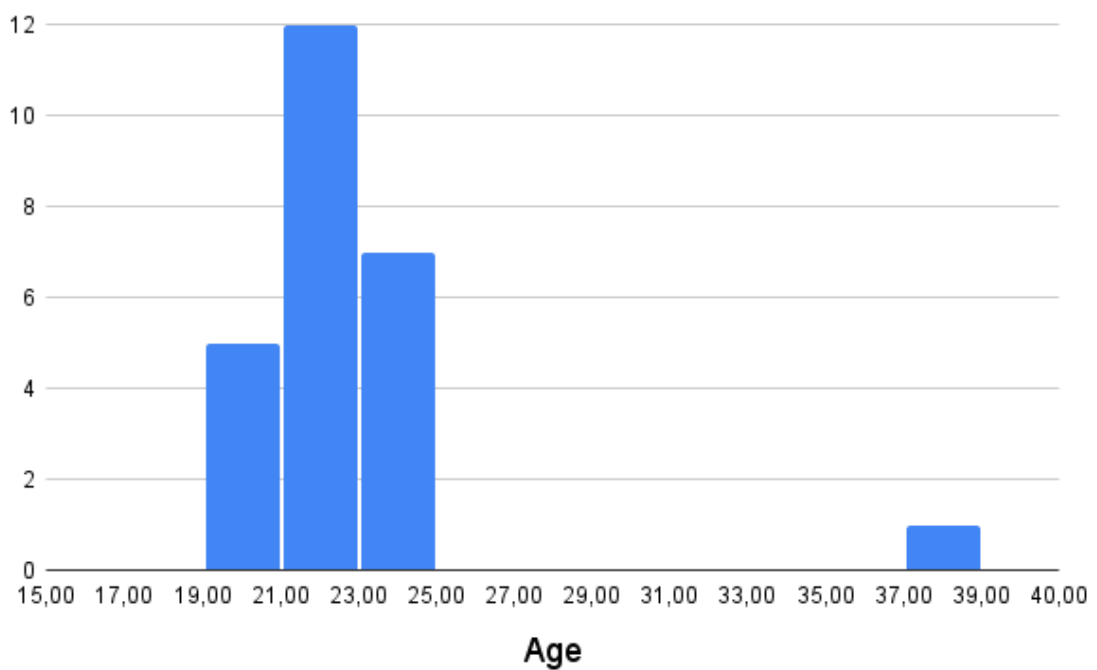


Figure C.5: Usability Questionnaire - Age - Student Response Distribution

Table C.1: Usability Questionnaire - "Could a prototype like this help you in your academic life?"
- Student Response Distribution

Could a prototype like this help you in your academic life?
I believe so, but it might give me a little anxiety.
Sim
will be very useful, but it still needs more UI improvements
Of course. It would help me to keep track the my current and target status.
Yes!
i think it would be very helpfull
Yes! especially if integrated with sigarra (more info). It is nice and useful to have some framework to help us track our goals and grades in a easier way
Possibly. Didn't quite understand how the predictions work.
Maybe
I think it could work in favor of it, but at the same time I think that if my predicted motivation for the course was low, I would start the course somewhat discouraged
Nao creio
Yes
Predicting grades looks nice. I didn't understand the point of the motivation but, depending on what it was, it could be helpful
Sim, bastante
Sim
Yes, I believe it could be useful to track my academic performance
yes
I dont think so
Yes
I have some difficulty deciding which subject is my priority and organising my studies in general.
Yes
Não sei, acho que é algo interessante mas precisava de ser mais intuitivo e dinâmico. É algo que eu procuro diariamente, por isso usaria.
Not sure
yes
Yes definitely

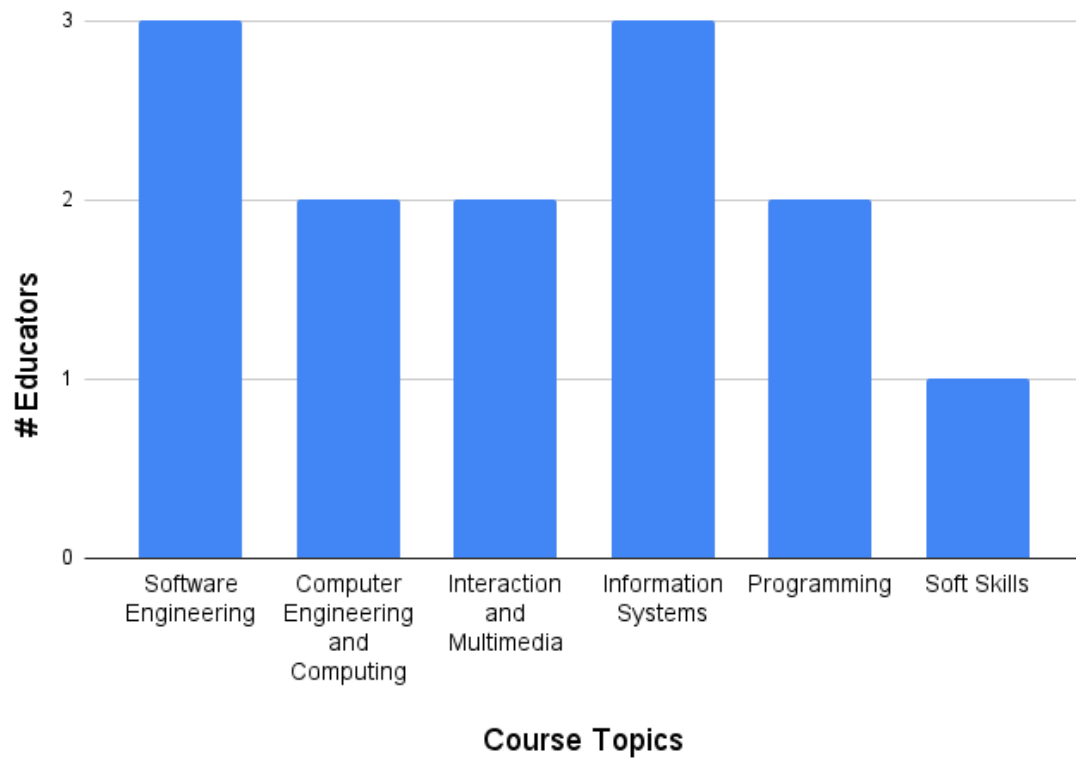


Figure C.6: Usability Questionnaire - "The courses you usually teach are framed in what topics?"
- Educator Response Distribution

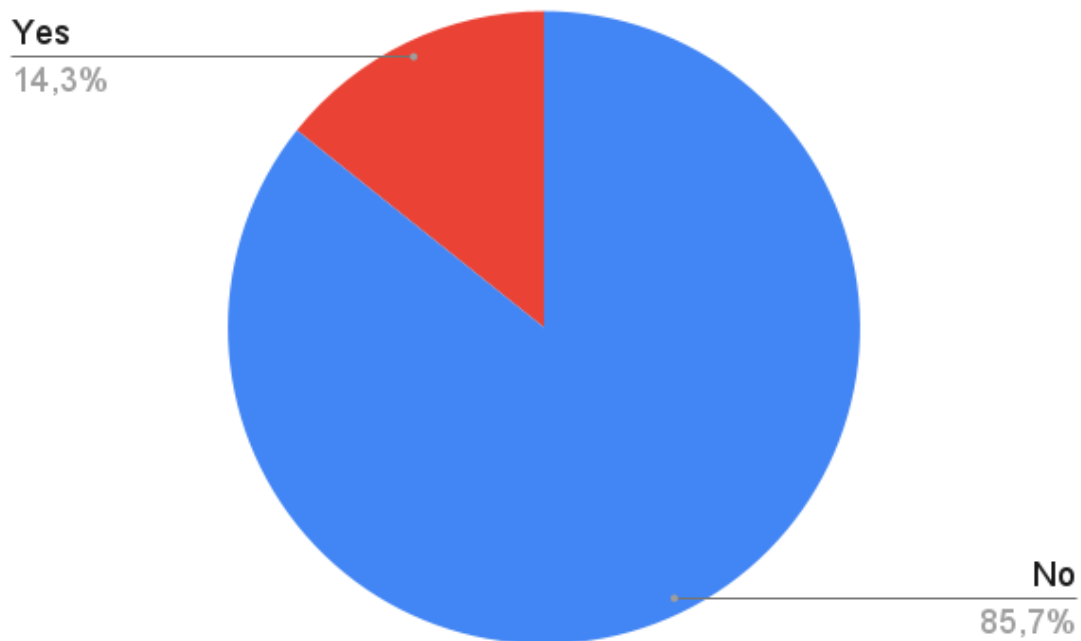


Figure C.7: Usability Questionnaire - "Have you had any experience with a Learning Analytics Dashboard? (Learning Management Systems such as Moodle and Canvas do not apply)" - Educator Response Distribution

Table C.2: Usability Questionnaire - "Could a prototype like this help you or your students in your / their academic life?" - Educator Response Distribution

Could a prototype like this help you or your students in your / their academic life?
Definitely. From an educator perspective this is quite useful. Keep improving it and expanding it with more metrics.
I would have to see it working on a real setting, but it looks like an idea that has some promise.
es
Yes.
It could be useful for me to follow the students' progress more closely/frequently. Not being a psychologist, I cannot really tell whether it would be effective in helping the students. Probably, it would.
I think it could help, although transparency on how the "predicted" data was generated would be critical
Possibly yes, but the UI needs much work. This prototype seems rather initial and to be more convincing, it would need much more attention to design. Regardless, if the features are working, the conceptual idea is promising.

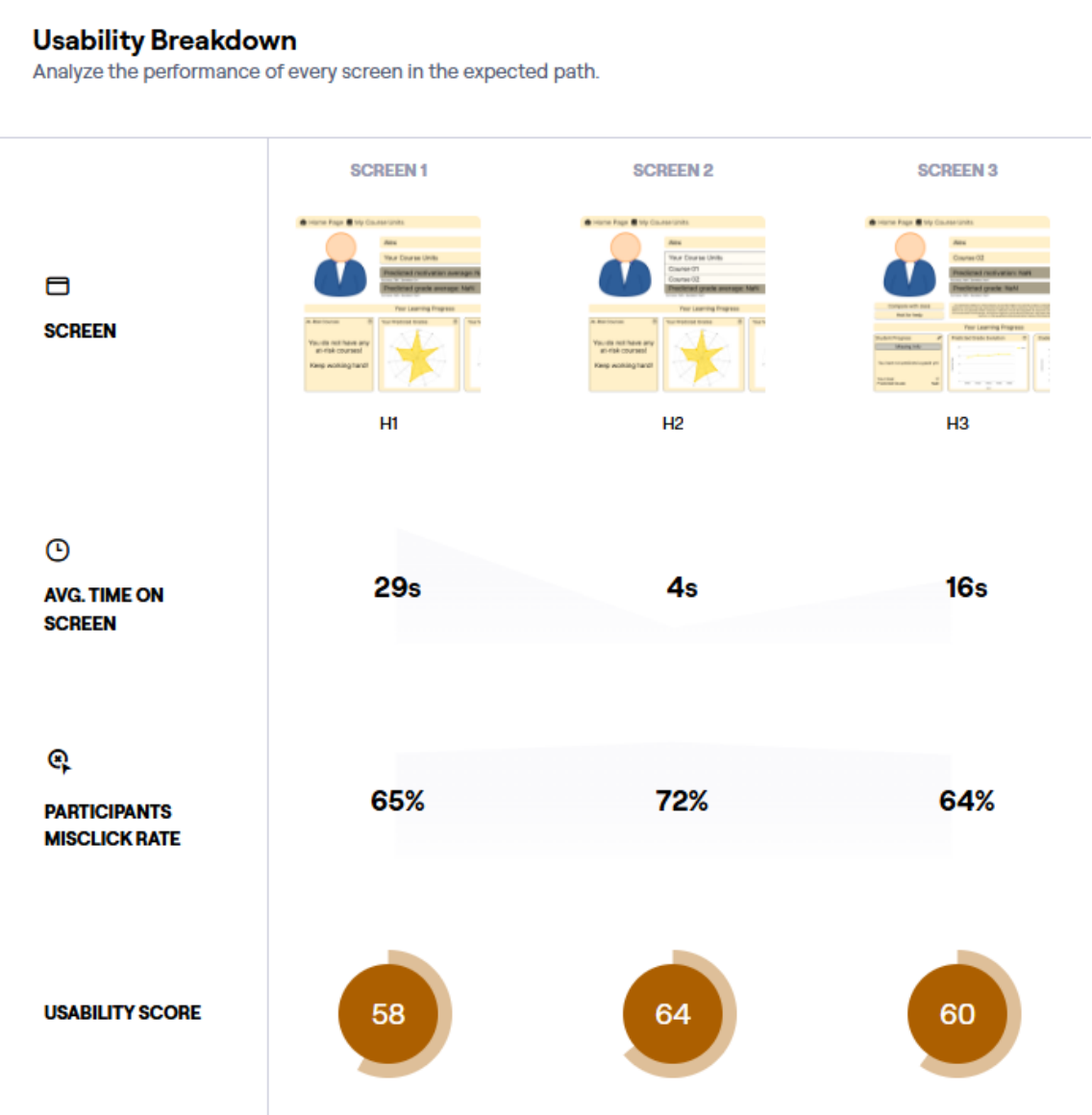


Figure C.8: Task 1 - Student Performance - Maze Breakdown

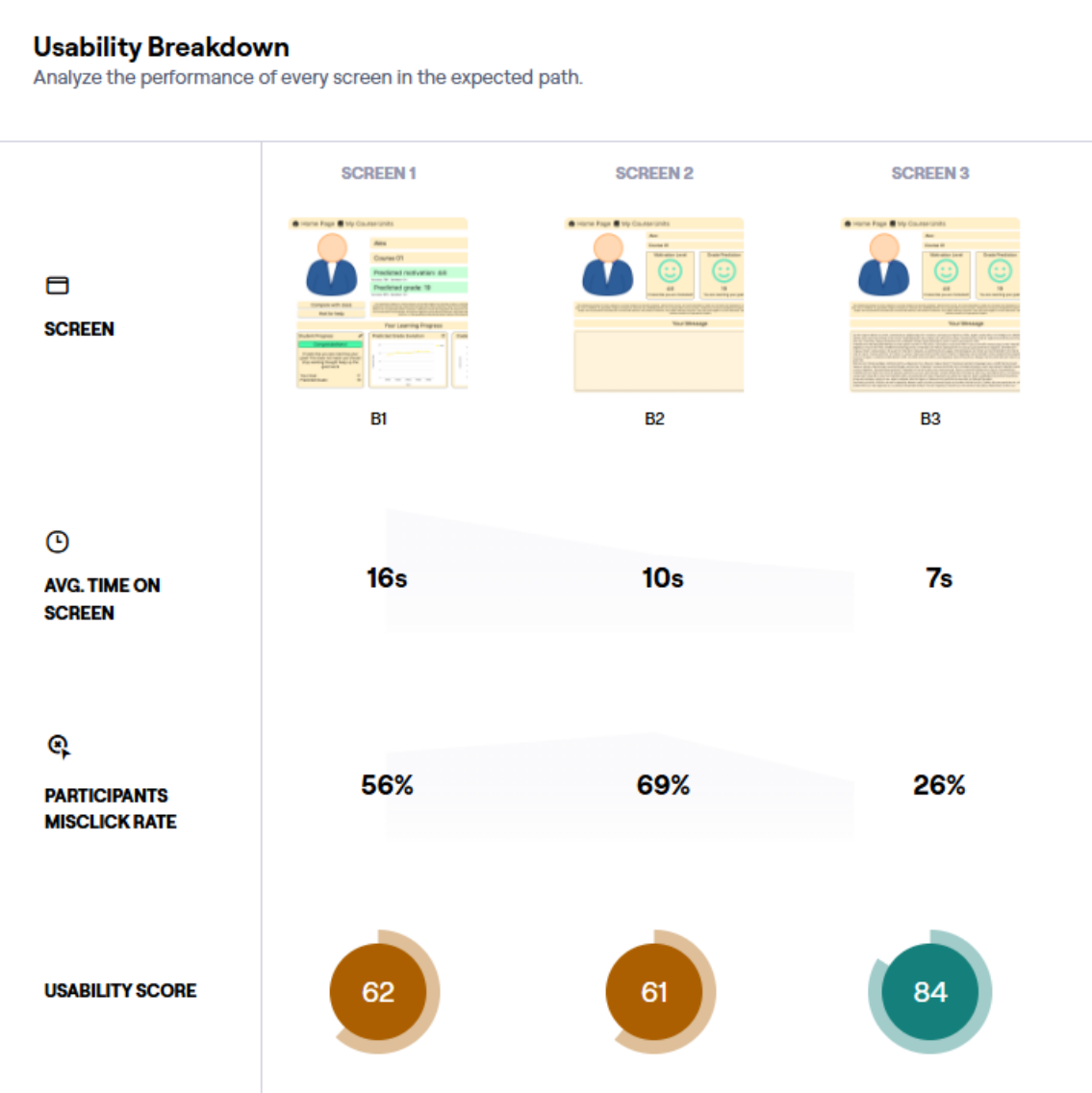


Figure C.9: Task 2 - Student Performance - Maze Breakdown

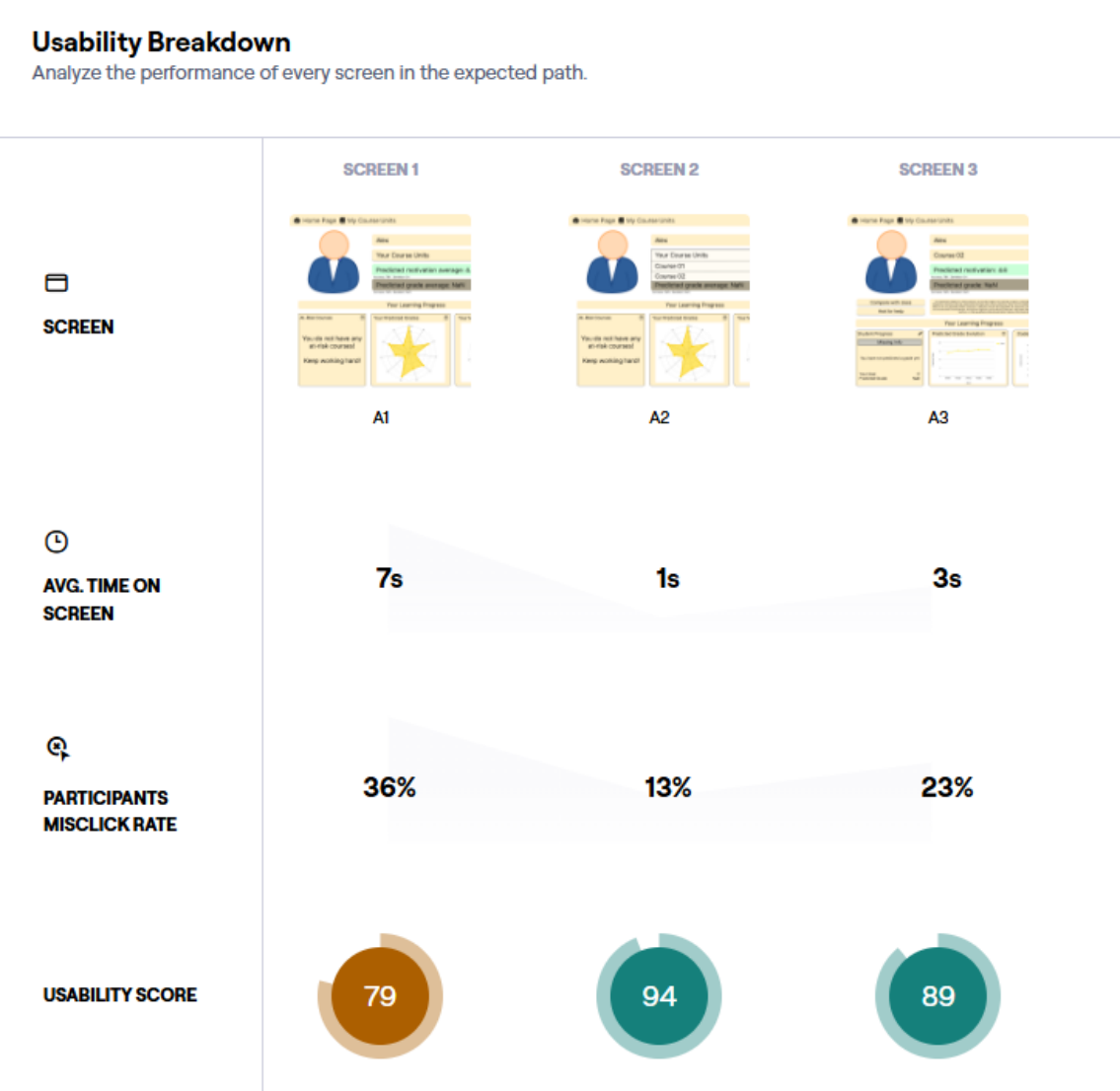


Figure C.10: Task 3 - Student Performance - Maze Breakdown

Usability Breakdown

Analyze the performance of every screen in the expected path.

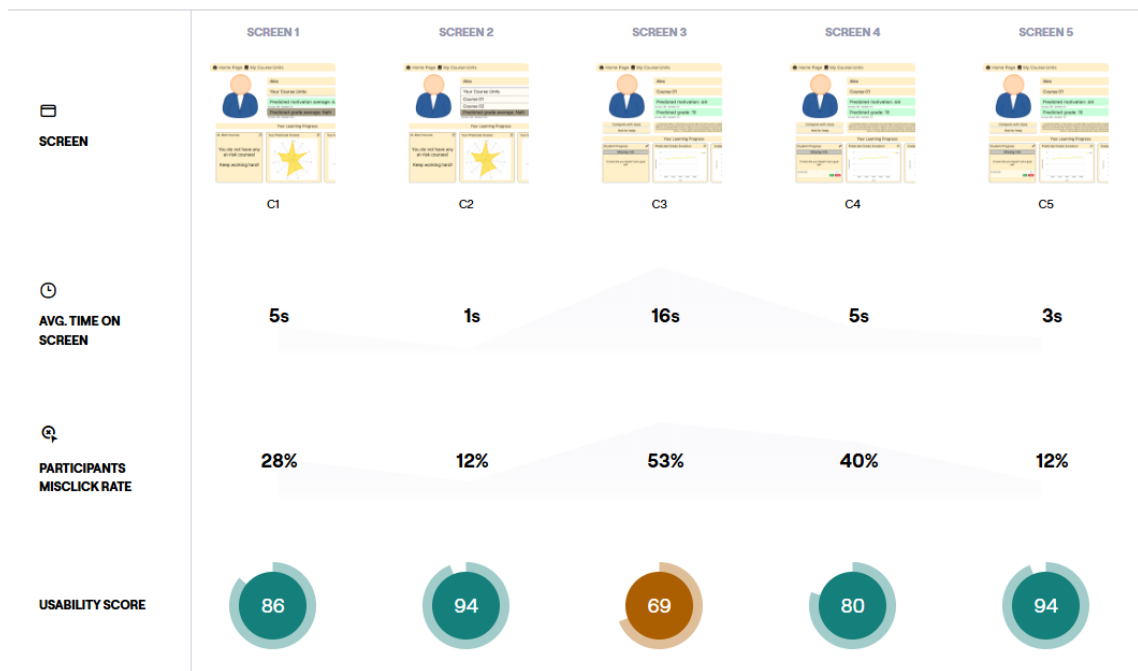


Figure C.11: Task 4 - Student Performance - Maze Breakdown

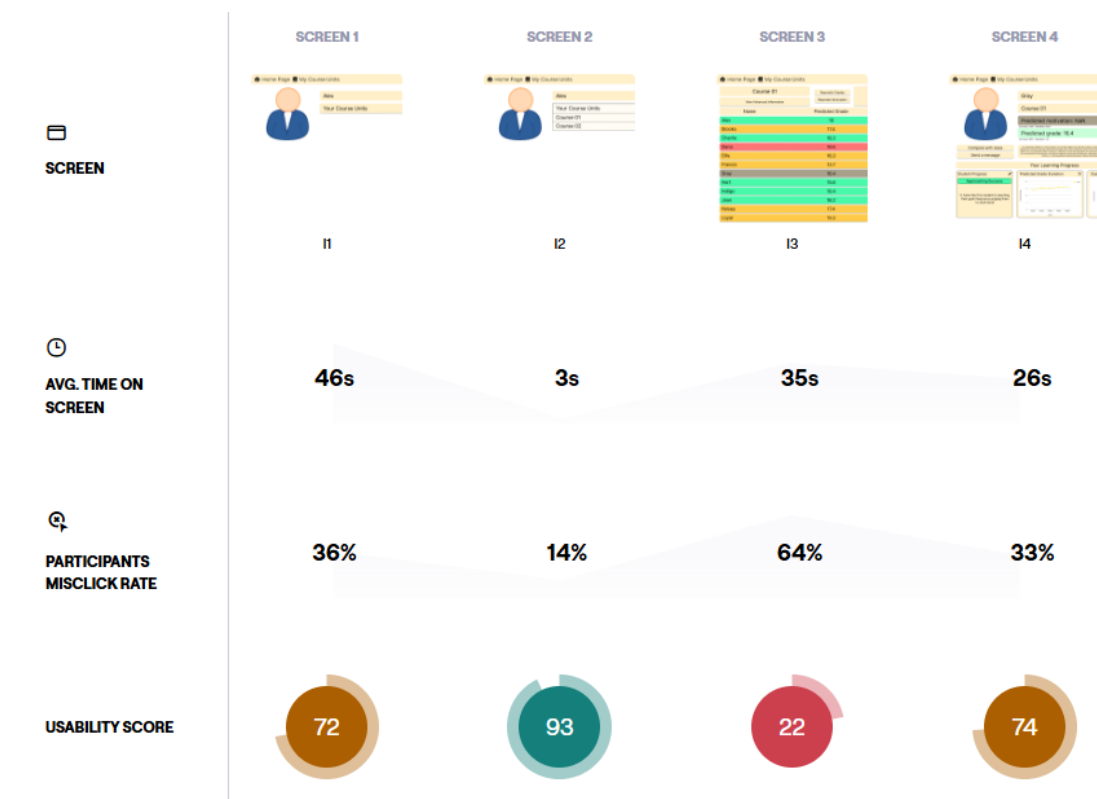


Figure C.12: Task 1 - Educator Performance - Maze Breakdown

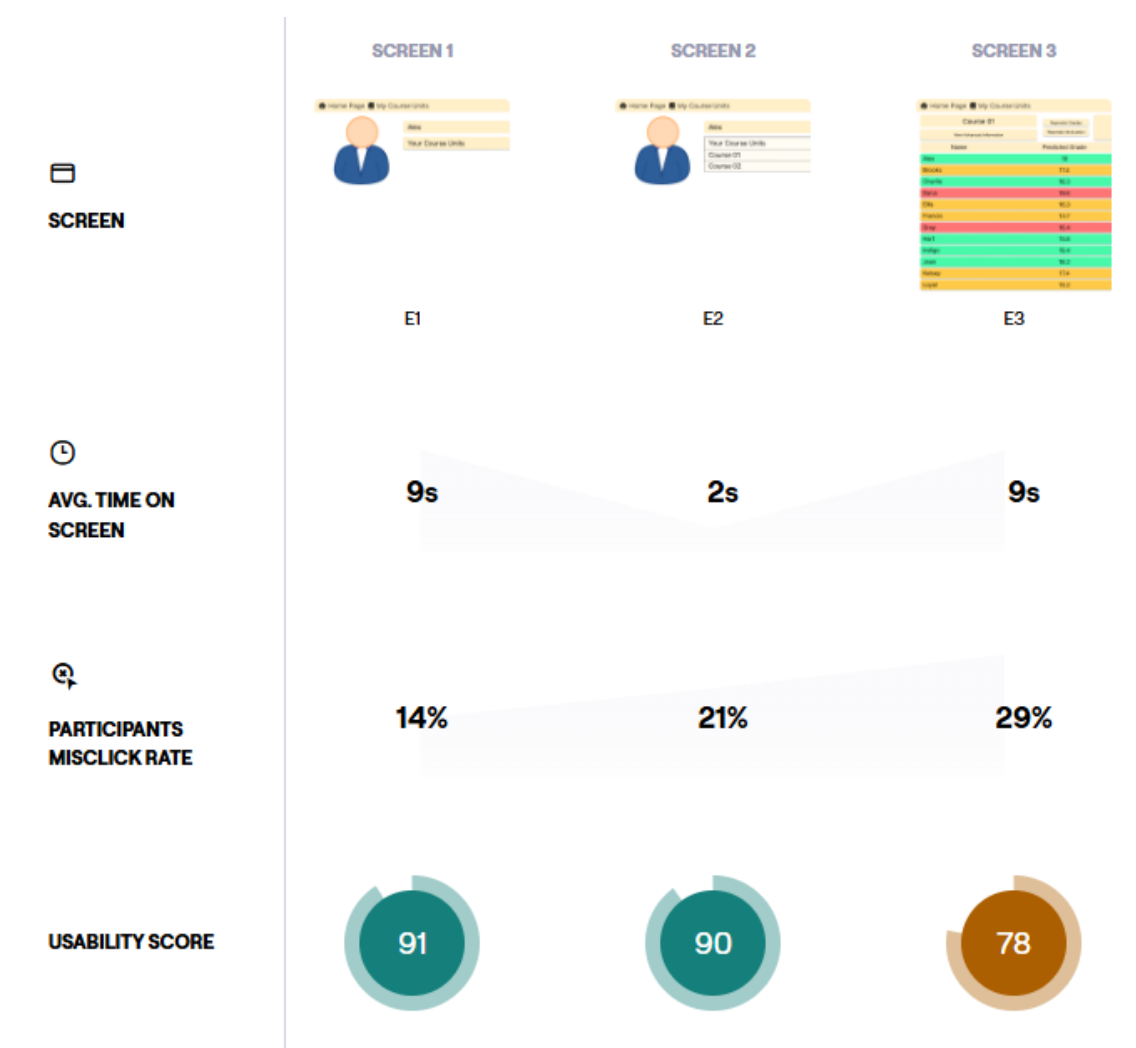


Figure C.13: Task 2 - Educator Performance - Maze Breakdown



Figure C.14: Task 3 - Educator Performance - Maze Breakdown



Figure C.15: Task 4 - Educator Performance - Maze Breakdown



Figure C.16: Task 5 - Educator Performance - Maze Breakdown