# Predicting student performance through cross-institutional learning analytics : development of the CISE model and ReflectMate tool

Dynil Duch

COLLEGE DOCTORAL PAYS DE LA LOIRE | ED MATHEMATIQUES STIC

Le Mans Université

# THÈSE DE DOCTORAT DE

LE MANS UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Informatique*

Par
**Dynil DUCH**

# Predicting Student Performance through Cross-Institutional Learning Analytics: Development of the CISE Model and ReflectMate Tool

**Thèse présentée et soutenue au Mans, le 21/05/2025**
**Unité de recherche : Laboratoire d'Informatique de l'Université du Mans (LIUM)**
**Cambodia Academy of Digital Technology (CADT)**
**Thèse N° : 2025LEMA1008**

## Rapporteurs avant soutenance :

| | |
|---|---|
| Armelle BRUN | Professeure des universités, Université de Lorraine |
| Agathe MERCERON | Professeure des universités, Université de Berlin |

## Composition du Jury :

| | | |
|---|---|---|
| Président : | Yvan PETER | Professeur des universités, Université de Lille |
| Examinateurs : | Armelle BRUN | Professeure des universités, Université de Lorraine |
| | Agathe MERCERON | Professeure des universités, Université de Berlin |
| | François BOUCHET | Maître de conférences, Sorbonne Université |
| | Yvan PETER | Professeur des universités, Université de Lille |
| Dir. de thèse : | Sébastien GEORGE | Professeur des universités, Le Mans Université |
| Co-enc. de thèse : | Madeth MAY | Maître de conférences, Le Mans Université |

# ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, **Madeth May**, for his unwavering support, guidance, and encouragement throughout my PhD journey. His insights, feedback, and continuous motivation have been invaluable to the development and completion of this dissertation. I am sincerely thankful for his mentorship, which has not only enriched my academic experience but also shaped me as a researcher.

I am also deeply grateful to my thesis director, **Sébastien George**, whose expertise, thoughtful advice, and persistent encouragement have been crucial to the successful completion of this work. His insightful comments and suggestions have significantly contributed to the quality of this dissertation, and I am fortunate to have had the opportunity to work under his direction.

I would like to extend my heartfelt thanks to the **Cambodia Academy of Digital Technology (CADT)** and the **Bourses du Gouvernement Français (BGF)** for providing the financial support that made this research possible. The scholarship, along with the allowance provided by **CADT**, has been instrumental in allowing me to focus on my studies and research without financial concerns.

Special thanks go to the **Laboratoire d'Informatique de l'Université du Mans (LIUM)**, where I am currently pursuing my PhD. The laboratory has provided a stimulating research environment, and I am grateful for the resources and support that have been made available to me.

I am also grateful for the collaboration and support from **École Nationale Supérieure d'Ingénieurs du Mans (ENSIM)**, where additional data collection took place. The

data gathered from **ENSIM**, **CADT**, and **Institute of Technology of Cambodia(ITC)** were critical to the empirical research conducted in this dissertation. I am deeply appreciative of the students at **CADT** and **ITC** who participated in the surveys. Their cooperation and willingness to contribute to this study are greatly appreciated. Without their participation, the empirical component of this dissertation would not have been possible.

I would like to acknowledge the support and collaboration of my colleagues and peers, who provided me with valuable feedback, assistance, and encouragement throughout the various stages of this research. Their camaraderie and shared experiences have made this journey both enriching and enjoyable.

Lastly, I thank my wife, family and friends for their constant support, patience, and understanding during this challenging yet rewarding process. Their encouragement and belief in my abilities have been a source of strength and motivation for me.

To everyone who has been a part of this journey, thank you.

# Contents

Chapter 1

# Introduction

## 1.1 Background and Motivation

In the digital transformation era, integrating technology into educational environments has dramatically changed the teaching and learning landscape (Costaa, Alvelosa, and Teixeiraa 2015). The proliferation of digital tools such as Learning Management Systems (LMS), online learning platforms, and educational applications has resulted in the generation of vast amounts of data (Susanto, Chen, and Almunawar 2018). This trend has been significantly accelerated by the global COVID-19 pandemic, which necessitated a rapid shift to remote and hybrid learning environments. During this period, educational institutions increasingly relied on digital technologies to deliver instruction, assess student performance, and maintain engagement. As a result, the volume and diversity of educational data have grown exponentially, providing new opportunities to gain insights into student behavior, learning outcomes, and the effectiveness of educational interventions (M. Khan et al. 2023; Madiah and Mohemad 2023). This data includes information about student interactions, assessments, discussion participation, and overall course content engagement. Analyzing this data can provide invaluable insights into student behavior (M. Khan et al. 2023), learning outcomes (Madiah and Mohemad 2023), and the effectiveness of educational interventions (Asif, Merceron, S. A. Ali, et al. 2017; Z. Pan et al. 2024).

Two key fields have emerged to harness these insights: **Educational Data Mining**

**(EDM)** and **Learning Analytics (LA)**. While these terms are often used interchangeably, they differ in focus and methodology. According to Calvet Liñán and Juan Pérez (Liñán and Pérez 2015), **EDM** is primarily concerned with developing computational methods to extract patterns and knowledge from educational datasets, often using machine learning and statistical techniques. In contrast, **LA** focuses on measuring, collecting, analyzing, and reporting data about learners and their contexts to optimize learning and the environments in which it occurs. LA emphasizes real-time feedback and actionable insights for educators and students. In contrast, EDM tends to prioritize predictive modeling and algorithmic discovery.

This dissertation bridges both fields by leveraging the strengths of **EDM** for predictive modeling and **LA** for reflective learning analytics. Specifically, the research develops a **Cross-Institutional Stacking Ensemble (CISE)** predictive model rooted in EDM methodologies to forecast student performance across diverse educational settings. Our research, **Student performance** refers to the measurable outcomes of a learner's academic activities, typically evaluated through assessments such as grades, test scores, course completion, or learning achievements. In this research, student performance serves as the primary **dependent variable (Grade A, B+, B, C+, C, etc..)** used to evaluate the effectiveness of predictive models and learning interventions, including the application of reflective learning tools (RLT) within the learning analytics (LA) framework. Complementing this, we introduce **ReflectMate**, a reflective learning analytics tool inspired by LA principles, which empowers students with personalized feedback and fosters self-regulation. The research demonstrates how EDM and LA can work synergistically to enhance educational outcomes by integrating these two approaches. Figure 1.1 illustrates the integration of EDM and LA in our research. On the left, EDM focuses on predictive modeling and computational techniques, exemplified by the CISE model. On the right, LA emphasizes real-time feedback and student empowerment through ReflectMate. The overlap demonstrates how predictive insights from EDM inform ReflectMate, fostering enhanced student outcomes and addressing challenges like the digital divide.

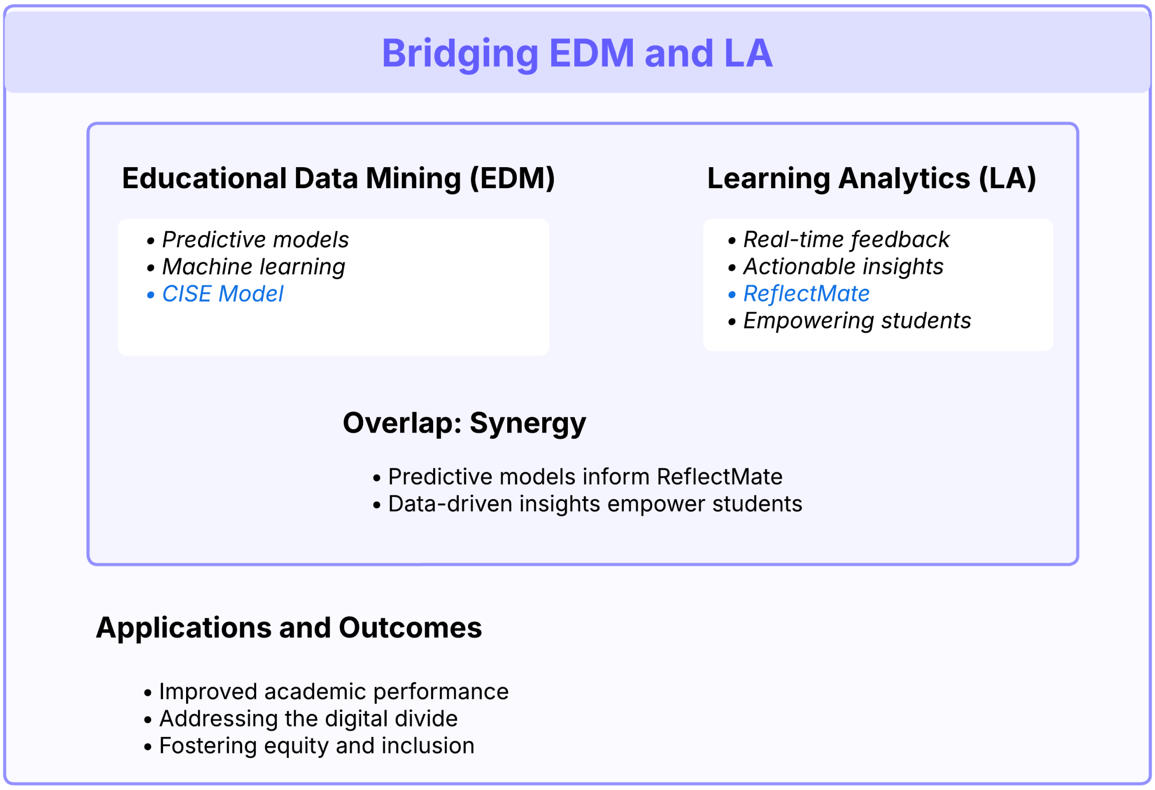

Bridging EDM and LA
Educational Data Mining (EDM)
• Predictive models
• Machine learning
• CISE Model
Learning Analytics (LA)
• Real-time feedback
• Actionable insights
• ReflectMate
• Empowering students
Overlap: Synergy
• Predictive models inform ReflectMate
• Data-driven insights empower students
Applications and Outcomes
• Improved academic performance
• Addressing the digital divide
• Fostering equity and inclusion

Figure 1.1 – Bridging EDM and LA

The transformative potential of EDM and LA in education is immense (Romero and Ventura 2020). These tools can uncover patterns and trends that traditional methods might miss. By leveraging machine learning algorithms, statistical models, and data visualization techniques, educators and researchers can predict student performance (Enughwure and Ogbise 2020; Asif, Merceron, and Pathan 2014), personalize learning experiences (Alam 2023), and develop targeted interventions for at-risk students (Al-Shabandar et al. 2019). The growing interest in predictive modeling in education is driven by its ability to forecast outcomes such as student grades, retention rates, and long-term career success.

However, applying predictive modeling in education has challenges because educational data is inherently complex, characterized by its high dimensionality, noise, and variability (Bai et al. 2021). Moreover, the educational context is highly dynamic, with student behavior influenced by many factors, including socioeconomic background, prior knowledge, motivation, and the quality of instruction. As a result, developing predictive models that are both accurate and generalizable across different educational settings remains a significant challenge (Gardner et al. 2023).

While the technical challenges of predictive modeling in education are significant, the ethical considerations demand our attention (Jones 2016). Using student data for predictive purposes raises important questions about privacy, consent, and potential bias and discrimination (W. Li et al. 2022). For instance, a predictive model trained on data from a predominantly affluent student population may perform poorly when applied to a more diverse group of students, potentially reinforcing existing inequalities. Such biases can lead to unfair treatment, disproportionately affecting underrepresented or disadvantaged groups. Furthermore, the collection and use of student data must comply with legal frameworks such as the **General Data Protection Regulation (GDPR)**, which emphasizes transparency, data minimization, and individuals' rights to control their personal information. Respecting these laws is essential to building trust and ensuring that predictive models are developed and deployed responsibly. Addressing these ethical concerns is crucial to realizing the benefits of predictive modeling in education without compromising students' rights, interests, and dignity.

Another critical aspect of modern education that has gained attention is the concept of reflective learning. LA is a powerful approach that substantially supports educators and content creators in enhancing the teaching and learning experiences (Banihashem et al. 2022; Hernández-de-Menéndez et al. 2022). However, while existing tools and services in LA are mostly dedicated to instructors, there is a lack of similar supports that directly empower students (Chatti et al. 2021; Arthars et al. 2019). Nonetheless, it has been demonstrated that students strongly need self-assessment throughout their learning process to gain motivation and higher achievement (McMillan and Hearn 2008; Andrade 2019). Thus, providing reflective tools that allow students to do so is crucial from a pedagogical standpoint and a significant research challenge. Employing reflective learning tools (RLT) as a LA approach offers a powerful combination for enhancing student engagement and outcomes (Duch, May, and George 2024). By providing students with data-driven insights into their learning behaviors, these tools empower them to take control of their learning journey, set goals, and monitor their progress (Joksimović, Kovanović, and S. Dawson 2019; Arthars et al. 2019). This personalized approach to learning aligns with the principles of constructivist learning theories, which emphasize the learner's active role in constructing knowledge through experience and reflection. This empowerment can inspire and motivate students to take charge of their learning.

Despite the potential benefits of predictive modeling and reflective learning, the digital divide remains a significant barrier to their widespread implementation (Bon, Saa-Dittoh, and Akkermans 2024; Vassilakopoulou and Hustad 2023; Nakayama et al. 2023). The digital divide refers to the gap between individuals with access to modern information and communication technologies (ICT) and those without it. This divide is often influenced by socioeconomic status, geographic location, and gender, leading to disparities in access to educational resources and opportunities (Farooqi, Khalid, and A. Khan 2022; Cheshmehzangi et al. 2023). However, the potential of predictive modeling and reflective learning offers hope and optimism for the future of education, inspiring us to address the digital divide and ensure that all students can benefit from the advancements in educational technology.

In many developing countries, the digital divide is a significant educational barrier. Students often lack access to reliable internet connectivity (Yu et al. 2016; West 2015), digital devices (Chetty et al. 2018; Santos, Azevedo, and Pedro 2013), and the necessary infrastructure for online learning (Carlson and Isaacs 2018; Calderón Gómez 2019). This lack of access affects students' academic performance. It limits the effectiveness of predictive models and RLT, as the data collected from these students may be incomplete or biased (Zilka et al. 2021). It is our responsibility and commitment to address the digital divide, ensuring that all students can benefit from the advancements in educational technology and have equal educational opportunities.

## 1.2 Research Problem

The primary research problem addressed in this PhD dissertation is the development of accurate, generalizable, and ethically sound predictive models, particularly in diverse and resource-constrained educational settings. While significant progress has been made in EDM, several challenges that limit the effectiveness and applicability of predictive models in education still need to be addressed.

One key challenge is the **generalizability of predictive models**. Many existing models are developed and tested in specific educational contexts, often within a single institution or a homogeneous group of students. As a result, these models may perform poorly when applied to different educational settings, where the student population, curricular structure, and teaching practices may differ significantly. This lack of generalizability limits the broader applicability of predictive models and their potential to improve educational outcomes on a larger scale.

Another challenge is the **lack of reflective tools** that empower students to engage actively with their learning data. While predictive models can provide valuable insights into student performance, there is a critical need for tools that generate predictions and facilitate student reflection on their learning behaviors and progress. Without such reflective learning tools (RLT), students may passively receive data about their performance without understanding how to adjust their strategies or set meaningful learning goals. Educators, too, face difficulties in fostering self-regulated learning if students are not equipped with tools that encourage reflection and active participation in their educational journey. This gap between predictive insights and actionable, reflective practices hinders the full potential of LA in promoting student empowerment and long-term academic success.

The **ethical implications of predictive modeling in education** also present a significant challenge. Using student data to predict academic performance or potential career paths raises important questions about privacy, consent, and the potential for bias and discrimination. Ensuring that predictive models are fair, transparent, and used responsibly is crucial to protecting students' rights and fostering trust in the use of technology in education.

In addition to these challenges, the **digital divide** poses a significant barrier to effectively implementing predictive models and reflective learning tools (RLT). This dissertation takes the opportunity to examine the digital divide through existing work and empirical data, shedding light on its profound impact on educational equity. For instance, students from low human development areas often lack reliable internet connectivity, digital devices, and the necessary infrastructure for online learning (Yu et al. 2016; West 2015). Similarly, female students and those from underrepresented groups face additional barriers in accessing and succeeding in ICT-related fields (Economic Co-operation and (OECD) 2018; Mumporeze and Prieler 2017). These disparities affect students' academic performance and limit the effectiveness of predictive models and RLT, as the data collected from these students may be incomplete or biased (Zilka et al. 2021).

This research leverages the strengths of **predictive modeling** and **reflective learning analytics** to address these challenges. The **Cross-Institutional Stacking Ensemble (CISE)** model demonstrates its ability to generalize across diverse educational datasets, ensuring that predictive insights are applicable even in resource-constrained environments. By identifying at-risk students, CISE enables targeted interventions to mitigate the digital divide's effects, such as providing additional support or resources to disadvantaged learners.

Complementing this, **ReflectMate** empowers students by translating predictive insights into actionable feedback. By offering personalized recommendations based on engagement metrics, ReflectMate ensures that students from underrepresented groups receive tailored guidance to improve their performance. For example, students with limited access to technology can benefit from ReflectMate's focus on optimizing their interactions with available resources. In contrast, female students in male-dominated fields can use the tool to build confidence and self-regulation skills.

Through this integrated approach, the dissertation demonstrates how predictive models and reflective tools can work synergistically to bridge the digital divide. By addressing disparities in access and engagement, these tools have the potential to create more inclusive and equitable educational environments, ensuring that all students—regardless of their background—have the opportunity to succeed.

Addressing the digital divide is an essential component of any effort to develop and implement effective and equitable predictive models. This dissertation aims to contribute to this goal by developing and validating predictive models that are generalizable across diverse educational contexts, examining the use of RLT as a LA approach, and investigating the impact of the digital divide on the effectiveness of such tools.

## 1.3 Research Objectives

The overarching objective of this dissertation is to develop predictive models and RLT that enhance student engagement and academic performance while addressing the challenges of generalizability, interpretability, and equity in education. The specific objectives of the research are as follows:

1. **To develop and validate generalizable predictive models across different educational institutions.** This objective focuses on developing predictive models that can be applied to diverse educational settings, considering variations in student demographics, curricular structures, and teaching practices. The research will involve cross-institutional validation of the models to ensure their robustness and applicability in various contexts.

2. **To explore integrating RLT as a learning analytics to enhance student engagement and academic performance.** This objective investigates the potential of RLT, supported by learning analytics, to improve students' self-regulation, critical thinking, and overall academic outcomes. The research will involve developing and evaluating RLT that provide students with data-driven insights into their learning behaviors.

3. **To examine the impact of the digital divide on the effectiveness of predictive models and RLT.** This objective seeks to understand how disparities in access to technology and digital literacy affect the implementation and effectiveness of predictive models and RLT. The research will explore strategies for mitigating the impact of the digital divide and ensuring that all students can benefit from these advancements.

4. **To consider the ethical implications of predictive modeling in education as a factor often neglected in research and practice.** While this dissertation does not directly address the ethical implications of predictive modeling, it acknowledges the importance of fairness, transparency, and the responsible use of student data as critical factors often overlooked in developing and deploying predictive models and RLT. By incorporating considerations such as bias, privacy,

and inclusivity into the design and evaluation of these tools, the research aims to contribute to the growing discourse on ethical practices in EDM and LA. This approach ensures that the tools developed are practical and aligned with principles of equity and trust.

## 1.4 Research Questions

The following key questions guide the research:

1. **How can a predictive model be designed to predict student performance across institutions with generalizability?** This question investigates how to build a predictive model that accurately forecasts student performance (e.g., final course grades) while maintaining robustness across diverse institutional contexts. Variations in grading scales, curricula, teaching styles, and data structures across institutions pose significant challenges to model generalizability. To address this, the Cross-Institutional Stacking Ensemble (CISE) model was developed. It integrates multiple machine learning classifiers into a meta-model, allowing the system to capture complex patterns across heterogeneous datasets. This question focuses on the design, training, evaluation, and cross-validation of CISE using datasets from different institutions to ensure its reliability and applicability beyond a single educational context.

2. **How can insights from student performance predictive models be leveraged to design a reflective learning dashboard to enhance student engagement and academic performance?** This question explores the connection between predictive analytics and student empowerment through reflective tools. While most learning analytics systems are instructor-focused, the proposed ReflectMate dashboard is designed specifically for students, using data insights

from models like CISE to provide personalized feedback. The dashboard encourages metacognitive practices—such as self-awareness, self-regulation, and goal setting—by offering indicators like progress tracking, performance trends, and tailored recommendations. Although ReflectMate is currently in prototype design, this research question assesses how its features can improve engagement, academic self-efficacy, and motivation when integrated into the LMS.

3. **How can predictive modeling and a reflective learning dashboard help reduce disparities based on gender and location by providing fair predictions and personalized support for all students?** This question addresses the digital and educational inequalities prevalent in developing countries, particularly those related to gender-based and geographic disparities. By examining how the CISE model performs across different demographic groups and how ReflectMate can offer equitable support, the research aims to assess the tools' potential to promote fairness and inclusivity. The goal is to ensure that underrepresented students, such as females or those from low-resource areas, receive predictions and feedback that are just as accurate and beneficial as those for better-represented groups. This question also considers ethical practices, such as anonymization and bias mitigation, in the development and deployment of both tools and their applications.

4. **What are the ethical considerations associated with the use of predictive models in education, and how can they be addressed to ensure fairness and transparency?** This question explores the ethical issues related to using predictive models in education, including concerns about privacy, bias, and discrimination. It will investigate how developing fair, transparent, and responsible models can mitigate these issues.

## 1.5 Significance and Contributions of the Study

This dissertation contributes to educational data mining and learning analytics by addressing key challenges and advancing our understanding of how predictive models and reflective learning tools (RLT) can enhance educational outcomes. The findings and methodologies presented in this research provide empirical evidence and offer practical solutions for improving student engagement, academic performance, and equity in diverse educational settings. Below, we outline this study's specific contributions and their significance from both scientific and technical standpoints.

1. **Development of Generalizable Predictive Models:** this research contributes to developing robust predictive models across different educational institutions. By focusing on cross-institutional validation, the study provides insights into the factors influencing model generalizability, such as dataset heterogeneity, class imbalance, and feature engineering. The CISE model demonstrates its ability to generalize across datasets with varying characteristics, achieving an F1 score of 78.25% on the ENSIM validation dataset. These results illustrate the potential of ensemble methods to address challenges in educational data mining, offering a technical framework for building scalable and reliable predictive systems.

2. **Leveraging the RLT with LA:** integrating ReflectMate—a reflective learning analytics tool—with predictive modeling provides a novel approach to fostering student empowerment. The first batch of empirical data from ReflectMate highlights its potential to inspire self-awareness, self-regulation, and proactive engagement among students. ReflectMate explains the phenomenon of metacognition in digital learning environments by translating predictive insights into actionable feedback. For instance, personalized progress indicators and recommendation engines enable students to monitor their learning behaviors and adjust their strategies accordingly.

From a technical standpoint, this integration bridges the gap between predictive analytics and pedagogical practices, offering a dual-layered system that supports institutional decision-making and individual student growth.

3. **Addressing the Digital Divide:** this study examines the impact of the digital divide on the effectiveness of predictive models and RLT, particularly in developing countries like Cambodia. Statistical insights reveal disparities in ICT education based on gender and geographic location, underscoring the need for targeted interventions. The CISE model's ability to identify systemic patterns in performance disparities across institutions contributes to a deeper understanding of how structural inequities are reflected in educational data. Likewise, ReflectMate's emphasis on individualized support ensures that students from underrepresented groups receive tailored guidance. Together, these tools demonstrate the potential of data-driven approaches to promote inclusion and equity in education. These contributions align with broader equity educational goals, providing a foundation for future research and policy development.

4. **Ethical Considerations in Predictive Modeling:** while this dissertation does not directly address ethical implications, it incorporates considerations such as fairness, transparency, and bias mitigation as critical factors often neglected in predictive modeling. By emphasizing the importance of ethical practices in designing and deploying predictive tools, the study contributes to the growing discourse on responsible AI in education. For example, techniques like oversampling and careful feature selection mitigate the risk of biased predictions, ensuring that models remain equitable and inclusive. This approach underscores the importance of balancing technical innovation with ethical responsibility.

5. **Empirical Evidence and Practical Recommendations:** the empirical evi-

dence generated through this research offers valuable insights into the effectiveness of predictive models and RLT in enhancing student outcomes. Survey results and self-reported engagement data demonstrate the potential positive impact of reflective tools like ReflectMate on student motivation and accountability, contributing to our understanding of how such tools can foster lifelong learning. From a technical perspective, the study provides actionable recommendations for educators, policymakers, and researchers on leveraging these tools to improve educational practices. For instance, integrating predictive analytics into learning management systems (LMS) can facilitate real-time interventions. At the same time, reflective tools can promote sustained engagement and self-directed learning.

This dissertation presents novel contributions at the intersection of predictive modeling and reflective learning analytics in education. It introduces the CISE model, a unique predictive framework designed to generalize student performance predictions across diverse institutional datasets, addressing the challenges of heterogeneity datasets. Additionally, it proposes ReflectMate, a prototype reflective learning tool that translates predictive insights into personalized feedback to foster student metacognition and engagement. The integration of fairness-oriented techniques and the consideration of gender and location disparities highlight the study's focus on educational equity. Together, these innovations contribute a dual-layered approach, predictive and reflective, that not only enhances academic forecasting but also promotes inclusive, student-centered learning support in digitally evolving educational contexts.

## 1.6 List of Research Articles

The research findings of this dissertation contributed to the following journal and conference articles.

- **Peer-reviewed Journal Articles**

  J1. **Dynil Duch**, Madeth May, Sébastien George, “Enhancing Predictive Analytics for Students’ Performance in Moodle: Insight from an Empirical Study,” in *Journal of Data Science and Intelligent Systems*, pp. 1–14, 2024, doi:10.47852/bonviewJDSIS42023777.

- **Journal Articles - In review**

  J4. **Dynil Duch**, Madeth May, Sébastien George, “ReflectMate: Empowering Students through Advanced Reflective Learning Analytics,” in *Springer Nature Computer Science*, 2024. [Under second review]

- **Peer-reviewed Conference Articles**

  C1. **Dynil Duch**, Madeth May, Sébastien George, “Students’ Performance in Learning Management System: An Approach to Key Attributes Identification and Predictive Algorithm Design,” in *13th International Conference on Data Science, Technology and Applications (Data 2024)*, pp. 285–292, 2024. doi:10.5220/0012754200003756.

  C2. **Dynil Duch**, Madeth May, Sébastien George, “Empowering Students: A Reflective Learning Analytics Approach to Enhance Academic Performance,” in *16th International Conference on Computer Supported Education (CSEDU 2024)*, pp. 385-396, 2024. doi:10.5220/0012634600003693.

  C3. **Dynil Duch**, Madeth May,Emmanuel G. Blanchard, Sébastien George, “On Digital Divide within a Developing Country: Investigating Gender and Location Influence in ICT Study Engagement in Cambodia ,” in *7th International Workshop on Culturally-Aware Tutoring Systems (CATS 2024)*, pp. 11–23, 2024. URL: https://hal.science/hal-04669404v1.

## 1.7 Structure of the Dissertation

The dissertation is organized into seven chapters, each addressing different aspects of the research. The structure of the dissertation is as follows:

**Chapter 1 Introduction** this chapter provides an overview of the research, including the background and motivation for the study, the research problem and objectives, the study’s significance, and the dissertation’s contributions. It also outlines the structure of the dissertation.

**Chapter 2 Literature Review** the literature review comprehensively analyzes the existing research in educational data mining, predictive modeling, reflective learning, and the digital divide. It identifies the key themes and challenges in the field, highlights the gaps in the current research, and positions the dissertation within the broader academic context.

**Chapter 3 Research Methodology** this chapter outlines the research methodology used in the study, including the research design, data collection methods, and data analysis techniques. It also discusses the ethical considerations associated with the research and the strategies used to address them.

**Chapter 4: Development of Predictive Models** this chapter focuses on developing and validating predictive models that are generalizable across different educational institutions. It provides detailed descriptions of the models, the methods used to train and validate them, and the results of the cross-institutional validation.

**Chapter 5: Reflective Learning Analytics for Student Empowerment** this chapter explores the leveraging of RLT within learning analytics, providing empirical evidence on the impact of this integration on student engagement and academic performance. It includes case studies and examples of how these tools have been implemented in educational settings.

**Chapter 6: Discussion and Implications** this chapter synthesizes the findings from the previous chapters, discussing their theoretical and practical implications for educational data mining and learning analytics. The chapter also considers the study's limitations and suggests directions for future research.

**Chapter 7: Conclusion** the final chapter summarizes the key findings of the research, discusses the implications for practice and policy, and provides recommendations for future research. It also reflects on the limitations of the study and the potential for future developments in the field.

## 1.8 Research Scope and Limitations

The scope of this research is broad, covering several critical areas within educational data mining, predictive modeling, reflective learning analytics, and the digital divide. However, it is important to acknowledge certain limitations inherent to the study's design and context.

One limitation lies in the choice of institutions included in this research. The study focuses on authentic data collected from educational institutions in Cambodia and France. This selection represents a starting point driven by the availability of real-world, contextualized data that aligns with our research environment. While these datasets provide valuable insights, expanding the research to include more institutions from diverse geographic and cultural contexts will be essential for further improving the generalizability and robustness of our findings. Future iterations of this work will aim to incorporate additional datasets to refine and enhance the predictive models and reflective tools developed here.

Another consideration relates to the nature of the data collected. The current research emphasizes engagement metrics and academic performance indicators central to understanding student behavior and outcomes. While these aspects form a strong foundation for predictive modeling and reflective learning, they do not encompass all potential variables that could influence student success—such as psychological or socio-emotional factors. We are mindful of the interconnectedness of these dimensions and their importance in shaping holistic educational experiences. However, the decision to focus on engagement and performance was guided by the specific research environment and the type of data readily accessible within our scope. Future research could explore integrating additional data sources to enrich the models and provide deeper insights.

Finally, the data quality plays a crucial role in determining the accuracy and reliability of the predictive models. While efforts have been made to preprocess and standardize the data, challenges such as missing values, class imbalances, and variability in grading scales remain. These issues highlight the need for ongoing data collection and processing methodologies refinement. Despite these constraints, the study provides a solid foundation for advancing predictive analytics and reflective learning tools in education.

In summary, this research represents an initial step toward addressing key educational data mining and learning analytics challenges. We aim to develop practical solutions tailored to real-world educational contexts by working with authentic data from Cambodian and French institutions. As the research evolves, incorporating broader datasets and exploring additional dimensions of student learning will further strengthen the applicability and impact of our work.

## 1.9 Summary

In summary, this chapter has introduced the key themes and objectives of the research, providing a comprehensive overview of the background, research problem, and significance of the study. The chapter has also outlined the structure of the thesis, providing a roadmap for the reader to follow. The next chapter, **Chapter 2: Literature Review**, will provide a detailed review of the existing literature in the field, situating the research within the broader academic context and highlighting the gaps that the thesis seeks to address situating the research within the broader academic context and highlighting the gaps that the thesis seeks to address.

Chapter 2

# Literature Review

## 2.1 Introduction

Integrating digital technologies into education has transformed the learning environment, creating opportunities for enhanced student engagement, personalized learning experiences, and data-driven decision-making (Haleem et al. 2022; Valverde-Berrocoso et al. 2021). This transformation has been primarily driven by Educational Data Mining (EDM) and Learning Analytics (LA), which focus on extracting meaningful insights from educational data to improve teaching and learning outcomes (S. Dawson et al. 2019; B. T.-M. Wong, K. C. Li, and Choi 2018; Papamitsiou and Economides 2014; Viberg et al. 2018; B. T.-m. Wong and K. C. Li 2020). These fields leverage computational techniques, including machine learning, statistical analysis, and natural language processing, to analyze large datasets from LMS, online educational platforms, and other digital tools.

This literature review provides a comprehensive analysis of the current research in EDM and LA, focusing on predictive modeling, ensemble learning, reflective LA, and the digital divide. The chapter begins by tracing the evolution of EDM and its applications in education, followed by a discussion of predictive modeling techniques and their challenges. It then explores the role of ensemble learning in improving the accuracy and generalizability of predictive models before examining the leveraging of Reflective Learning Tools (RLT) within LA to empower students. The chapter concludes with a discussion of the digital divide and its implications for educational equity, highlighting the importance of addressing this issue in the context of EDM.

## 2.2 Evolution of Educational Data Mining and Learning Analytics

The fields of Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as vital interdisciplinary research areas, leveraging techniques from data mining, machine learning, statistics, and education to analyze educational data and improve learning outcomes (AlShammari, Aldhafiri, and Al-Shammari 2013). While EDM focuses on extracting patterns and insights from large datasets to inform educational practices, LA emphasizes using data to support real-time decision-making and enhance the learning experience for both educators and students. These fields represent complementary approaches to understanding and improving teaching and learning in the digital age.

### 2.2.1 Evolution of Educational Data Mining

The roots of EDM can be traced back to the early 2000s when researchers began applying data mining techniques to educational data to uncover patterns and insights that could inform educational practice (Scholz, T. E. Kolb, and Neidhardt 2024; Feng and Fan 2024; Sarker et al. 2024). These early efforts laid the groundwork for a rapidly growing field driven by the increasing availability of digital data from educational institutions. Initially, the focus was on applying traditional data mining techniques, such as clustering, classification, and association rule mining, to educational datasets (Félix et al. 2018; Namoun and Alshanqiti 2020). These techniques were used to identify patterns in student behavior, predict academic performance, and optimize instructional strategies.

As the field matured, researchers began to explore more advanced techniques, including machine learning algorithms (Barakeh, Mezher, and Alharbi 2024), natural language processing (NLP) (Lukwaro, Kalegele, and Nyambo 2024), and network analysis, to analyze increasingly complex educational data (Froehlich, Van Waes, and Schäfer 2020). The widespread adoption of Learning Management Systems (LMS) and other digital tools in education has generated vast amounts of data, including clickstream data (Yürüm, Taşkaya-Temizel, and Yıldırım 2023), forum posts (Pong-inwong and Rungworawut 2012), quiz results (Juhaňák, Zounek, and Rohlíková 2019), and more, providing a rich source of information for researchers (Okike and Mogorosi 2020).

Today, EDM has expanded to include a broader range of applications, such as supporting adaptive learning (Demartini et al. 2024; Barbosa et al. 2024), detecting academic dishonesty (Masrom et al. 2024; Zhen and Zhu 2024), analyzing social networks within learning environments (Alshareef et al. 2020), and assessing the effectiveness of educational interventions (S. Alturki and N. Alturki 2021). This expansion has been facilitated by integrating EDM with other fields, such as LA, which uses data to inform and optimize learning experiences in real time.

### 2.2.2 Evolution of Learning Analytics

While EDM focuses on extracting patterns from data, LA emphasizes using data to enhance learning and teaching processes in real time. The term "Learning Analytics" was first introduced in the early 2010s, reflecting the growing recognition of the potential of data-driven approaches to support personalized and adaptive learning (Siemens and Long 2011). LA builds on principles from EDM but emphasizes actionable insights, user-centered design, and analytics integration into educational workflows.

The evolution of LA has been shaped by advancements in technology, particularly the proliferation of LMS platforms, online learning environments, and mobile technologies.

These systems generate vast amounts of data about student interactions, engagement, and performance, enabling educators to monitor learning progress, identify at-risk students, and provide timely interventions (S. Dawson et al. 2019; B. T.-M. Wong, K. C. Li, and Choi 2018). Early applications of LA focused on descriptive analytics, such as dashboards and visualizations that provided educators with summaries of student performance and engagement metrics (Papamitsiou and Economides 2014; Viberg et al. 2018).

Over time, LA has evolved to incorporate more sophisticated techniques, such as predictive and prescriptive analytics. Predictive analytics uses historical data to forecast future outcomes, such as student success or dropout rates. In contrast, prescriptive analytics provides recommendations for improving learning outcomes (B. T.-m. Wong and K. C. Li 2020).

### 2.2.3 Integration of EDM and LA

The integration of EDM and LA has created a powerful synergy, combining the strengths of both fields to address complex educational challenges. EDM provides the technical foundation for analyzing large-scale educational datasets. At the same time, LA focuses on translating these insights into actionable strategies for educators and students. These fields enable a holistic approach to understanding and improving learning outcomes.

For instance, predictive models developed through EDM can be integrated into LA platforms to provide real-time feedback to students and educators. Tools like Reflect-Mate, discussed later in this dissertation, exemplify this integration by combining predictive insights with reflective learning analytics to empower students to take control of their learning journeys (Arthars et al. 2019). Similarly, cross-institutional validation techniques, such as those used in the CISE model, demonstrate how EDM and LA can work together to ensure the generalizability and applicability of predictive models across diverse educational contexts.

### 2.2.4 Challenges and Opportunities

Despite significant advancements, both EDM and LA face several challenges. One critical challenge is the need for more sophisticated models that can handle the complexity and variability of educational data. Educational data is often noisy, incomplete, and highly contextual, making it challenging to develop accurate and generalizable models (Baker et al. 2019). Another challenge is the need for greater interpretability of models, particularly in educational settings where the decisions made based on model predictions can have significant implications for students' academic trajectories.

Ethical considerations also play a crucial role in developing and deploying EDM and LA tools. Using student data must comply with legal and ethical standards to protect student privacy and ensure data is used responsibly (Mengash 2020). Addressing these concerns requires a careful balance between the benefits of predictive modeling and the need to protect students' rights and interests.

In conclusion, the evolution of EDM and LA reflects a growing recognition of the transformative potential of data-driven approaches in education. By leveraging advanced computational techniques and focusing on actionable insights, these fields offer promising solutions to some of the most pressing challenges in education today, from predicting student outcomes to fostering equity and inclusion.

## 2.3 Educational Data Mining and Predictive Modeling

### 2.3.1 Predictive Modeling in Education

Predictive modeling has become a cornerstone of EDM, potentially improving educational outcomes by identifying at-risk students with poor academic performance or dropping out (Quinn and Gray 2019). Predictive models use historical data to forecast future events, such as a student's likelihood of success in a course or program (Arizmendi et al. 2022). These models are built using various statistical and machine-learning techniques, each with strengths and limitations.

One of the most common techniques used in predictive modeling is Decision Trees, which are favored for their simplicity and interpretability. Decision Trees work by recursively partitioning the data into subsets based on the values of input variables, resulting in a tree-like structure where each branch represents a decision or classification (Patidar, Dangra, and Rawar 2015). Random Forests, an extension of Decision Trees, involve constructing multiple trees and averaging their predictions to improve accuracy and reduce the risk of overfitting (Assegie et al. 2024).

Support Vector Machines (SVMs) are another popular technique in predictive modeling. SVMs are particularly effective in classification tasks, where they find the hyperplane that best separates different classes in the data (Manu 2016). However, SVMs can be challenging to interpret, which has led to their limited use in educational settings where transparency is crucial.

Deep learning models, especially Neural Networks, have also gained traction in educational data mining (Merceron and Tato 2023). These models can capture complex, nonlinear relationships in large datasets, making them suitable for predicting student performance, understanding student engagement, and recommending personalized learning resources (Okewu et al. 2021). However, the "black box" nature of deep learning models raises concerns about their transparency and the ability of educators to understand and trust their predictions.

Despite the effectiveness of these models, several challenges still need to be solved in applying predictive modeling to educational data. One of the main challenges is dealing with the inherent noisiness and imbalance in educational datasets (Ashfaq, Booma, and

Mafas 2020; Krawczyk 2016). Educational data is often incomplete, with missing values and irregular data points, making it difficult to build accurate models. Additionally, the imbalanced nature of educational outcomes—where most students pass while a minority fail—can lead to biased models towards the majority class (H. Ali et al. 2019). To address this issue, researchers often employ techniques such as oversampling, undersampling, and synthetic data generation to balance the data before training the model (Wongvorachan, S. He, and Bulut 2023).

Another significant challenge in predictive modeling is the generalizability of models (Baker et al. 2019). Educational institutions vary widely in their curricular structures, student demographics, and teaching practices, which means that models trained on data from one institution may need to improve when applied to another. This lack of generalizability limits the broader applicability of predictive models in education. Cross-institutional validation, where models are tested on data from multiple institutions, is essential to ensure that predictive models are robust and applicable in diverse educational environments.

In recent years, there has been growing interest in using ensemble learning techniques to improve the performance and generalizability of predictive models in education. Ensemble learning combines multiple models to produce a more accurate prediction. By aggregating the strengths of different models, ensemble methods can often outperform individual models, particularly in complex and noisy datasets (Ashraf, Zaman, and M. Ahmed 2020; Niyogisubizo et al. 2022).

### 2.3.2 Challenges in Predictive Modeling

The challenges associated with predictive modeling in education require careful consideration. One of the most significant challenges is overfitting, where models perform exceptionally well on training data but must generalize to new data. Overfitting is particularly problematic in educational contexts, where datasets are often small and vary greatly

(Ashraf, Zaman, and M. Ahmed 2020). Researchers commonly use cross-validation, regularization, and ensemble methods to mitigate overfitting.

Cross-validation is a powerful technique for assessing the performance of predictive models, particularly in settings where data is scarce. In k-fold cross-validation, the data is divided into k subsets, and the model is trained on k-1 subsets while being tested on the remaining subset. This process is repeated k times, with each subset serving as the test set once, and the results are averaged to produce a more reliable estimate of model performance (Yağcı 2022a). Cross-validation helps to ensure that the model's performance is not overly dependent on a particular subset of the data, reducing the risk of overfitting.

Regularization is another technique for preventing overfitting. It adds a penalty term to the model's objective function. This penalty discourages the model from fitting the noise in the data, leading to simpler models that are less likely to overfit. L1 regularization (Lasso) and L2 regularization (Ridge) are two common types of regularization used in predictive modeling (Kernbach and Staartjes 2022).

Despite these techniques, interpretability remains a significant challenge in predictive modeling, mainly when using complex models like Neural Networks. Educators and stakeholders must understand how predictions are made in educational settings, as this transparency builds trust and facilitates the integration of predictive models into decision-making processes (Abdelqader et al. 2022). To address this issue, researchers have developed various techniques to enhance the interpretability of models, including feature importance measures, decision rules, and model simplification methods.

Feature importance measures, such as Gini and permutation importance, rank the input variables based on their contribution to the model's predictions (Young and Caballero 2021). These measures provide insights into which features drive the model's decisions, helping educators understand the factors influencing student outcomes. On the other hand, decision rules involve extracting simple "if-then" rules from the model, which educators can easily understand and apply. Model simplification methods, such as pruning and rule extraction, aim to reduce the complexity of the model without significantly compromising its accuracy.

Another essential consideration in predictive modeling is the ethical use of data. The use of student data must comply with legal and ethical standards to protect student privacy and ensure that data is used responsibly. Ethical concerns also arise when predictive models are used to make decisions that can significantly impact a student's educational trajectory, such as identifying students for remedial programs or determining admissions criteria (Mengash 2020). Addressing these concerns requires a careful balance between the benefits of predictive modeling and the need to protect students' rights and interests.

The bias in predictive models is also a critical concern, particularly in educational contexts where decisions based on model predictions can have far-reaching consequences (Hu and Rangwala 2020). Bias can enter models through the data used for training or through the design of the model itself. For instance, if a model is trained on data from a predominantly affluent school, it may perform poorly when applied to a more diverse population, potentially disadvantaging students from underrepresented groups. Ensuring that models are fair and unbiased requires careful attention to the selection of training data, the design of the model, and the evaluation of model outputs.

Recently, there has been increasing recognition of the need to develop fair and transparent predictive models in education. Researchers are exploring various techniques to address bias and ensure that models are equitable and just. These techniques include fairness-aware algorithms (Le Quy 2024), designed to minimize disparities in model predictions across different groups, and post hoc adjustment methods (Palacios et al. 2021), which involve modifying the model's outputs to correct for bias.

## 2.4 Identifying Key Attributes for Predicting Student Performance

Predicting student performance has become crucial to educational data mining and learning analytics (Hirokawa 2018). A growing body of literature has identified various attributes indicative of academic success or challenges, ranging from cognitive abilities and prior academic achievements to behavioral patterns and socio-demographic factors. This section provides a comprehensive review of these critical attributes, synthesizing findings from existing studies to guide the selection of variables for predictive modeling in this research.

### 2.4.1 Overview of Predictive Factors in Education

Educational data mining research consistently highlights the importance of identifying and selecting appropriate attributes for accurately predicting student performance (Umer et al. 2023). These attributes are generally categorized into cognitive, behavioral, demographic, psychological, and teacher pedagogy factors. Each of these categories contributes uniquely to understanding the complex dynamics of student learning and performance.

The literature reveals that attributes derived from LMS are critical in predictive models. These include interaction metrics such as frequency of logins, time spent on course materials, and participation in online discussions, which are often strong indicators of student engagement and academic performance. Other factors, such as prior academic history, grades, and test scores, are significant predictors.

### 2.4.2 Academic and Cognitive Attributes

Academic performance in previous courses and standardized tests has long been recognized as a robust predictor of future success. Research by Yağcı (2022b) suggests that prior GPA, course grades, and exam scores are among academic achievement's most significant cognitive predictors. These attributes reflect a student's foundational knowledge and academic abilities, which are critical for understanding their capacity to succeed in future academic endeavors.

Moreover, cognitive load, as discussed in Gorbunova et al. (2024), influences student performance. Cognitive load refers to the mental effort required to learn new information, and students who can effectively manage this load tend to perform better academically. This attribute is often assessed through metrics related to task complexity and the student's ability to process and understand course content.

### 2.4.3 Behavioral and Engagement Metrics

Behavioral data captured through LMS platforms has become increasingly important in educational research. Metrics such as login frequency, session duration, and clickstream data are commonly used to gauge student engagement. Studies like those by Smith, Cobham, and Jacques (2022) have shown that higher levels of engagement, as measured by these metrics, are positively correlated with better academic outcomes.

For instance, Brahim (2022) demonstrated that students who regularly log into the LMS, spend more time on learning activities, and actively participate in discussion forums are more likely to achieve higher grades. These behavioral attributes provide insight into the student's commitment to the course and their interaction with the learning materials, which are crucial for predicting academic success.

### 2.4.4 Demographic and Contextual Factors

Demographic attributes, including age, gender, socio-economic status, and geographic location, have been shown to significantly influence student performance. According to Segura-Morales and Loza-Aguirre (2017), socio-economic status is a critical determinant of access to educational resources, affecting academic outcomes. Similarly, Nguyen, Allen, and Fraccastoro (2005) found that gender differences in learning styles and subject preferences could lead to variations in academic performance.

The digital divide, particularly in developing countries, adds another layer of complexity to the predictive modeling of student performance. Students from lower socio-economic backgrounds or rural areas often face challenges related to limited access to technology, as highlighted by Farooqi, Khalid, and A. Khan (2022) and Cheshmehzangi et al. (2023). These challenges can hinder their ability to fully engage with online learning platforms, impacting their academic performance.

### 2.4.5 Psychological and Affective Attributes

Psychological factors such as motivation, self-efficacy, and stress are increasingly recognized as significant predictors of student performance. Research by S. P. Dawson, L. Macfadyen, and Lockyer (2009) and Purwaningsih and Suwarno (2016) has demonstrated that students with higher motivation and self-efficacy are likelier to persist through challenges and achieve academic success. These attributes are often assessed through self-reported surveys or inferred from behavioral data, such as the consistency and intensity of study sessions.

Additionally, affective states like anxiety and stress, as discussed by Zhang and Henderson (2019), can negatively impact academic performance. For example, high-stress levels during exam periods have been linked to lower performance, making it a critical attribute to consider in predictive models.

### 2.4.6 Teacher Pedagogy

Teacher pedagogy is a crucial factor influencing student performance and must be considered in predictive models and reflective tools (Trindade and Ferreira 2021). Teaching methods, such as active learning, improve engagement and outcomes compared to traditional lectures (Freeman et al. 2014). Teacher-student interactions, including feedback frequency, are also critical for student success (Hattie 2009). Assessment practices, especially formative feedback, support learning and can inform predictive models (Black and Wiliam 1998). Teacher expectations significantly shape student performance (Rosenthal and Jacobson 1968). Incorporating these pedagogical attributes enhances the predictive power of models and the effectiveness of reflective learning tools.

### 2.4.7 Synthesis of Key Attributes

The preceding sections have systematically reviewed the diverse attributes that influence student performance, drawing on existing literature to identify predictors across cognitive, behavioral, demographic, psychological, and pedagogical dimensions, as shown in Fugure 2.1. These attributes have been extensively studied in educational data mining (EDM) and learning analytics (LA), with prior research demonstrating their potential to enhance the accuracy and interpretability of predictive models. However, synthesizing these findings reveals critical insights into how these attributes have been operationalized, their contributions to predictive modeling, and the gaps our research seeks to address.

Research has consistently shown that academic and cognitive attributes—such as prior GPA, standardized test scores, and cognitive load metrics—are among the strongest predictors of student success (Yağcı 2022b; Gorbunova et al. 2024). These attributes are often

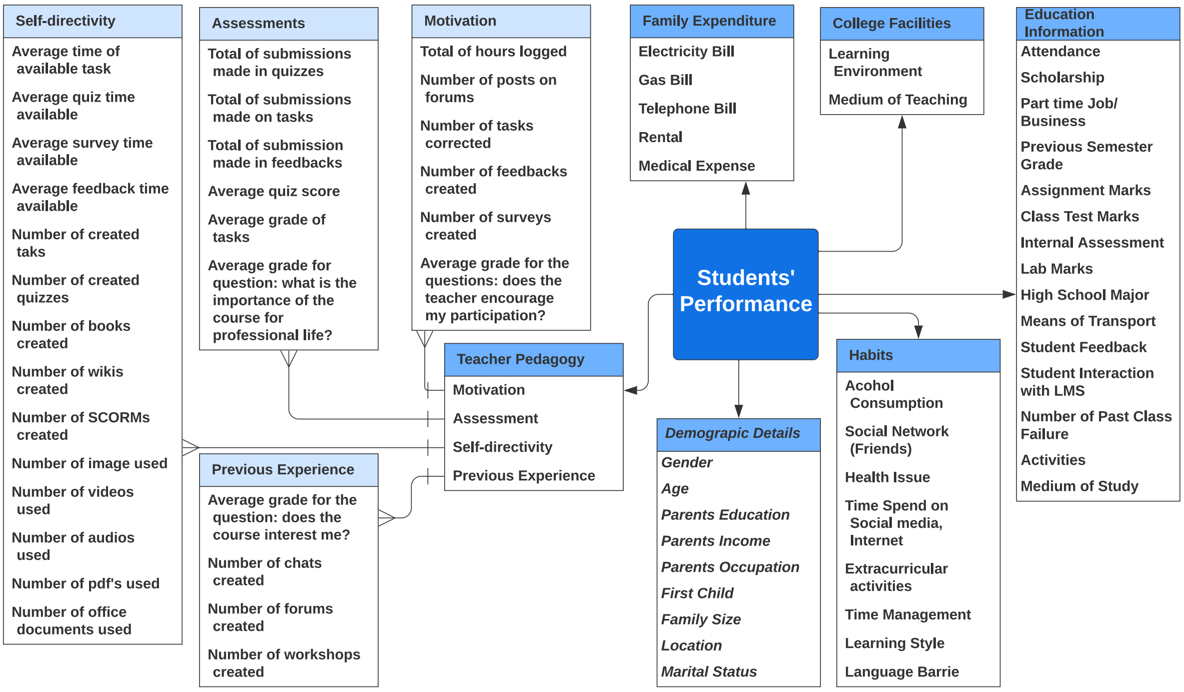


Figure 2.1 – Synthesis of the attributes for predicting students' performance

derived from institutional records or LMS data and have been validated across numerous studies for their ability to forecast academic outcomes. Similarly, behavioral metrics like login frequency, time spent on tasks, and participation in online discussions have emerged as reliable indicators of engagement and performance (Smith, Cobham, and Jacques 2022; Brahim 2022). These attributes form the backbone of many predictive models, providing a foundation for understanding student behavior and identifying at-risk learners. Demographic and contextual factors, such as socio-economic status, geographic location, and gender, add another layer of complexity to predictive modeling. While these attributes are less directly tied to individual performance, they help account for systemic inequities and external barriers that influence educational outcomes (Segura-Morales and Loza-Aguirre 2017; Farooqi, Khalid, and A. Khan 2022). Psychological attributes, including motivation, self-efficacy, and stress levels, further enrich predictive models by capturing the affective dimensions of learning (S. P. Dawson, L. Macfadyen, and Lockyer 2009; Zhang and

Henderson 2019). Finally, teacher pedagogy attributes highlight the role of instructional practices in shaping student success, underscoring the importance of aligning predictive tools with teaching strategies (Freeman et al. 2014; Hattie 2009).

Integrating these diverse attributes into predictive models has yielded significant advancements in forecasting student outcomes. For example, ensemble learning techniques have been used to combine multiple models, improving model accuracy and generalizability (Ashraf, Zaman, and M. Ahmed 2020; Niyogisubizo et al. 2022). Similarly, reflective learning tools (RLT) have demonstrated the potential to enhance predictive models by incorporating self-regulatory behaviors and metacognitive strategies into the analysis (Arthars et al. 2019). Despite these advances, challenges remain in ensuring that predictive models are accurate and equitable, particularly in resource-constrained settings where data availability and quality may vary.

Building on this foundation, our research takes a novel approach to attribute selection by focusing on the availability and relevance of these attributes within the specific datasets under investigation. Unlike prior studies that often rely on predefined sets of predictors, we adopt a dynamic methodology that prioritizes attributes based on their presence—or absence—in the data. It ensures that our predictive models are grounded in real-world data while remaining adaptable to variations across institutions and educational contexts.

In summary, identifying and utilizing key attributes in predictive modeling represent a synthesis of decades of research in EDM and LA. Our contribution lies in refining this process to align with the unique characteristics of our datasets. This approach seeks to advance the field by demonstrating how thoughtful attribute selection can enhance predictive models' theoretical grounding and practical applicability.

### 2.4.8 Gaps in the Literature and Justification for Attribute Selection

While the literature provides a solid foundation for identifying key attributes, several gaps remain. For example, limited research exists on the integration of RLT with traditional behavioral metrics, which could provide deeper insights into student engagement. Furthermore, the impact of the digital divide on predictive modeling has not been extensively explored, particularly in the context of developing countries.

This study addresses these gaps by incorporating traditional and novel attributes related to reflective learning and digital equity. The selection of attributes is guided by the existing literature and the study's specific context, ensuring that the models developed are theoretically grounded and practically relevant.

## 2.5 Ensemble Learning and Meta-Modeling

### 2.5.1 Introduction to Ensemble Learning

Ensemble learning is a powerful machine learning technique that involves combining multiple models to improve overall predictive performance. The central idea behind ensemble learning is that by aggregating the predictions of several models, the ensemble can reduce variance and bias, leading to more accurate and robust predictions. This approach has proven particularly effective in educational contexts, where data can be noisy, incomplete, and complex (Dietterich 2000).

There are several types of ensemble methods, each with its unique advantages. Bagging, or Bootstrap Aggregating, is a popular ensemble technique that involves training multiple models on different subsets of the training data and then averaging their predictions. Bagging is particularly effective at reducing variance and preventing overfitting, making it widely used in educational data mining (Breiman 1996). Random Forests, a bagging method, involves constructing multiple Decision Trees and averaging their predictions to improve accuracy and stability (Breiman 2001).

Boosting is another widely used ensemble technique that focuses on training models sequentially, with each model correcting the errors of its predecessor. Boosting methods, such as AdaBoost and Gradient Boosting, have improved the accuracy of predictive models in various educational tasks (Friedman 2001). The key idea behind boosting is to assign higher weights to data points misclassified by previous models, forcing subsequent models to focus on these "hard" cases. This iterative process can significantly improve model accuracy, particularly in datasets with complex decision boundaries.

Stacking, also known as stacked generalization, is another ensemble method that involves training a meta-learner to combine the predictions of multiple base learners. This method can capture the strengths of different models and has been used successfully in educational data mining to improve the performance of predictive models (Wolpert 1992). Stacking is particularly advantageous when the base learners have different strengths and weaknesses; for example, one model might excel at identifying high-achieving students, while another is better at detecting students at risk of failure. Combining these models allows the stacked ensemble to provide a more nuanced and accurate prediction.

### 2.5.2 Applications of Ensemble Learning in Predictive Modeling

Ensemble learning has been applied to various predictive tasks in education, from predicting student performance to identifying at-risk students and optimizing curriculum design. One of the most common applications of ensemble learning in education is predicting student success, where models such as Random Forests and Gradient Boosting Machines have been used to predict final grades, dropout rates, and standardized test scores (Breiman 2001; Kotsiantis and Kanellopoulos 2012).

For example, Random Forests, an ensemble method based on Decision Trees, is widely used for predicting student outcomes due to its ability to handle large datasets with many features and its resilience to overfitting. Studies have demonstrated that Random Forests often outperform individual machine learning models, particularly in datasets with complex, nonlinear relationships between variables. A study by Nafea et al. (2023) developed an ensemble model combining Random Forest, AdaBoost, Decision Trees, and Support Vector Machines, achieving an accuracy of 95% in classifying student performance. Similarly, Kamal and Ahuja (2019) created an ensemble-based model integrating Decision Tree, Gradient Boosting, and Naïve Bayes techniques, which achieved a 99% accuracy rate in predicting academic performance.

Boosting methods, such as AdaBoost and Gradient Boosting, have also been used to enhance the performance of predictive models in education. These methods work by sequentially training models, with each model focusing on the errors made by the previous one. The resulting ensemble is typically more accurate and less prone to overfitting than individual models, making boosting an effective technique for improving the accuracy of predictive models in educational contexts (Freund and Schapire 1997; Friedman 2001). A study by Y. Wang et al. (2021) introduced a graph-based ensemble machine-learning method that outperformed traditional algorithms by up to 14.8% in prediction accuracy, demonstrating the effectiveness of ensemble-based approaches in predicting student performance.

Stacking, or stacked generalization, is another ensemble technique used in educational data mining. In stacking, multiple base learners are trained on the same dataset, and a meta-learner is used to combine their predictions. This approach allows the ensemble to leverage the strengths of different models. It has been shown to improve the accuracy of predictive models in educational settings. Smirani et al. (2022) proposed a stacked generalization model that combined Light Gradient Boosting Machine, eXtreme Gradient

Boosting, and Random Forest algorithms, resulting in a sensitivity average of 97.3% and a precision average of 97.2%. In the context of this dissertation, stacking combines the predictions of different models trained on data from multiple institutions, enhancing the generalizability of the final model.

Ensemble learning has also been applied to specific educational tasks such as predicting student dropout, detecting learning disabilities, and recommending personalized learning resources. For instance, an ensemble of models might be used to predict whether a student is likely to drop out based on their attendance records, grades, and participation in online activities. By combining the predictions of multiple models, the ensemble can provide a more accurate and reliable prediction, allowing educators to intervene early and support at-risk students (Márquez-Vera et al. 2016).

Another application of ensemble learning in education is the development of adaptive learning systems, which use predictive models to tailor instructional content to individual students' needs. Ensemble methods can improve the accuracy of these models, ensuring that the content delivered to each student is appropriately challenging and supportive. For example, an adaptive learning system might use an ensemble of models to predict which topics a student is struggling with and then adjust the curriculum to provide additional practice or resources (Wu 2025; Duan and W. Wang 2024; Koedinger et al. 2013).

### 2.5.3 Meta-Modeling and Cross-Institutional Validation

Meta-modeling, or stacking, is an advanced ensemble technique that involves training a meta-learner to combine the predictions of multiple base models. The meta-learner is typically a simple model, such as linear regression, that takes the predictions of the base models as input and produces a final prediction. This approach can capture the strengths of different base models, leading to improved predictive performance (Wolpert 1992).

In this dissertation, meta-modeling integrates predictive models trained on data from multiple institutions. The goal is to develop a final accurate and generalizable model across different educational settings. Cross-institutional validation is critical to this research, as it ensures that the models developed are robust and applicable in diverse contexts.

Cross-institutional validation is critical in educational data mining, where there is significant variation in curricular structures, student demographics, and teaching practices across institutions. Traditional predictive models often need to generalize better when applied to data from different institutions, leading to poor predictive performance. By validating models across multiple institutions, this research aims to ensure that the predictive models developed are robust and applicable in various educational settings (Porras et al. 2023).

One approach to cross-institutional validation is to use data from multiple institutions to train the base models in the ensemble. It ensures the ensemble is exposed to various educational contexts, making it more likely to generalize well to new data. Another approach is to use transfer learning, where a model trained on data from one institution is fine-tuned on data from another institution. This approach can be efficient when the institutions have similar characteristics, such as similar student demographics or curricular structures (S. J. Pan and Yang 2009).

The importance of cross-institutional validation cannot be overstated. It addresses one of the most significant challenges in educational data mining: the lack of generalizability of predictive models. By developing robust models across different educational settings, this research contributes to the broader goal of making educational data mining tools more widely applicable and effective in diverse contexts.

In addition to improving generalizability, cross-institutional validation also provides valuable insights into the factors that influence the success of predictive models. For example, by comparing a model's performance across different institutions, researchers can identify which features are most important for predicting student outcomes and which aspects of the educational context are most influential. This knowledge can refine the models and develop more effective interventions for supporting students (Porras et al. 2023).

Another benefit of cross-institutional validation is that it helps to mitigate the risk of bias in predictive models. By training and validating models on data from multiple institutions, researchers can ensure that the models are balanced by the characteristics of a single institution, such as its student population or teaching practices. It reduces the likelihood that the models will be biased toward specific groups of students or educational contexts, making them more equitable and fair (Khademi and Honavar 2020).

## 2.6 Reflective Learning Analytics and Student Empowerment

### 2.6.1 Concept of Reflective Learning

Reflective learning is an educational approach that emphasizes the importance of students actively engaging in reviewing and analyzing their learning experiences. This process fosters profound understanding, critical thinking, and self-regulation, essential lifelong learning skills (Öz and Şen 2021). Reflective learning encourages students to think about what they have learned, how they have learned it, and how they can apply it in the future.

Reflective learning has its roots in the work of educational theorists such as Dewey (1933) and Schon (2008), who argued that reflection is a critical component of the learning process. Dewey defined reflective thinking as "active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it". Schon expanded on this idea, introducing the concept of the "reflective practitioner," who engages in a continuous process of reflection in action and reflection on action.

In the context of this dissertation, reflective learning is supported by LA, which provides students with data-driven insights into their learning behaviors. Using these insights, students can make informed decisions about their study practices, set goals, and monitor their progress. Leveraging reflective learning within LA represents a shift towards more personalized and student-centered education (Schon 2008).

Reflective learning is about looking back on what has been learned and using that reflection to guide future actions. This continuous improvement process is central to the concept of lifelong learning, where individuals are encouraged to take responsibility for their own learning and seek out opportunities for growth throughout their lives (D. A. Kolb 2014). Reflective learning also aligns with the principles of constructivist learning theories, which emphasize the learner's active role in constructing knowledge through experience and reflection (Piaget 1952).

Reflective learning is also closely related to metacognition, which involves thinking about one's thinking processes. Metacognitive strategies such as planning, monitoring, and evaluating one's learning are critical components of reflective learning and have been shown to improve academic performance and self-regulation (Flavell 1979). By engaging in reflective learning, students develop metacognitive skills that enable them to control their learning and become more effective and independent learners (Zimmerman 2002).

### 2.6.2 RLT in Education

Before delving into the role of Reflective Learning Tools (RLT) in education, it is important to clarify what these tools are. RLT refers to digital or analog instruments designed to support students in actively reviewing, analyzing, and reflecting on their learning experiences. These tools encourage learners to think critically about their progress, identify strengths and areas for improvement, and take ownership of their educational journey. By integrating data-driven insights—such as engagement metrics, performance trends, and personalized feedback—RLT empowers students to make informed decisions about their study habits and academic strategies.

In online learning, RLT have gained significant attention due to their potential to address challenges such as low engagement, lack of self-regulation, and limited opportunities for meaningful reflection. Research within Metacgnition Online Learning Environments (MOLE) has identified core issues related to students and the integration of reflective tools in online learning. Ndukwe and Daniel (2020) conducted a study exploring the expectations of students regarding LA tools in online courses. Their findings indicated that students desired more user progress feedback and a greater emphasis on real-time progress tracking. These insights shed light on specific areas for improvement in LA tools, suggesting a need for enhancements in features related to user progress learning experiences (Fatma Gizem Karaoglan Yilmaz 2020; Hegde, Pai, and Shastry 2022; Fatma Gizem Karaoglan Yilmaz 2022; Karaoglan Yilmaz 2022) and continuous monitoring of academic progress (QAZDAR et al. 2022). Silvola et al. (2021) examined the expectations of educators and online learners concerning LA dashboards, emphasizing the need for user progress insights in virtual classrooms.

In the context of online learning, the existing body of literature unveils the potential of LA to provide valuable understanding of student engagement and performance. It emphasizes the need for tailored support and technological solutions in online education. Furthermore, the research mentioned core issues related to students and integrating reflective tools in virtual classrooms. Despite all that, a significant gap persists in developing reflective tools that empower students within online learning. Not to mention that most existing supports are often designed to assess the final outcomes of learning activities. Accordingly, the data indicators provided are not exactly exploited by the students as reflective tools during their learning process, but are mainly used at the end as feedback or report on their final academic outcomes.

### 2.6.3 Impact of Reflective Learning on Academic Performance

The impact of reflective learning on academic performance has been the subject of considerable research. Studies have shown that students who engage in reflective practices tend to achieve higher academic outcomes, as they are more likely to develop a deep understanding of the subject matter and to transfer their learning to new contexts (Zimmerman 2002). Reflective learning also promotes the development of critical thinking skills, which are essential for success in higher education and beyond (Kreber 2012).

One of the literature's key findings is that when leveraged within LA, RLT can significantly enhance student performance. For example, a study by L. P. Macfadyen and S. Dawson (2010) found that students who regularly used RLT in conjunction with analytics features showed more significant improvements in their grades than those who did not. Reflective learning, supported by analytics, can give students the insights and motivation to succeed academically.

Another study by Boud, Keogh, and Walker (2013) demonstrated the positive effects of reflective learning on student engagement and self-regulation. They found that students who engaged in regular reflection were likelier to set and achieve academic goals, manage their time effectively, and seek additional resources when needed. These findings support the idea that reflective learning improves academic performance and helps students develop the skills needed for lifelong learning.

Integrating reflective learning with learning analytics has also been shown to enhance the effectiveness of personalized learning. By providing students with data-driven insights into their learning behaviors, learning analytics platforms can help students identify areas for improvement and make informed decisions about their study practices. This personalized approach to learning has been shown to improve student outcomes, particularly in online and blended learning environments (Gašević, S. Dawson, and Siemens 2015).

This dissertation explores the impact of RLT on student engagement and academic performance through a case study. Students are provided with access to a reflective learning analytics platform. The study aims to provide empirical evidence on the effectiveness of these tools, contributing to the growing body of knowledge on reflective learning and its role in education.

The benefits of reflective learning extend beyond academic performance. Research has shown that students who engage in reflective learning are more likely to develop a positive attitude toward learning, take responsibility for it, and develop the skills needed for lifelong learning. Reflective learning also promotes the development of critical thinking, problem-solving, and decision-making skills, which are essential for success in higher education and beyond (Öz and Şen 2021).

## 2.7 Digital Divide in Education

### 2.7.1 Understanding the Digital Divide

The digital divide, a significant global issue characterized by unequal access to ICT, has far-reaching consequences (Afzal et al. 2023; Farooqi, Khalid, and A. Khan 2022; A. Ahmed 2007). Studies have shown this divide is prevalent across various countries, leading to disparities in education, employment, and overall quality of life. For instance, limited access to the internet and digital devices hampers educational opportunities in many developing nations, affecting students' ability to engage with modern learning resources and tools (Moore, Vitale, and Stawinoga 2018).

### 2.7.2 Impact of the Digital Divide on Educational Outcomes

The digital divide particularly impacts women, who are underrepresented in STEM fields, including ICT (Economic Co-operation and (OECD) 2018; Marzano and Lubkina 2019; Larsson and Viitaoja 2019). Research indicates that women face multiple barriers, such as societal norms and gender biases, which discourage their participation in these fields. This underrepresentation is a global issue (Kerras et al. 2020) observed in developed (Pérez-Castro, Mohamed-Maslouhi, and Montero-Alonso 2021) and developing (Antonio and Tuffley 2014; Mumporeze and Prieler 2017) countries. It contributes to the broader gender gap in the technology and engineering sectors (Mariscal et al. 2019).

Within this discourse, a research by Pérez-Castro, Mohamed-Maslouhi, and Montero-Alonso (2021) has extensively discussed the unequal distribution of the digital divide among countries and its potential impact on future populations. Their insights highlight the stark differences between wealthy Western countries and the rest of the world, underscoring how access to ICT can influence economic and social development. This disparity not only affects individual opportunities but also has broader implications for national progress and competitiveness in the global market.

Similarly, Blanchard (2015), drawing on research from psychology, has explored similar imbalances within the field of artificial intelligence in education. Their work suggests that psychological factors, including self-efficacy and societal expectations, significantly shape individuals' engagement with ICT.

### 2.7.3 Addressing the Digital Divide through Educational Data Mining

Addressing the digital divide is critical for ensuring that all students have the opportunity to succeed in education. Educational data mining can play a role in mitigating the effects of the digital divide by identifying at-risk students and providing targeted interventions to support their learning. For example, predictive models can be used to identify students struggling due to a lack of access to technology and develop strategies to help them overcome these barriers (Reich 2020).

One approach to addressing the digital divide is to use predictive models to identify students at risk of falling behind due to limited access to technology. These models can consider factors such as internet connectivity, device availability, and digital literacy, allowing educators to target interventions where they are most needed. For example, schools might provide loaner devices or offer internet subsidies to students without access to the necessary technology (Warschauer 2004).

Another approach is to design learning analytics platforms accessible to all students, regardless of their level of digital literacy. It might involve creating user-friendly interfaces, providing digital literacy training, or offering alternative ways to access learning materials (e.g., offline access or mobile-friendly formats). By making learning analytics platforms more accessible, educators can help bridge the digital divide and ensure that all students can benefit from data-driven insights into their learning (Gašević, S. Dawson, and Siemens 2015).

In addition to addressing the digital divide, educational data mining can promote digital inclusion by identifying and supporting underrepresented students in digital learning environments. For example, predictive models can identify students less likely to engage with digital learning resources, and targeted interventions can be developed to encourage their participation. It might involve providing additional support, such as tutoring or mentoring, or creating culturally relevant content that resonates with diverse student populations (Reich 2020).

In this dissertation, the digital divide is considered a contextual factor that influences the implementation and effectiveness of predictive models and RLT. By understanding how the digital divide affects student engagement and academic outcomes, this research aims to develop inclusive and applicable solutions across diverse educational settings.

## 2.8 Synthesis of Literature

### 2.8.1 Identifying Gaps in the Literature

Reviewing the existing literature on educational data mining, predictive modeling, and learning analytics reveals several critical gaps this dissertation seeks to address.

First, while extensive research has been conducted on individual attributes that predict student performance—such as prior academic achievement, engagement metrics, and demographic factors—comprehensive models that integrate a broad spectrum of attributes are lacking. Most studies focus on isolated predictors, failing to consider how cognitive, behavioral, demographic, psychological, and teacher pedagogy factors might interact to influence student outcomes.

Moreover, a significant gap exists in the research on the generalizability of these predictive models across different educational contexts. Many studies develop models based on data from a single institution or a homogenous student population, which limits their applicability in more diverse settings. This lack of generalizability raises concerns about the validity of these models when applied to different educational environments, especially those with varying socio-economic conditions, academic structures, and student demographics.

Additionally, more needs to be explored about how RLT, which encourage student self-regulation and critical thinking, can be effectively integrated into predictive models. The current literature treats these tools as separate from traditional predictive metrics, and more opportunities need to be created to enhance model accuracy and student empowerment through their inclusion.

Another significant gap pertains to the digital divide, especially in developing countries, where disparities in access to technology and digital literacy profoundly affect student engagement and performance. Existing predictive models often overlook these contextual factors, leading to potential biases and reduced applicability in diverse educational settings. Addressing these gaps is crucial for developing predictive models that are accurate but also equitable and inclusive.

### 2.8.2 Positioning Our Research

This research is strategically positioned to address the identified gaps by developing a comprehensive framework for predictive modeling that draws on a broad spectrum of attributes—cognitive, behavioral, demographic, psychological—identified, and teacher pedagogy in the literature. While the review provides a thorough foundation for understanding the potential predictors of student performance, it is essential to note that not all identified attributes will be used directly in our predictive models. Instead, the literature serves as a guide for selecting the most relevant and impactful attributes that align with this research's specific goals and context.

Identifying these attributes based on the literature is crucial for defining a global set of potential predictors. However, our research will focus on selecting and refining attributes with the most significant predictive power in the context of the educational environments studied. This selective approach ensures that the models developed are both parsimonious and effective, avoiding the pitfalls of overfitting while still capturing the key factors that influence student performance.

In addition to this strategic selection of attributes, the research will pioneer including RLT within predictive models. By examining how attributes related to self-regulation, motivation, and other reflective practices interact with traditional metrics, the study seeks to enhance the predictive power of the models and contribute to a deeper understanding of student learning processes.

Furthermore, the research is particularly attentive to the generalizability of the predictive models across different educational contexts. The study addresses a significant gap in the current literature by testing the models across multiple institutions with varying student demographics and educational structures. This focus on generalizability ensures that the models are accurate within a single context and applicable and reliable across diverse educational settings.

Moreover, this research acknowledges the influence of the digital divide on student performance, particularly in developing countries. The CISE model addresses this challenge by providing fair and generalizable performance predictions across diverse datasets, including those from underrepresented regions. ReflectMate complements this by offering personalized, student-centered feedback and guidance that supports learners from various backgrounds. Together, they contribute to reducing gender and location-based disparities, fostering more equitable and inclusive educational opportunities.

Through these contributions, this dissertation aims to advance the field of educational data mining by providing predictive models that are both theoretically grounded and practically relevant from this selective approach to attribute integration, including RLT, and the emphasis on generalizability and equity position this research as a significant step forward in developing robust and applicable predictive models in education.

## 2.9 Conclusion

The literature review has comprehensively explored the various factors that influence student performance and the methodologies used to predict educational outcomes. Through an in-depth examination of cognitive, behavioral, demographic, psychological, and teacher pedagogy attributes, the review has highlighted the diverse range of predictors that can be considered when developing predictive models. However, it is essential to note that while identifying these attributes based on the literature offers a broad perspective, not all attributes will be utilized directly in the predictive models developed in this research. Instead, this review serves as an insightful foundation, guiding the selective integration of our context's most relevant and impactful attributes.

A critical gap identified in the literature is the limited research on the generalizability of predictive models across different educational contexts. Many existing models are designed and validated within single institutions or homogenous student populations, raising concerns about their applicability in more diverse settings. This dissertation makes a unique and significant contribution by addressing this gap. It focuses on the development of predictive models that are generalizable and reliable across multiple educational environments, ensuring that they are both accurate and applicable in various contexts.

Additionally, the literature review highlights the importance of incorporating RLT and addressing gender and location-based disparities, especially within developing countries. This research responds to these needs by designing ReflectMate, a reflective tool focused on student empowerment—and developing the CISE model to ensure fair and generalizable predictions. Together, these contributions aim to improve the relevance of predictive analytics while advancing educational equity and inclusion.

In summary, the insights gained from this literature review will play a pivotal role in shaping the methodological approach of this study. It ensures that the predictive models developed are not only grounded in existing research but also tailored to meet the specific needs of diverse educational settings. The following chapters will build upon this foundation, detailing the research design and methodological approaches used to select, operationalize, and validate the predictive models that are at the heart of this dissertation.

The next Chapter, Chapter 3: Research Methodology, will outline the methods and approaches used in this research to address the research questions and achieve the objectives identified in this literature review.

Chapter 3

# Research Methodology

## 3.1 Introduction

This chapter outlines the methodological approach of this research, detailing the general workflow and stages undertaken to develop a predictive modeling and reflective learning analytics system. The research methodology is structured into four primary phases: (1) Data Extraction, (2) Data Preprocessing, (3) Predictive Modeling via the CISE (Cross-Institutional Stacking Ensemble) Model, and (4) ReflectMate Development (reflective tool). These phases were systematically executed to ensure robust data processing, accurate predictive modeling, and effective reflective learning support.

## 3.2 Research Phases Overview

The research methodology is divided into four key phases, each designed to address specific requirements for predictive modeling and reflective learning support. The phases include general data handling across multiple datasets, model training and evaluation, and ReflectMate functionality development.

Figure 3.1 illustrates the workflow of the research methodology, including the data extraction, data preprocessing, CISE model, and ReflectMate.

1. **Phase 1: Data Extraction**—in this phase, data was extracted from multiple

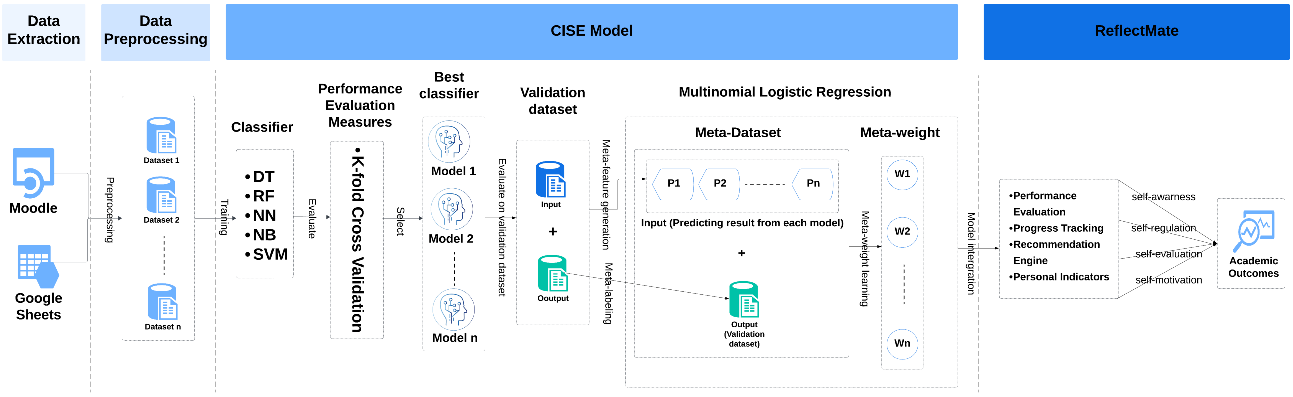


Figure 3.1 – The research methodology workflow

sources, including Learning Management Systems (LMS) such as Moodle, and student grades recorded in institutional data. Various LMS engagement metrics were collected, such as assignment submissions, quiz attempts, and interaction logs, which formed the basis of the predictive model's features.

2. **Phase 2: Data Preprocessing**—this phase involved cleaning and standardizing data for consistency across datasets. The grading scales, engagement logs, and other performance metrics were normalized and structured into feature sets. Data preprocessing included handling missing data, and normalizing scores across different grading scales to predicting academic outcomes.

3. **Phase 3: CISE Model Development**—the Cross-Institutional Stacking Ensemble (CISE) model, developed during this phase, aimed to provide robust performance predictions across varied datasets. The model leveraged multiple classifiers identified from the literature (Decision Trees, Random Forest, Neural Network, Naive Bayes, and Support Vector Machine) commonly used in predicting student performance. Each classifier was trained and used k-fold cross-validation to select the best-performing models as base classifiers. A multinomial logistic regression model was then applied as a meta-model, using the predictions from the base mod-

els as meta-features. This phase also included performance evaluation measures to determine model accuracy and reliability.

4. **Phase 4: ReflectMate Development**—ReflectMate was designed to support reflective learning by providing personalized indicators and recommendations for students. This tool facilitates self-awareness, self-regulation, self-evaluation, and self-motivation, aiming to enhance students' academic outcomes through features such as performance evaluation, progress tracking, and a recommendation engine. (Detailed descriptions of ReflectMate's functionality will be covered in Chapter 5.)

## 3.3 Phase 1: Data Extraction

The first phase of this research involved extracting data from multiple sources to create a comprehensive dataset for predictive modeling and reflective learning analytics. The primary data sources were Learning Management Systems (LMS), Moodle, and institutional data. These two systems served complementary roles in capturing different student behavior and performance aspects, ensuring a holistic view of the learning process.

- **Moodle Logs:** as the primary platform for course delivery, Moodle provided detailed logs of student interactions with course materials. These metrics included, A File Has Been Uploaded, A Submission Has Been Submitted, Attendance Taken by Student, Course Activity Completion Updated, Course Module Viewed, Course Viewed, Quiz Attempt Viewed, Session Report Viewed, Submission Created, Submission Form Viewed, Submission Status Viewed, and User Accepted Statement of Submission. These engagement metrics are critical predictors of student performance and form the foundation of the predictive models developed in this study.
- **Institutional Data:** while Moodle captured interaction data, students' final grades or scores were recorded in institutional data. It was necessary because Moodle's

grading system was either not fully utilized by instructors or did not align with the institution's preferred methods for storing academic records. By integrating institutional data data containing the final scores assigned by instructors, we ensured that the predictive models could correlate students' LMS engagement patterns with their actual academic outcomes, providing a complete picture of their performance.

The combination of Moodle and institutional data highlights an important distinction between learning behaviors and academic results. Moodle excels at tracking how students interact with course content—when they log in, what resources they access, and how frequently they engage with assessments. However, these interactions alone do not provide insight into whether those efforts translate into tangible academic success. For instance:

— Students may spend significant time on quizzes but still perform poorly if they misunderstand key concepts.

— Another student might have minimal LMS activity yet achieve high grades due to external preparation or prior knowledge.

We bridge this gap by incorporating institutional data, which contains the final grades assigned by instructors. The integration allows us to connect behavioral indicators from Moodle with concrete outcomes stored in institutional data, enabling more accurate predictions about student performance. This dual-source approach ensures that the predictive model accounts for effort (engagement) and achievement (grades).

### 3.3.1 Data Extraction Pipeline

To streamline the extraction process while ensuring privacy protection, a robust pipeline was established to integrate data from Moodle logs and institutional data efficiently. Below is a step-by-step breakdown of the pipeline:

1. **Extracting Raw Data from Moodle :**

- Interaction logs were pulled directly from Moodle's database using APIs or manual export functions. These logs included timestamps for activities like logging in, submitting assignments, attempting quizzes, and viewing course modules.
- To ensure privacy compliance, each log entry was tagged with an encrypted identifier. This encryption process replaced identifiable information, such as student names or IDs, with unique, anonymized codes that could not be traced back to individuals.

2. **Retrieving Final Grades from institutional data :**
   - Instructors maintained final grade records in institutional data, often organized by course and semester. These sheets were exported as CSV files and cleaned to remove irrelevant columns (e.g., comments or notes).
   - Like Moodle data, each record in institutional data was matched to the corresponding encrypted identifier used in Moodle. It ensured consistency during data merging without compromising student privacy.
3. **Merging Datasets :**
   - Once extracted, the Moodle and institutional data datasets were merged based on the shared encrypted identifiers. This created a unified dataset in which each row represented a single student, with columns detailing their LMS interactions and final grades.
   - Any discrepancies in formatting or labeling were resolved during this stage to ensure seamless analysis downstream.
4. **Labeling Datasets :**
   - Each dataset was labeled uniquely to maintain generalizability across institutions (e.g., Dataset 1, Dataset 2). This labeling allowed researchers to track the origin of the data while applying consistent preprocessing techniques across all

datasets.

### 3.3.2 Example Scenario: Connecting Engagement Metrics with Academic Outcomes

Consider a hypothetical scenario involving a student named Alex enrolled in an online programming course:

— **Moodle Logs:** Over the semester, Alex accessed the course module 15 times, attempted quizzes three times, submitted two out of four assignments, and participated in one discussion forum thread.

— **Institutional Data Record:** Despite moderate engagement in Moodle, Alex received a final grade of B+ (87%), according to the instructor's evaluation in institutional data. In this case, the combined dataset would show that Alex's limited LMS activity did not prevent them from achieving a substantial grade. Upon further investigation, it might be revealed that Alex supplemented their learning through offline study groups or external coding tutorials. Such insights underscore the importance of integrating diverse data sources; without institutional data data, the predictive model might incorrectly assume that Alex was underperforming based solely on their Moodle activity.

### 3.3.3 Challenges Addressed During Data Extraction

Several challenges arose during the data extraction phase, particularly related to data quality and compatibility:

— **Missing Data:** Some students had incomplete Moodle logs due to technical issues or non-participation. Similarly, specific entries in institutional data lacked final grades, requiring careful handling during preprocessing.

— **Disparate Formats:** Moodle logs and institutional data often followed different structures and formats, necessitating standardization before merging.

— **Privacy Concerns:** Encrypting all student identifiers before proceeding with analysis ensured compliance with ethical guidelines. This ensured that no personally identifiable information (PII) was exposed during the research process.

These challenges were addressed systematically during subsequent phases of data preprocessing, as described in Section 3.4.

## 3.4 Phase 2: Data Preprocessing

Data preprocessing involved preparing the extracted datasets for analysis. This phase included handling missing values, normalizing grading scales, and engineering features.

### 3.4.1 Normalization of Grading Scales

Grading scales varied across datasets, necessitating normalization for cross-dataset compatibility. Each grading score was normalized using the formula:

$$\text{Normalized Grade} = \frac{\text{Grade}}{\text{Max Grade}} \tag{3.1}$$

Once normalized, grades were categorized (e.g., A, B+, B, etc.) using the academic standard grading system, as shown in Appendix A, to frame the prediction task as a classification problem, thereby improving interpretability and aligning with educational assessment practices.

### 3.4.2 Handling Missing Data

Missing data was managed using multiple imputation strategies. Median imputation was applied to engagement metrics (e.g., time spent), while zero imputation was used for features indicating non-participation (e.g., assignment completion). Rows with missing target values (grades) were removed to maintain consistency.

### 3.4.3 Feature Engineering

Feature engineering was a critical step in transforming raw data into meaningful variables that could be used to train predictive models. This process was essential because raw data, such as LMS logs, often lack the structure and clarity needed for machine learning algorithms to make accurate predictions. Without proper feature engineering, the models would struggle to identify student behavior and performance patterns, leading to less reliable outcomes.

The main challenge in feature engineering was determining which aspects of the data were most relevant to predicting student performance. For example, raw metrics like *time spent on the platform* or *number of clicks* would not provide enough context for the model to understand how these behaviors relate to academic success. Instead, we needed to create features that captured the essence of student engagement and learning habits. We focused on key features such as *Quiz Attempts*, *Assignment Submissions*, and *Module Views* to address this. These features were chosen because they directly reflect how students interact with course materials and assessments, providing a clearer picture of their learning behaviors.

For instance, *Quiz Attempts* were engineered to measure whether a student attempted a quiz and how many times they tried, indicating persistence and effort. Similarly, *Assignment Submissions* tracked whether an assignment was submitted and whether it was done on time, offering insights into time management skills. *Module Views* counted how often students accessed different parts of the course content, helping to identify areas where they might be struggling or excelling. By focusing on these features, we ensured that the predictive models had access to rich, actionable data that could reliably capture students' learning behaviors.

This approach allowed us to build a foundational dataset that could effectively train the models to predict student outcomes. This careful selection of features ensured that the models could interpret the data accurately and provide meaningful predictions. In short, feature engineering was necessary to bridge the gap between raw, unstructured data and the structured inputs required for machine learning, enabling the models to understand student performance better and predict it.

## 3.5 Phase 3: CISE Model Development

The Cross-Institutional Stacking Ensemble (CISE) model is designed to predict student performance using a meta-model approach, combining multiple classifiers to achieve higher predictive accuracy. This section details the training of base classifiers, evaluation metrics, and the development of the meta-model.

### 3.5.1 Training Multiple Classifiers

The CISE model incorporates five classifiers—Decision Tree (DT), Random Forest (RF), Neural Network (NN), Naive Bayes (NB), and Support Vector Machine (SVM)—selected based on their effectiveness in similar studies on student performance prediction. Using k-fold cross-validation, each classifier was trained on the various datasets (Dataset 1, Dataset 2, ..., Dataset n) to ensure robustness across different data samples.

### 3.5.2 Evaluation of Base Classifiers

The performance of each base classifier was evaluated using k-fold cross-validation. This approach provided insights into the consistency and generalizability of each classifier across datasets. After cross-validation, the best-performing classifiers were selected as base models for the CISE ensemble.

### 3.5.3 Meta-Model Development

The meta-model development approach described in this research—using multinomial logistic regression as a meta-model to integrate predictions from multiple base classifiers—is not entirely novel but has been successfully applied in various domains, including educational data mining (EDM) and other fields involving ensemble learning. However, its application in cross-institutional predictive modeling for student performance is relatively innovative. It aligns well with the challenges of generalizability in education.

Meta-modeling, or stacking, is a well-established technique in machine learning that combines predictions from multiple base models to improve overall performance. Logistic regression as a meta-model is particularly common because it is interpretable, computationally efficient, and combines outputs from diverse classifiers effectively. For example:

— Dietterich (2000) introduced stacking as a method to combine the strengths of different models, emphasizing its ability to reduce bias and variance while improving generalization.

— Wolpert (1992) first formalized stacked generalization, demonstrating its effectiveness in scenarios where individual models have complementary strengths.

In educational contexts, stacking has been used to enhance predictive accuracy in tasks such as predicting student dropout rates and academic success:

— Kotsiantis and Kanellopoulos (2012) applied stacking to predict student performance using datasets from multiple institutions, achieving higher accuracy than single-model approaches.

— Porras et al. (2023) explored cross-institutional validation of predictive models in education, highlighting the importance of ensemble methods like stacking to address variability across datasets.

These studies underscore the utility of stacking in educational data mining, mainly when dealing with heterogeneous datasets from different institutions or populations.

The specific challenge addressed in this research—developing predictive models that generalize across diverse educational settings—is a growing area of interest in EDM. While ensemble learning techniques like Random Forests and Gradient Boosting are widely used, fewer studies have explicitly focused on stacking as a solution for cross-institutional generalizability. However, there are notable examples:

— Reich (2020) discussed the need for equitable and inclusive predictive models in education, advocating for ensemble methods that account for contextual differences between institutions. Their work highlights the potential of meta-models to mitigate biases arising from institutional disparities.

— Gašević, S. Dawson, and Siemens (2015) proposed a framework for learning analytics that integrates data from multiple sources, emphasizing the role of ensemble methods in creating robust and interpretable models. Their approach aligns closely with the CISE model's focus on integrating diverse datasets through a meta-model.

The choice of multinomial logistic regression as the meta-model is supported by its suitability for classification tasks with multiple classes (e.g., predicting letter grades). This approach offers several advantages:

1. **Interpretability:** Unlike complex models like Neural Networks, logistic regression provides insight into how each base classifier contributes to the final prediction. This transparency is crucial in educational contexts where stakeholders must trust

and understand the model's decisions.

2. **Flexibility:** Multinomial logistic regression can handle multi-class problems directly, making it ideal for predicting discrete grade categories (A, B+, B, etc.) rather than relying on binary classifications.

3. **Efficiency:** Logistic regression is computationally lightweight, allowing faster training and evaluation than more complex meta-models.

Incorporating meta-modeling into reflective learning tools like ReflectMate further enhances their utility. By leveraging predictions from diverse classifiers trained on cross-institutional data, ReflectMate can provide students with more accurate and personalized feedback. For instance:

1. L. P. Macfadyen and S. Dawson (2010) demonstrated that combining multiple data sources and models improved the accuracy of learning analytics tools, enabling better support for student self-reflection and goal-setting.

2. Boud, Keogh, and Walker (2013) highlighted the importance of actionable feedback in reflective learning, emphasizing the need for tools that integrate diverse data streams to provide holistic insights.

While using multinomial logistic regression as a meta-model in ensemble learning is not new, its application in cross-institutional predictive modeling for education represents a significant advancement. This approach addresses key challenges in EDM, such as generalizability, while supporting the development of reflective learning tools like ReflectMate. The references cited above validate the choice of this methodology and highlight its potential to transform educational practices by empowering educators and students with reliable, data-driven insights. By building on existing research and adapting proven techniques to the unique demands of educational data, this study contributes to the growing body of knowledge on ensemble methods in EDM. It lays the groundwork for future innovations in reflective learning analytics.

## 3.6 Phase 4: ReflectMate Development

ReflectMate was designed as an innovative tool to empower students by providing personalized, data-driven insights into their learning behaviors and academic progress. The primary goal of ReflectMate is to foster self-awareness, self-regulation, self-evaluation, and self-motivation among students, enabling them to take ownership of their learning journey. By integrating predictive analytics with reflective learning principles, ReflectMate aims to enhance students' academic outcomes while promoting lifelong learning skills. ReflectMate operates through four core functionalities, each tailored to support students in different aspects of their learning process:

1. **Performance Evaluation:** This feature allows students to clearly understand their academic performance relative to their peers. Students can identify areas where they excel or need improvement by comparing metrics such as quiz scores, assignment submissions, and overall engagement. This comparative analysis helps students contextualize their progress and motivates them to achieve better results.
2. **Progress Tracking:** ReflectMate provides students with real-time tracking of their engagement metrics, such as the number of quiz attempts, time spent on course materials, and assignment completion rates. These metrics are visualized intuitively, enabling students to monitor their progress over time. ReflectMate encourages students to reflect on their study patterns and make necessary adjustments by offering a transparent view of their learning habits.
3. **Recommendation Engine:** One of ReflectMate's standout features is its ability to offer personalized recommendations based on individual student data. Using insights derived from predictive models (developed in Phase 3), ReflectMate suggests

actionable strategies to improve learning outcomes. For example, suppose a student shows low engagement with quizzes. In that case, ReflectMate might recommend increasing quiz attempts or revisiting specific modules to reinforce understanding. These tailored suggestions help students focus their efforts effectively and address weaknesses proactively.

4. **Personal Indicators:** ReflectMate also includes personal indicators that display trends and patterns in students' learning behaviors. These indicators highlight key aspects of their engagement, such as consistency in completing tasks, time management, and interaction with course content. By presenting these insights in an accessible format, ReflectMate supports students in developing self-awareness and making informed decisions about their learning strategies.

The development of ReflectMate aligns closely with the principles of reflective learning, which emphasize the importance of reviewing and analyzing one's learning experiences to foster deeper understanding and critical thinking. ReflectMate bridges the gap between traditional feedback systems and student-centered tools by combining these principles with advanced learning analytics. Unlike conventional learning analytics platforms, which primarily serve educators, ReflectMate focuses on students, empowering them to actively engage with their data and take charge of their academic growth.

The detailed functionalities of ReflectMate, including its impact on students' academic outcomes and its role in fostering reflective learning practices, will be explored further in Chapter 5. This chapter will also examine how ReflectMate integrates with the predictive models developed earlier in the research, creating a seamless connection between data-driven insights and actionable student empowerment.

## 3.7 Conclusion

The research methodology outlined in this chapter encompasses four key phases aimed at developing predictive models and reflective learning tools. The first phase involved extracting data from multiple sources, including LMS like Moodle and performance records from institutional data, to gather a wide range of engagement metrics and assessment scores. It was followed by the second phase, which focused on preprocessing the data to handle inconsistencies, normalizing grading scales, and engineering meaningful features for model training. In the third phase, the CISE model was developed, leveraging five machine learning classifiers—Decision Tree, Random Forest, Neural Network, Naive Bayes, and Support Vector Machine—trained and evaluated using k-fold cross-validation. The best-performing classifiers were selected as base models, and a multinomial logistic regression meta-model was applied to integrate their predictions. Finally, the fourth phase involved the development of ReflectMate, a tool designed to empower students through personalized insights into their learning progress. ReflectMate supports self-awareness, self-regulation, self-evaluation, and self-motivation by offering performance evaluation, progress tracking, and a recommendation engine. This structured approach ensures comprehensive data handling and model development, emphasizing personalized, reflective learning support. Further details on the CISE model are provided in Chapter 4, while Chapter 5 explores ReflectMate's role in promoting student empowerment.

Chapter 4

# Development of Predictive Models

## 4.1 Introduction

Predicting student performance is a cornerstone of modern educational data mining (EDM), enabling educators to identify at-risk students and develop targeted interventions to improve academic outcomes (Romero and Ventura 2020). However, the complexity and variability of educational datasets pose significant challenges for predictive modeling. Traditional approaches often rely on single classifiers trained on institution-specific data, which limits their generalizability across diverse educational contexts (Porras et al. 2023). To address these limitations, this study introduces the **Cross-Institutional Stacking Ensemble (CISE)** model, an innovative approach designed to enhance predictive accuracy and generalizability in student performance prediction.

The development of the CISE model builds upon the foundational principles of ensemble learning, which combines multiple machine learning algorithms to improve overall performance (Dietterich 2000). Initially, individual classifiers—such as Decision Trees (DT), Random Forests (RF), Neural Networks (NN), Naive Bayes (NB), and Support Vector Machines (SVM)—were trained and evaluated on datasets from multiple institutions. These classifiers were selected based on their demonstrated effectiveness in similar studies on student performance prediction (Kotsiantis and Kanellopoulos 2012). However, while individual classifiers provide valuable insights, they often struggle to generalize across datasets with varying grading systems, feature distributions, and class imbalances (Baker et al. 2019).

To overcome these challenges, the CISE model integrates a meta-modeling approach, leveraging the strengths of multiple base classifiers through stacking. Stacking involves training a meta-model to combine the predictions of base models, creating a unified framework that captures complementary patterns in the data (Wolpert 1992). In this study, multinomial logistic regression was employed as the meta-model, dynamically weighing the contributions of each base classifier to optimize predictive accuracy. This approach enhances generalizability and addresses issues such as overfitting and class imbalance, prevalent in educational datasets (Ashraf, Zaman, and M. Ahmed 2020).

The design of the CISE model follows a systematic step-by-step process:

1. **Training Individual Classifiers:** Five machine learning algorithms—decision Tree, Random Forest, Neural Network, naive bayes, and Support Vector Machine—were trained on three distinct institutional datasets: the Institute of Digital Technology (IDT) and the Institute of Digital Governance (IDG), both of which are part of the Cambodia Academy of Digital Technology (CADT), and the Institute of Technology of Cambodia (ITC), which is a separate institution. These datasets represent diverse academic environments, contributing to digital education and technology research.

2. **Selecting Base Models:** Selecting Base Models: The best-performing classifiers from each dataset were identified and validated on an unseen dataset from École Nationale Supérieure d'Ingénieurs du Mans (ENSIM) to ensure robustness and generalizability.

3. **Constructing the Meta-Model:**A multinomial logistic regression model was developed to integrate the predictions of the selected base models, forming the

final CISE model.

The introduction of the CISE meta-model represents a significant advancement in predictive modeling for education. Unlike traditional models that rely on a single algorithm or dataset, the CISE model leverages the collective strengths of multiple classifiers, enabling it to adapt to variations in data structure and context. This flexibility is significant in cross-institutional settings, where differences in curricula, teaching practices, and student demographics can significantly impact model performance (Gardner et al. 2023).

The results demonstrate the effectiveness of ensemble methods in achieving high predictive accuracy and generalizability across institutional boundaries. The chapter structure is as follows:

— **Section 4.2** outlines data preprocessing and feature engineering.

— **Section 4.3** evaluates individual classifiers on training datasets.

— **Section 4.5** constructs the Meta-Model.

— **Section 4.6** presents validation results for the base models and meta-models.

— **Section 4.7** discusses the results and their implications.

— **Section 4.8** concludes the chapter by summarizing key findings.

The CISE model offers a robust solution for predicting student performance in diverse educational contexts by systematically addressing the challenges of variability, generalizability, and interpretability. This chapter aims to comprehensively understand the model's architecture, underlying mechanisms, and potential applications in real-world educational settings.

## 4.2 Data Preprocessing and Feature Engineering

The datasets from IDT, IDG, ITC, and ENSIM varied in structure, grading scales, and features related to student engagement. Thus, data preprocessing and feature engineering were crucial in preparing the data for model training. This section outlines the steps to clean, normalize, and transform the raw data into meaningful input features for the machine learning models.

### 4.2.1 Normalization of Grading Scales

One of the key challenges was the discrepancy in grading scales across the different datasets. For instance, the IDT, IDG, and ITC datasets used a grading scale ranging from 0 to 100, while the ENSIM dataset used a grading scale of 0 to 30. Normalization was applied to standardize the grades to ensure the model's consistency and comparability across institutions.

After normalization, grades were converted into discrete categories (e.g., A, B+, B, C+, C, D, E, F) to frame the prediction task as a classification problem. These categories enabled the machine learning models to predict letter grades, improving interpretability and aligning the output with academic grading standards.

Table 4.1 illustrates the original and normalized grading scales for each dataset.

Table 4.1 – Original and Normalized Grading Scales Across Datasets

| **Institution** | **Original Scale** | **Normalized Scale (0-1)** |
|---|---|---|
| IDT | 0-100 | 0-1 |
| IDG | 0-100 | 0-1 |
| ITC | 0-100 | 0-1 |
| ENSIM | 0-30 | 0-1 |

### 4.2.2 Handling Missing Data

More data is needed in educational datasets, especially for engagement metrics like LMS interaction logs. Various techniques were employed to address missing values:

— **Median Imputation**: for numerical features such as *Attendance*, missing values were replaced with the median of the feature to prevent outliers from skewing the data.

— **Zero Imputation**: for features indicating non-participation (e.g., *Quiz Attempts* or *Tasks Submitted*), missing values were replaced with 0. This distinction allowed the models to differentiate between students who participated in specific activities and those who did not engage at all.

— **Dropping Rows with Missing Targets**: rows with missing values for the target variable (final grade) were removed from the dataset to ensure that only complete records were used for training and evaluation.

Table 4.2 shows the percentage of missing values handled in each dataset.

Table 4.2 – Percentage of Missing Data Imputed in Each Dataset

| **Institution** | **Numerical Features** | **Target (Grade)** |
|---|---|---|
| IDT | 0.0% | 0.0% |
| IDG | 0.0% | 0.0% |
| ITC | 67.48% | 0.12% |
| ENSIM | 0.0% | 0.0% |

### 4.2.3 Feature Engineering

Feature engineering was critical in transforming raw LMS log data into meaningful features for machine learning models. The key features included in the datasets were derived primarily from student engagement with course materials, as recorded by LMS systems. These features captured how students interacted with the online learning environment and how these interactions related to their final academic performance.

The key features engineered from the raw datasets were as follows:

— **Previous Grade**: a strong predictor of future performance, this feature provided context for each student's academic trajectory, using their prior term or semester grade.

— **Engagement Metrics**: features such as *Attendance*, *Quiz Attempts*, *Tasks Submitted*, and *Course Module Views* were extracted from LMS logs to quantify student engagement with online content. These engagement metrics helped the model assess how actively students participated in their coursework.

These features provided the machine learning models with a rich set of variables to predict student performance, focusing on how students' behaviors and interactions with LMS systems affected their academic outcomes.

Table 4.3 presents a summary of the key features engineered from the raw datasets.

| **Feature** | **Description** | **Source** |
|---|---|---|
| Previous Grade | Final grade from the previous semester | Academic records |
| A File Has Been Uploaded | Number of file was uploaded by a student | LMS logs |
| A Submission Has Been Submitted | Number of a student submitted an assignment | LMS logs |
| Attendance Taken by Student | Number of a student attended a class or session | LMS logs |
| Course Activity Completion Updated | Number of a student completes a course activity | LMS logs |
| Course Module Viewed | Number of course modules viewed by the student | LMS logs |
| Course Viewed | Number of a student views a course | LMS logs |
| Quiz Attempt Viewed | Indicator of whether the student viewed a quiz attempt | LMS logs |
| Session Report Viewed | Number of a student viewed the session report | LMS logs |
| Submission Created | Indicator when a submission is created by the student | LMS logs |
| Submission Form Viewed | Number of a student views the submission form | LMS logs |
| Submission Status Viewed | Number of a student has viewed the status of their submission | LMS logs |
| User Accepted Statement of Submission | Number of a student accepted the statement for submission | LMS logs |

Table 4.3 – Key Features Engineered from Raw Datasets

The features listed in Table 4.3 were selected based on the following considerations:

1. **Alignment with Predictive Goals:** The primary objective of the predictive model is to forecast student performance, which is influenced by academic history and engagement behaviors. Features such as *Previous Grade* and *Course Module Viewed* directly reflect these factors. For instance, prior academic performance (*Previous Grade*) is a well-established predictor of future success (Yağcı 2022b). At the same time, engagement metrics like *Course Module Viewed* capture the extent to which students interact with course materials, a strong indicator of commitment and effort (Brahim 2022).

2. **Empirical Evidence from Literature:** The selection of features was guided by

findings from existing research. Studies have consistently shown that engagement metrics—such as attendance, assignment submissions, and quiz participation—are significant predictors of academic outcomes (Umer et al. 2023). For example, Smith, Cobham, and Jacques (2022) demonstrated that higher levels of engagement, as measured by these metrics, are positively correlated with better academic performance.

3. **Availability and Consistency Across Datasets:** Not all potential features could be included due to limitations in data availability and consistency across datasets. For instance, while socio-economic factors (e.g., parental income) are known to influence student performance (Segura-Morales and Loza-Aguirre 2017), such data was unavailable in the current datasets. Similarly, some features, such as psychological attributes (e.g., motivation), are challenging to quantify objectively without additional surveys or self-reported data.

4. **Practical Relevance and Actionability:** The selected features are actionable and can inform interventions to improve student outcomes. For example, if a student's *Quiz Attempt Viewed* metric is low, educators can encourage them to engage more with quizzes to reinforce their understanding of key concepts. This practical relevance ensures that the predictive model provides accurate predictions and actionable insights.

5. **Reduction of Redundancy:** Features highly correlated or redundant were excluded to avoid multicollinearity, which can degrade model performance. For instance, while *Course Viewed* and *Course Module Viewed* measure engagement, they were retained because they capture different aspects of interaction with the LMS.

6. **Focus on Digital Learning Environments:** Given the increasing reliance on digital tools in education, the selected features emphasize interactions within the

LMS. These include indicators such as *A Submission Has Been Submitted* and *Session Report Viewed*, which reflect students' active participation in the digital learning environment. This focus aligns with the broader trend toward online and blended learning (Valverde-Berrocoso et al. 2021).

While other features, such as demographic attributes (e.g., age, gender) and psychological factors (e.g., stress, self-efficacy), have been identified in the literature as potential predictors of student performance (S. P. Dawson, L. Macfadyen, and Lockyer 2009), they were not included in this study for several reasons:

— **Data Limitations:** Demographic and psychological data were unavailable or inconsistent across datasets.

— **Ethical Considerations:** Including sensitive attributes like gender or socio-economic status raises ethical concerns about bias and fairness in predictive modeling (Mengash 2020).

— **Focus on Engagement:** The study prioritized features that directly reflect student engagement and behavior, as these are more actionable and less prone to bias than static demographic attributes.

By carefully selecting features that align with the predictive goals, are empirically supported, and are practically relevant, the engineered dataset provides a robust foundation for training the predictive models. This approach ensures that the models are interpretable, generalizable, and capable of driving meaningful interventions to support student success.

### 4.2.4 Handling Class Imbalance

One of the challenges faced during model training was class imbalance, where most students achieved middle-range grades (e.g., B, C). In contrast, fewer students achieved the highest (e.g., A) or lowest grades (e.g., F). Class imbalance is a common issue in educational datasets. It can significantly impact the performance of machine learning models, particularly their ability to accurately predict outcomes for underrepresented classes (H. He and Garcia 2009). To address this issue, the **RandomOverSampler** method was applied to balance the dataset by generating synthetic examples for underrepresented grades. This technique ensured that the model had sufficient data to learn from all grade categories, improving its ability to predict student outcomes across the full spectrum of academic performance. Random oversampling has been widely used in imbalanced classification tasks due to its simplicity and effectiveness in mitigating bias toward majority classes (Chawla et al. 2002).

## 4.3 Classifier Performance on Training Datasets

Five classifiers' performance was evaluated on the IDT, IDG, and ITC datasets using metrics such as accuracy, precision, recall, and F1 score. The results provide insights into each algorithm's strengths and weaknesses.

### 4.3.1 Performance on the IDT Dataset

The IDT dataset results, summarized in Table 4.4, highlight the superior performance of Decision Tree, achieving the highest F1 score (77.03%).

Table 4.4 – Performance of Classifiers on the IDT Dataset

| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Decision Tree | 79.90% | 77.83% | 79.90% | **77.03%** |
| Random Forest | 79.98% | 78.27% | 79.98% | 76.95% |
| Naive Bayes | 36.06% | 30.65% | 36.06% | 29.71% |
| Neural Network | 41.43% | 37.43% | 41.43% | 37.38% |
| SVM | 41.97% | 36.10% | 41.97% | 37.70% |

**Analysis**:

- Decision Tree emerged as the top performer with an F1 score of 77.03%. Its ability to handle categorical splits and its straightforward structure allowed it to effectively model the relationships between features in the IDT dataset, which included LMS engagement logs and academic performance metrics.
- Random Forest, although slightly behind the Decision Tree in F1 score, maintained competitive results across all metrics. Its ensemble approach provided resilience to noise and variability in the dataset, making it a robust alternative for applications requiring higher accuracy and recall.
- Naive Bayes underperformed significantly, with an F1 score of 29.71%. Its reliance on the independence assumption rendered it ill-suited for the interdependent features in the IDT dataset.
- Neural Network and SVM achieved moderate results, with F1 scores of 37.38% and 37.70%, respectively. Neural Network struggled with convergence due to the relatively small size of the IDT dataset. SVM's performance was hindered by the dataset's imbalance across grade categories, limiting its recall.

### 4.3.2 Performance on the IDG Dataset

Table 4.5 presents the IDG dataset results. Random Forest achieved the highest F1 score (87.59%), marginally outperforming Decision Tree.

Table 4.5 – Performance of Classifiers on the IDG Dataset

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 88.47% | 88.63% | 88.47% | 87.43% |
| Random Forest | 88.41% | 88.38% | 88.41% | **87.59%** |
| Naive Bayes | 31.16% | 29.78% | 31.16% | 28.65% |
| Neural Network | 29.68% | 26.35% | 29.68% | 26.48% |
| SVM | 30.55% | 30.14% | 30.55% | 29.36% |

**Analysis**:

- Random Forest excelled due to its ensemble nature, which combines multiple decision trees to reduce variance and overfitting. Its robustness in handling missing values and noisy data was particularly beneficial for the IDG dataset, which contained engagement metrics and performance records from digital governance courses.
- Decision Tree demonstrated competitive performance with an F1 score of 87.43%. Its interpretability makes it a valuable tool for understanding the contributions of various features to student performance. However, it was slightly less accurate than Random Forest, likely due to its tendency to overfit on complex datasets.
- Naive Bayes performed poorly, with an F1 score of 28.65%. This underperformance stemmed from the classifier's reliance on the independence assumption, which is often violated in datasets with interdependent features like LMS engagement metrics.
- Neural Network and SVM also struggled, with F1 scores of 26.48% and 29.36%, respectively. The Neural Network faced challenges due to the dataset size, which was insufficient for effective learning in its multi-layer architecture. SVM demonstrated reasonable precision but lower recall, reflecting difficulties in balancing predictions across class boundaries.

### 4.3.3 Performance on the ITC Dataset

Random Forest outperformed all classifiers on the ITC dataset, as shown in Table 4.6.

**Analysis:**

- Random Forest excelled on the ITC dataset due to its ensemble nature, which allowed it to capture intricate relationships between input features. Its ability to handle non-linear interactions and mitigate overfitting contributed to its strong performance, particularly in achieving the highest F1 score (94.02%).

Table 4.6 – Performance of Classifiers on the ITC Dataset

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 93.74% | 94.02% | 93.74% | 93.70% |
| Random Forest | 94.06% | 94.36% | 94.06% | **94.02%** |
| Naive Bayes | 39.67% | 60.55% | 39.67% | 32.82% |
| Neural Network | 50.06% | 52.73% | 50.06% | 45.43% |
| SVM | 50.97% | 60.60% | 50.97% | 49.72% |

- Decision Tree closely followed Random Forest in accuracy (93.74%), precision (94.02%), and recall (93.74%). While slightly less robust than Random Forest, its simplicity and interpretability make it a valuable alternative for analyzing educational data.
- Naive Bayes exhibited significant limitations, with an F1 score of only 32.82%. This performance was hindered by violating its feature independence assumption, which does not align with the complex interdependencies in LMS data.
- Neural Network and SVM demonstrated moderate performance. The Neural Network achieved a balanced precision-recall trade-off but struggled with overfitting due to the relatively small size of the ITC dataset. SVM showed strong precision but lower recall, indicating its sensitivity to the class distribution.

## 4.4 Classifier Performance on Validation Dataset

To assess the standalone performance of individual machine learning algorithms, an experiment was conducted using the ENSIM dataset for both training and validation. This intra-dataset evaluation helps determine the effectiveness of each model in a controlled environment, serving as a baseline for later cross-institutional comparisons.

Five classifiers were trained and validated using k-fold cross-validation on the ENSIM dataset: Decision Tree, Random Forest, Naïve Bayes, Neural Network, and Support Vector Machine (SVM). Table 4.7 summarizes the evaluation metrics.

Table 4.7 – Performance of Classifiers Trained and Validated on the ENSIM Dataset

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Decision Tree | 83.58% | 72.15% | 83.58% | 77.11% |
| Random Forest | 83.58% | 72.15% | 83.58% | 77.11% |
| Naïve Bayes | 83.96% | 74.68% | 83.96% | 78.23% |
| Neural Network | **85.07%** | **72.70%** | **85.07%** | **78.40%** |
| SVM | 84.33% | 71.89% | 84.33% | 77.51% |

**Analysis:**

- The **Neural Network** achieved the highest overall performance across all metrics, suggesting its strong ability to model complex relationships in the dataset.
- **Naïve Bayes** performed surprisingly well, particularly in F1 Score, despite its assumption of feature independence.
- **Decision Tree** and **Random Forest** exhibited identical performance, indicating that additional trees in Random Forest did not significantly improve predictive capability for this dataset.
- **SVM** maintained competitive results, with balanced performance across precision, recall, and F1.

These results confirm that multiple classifiers are suitable for modeling student performance on the ENSIM dataset and highlight the value of combining their strengths in an ensemble framework like the CISE model.

## 4.5 Construction of the Meta-Model

The **meta-model** is the backbone of the Cross-Institutional Stacking Ensemble (CISE), integrating the strengths of multiple base classifiers to produce a unified and robust prediction. This section provides a detailed explanation of the meta-model's design, implementation, and rationale.

### 4.5.1 Purpose of the Meta-Model

The primary purpose of the meta-model is to combine the predictions of the base classifiers into a single, cohesive output. By leveraging the complementary strengths of Decision Trees, Random Forests, Neural Networks, Naive Bayes, and Support Vector Machines, the meta-model addresses the limitations of individual classifiers, such as overfitting, sensitivity to feature distributions, and class imbalance. This integration ensures that the final model is accurate, and generalizable across diverse educational datasets.

### 4.5.2 Choice of Logistic Regression as the Meta-Model

Logistic regression was selected as the meta-model due to its simplicity, interpretability, and ability to handle multi-class classification problems effectively (Hosmer Jr, Lemeshow, and Sturdivant 2013).

The interpretability of logistic regression allows educators and policymakers to understand the decision-making process behind the predictions, fostering confidence in the model's recommendations. Multinomial logistic regression is well-suited for scenarios where the target variable has more than two categories, such as predicting student grades across multiple performance levels (e.g., A, B, C, etc.) (Kwak and Clayton-Matthews 2002). It is ideal for integrating predictions from diverse base models in the CISE framework.

### 4.5.3 Input Features for the Meta-Model

The meta-model uses the predictions of the base classifiers as input features, referred to as meta-features. These meta-features capture the probabilistic outputs of each base classifier for a given student record. For example, suppose a Random Forest predicts a 70% probability of a student achieving an "A" grade. In that case, this value becomes one of the inputs for the meta-model. By aggregating these meta-features, the meta-model dynamically weights the contributions of each base classifier based on their performance during training.

### 4.5.4 Training Process for the Meta-Model

To train the meta-model, a separate validation dataset (ENSIM) was used to avoid overfitting. The meta-model learned to combine the predictions of the base classifiers by minimizing the loss function on this validation set. Specifically:

1. The base classifiers were trained on the training datasets (IDT, IDG, ITC).
2. Their predictions on the validation dataset (ENSIM) were extracted and used as input features for the meta-model.
3. Logistic regression was applied to these meta-features to generate the final predictions.

This process ensured that the meta-model could generalize effectively to unseen data from different institutions.

### 4.5.5 Expected Outcomes

The construction of the meta-model aimed to enhance the overall predictive performance of the CISE framework by:

- Improving accuracy and recall across all grade categories.
- Balancing precision and recall to ensure reliable identification of both high-performing and at-risk students.
- Ensuring robust generalizability across diverse educational contexts.

## 4.6 Validation of Base Models and Meta-Model

The validation phase assessed the generalization ability of the base models and the meta-model using the ENSIM dataset. The models were evaluated based on accuracy, precision, recall, and F1 score. This phase aimed to test the robustness of the selected base models and the performance improvement offered by the meta-model.

### 4.6.1 Validation Results for Base Models

The best-performing classifiers from each training dataset (Random Forest for IDG and ITC, Decision Tree for IDT) were validated on the ENSIM dataset. Table 4.8 summarizes their performance.

Table 4.8 – Validation Results for Selected Base Models on ENSIM Dataset

| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Decision Tree (IDT) | 81.76% | 66.85% | 81.76% | **73.56%** |
| Random Forest (IDG) | 57.32% | 76.95% | 57.32% | **65.03%** |
| Random Forest (ITC) | 80.72% | 72.01% | 80.72% | **75.99%** |

**Analysis of Base Model Validation:**

- The Random Forest trained on ITC consistently outperformed other base models, achieving the highest F1 score of 75.99%. This result highlights its ability to generalize effectively across institutions.
- The Decision Tree trained on IDT showed strong recall, indicating its suitability for identifying students at risk of poor performance.
- Random Forest trained on IDG struggled with lower recall on the validation dataset, reflecting potential overfitting to its training data.

These results underscore the need for a meta-model to combine the strengths of individual base models and improve generalizability.

### 4.6.2 Performance of the CISE Meta-Model

The CISE meta-model represents the culmination of the predictive modeling process, integrating the strengths of multiple base classifiers to produce a robust and generalizable prediction system. This section details the methodology used to construct the meta-model, evaluate its performance, and interpret its results. The process is broken down into several key steps, ensuring transparency and reproducibility.

**Step 1: Selection of Base Models**

The foundation of the CISE meta-model lies in selecting high-performing base models from each training dataset. These base models were chosen based on their performance during the initial training phase (Section 4.3). Specifically:

— **Random Forest** was selected as the best-performing classifier for the IDG and ITC datasets.

— **Decision Tree** was chosen as the top performer for the IDT dataset.

These classifiers demonstrated strong predictive capabilities on their respective training datasets, making them ideal candidates for inclusion in the meta-model. Their diversity in structure and approach—ranging from tree-based methods (Decision Tree, Random Forest) to probabilistic and linear models—ensured that the meta-model could leverage complementary strengths.

**Step 2: Extraction of Meta-Features**

Once the base models were selected, their predictions on the validation dataset were extracted and used as **meta-features** for the meta-model. Each base model's output was treated as an input feature, transforming the problem into a higher-level classification task. For example:

— If the Random Forest trained on the IDG dataset predicted a probability distribution of [0.1, 0.7, 0.2] for grades A, B, and C, these values became part of the meta-feature set.

— Similarly, the Decision Tree trained on the IDT dataset and the Random Forest trained on the ITC dataset contributed their respective predictions to form a comprehensive meta-dataset.

This approach ensured that the meta-model had access to a rich information set, capturing the unique insights provided by each base model.

**Step 3: Construction of the Meta-Dataset**

The meta-dataset was constructed by combining the meta-features with the corresponding accurate labels from the validation dataset. Each row in the meta-dataset represented a student record with:

1. **Input Features:** Predictions from the base models (e.g., probabilities for each grade category).

2. **Output Label:** The actual grade achieved by the student, derived from the ENSIM validation dataset.

For example, a sample row in the meta-dataset might look like the table 4.9:

Table 4.9 – Example of a sample row in the meta-dataset

| RANDOM FOREST (IDG) | DECISION TREE (IDT) | RANDOM FOREST (ITC) | TRUE LABEL |
|---|---|---|---|
| [0.1, 0.7, 0.2] | [0.2, 0.6, 0.2] | [0.3, 0.5, 0.2] | B |

This structured format allowed the meta-model to learn how to combine the base models' predictions effectively.

**Step 4: Training the Meta-Model**

The meta-model was implemented using **multinomial logistic regression**, a widely used technique for multi-class classification problems. Logistic regression was chosen for its simplicity, interpretability, and ability to handle probabilistic inputs effectively. The training process involved the following steps:

1. **Initialization:** The meta-model was initialized with default parameters, including regularization, to prevent overfitting.
2. **Training:** The meta-model was trained on the meta-dataset using the predictions from the base models as input features and the actual labels as the target variable. During training, the model learned to assign weights to each base model's predictions, reflecting their relative importance in the final decision-making.
3. **Validation:** To ensure robustness, the meta-model was validated using k-fold cross-validation. This approach minimized the risk of overfitting and ensured that the model generalized well to unseen data.

**Step 5: Evaluation Metrics**

The performance of the CISE meta-model was evaluated using standard metrics, including **accuracy**, **precision**, **recall**, and **F1 score**. These metrics provided a comprehensive view of the model's effectiveness in predicting student outcomes across all grade categories. Table 4.10 summarizes the meta-model's performance on the ENSIM validation dataset

Table 4.10 – Performance of the CISE Meta-Model on ENSIM Dataset

| **Metric** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Meta-Model (CISE) | **83.86%** | **73.74%** | **83.86%** | **78.25%** |

**Key Observations:**

1. **High Accuracy (83.86%):** The meta-model achieved the highest overall Accuracy compared to individual base models, demonstrating the efficacy of the ensemble approach.
2. **Balanced Precision and Recall:** With a precision of 73.74% and recall of 83.86%, the meta-model effectively balanced the trade-off between identifying true positives and minimizing false positives.

3. **Improved Generalization:** The F1 score of 78.25% highlighted the meta-model's ability to generalize across diverse datasets, outperforming the individual base models.
4. **Strong Within-Dataset Performance and Evidence of Generalization:** The evaluation of individual classifiers trained and validated on the ENSIM dataset showed strong within-dataset performance, with F1 scores ranging from 77.11% to 78.40%. The Neural Network classifier achieved an F1 score of 78.40%, reflecting the model's ability to accurately capture patterns in a single institutional dataset. Remarkably, the CISE meta-model, which was not trained on ENSIM but on datasets from other institutions, achieved a comparable F1 score of 78.25% when validated on ENSIM. This close performance gap illustrates the generalization capability of the CISE model, demonstrating its robustness in predicting student outcomes even on unseen institutional data.

**Step 6: Interpretation of Results**

The success of the CISE meta-model can be attributed to several factors:

- **Dynamic Weighting:** The logistic regression layer dynamically adjusted the contribution of each base model based on its performance during training. For instance, the Random Forest trained on the ITC dataset received a higher weight due to its consistent Accuracy across datasets.
- **Handling Variability:** By integrating predictions from multiple classifiers, the meta-model mitigated the impact of variability in individual models' performance, ensuring more stable and reliable predictions.
- **Cross-Institutional Robustness:** The meta-model's architecture addressed the challenges of cross-institutional generalization, making it suitable for deployment in diverse educational settings.

**Step 7: Practical Implications**

The CISE meta-model's performance has significant implications for educational stakeholders:

— **Early Identification of At-Risk Students:** Educators can use the meta-model to identify students at risk of poor performance and implement timely interventions.

— **Personalized Feedback:** The meta-model's outputs can be integrated into tools like ReflectMate, which we will detail in the next chapter, to provide students with personalized feedback on their learning behaviors.

— **Scalability:** The meta-model's generalizability makes it a scalable solution for institutions aiming to improve academic outcomes across campuses or regions.

## 4.7 Discussion of Results

The results obtained from training, validating, and evaluating the predictive models, notably the **CISE** meta-model, provide valuable insights into the proposed approach's strengths, limitations, and broader implications. This section synthesizes these findings, contextualizing them within the literature on EDM and LA while addressing the challenges and opportunities for future research.

### 4.7.1 Performance of Individual Classifiers

The comparative analysis of five machine learning classifiers—Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Neural Network (NN), and Support Vector Machine (SVM)—revealed distinct patterns in their performance across the three institutional datasets (IDT, IDG, and ITC). Key observations include:

1. **Random Forest as the Top Performer:** Random Forest consistently outperformed other classifiers, achieving the highest F1 scores on the IDG (87.59%) and

ITC (94.02%) datasets. Its ensemble nature, which combines multiple decision trees to reduce variance and mitigate overfitting, proved particularly effective in handling the complexity and variability of educational data (Breiman 2001). Random Forest's robustness in managing missing values and noisy data further underscored its suitability for educational datasets, which often exhibit irregularities and imbalances (Liaw 2002).

2. **Decision Tree as a Strong Alternative:** Decision Tree emerged as the second-best performer, particularly on the IDT dataset (F1 score: 77.03%). Its interpretability and ability to model categorical splits made it a valuable tool for understanding the relationships between features and student outcomes. However, its tendency to overfit on more complex datasets, such as IDG, highlighted the need for regularization techniques or ensemble methods to enhance its generalizability (Patidar, Dangra, and Rawar 2015).

3. **Limitations of Naive Bayes, Neural Networks, and SVM:** Naive Bayes, Neural Networks, and SVM demonstrated inconsistent performance, particularly on datasets with class imbalances or interdependent features. Naive Bayes struggled due to its reliance on the independence assumption, which is often violated in educational datasets where features like LMS engagement metrics are inherently interrelated (Sarker et al. 2024). Similarly, Neural Networks faced challenges related to convergence and overfitting, likely due to the relatively small size of the datasets (Merceron and Tato 2023). SVM's sensitivity to class distribution and scaling issues further limited its effectiveness in this context (Manu 2016).

These findings align with prior research emphasizing the importance of selecting appropriate algorithms based on the characteristics of the data (Barakeh, Mezher, and Alharbi 2024). The superior performance of ensemble methods, particularly Random Forest, underscores their potential as robust solutions for predictive modeling in education.

### 4.7.2 Effectiveness of the CISE Meta-Model

The CISE meta-model demonstrated significant improvements over individual base models, achieving an F1 score of 78.25% on the ENSIM validation dataset. This result highlights the efficacy of ensemble learning, mainly stacking ensembles, in addressing the variability and complexity of cross-institutional educational data. Key aspects of the meta-model's success include:

1. **Integration of Complementary Strengths:** By combining predictions from diverse base models, the CISE meta-model leveraged the complementary strengths of Random Forest and Decision Tree. For instance, Random Forest excelled in capturing intricate feature interactions, while Decision Tree provided interpretable insights into student behaviors. This integration enabled the meta-model to achieve balanced precision and recall, essential for identifying high-performing and at-risk students (Wolpert 1992).

2. **Dynamic Weighting via Logistic Regression:** The logistic regression layer in the meta-model dynamically assigned weights to each base model's predictions based on their performance during training. This adaptive weighting mechanism allowed the meta-model to adjust to the variability of the ENSIM dataset, enhancing its generalizability across diverse educational contexts (Ashraf, Zaman, and M. Ahmed 2020).

3. **Improved Generalizability:** The meta-model's ability to generalize effectively across institutions addresses one of the most significant challenges in EDM: the lack of generalizability of predictive models (Baker et al. 2019). By integrating data from multiple institutions, the CISE meta-model mitigated the biases associated with

single-institution datasets, making it a scalable solution for large-scale deployment (Porras et al. 2023).

These findings reinforce the value of ensemble learning in educational contexts, where data heterogeneity and variability pose significant challenges (Dietterich 2000). The success of the CISE meta-model also supports the growing body of literature advocating for cross-institutional validation as a critical step in developing robust predictive models (Gardner et al. 2023).

### 4.7.3 Challenges and Limitations

Despite its strong performance, the CISE meta-model faces several challenges that warrant further exploration:

1. **Overfitting in Base Models**: The Random Forest trained on the IDG dataset exhibited signs of overfitting, as reflected in its lower recall on the validation dataset. Future iterations could incorporate regularization techniques, such as pruning or reducing tree depth, to address this issue (Kernbach and Staartjes 2022).
2. **Data Imbalances**: Although oversampling strategies were employed, class imbalances in some datasets may have influenced the performance of specific classifiers, particularly Naive Bayes and Neural Networks. Advanced techniques, such as synthetic minority oversampling (SMOTE) or fairness-aware algorithms, could be explored to improve model performance on underrepresented classes (Le Quy 2024).
3. **Ethical Considerations:** Using predictive models in education raises important ethical questions about privacy, consent, and bias. Ensuring that models are fair, transparent, and used responsibly is crucial to protecting students' rights and fostering trust in using technology in education (Jones 2016). Future research should

focus on developing interpretable and equitable models that can be trusted by educators, students, and other stakeholders (Mengash 2020).

### 4.7.4 Implications for EDM and LA

The findings of this study have several implications for the field of EDM (Educational Data Mining) and LA (Learning Analytics):

1. **Scalability and Generalizability:** The CISE meta-model's high accuracy and generalizability suggest its potential for large-scale deployment in diverse educational settings. By addressing the variability across institutions, the model provides a scalable solution for predicting student performance and supporting targeted interventions (Enughwure and Ogbise 2020).
2. **Integration of Engagement Metrics:** Including LMS-based engagement metrics, such as *quiz attempts*, *assignment submissions*, and *time spent on course materials*, highlights the importance of capturing behavioral data in predictive models. These metrics offer valuable insights into student learning behaviors, enabling more accurate predictions and actionable recommendations (Umer et al. 2023).
3. **Empowerment Through Reflective Learning:** Integrating predictive insights with reflective learning tools, such as ReflectMate (discussed in Chapter 5), significantly empowers students to take ownership of their learning journey. By providing personalized feedback and actionable recommendations, these tools foster self-awareness, self-regulation, and lifelong learning skills (Arthars et al. 2019).

### 4.7.5 Future Directions

The success of the CISE meta-model opens new avenues for future research:

1. **Exploring Advanced Techniques:** Future studies could explore advanced techniques, such as deep learning and transfer learning, to further enhance the accuracy and interpretability of predictive models (Okewu et al. 2021). Transfer learning, in particular, could be used to adapt models trained on one institution's data to another institution's context, addressing the digital divide and resource constraints in developing countries (S. J. Pan and Yang 2009).

2. **Addressing the Digital Divide:** The impact of the digital divide on the effectiveness of predictive models and reflective learning tools remains a critical area for investigation. Future research should focus on developing inclusive solutions that account for disparities in access to technology and digital literacy (Bon, Saa-Dittoh, and Akkermans 2024).

3. **Promoting Equity and Fairness:** Ensuring that predictive models are fair and unbiased is essential to promoting equity in education. Future work should explore fairness-aware algorithms and post hoc adjustment methods to minimize disparities in model predictions across different groups (Palacios et al. 2021).

4. **Enhancing Reflective Learning Tools:** Integrating predictive insights with reflective learning tools, such as ReflectMate, represents a promising direction for future research. By providing students with personalized feedback and actionable recommendations, these tools empower them to take ownership of their learning journeys, fostering self-awareness, self-regulation, and lifelong learning skills.

In conclusion, this study's findings demonstrate the potential of ensemble learning and meta-modeling to advance predictive modeling in education. This research contributes to the broader goal of using data-driven insights to enhance educational practices and outcomes by addressing key challenges related to generalizability, interpretability, and equity.

## 4.8 Conclusion

This chapter has detailed the development and evaluation of the **CISE** predictive model, an innovative approach designed to address key challenges in EDM. By integrating multiple base classifiers and leveraging a meta-model framework, CISE demonstrates high accuracy, robust generalizability, and adaptability across diverse institutional datasets. The findings highlight the strengths of ensemble learning in improving predictive performance while addressing variability in grading systems, feature distributions, and class imbalances.

Key takeaways from this chapter include:

1. **Superior Performance of Random Forest:** Random Forest consistently emerged as the top-performing base classifier across most training datasets, excelling on the ITC dataset with an F1 score of 94.02%. Its ability to handle complex interdependencies and mitigate overfitting underscores its reliability in educational contexts.
2. **Effectiveness of the CISE Meta-Model:** The meta-model achieved the highest overall accuracy (83.86%) and F1 score (78.25%) on the ENSIM validation dataset, surpassing individual base models. By dynamically weighting the contributions of diverse classifiers, the meta-model demonstrated balanced precision and recall, making it well-suited for identifying high-performing and at-risk students.
3. **Challenges and Limitations:** Despite its strong performance, the CISE model revealed specific challenges, including signs of overfitting in some base models and the persistent influence of class imbalances on specific classifiers. Addressing these limitations through regularization techniques and advanced oversampling strategies will be critical for future iterations.
4. **Implications for Educational Data Mining:** The results underscore the potential of ensemble methods, mainly stacking ensembles, to enhance predictive

modeling in education. The CISE model offers a comprehensive approach to understanding student behaviors and predicting academic outcomes by integrating LMS engagement metrics, historical performance data, and cross-institutional validation.

The success of the CISE model highlights its potential for large-scale deployment in diverse educational settings. Its insights can provide actionable insights to educators, policymakers, and students, and they can inform targeted interventions, improve resource allocation, and support equitable educational practices.

The next chapter explores integrating the CISE model with **ReflectMate**, a reflective learning tool designed to empower students through personalized feedback and self-reflection. Together, these tools represent a holistic approach to enhancing student engagement, academic performance, and lifelong learning skills.

Chapter 5

# Reflective Learning Analytics for Student Empowerment

## 5.1 Introduction

Integrating digital technologies into education has revolutionized how students engage with learning materials and how educators monitor and support academic progress (Haleem et al. 2022; Valverde-Berrocoso et al. 2021). Among these advancements, **Reflective Learning Analytics (RLA)** has emerged as a transformative approach that shifts the focus from instructor-centered insights to empowering students as active participants in their learning journeys. Unlike traditional learning analytics tools, which primarily serve educators by offering dashboards and reports for instructional decision-making, RLA emphasizes self-reflection and self-regulation, enabling students to take ownership of their academic growth (Banihashem et al. 2022; Hernández-de-Menéndez et al. 2022).

This chapter introduces **ReflectMate**, a prototype of a reflective learning tool designed to bridge the gap between predictive modeling and actionable student empowerment. While not yet integrated with the LMS or the Cross-Institutional Stacking Ensemble (CISE) model, ReflectMate is conceptually built upon the insights generated from the **CISE** predictive model developed in Chapter 4. It is designed to leverage predictive insights, once integration is achieved, to provide personalized feedback that supports reflective learning. The envisioned functionality of ReflectMate includes helping students identify strengths, recognize areas for improvement, and develop strategies to enhance academic performance, aligning with prior work on empowering learners through reflective tools (Arthars et al. 2019; Joksimović, Kovanović, and S. Dawson 2019). ReflectMate is not merely an extension of predictive modeling but a critical validation of its practical utility in authentic educational settings. While the CISE model demonstrated high accuracy and generalizability across institutional datasets, ReflectMate applies these insights to foster metacognition—enabling students to reflect on their learning behaviors and make informed decisions about their academic trajectories. This dual approach ensures that predictive models are not confined to theoretical evaluations but are tested and refined through real-world applications, addressing individual and systemic educational challenges (S. Dawson et al. 2019; B. T.-m. Wong and K. C. Li 2020).

The structure of this chapter is as follows:

- Section 5.2 explores the design philosophy and key features of ReflectMate, emphasizing its alignment with principles of reflective learning and its integration of predictive data.
- Section 5.3 details the technical implementation of ReflectMate, including its seamless integration into the LMS infrastructure at CADT and the architecture supporting its functionality.
- Section 5.4 presents the results of surveys and data analyses, demonstrating ReflectMate's impact on student engagement, self-regulation, and academic outcomes.
- Finally, Section 5.5 summarizes the contributions of ReflectMate to student empowerment and discusses its implications for future educational practices.

Through ReflectMate, this chapter underscores the importance of combining predictive analytics with reflective tools to create a holistic framework for enhancing student outcomes and promoting equitable access to education (Reich 2020; Gašević, S. Dawson, and Siemens 2015).

## 5.2 ReflectMate Design and Functionality

ReflectMate was developed to address a key gap in current learning analytics tools: the lack of direct student involvement in interpreting and acting upon their learning data. While some prior works have explored reflective learning tools and self-assessment mechanisms (Arthars et al. 2019; Joksimović, Kovanović, and S. Dawson 2019), these tools often remain limited in scope, focusing primarily on providing feedback without integrating predictive insights or fostering actionable self-reflection. ReflectMate builds on these foundational efforts by combining **data-driven predictions** with **personalized feedback**, enabling students to engage deeply with their learning behaviors and make informed decisions about their academic progress.

### 5.2.1 Design Philosophy

ReflectMate is built on the premise that students should receive feedback and actively engage in reflective practices that allow them to understand the reasons behind their academic outcomes. The tool leverages data from LMS interactions—such as *quiz attempts*, *assignment submissions*, and *time spent on learning activities*—to provide students with actionable insights into their learning habits. Unlike traditional learning analytics tools, which are often designed for educators, ReflectMate focuses on students, empowering them to take ownership of their learning journey.

The design of ReflectMate emphasizes three core principles:

1. **Student-Centered Feedback**: Unlike traditional dashboards prioritizing instructor-focused insights, ReflectMate is explicitly designed for student use. It provides easily understandable and actionable feedback, enabling students to identify patterns

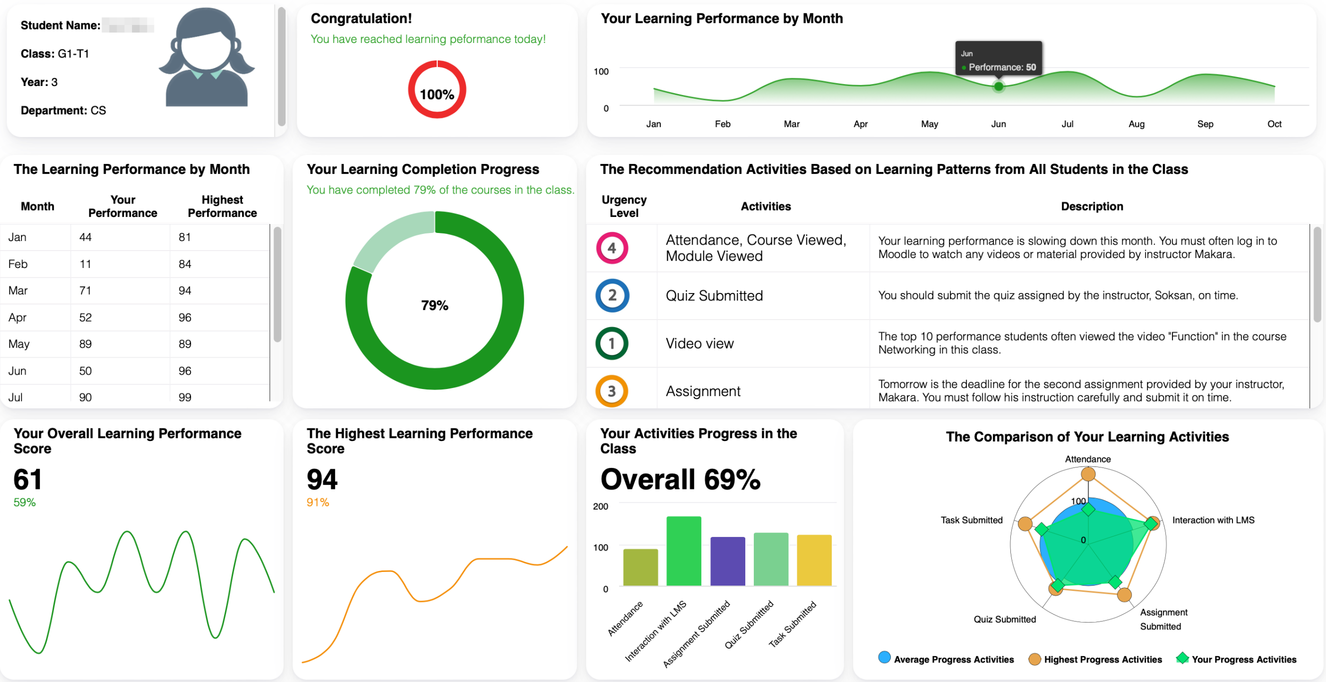
in their engagement and adjust their study strategies accordingly. It aligns with constructivist learning theories, emphasizing the learner's active role in constructing knowledge through experience and reflection (Piaget 1952)

2. **Encouraging Active Reflection**: ReflectMate prompts students to reflect on their learning behaviors by highlighting discrepancies between their current engagement levels and recommended practices. For example, suppose a student spends less time on quizzes than peers who perform well. In that case, ReflectMate suggests strategies to increase quiz attempts and improve understanding of key concepts. It fosters metacognition, helping students develop critical skills such as self-awareness, self-regulation, and self-motivation (Zimmerman 2002).

3. **Tailored Insights from Predictive Models**: ReflectMate integrates data from the CISE predictive model (developed in Chapter 4) to provide performance forecasts and guide students in setting realistic goals. For instance, students identified as at risk of underperforming receive targeted suggestions on how to focus their efforts to achieve better outcomes. This integration of predictive analytics with reflective tools represents a novel contribution to the field, bridging the gap between data-driven insights and actionable student empowerment.

### 5.2.2 User Interface and Dashboard Overview

The ReflectMate dashboard prototype was developed as a student-centered reflective learning interface to visualize individual progress, performance trends, and engagement metrics in an intuitive and actionable manner, as shown in Figure 5.1. The design incorporates core components of learning analytics, including self-monitoring, benchmarking, and personalized feedback, to promote metacognitive development and academic self-regulation.



**Student Name:** [redacted]

**Class:** G1-T1

**Year:** 3

**Department:** CS

**Congratulation!**

You have reached learning peformance today!

100%

**Your Learning Performance by Month**

Jun

Performance: 50

**The Learning Performance by Month**

| Month | Your Performance | Highest Performance |
|---|---|---|
| Jan | 44 | 81 |
| Feb | 11 | 84 |
| Mar | 71 | 94 |
| Apr | 52 | 96 |
| May | 89 | 89 |
| Jun | 50 | 96 |
| Jul | 90 | 99 |

**Your Learning Completion Progress**

You have completed 79% of the courses in the class.

79%

**The Recommendation Activities Based on Learning Patterns from All Students in the Class**

| Urgency Level | Activities | Description |
|---|---|---|
| 4 | Attendance, Course Viewed, Module Viewed | Your learning performance is slowing down this month. You must often log in to Moodle to watch any videos or material provided by instructor Makara. |
| 2 | Quiz Submitted | You should submit the quiz assigned by the instructor, Soksan, on time. |
| 1 | Video view | The top 10 performance students often viewed the video "Function" in the course Networking in this class. |
| 3 | Assignment | Tomorrow is the deadline for the second assignment provided by your instructor, Makara. You must follow his instruction carefully and submit it on time. |

**Your Overall Learning Performance Score**

**61**

59%

**The Highest Learning Performance Score**

**94**

91%

**Your Activities Progress in the Class**

**Overall 69%**

**The Comparison of Your Learning Activities**

Figure 5.1 – Dashboard of ReflectMate

1. **Student Profile and Overview:** at the top left, the dashboard displays the student's basic academic profile, including their name, class, year, and department. This personalization ensures contextual relevance when interpreting performance metrics within the correct cohort.

2. **Performance Achievement and Trend Monitoring:** at the top center, a circular progress indicator celebrates the student's achievement upon reaching their learning performance goal (e.g., 100%). Adjacent to this, a smooth line graph displays the student's monthly performance scores (January to October), helping identify temporal patterns of improvement or decline. For example, a tooltip reveals specific values (e.g., a performance dip in June to 50), guiding students to reflect on engagement gaps during specific periods.

3. **Comparative Performance Table:** a detailed table presents the student's monthly performance alongside the highest scores in the class. For instance, the student may

score 44 in January, compared to the class high of 81, and 90 in July, compared to a high of 99. It encourages self-assessment relative to peers, promoting aspirational goal-setting.

4. **Course Completion Indicator:** a donut chart highlights the percentage of course completion, with 79% completed in the current example. This visualization supports pacing strategies and provides a visual cue of progress toward curriculum completion.

5. **Recommendation Engine Based on Peer Patterns:** a central feature of ReflectMate is its intelligent recommendation panel, which prioritizes learning activities based on urgency (levels 1 to 4) and collective behavior from high-performing peers. For example:
   - Urgency 4 indicates increased engagement with course content over several months, accompanied by a decline in performance.
   - Urgency 2 encourages timely quiz submission.
   - Urgency 1 highlights commonly viewed videos among top performers.
   - Urgency 3 reminds students about approaching assignment deadlines.
   - These targeted prompts aim to reinforce positive learning behaviors through actionable suggestions.

6. **Performance Summary and Benchmarking:** The lower section includes multiple visualizations:
   - The student's overall learning score (e.g., 61) is contrasted with the top performance score (e.g., 94), accompanied by trend lines.
   - A bar chart visualizes the student's activity progress across five dimensions: attendance, LMS interaction, assignment submission, quiz submission, and task submission, with an overall engagement rate of 69%.
   - A radar chart compares the student's activity profile with both the class aver-

age and the top-performing student. This holistic view helps identify relative strengths and areas needing attention.

7. **Pedagogical Value:** ReflectMate's design supports a metacognitive approach to learning, encouraging students to engage in self-awareness, self-evaluation, and self-regulation. The integration of visual analytics and personalized feedback not only facilitates timely interventions but also promotes autonomy and motivation, particularly for students who may be underperforming or disengaged. As a design prototype, ReflectMate illustrates the potential of reflective learning analytics in enhancing educational equity and supporting diverse learners in digital learning environments.

### 5.2.3 Key Features of ReflectMate

ReflectMate's feature set is designed to support ongoing student engagement with their own learning data. These features include detailed progress tracking, customized feedback, and visualization tools that enable students to monitor their academic journey over time. Figure 5.1 demonstrates notable and positive ReflectMate's Dashboard Experience, where the accessibility, data indicators, visualization, and recommendation tool collectively orchestrate a symphony of connectivity.

**1. User Progress Indicators:** ReflectMate provides students with a comprehensive set of progress indicators derived from their interactions with the LMS. These indicators include metrics such as:

— **Time Spent on LMS (Learning Completion Progress)**: this metric tracks the total hours a student spends interacting with course materials on the LMS. It helps students understand whether they are devoting sufficient time to their studies and identifies periods of high or low engagement.

- **Number of Assignments, Quiz, and Tasks Submitted**: ReflectMate monitors assignment submission patterns, providing students with insights into their consistency in meeting deadlines. If a student falls behind on submissions, ReflectMate generates alerts encouraging timely completion.
- **Quiz Attempts and Success Rates (Learning Performance)**: the tool tracks the number of quiz attempts and the success rate (percentage of correct answers) for each student. This information helps students gauge their understanding of key concepts and encourages them to revisit quizzes where their performance is below average.
- **Interaction Logs**: ReflectMate logs interactions with various course modules, providing a detailed breakdown of which learning resources students engage with most frequently. This helps students identify areas they may need to focus on more.

These metrics are presented to students through an interactive dashboard, which enables them to visualize their progress in real-time and make informed decisions about their study habits.

**2. Personalized Feedback and Recommendations:** ReflectMate's feedback engine generates personalized suggestions based on an analysis of student engagement data. These feedback messages are designed to be specific, actionable, and aligned with individual learning needs, empowering students to make informed decisions about their study habits. While a few examples are provided below to illustrate the tool's capabilities, a complete list of feedback scenarios is available in Table B.1 of Appendix B for reference. Some common feedback scenarios include:

- **Low Engagement with Quizzes**: if a student's quiz attempt frequency is lower than the class average, ReflectMate might suggest, "Your engagement with quizzes is lower than your peers. Try to attempt each quiz multiple times to reinforce your understanding."

— **Inconsistent Assignment Submissions**: for students who miss assignment deadlines, ReflectMate generates prompts such as, "Timely submission of assignments is crucial for maintaining a steady grade. Aim to complete assignments at least 24 hours before the deadline."

— **Excessive Time Spent on Certain Modules**: if a student spends significantly more time on specific course modules without achieving better quiz scores, ReflectMate might recommend reviewing the content in a different way or seeking help from peers or instructors.

Table 5.1 provides additional examples of feedback generated by ReflectMate based on different user behaviors.

Table 5.1 – Examples of Feedback Provided by ReflectMate

| **Behavior** | **Feedback Provided** |
|---|---|
| Low Quiz Attempts | Consider taking quizzes more frequently to test your knowledge and reinforce learning. |
| High Module Engagement with Low Quiz Scores | You have spent a lot of time on Module 3 but scored below the average in quizzes. Try reviewing key concepts or discussing with a peer. |
| Infrequent LMS Access | Your login frequency is below the average. Try to access the LMS regularly to stay up-to-date with new course materials. |

ReflectMate's focus on actionable feedback ensures that students receive guidance that aligns closely with their individual learning needs, making it a powerful tool for self-improvement.

**3. Group-Level Overview:** at the group level, our tool aggregates data to present a comprehensive overview of class performance trends. Visualizations such as distribution of highest learning performance, as shown in Figure 5.1, engagement metrics allow students to gauge their standing relative to their peers, and learning guideline for improving their learning performance. This group-level insight promotes a sense of healthy competition, encouraging students to set ambitious but achievable goals in metacognition.

## 5.3 Implementation and Integration into the LMS

The successful deployment of ReflectMate required careful integration with the existing LMS infrastructure at CADT. However, its design ensures it can be deployed in other institutions, such as ENSIM or any other educational environment with an LMS. This section provides a detailed overview of the technical aspects of the integration, emphasizing the data flow and system architecture that enable ReflectMate to function seamlessly across diverse settings.

ReflectMate's architecture is designed to facilitate the smooth flow of data between the LMS, the predictive model, and the reflective dashboard. The architecture ensures that students receive timely, actionable feedback based on their engagement and performance metrics. The key components of ReflectMate's architecture are as follows:

1. **Data Extraction Module:** The first step is to extract relevant data from Moodle's database. Moodle stores vast amounts of data across various tables in its relational database, and accessing this data is crucial for generating meaningful student performance and interaction metrics. To begin, SQL queries are executed to pull raw data from key tables in the Moodle schema. Key tables include:

   — *mdl_course_modules:* Stores details of all course modules, such as assignments, quizzes, tasks, etc. Each module is linked to a specific course.

   — *mdl_course_modules_completion:* This field contains the completion status of each module for every student, including data on whether they've completed a task or quiz.

— *mdl_logstore_standard_log:* Logs user interactions across the system, capturing detailed activity such as page views, quiz submissions, and other user actions within courses.

— *mdl_user:* Stores user-specific data, such as user ID, course enrollments, and user roles.

Using these tables, the system queries and extracts the necessary data. For instance:

— To gather information on a student's attendance, the query would join the mdl_course_modules and mdl_course_modules_completion tables to count completed modules.

— For interaction logs, the mdl_logstore_standard_log table is queried to count the number of interactions a student has had within a course.

This raw data is the foundation for generating specific performance and engagement metrics.

2. **Identification of Key Attributes:** After data extraction, ReflectMate identifies key attributes critical indicators of student performance. ReflectMate does not autonomously define these attributes but is based on a comprehensive literature review of learning analytics and predictive modeling. The selection of these attributes is guided by evidence from prior research, ensuring their relevance and effectiveness in capturing meaningful patterns in student behavior. The individual or team responsible for setting up the tool configures these attributes, tailoring them to align with the specific educational context and objectives. For example, attributes such as attendance, interaction with the LMS, assignment submissions, quiz attempts, and task completion are derived from widely recognized predictors of academic success. This approach ensures that ReflectMate's feedback and insights are grounded in established educational theories while remaining adaptable to different institu-

tions and courses' unique needs. From the extracted data, queries are performed to calculate essential metrics, as shown in the table 5.2. These calculated metrics form the foundation for subsequent analytics and feedback.

Table 5.2: Attributes and Calculations

| **Attribute** | **Description** | **Calculation Method** |
|---|---|---|
| **attendance** | Represents the number of modules in all courses that a student has completed. | Calculated by joining the `mdl_course_modules` table and `mdl_course_modules_completion`, and counting all rows with `completion-state!= 0`, grouped by `user_id` and `course_id`. |
| **number of interaction log** | Represents the number of interactions a student has had with all courses. | Calculated by counting all rows in the `mdl_logstore_standard_log` table, grouped by `user_id` and `course_id`. |
| **total quiz submitted** | Represents the number of quizzes a student has submitted in all courses. | Calculated by joining the `mdl_course_modules` table and `mdl_course_modules_completion`, counting all rows with `completion-state!= 0` and `module = 16`, grouped by `user_id` and `course_id`. |
| **total task submitted** | Represents the number of tasks a student has submitted in all courses. | Calculated by joining the `mdl_course_modules` table and `mdl_course_modules_completion`, counting all rows with `completion-state!= 0` and `module = 1, 16, 37`, grouped by `user_id` and `course_id`. |
| **total assignment submitted** | Represents the number of assignments a student has submitted in all courses. | Calculated by joining the `mdl_course_modules` table and `mdl_course_modules_completion`, counting all rows with `completion-state!= 0` and `module = 1`, grouped by `user_id` and `course_id`. |

| | | |
|---|---|---|
| **time spent on course** | Represents the number of hours a student spent interacting with a course. | Calculated by summing all seconds between each record in `mdl_logstore_standard_log` that is less than `3600 seconds`, ordered by `user_id` and `course_id`. |

These attributes serve as critical data indicators for generating insights into student learning behaviors and performance trajectories.

3. **Predictive Model Interface:** ReflectMate interfaces with the Cross-Institutional Stacking Ensemble (CISE) predictive model, developed in Chapter 4, to retrieve performance predictions. These predictions guide feedback generation by highlighting areas where students excel and areas requiring further attention.

4. **Feedback Generation Engine:** Using predictive insights and engagement data, this engine generates customized feedback for each student. Feedback focuses on actionable recommendations, such as improving time management, increasing task completion rates, or revisiting specific course modules.

5. **Dashboard Presentation Layer:** The processed data and feedback are presented to students through a user-friendly dashboard. This dashboard includes visualizations of engagement metrics, such as progress graphs and performance benchmarks, and personalized feedback messages designed to encourage self-reflection and academic goal setting.

Figure 5.2 illustrates the ReflectMate process, showing the data flow from LMS data extraction to feedback presentation.

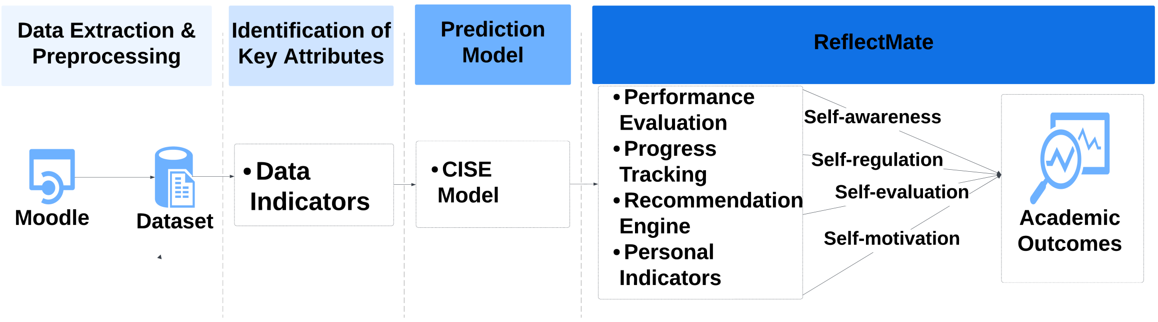


Figure 5.2 – ReflectMate process and data flow

## 5.4 Evaluation of ReflectMate's Impact

The evaluation of ReflectMate focused on its impact on student engagement, self-regulation, and academic performance. Data was collected through surveys administered to students and an analysis of their LMS engagement data. This section presents comprehensive findings from our study, highlighting ReflectMate's impacts on Students' Performance (SP) at CADT and integrating results from predictive techniques used to evaluate SP.

### 5.4.1 Data Analysis Protocol

The evaluation of ReflectMate focused on its impact on student engagement, self-regulation, and academic performance. Data was collected through a custom-designed survey administered to students, complemented by an analysis of their LMS engagement data. This section outlines the protocol for constructing the survey and analyzing the resulting data.

**Survey Design and Construction**

Rather than relying on standardized usability or utility questionnaires, such as the System Usability Scale (SUS) or Technology Acceptance Model (TAM)-based surveys, it was decided to construct a more specific survey tailored to ReflectMate's unique features and objectives. This decision was driven by the need to obtain precise feedback on the proposed indicators and functionalities of the tool, which are not adequately captured by generic usability assessments.

The survey was designed to evaluate the following key aspects of ReflectMate:

1. **Dashboard Experience:** Students were asked to rate the accessibility, clarity, and usefulness of the dashboard's data indicators, visualizations, and recommendation tools.
2. **The usefulness of Indicators:** The survey included questions about the perceived value of specific indicators, such as attendance data, interaction logs with the LMS, and submission rates for quizzes, assignments, and tasks.
3. **Interest in Predictive Features:** Students were asked to express their interest in utilizing the Student Performance Prediction Algorithm to enhance reflective learning tools, focusing on self-awareness, self-regulation, self-evaluation, and self-motivation.
4. **Open-Ended Feedback:** An open-ended question allowed students to provide additional comments or suggestions, enabling them to highlight strengths, weaknesses, or desired improvements in ReflectMate.

This tailored approach ensured that the survey captured detailed insights into how students interacted with ReflectMate and perceived its impact on their learning processes. By focusing on the specific indicators and features of ReflectMate, the survey provided actionable feedback to refine the tool and align it more closely with student needs.

**Data Collection and Analysis**

The survey was conducted over seven days in December 2023, with participation from 125 students at CADT out of a total of 160 engaged in the study, yielding a substantial response rate. Descriptive statistics were used to summarize quantitative responses. At the same time, qualitative data from open-ended questions was analyzed thematically to identify recurring patterns and trends.

Figure 5.3 illustrates the overall positive feedback and student perceptions of ReflectMate, highlighting its effectiveness in enhancing the learning experience. Students consistently expressed satisfaction with the user progress insights and tools designed to support their academic performance. The recommendation engine, in particular, was praised for its motivational impact and alignment with individual learning patterns.

### 5.4.2 Unveiling Novel Findings

As we take a closer look at the data from our study, three significant findings have emerged and will be discussed. These findings unveil the relevance of ReflectMate in SP enhancement and the correlation among learning components, including students' interactions, ReflectMate-oriented data indicators, and the impacts of ReflectMate on both individuals and the community.

**1. Integrated Dashboard Experience - A Symphony of Connectivity:** interactions are a cornerstone of the learning process, and a well-integrated tool like ReflectMate plays a pivotal role in encouraging these interactions, fostering active learning, and ultimately enhancing academic performance. To evaluate the effectiveness of ReflectMate's dashboard, participants were asked to rate their experiences across four key dimensions—accessibility, data indicators, visualization, and recommendation tools—on a scale ranging from "Excellent" to "Poor." The results, illustrated in Figure 5.4, highlight how ReflectMate's design positively influences student engagement with the LMS. It demonstrates a notably positive ReflectMate Dashboard Experience, where accessibility, data indicators, visualizations, and the recommendation engine create a cohesive and engaging user interface. These elements work in harmony to form what can be described as a "symphony of connectivity," providing students with a seamless and intuitive experience

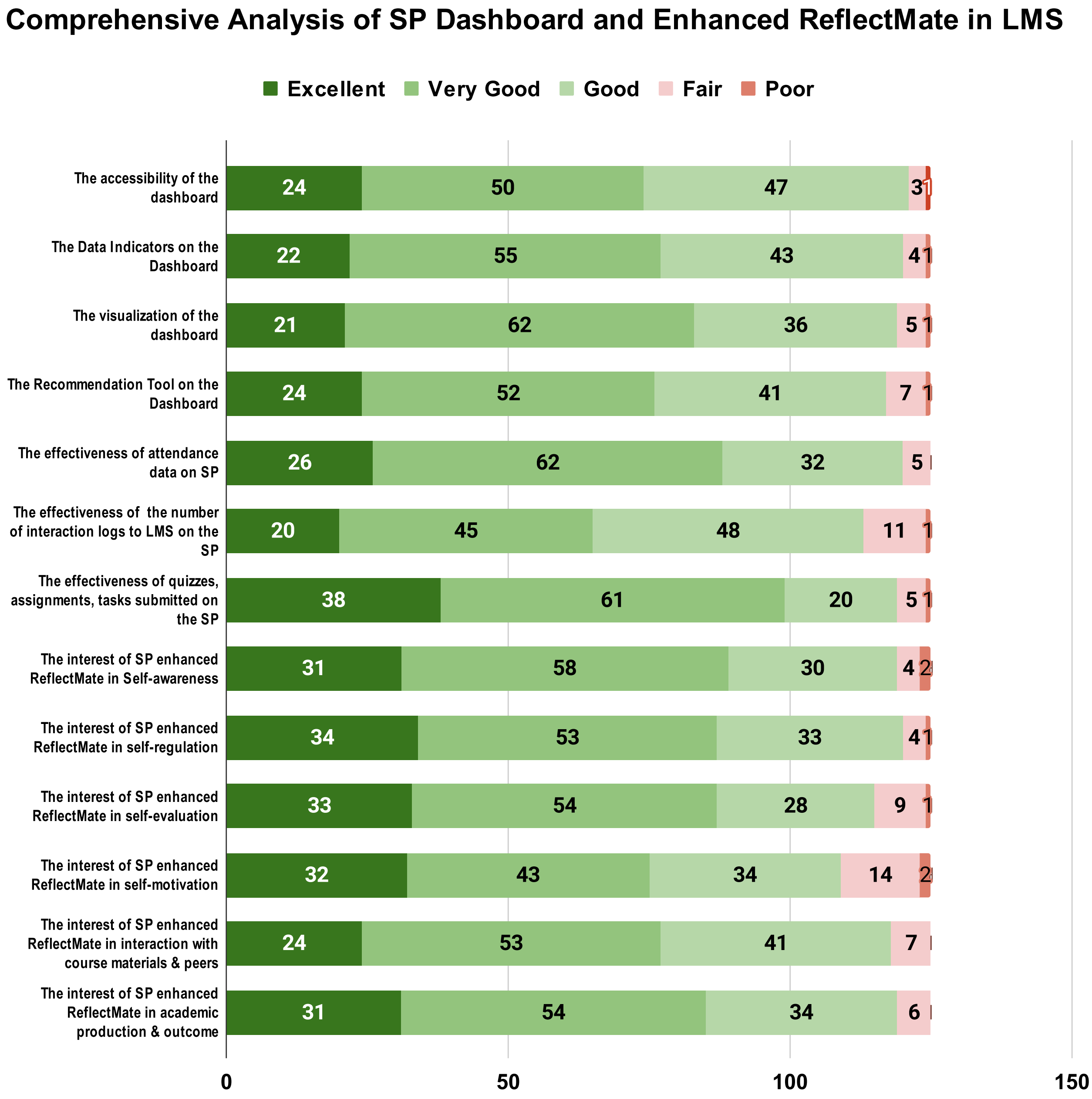


Figure 5.3 – The survey results of SP dashboard and enhanced ReflectMate in LMS

that encourages deeper interaction with their learning environment.

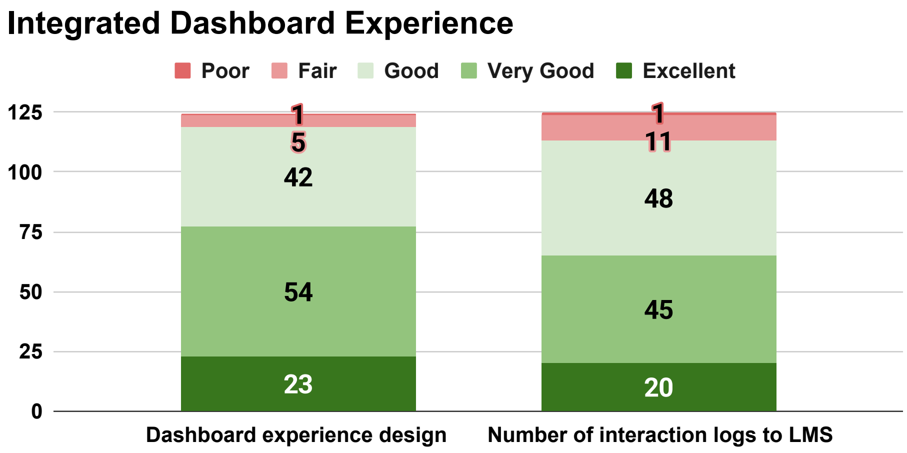


Figure 5.4 – The advantages of dashboard design encourage students to interact with LMS

Our findings reveal a clear correlation between positive dashboard experiences and increased LMS interaction logs. When students perceive the dashboard as accessible, informative, and actionable, their engagement with the LMS grows proportionally. For instance, students who rated the dashboard highly were observed to log in more frequently, spend more time on course materials, and participate more actively in quizzes and assignments. A well-designed dashboard enhances individual components of the learning process and creates a ripple effect, amplifying overall engagement within the learning environment. This finding underscores the critical role of a cohesive and self-reflective dashboard in shaping student behavior. By integrating meaningful data indicators and actionable insights, ReflectMate empowers students to take ownership of their learning journey, reinforcing the importance of thoughtful design in educational technology. Ultimately, ReflectMate's dashboard catalyzes increased interaction, demonstrating its potential to transform passive learners into active participants in their education.

**2. Performance Analytics and Engagement Mastery - A Virtuous Cycle:** our research also uncovered a virtuous cycle in performance analytics and engagement mastery. When students actively participate by attending classes, engaging in more interactions with the LMS, and diligently completing quizzes, assignments, and tasks, a cascade of positive outcomes follows, as shown in Figure 5.5. The academic production and overall outcome align with these engaged behaviors. This intrinsic connection illustrates that student engagement is a catalyst for immediate academic tasks and a predictor of broader academic success. ReflectMate plays a crucial role in helping students become aware of their engagement, thus inciting them to participate even more.

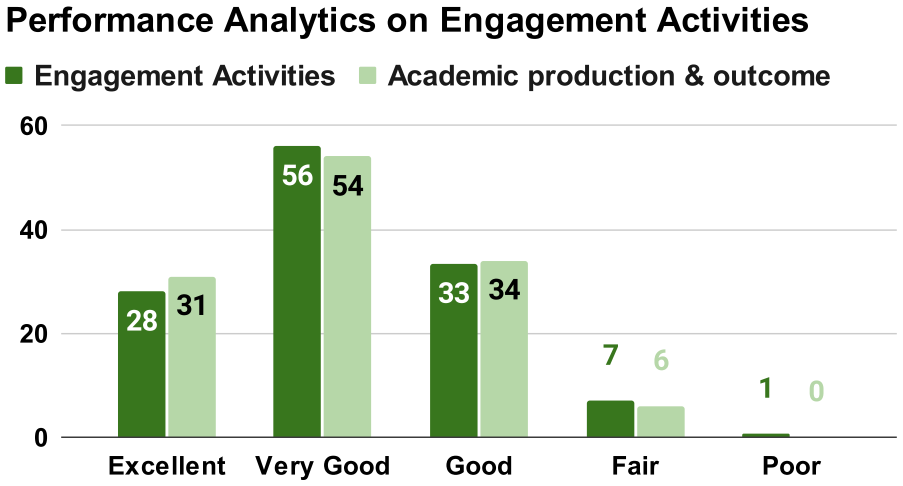


Figure 5.5 – The effectiveness of engagement activities on academic production and outcome

**3. Cognitive Self-Regulation Hub - Empowering Holistic Growth:** From a broader perspective, particularly within the domain of cognitive self-regulation, our findings highlight the critical role of ReflectMate in cultivating metacognitive skills such as self-awareness, self-regulation, self-evaluation, and self-motivation. These skills are foundational to fostering both individual growth and collaborative learning, creating a dynamic educational ecosystem where students can thrive.

The first column of **Figure 5.6** presents student ratings of their perceived development in metacognition after using ReflectMate. The ratings were collected through a survey where students evaluated their progress on a scale ranging from "Excellent" to "Poor." These ratings indicate that a majority of students (85 out of 125, or approximately **68%** ) rated their metacognitive development as either "Excellent" or "Very Good." ReflectMate effectively supports students in developing critical self-regulatory skills, enabling them to reflect on their learning behaviors and make informed decisions about their academic strategies.

The second column of **Figure 5.6** focuses on students ’ perceptions of their interactions with course materials and peers. Here, **61.6% of students** (77 out of 125) rated their interactions as either "Excellent" or "Very Good." This finding underscores ReflectMate's ability to foster meaningful engagement with course materials and promote collaborative learning among peers. For example, students who actively used ReflectMate reported increased participation in discussion forums, more frequent access to learning resources, and improved collaboration with classmates.

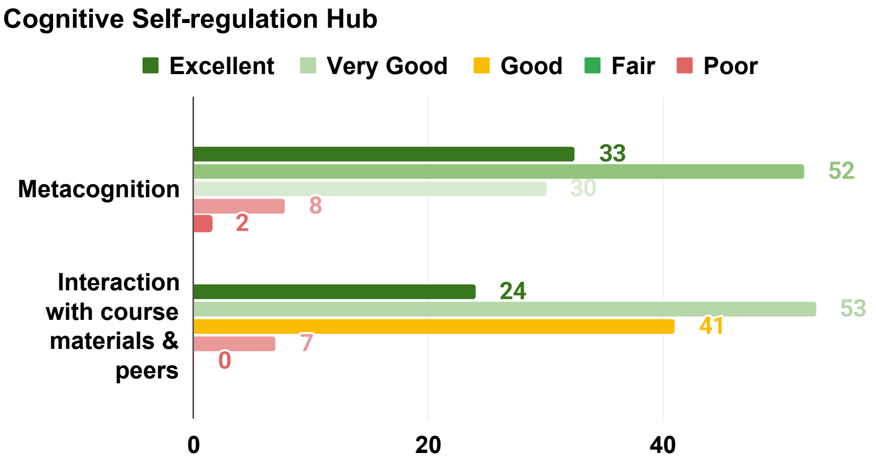


Figure 5.6 – The empowering of metacognition on interaction with course materials and peers

The ratings presented in **Figure 5.6** are based on declarative feedback from students obtained through surveys administered at the end of the study period. To ensure the reliability of these self-reported ratings, we cross-referenced them with objective engagement logs from the LMS. For instance:

- Students who rated their metacognitive development as "Excellent" or "Very Good" demonstrated higher engagement with course materials, as evidenced by increased quiz attempts, assignment submissions, and time spent on learning activities.
- Similarly, students who rated their interactions with peers positively showed higher participation rates in discussion forums and collaborative assignments.

This dual-method approach—combining declarative feedback with objective engagement data—ensures that the findings are grounded in subjective perceptions and observable behavioral patterns.

ReflectMate's impact extends beyond individual growth to foster a collaborative learning community. By encouraging cognitive self-regulation, the tool empowers students to take ownership of their learning while promoting opportunities for peer interaction. For example:

- Students engaged with one another by sharing best practices, comparing learning strategies, and offering mutual support, thus enriching the overall learning experience.
- Collaborative activities facilitated by ReflectMate led to richer discussions and a deeper understanding of course content.

This dynamic interplay between individual self-regulation and group collaboration underscores ReflectMate's potential to transform traditional educational ecosystems into thriving communities of learners.

ReflectMate positions itself not only as a tool for personal development but also as a critical facilitator in fostering a collaborative and interactive learning environment. By empowering students with actionable insights into their learning behaviors, ReflectMate promotes individual growth while offering opportunities for peer interaction. This dual focus on self-regulation and collaboration marks a paradigm shift, positioning ReflectMate as a cornerstone of modern education.

## 5.5 Conclusion

ReflectMate represents a significant advancement in reflective learning analytics, offering students a powerful tool to take ownership of their academic journey. By integrating predictive insights with personalized feedback, ReflectMate fosters self-awareness, self-regulation, self-evaluation, and self-motivation—core metacognition components essential for lifelong learning. The positive outcomes observed in student surveys and engagement data underscore its potential to transform traditional learning experiences into dynamic, student-centered environments.

The study highlights three key contributions of ReflectMate:

1. **Enhanced Engagement Through a Cohesive Dashboard Experience:** ReflectMate's well-designed dashboard, featuring intuitive data indicators, visualizations, and actionable recommendations, creates a "symphony of connectivity" that encourages increased interaction with course materials and peers. This interconnected design enhances individual learning behaviors and fosters a collaborative learning community.

2. **A Virtuous Cycle of Performance Analytics and Engagement Mastery:** ReflectMate demonstrates how active participation—such as attending classes, engaging with the LMS, and completing quizzes and assignments—improves academic outcomes. By making students aware of their engagement patterns, ReflectMate motivates them to adopt more proactive learning strategies, creating a positive feedback loop that drives continuous improvement.

3. **Empowerment Through Cognitive Self-Regulation:** ReflectMate positions

itself as more than just a tool for academic performance enhancement; it catalyzes holistic growth. By empowering students to reflect on their learning behaviors and progress, ReflectMate promotes individual development while encouraging peer interaction and collaboration. This dual focus on self-regulation and group dynamics lays the foundation for a thriving educational ecosystem.

The success of ReflectMate in fostering metacognition and improving student outcomes underscores its value as both a personal development tool and a facilitator of collaborative learning. By bridging the gap between predictive analytics and reflective practices, ReflectMate equips students with the skills they need to navigate their academic journeys effectively and prepares them for future challenges beyond the classroom.

As education continues to evolve in the digital age, tools like ReflectMate exemplify technology's transformative potential when designed with a deep understanding of student needs. ReflectMate's integration into broader educational ecosystems holds promise for addressing persistent challenges such as equity, accessibility, and the digital divide, ensuring that all learners can benefit from data-driven, reflective learning experiences. Building on these insights, **Chapter 6** synthesize the broader implications of this research, exploring how tools like CISE and ReflectMate can address systemic challenges such as the digital divide and educational inequities. Additionally, **Chapter 7** provides a comprehensive summary of the dissertation's contributions, reflecting on the significance of the research for educational practice and policy while outlining directions for future work. These final chapters underscore the potential of data-driven approaches to create more inclusive, equitable, and effective educational systems.

Chapter 6

# Discussion and Implications

## 6.1 Introduction

The rapid evolution of educational technologies has fundamentally reshaped how institutions understand and address student performance. This dissertation has explored two innovative tools—**CISE (Cross-Institutional Stacking Ensemble)** and **ReflectMate—designed** to enhance learning outcomes and promote equitable access to education. While CISE leverages advanced machine learning techniques to predict academic performance across diverse educational settings, ReflectMate empowers students through reflective learning analytics, fostering self-awareness, self-regulation, and proactive engagement in their academic journey.

Looking back at the objectives outlined in the **Introduction**, this chapter revisits the contributions of these tools from a broader perspective, critically examining their strengths, and implications for future research and practice. The discussion begins by synthesizing the key findings from the preceding chapters, highlighting how these tools address the challenges identified in the initial stages of this research. It then evaluates the rationale behind the choices in designing and implementing these tools, considering alternative approaches and their potential impact.

From a broader perspective, integrating predictive modeling and reflective analytics represents a significant step forward in addressing the dual goals of improving academic outcomes and fostering equity in education. However, this chapter also reflects on areas where the approaches could have been refined or expanded, acknowledging the inherent trade-offs in any research endeavor. This discussion aims to provide a balanced assessment of what has been achieved and what remains to be explored in pursuing inclusive, data-driven educational systems.

## 6.2 Performance of Predictive Models and ReflectMate

The development and implementation of **CISE** and **ReflectMate** represent significant advancements in addressing the challenges of predicting student performance and fostering self-regulated learning. These tools exemplify how predictive modeling and reflective analytics can synergistically enhance institutional decision-making and individual student empowerment. This section revisits the performance of these tools from a broader perspective, highlighting their contributions, strengths, and areas for improvement.

### 6.2.1 CISE: Generalization Across Institutional Boundaries

The **CISE model** was designed to address one of the most persistent challenges in educational data mining: the lack of generalizability across diverse datasets (Baker et al. 2019). By integrating multiple base models through a meta-model layer, CISE demonstrated its ability to handle variability in grading systems, feature distributions, and class imbalances across datasets from three institutions—IDT, IDG, and ITC—and validated its robustness on the ENSIM dataset.

Key outcomes of CISE model include:

— **High Predictive Accuracy:** The meta-model achieved an F1 score of **78.25%**

on the ENSIM validation dataset, outperforming all individual base models. This result underscores the efficacy of ensemble methods in improving predictive performance by leveraging the complementary strengths of diverse classifiers such as Decision Trees, Random Forests, Neural Networks, Naive Bayes, and Support Vector Machines (Dietterich 2000).

— **Cross-Institutional Robustness:** CISE demonstrated its capacity to generalize across diverse academic environments by training on datasets from multiple institutions. This cross-institutional validation is particularly significant given the variability in curricular structures, student demographics, and teaching practices across institutions (Porras et al. 2023).

— **Addressing Class Imbalance:** The model effectively mitigated the challenges of imbalanced grade distributions through oversampling techniques and careful feature engineering. For example, underrepresented categories such as top and bottom performers were better captured, ensuring that predictions were not biased toward the majority class (Wongvorachan, S. He, and Bulut 2023).

These findings highlight the potential of ensemble methods like CISE to provide educators with reliable tools for early identification of at-risk students, enabling timely interventions to improve academic outcomes. However, it is important to acknowledge certain limitations. For instance, while CISE performed well across institutions, its reliance on high-quality, standardized datasets may limit its applicability in settings where data collection practices are inconsistent or incomplete. Future iterations could explore transfer learning techniques to enhance generalizability (S. J. Pan and Yang 2009) further.

### 6.2.2 ReflectMate: Bridging Insights and Action

**ReflectMate** builds on CISE's predictive power by translating its outputs into actionable student feedback. By integrating directly into the CADT LMS, ReflectMate provides students real-time insights into their learning behaviors, fostering metacognition and proactive engagement. This dual focus on prediction and reflection represents a paradigm shift in how educational technologies can support institutional and individual needs.

Core benefits of ReflectMate include:

— **Enhanced Student Engagement:** ReflectMate's interactive dashboard allows students to visualize their progress, fostering a sense of accountability and motivation. Features such as user progress indicators, personalized feedback, and group-level overviews empower students to take ownership of their learning journey (Zimmerman 2002).

— **Actionable Feedback:** By providing specific recommendations—such as improving quiz engagement or submitting assignments on time—ReflectMate empowers students to make data-driven decisions about their learning strategies. For example, students with low quiz attempts receive targeted suggestions to reinforce their understanding of key concepts (Flavell 1979).

— **Promoting Equity:** ReflectMate's focus on individual progress indicators ensures that all students receive targeted support to improve their outcomes regardless of their starting point. This inclusivity is critical in addressing disparities caused by the digital divide, which disproportionately affects students from underrepresented groups (Farooqi, Khalid, and A. Khan 2022).

While ReflectMate has demonstrated significant potential, its effectiveness depends on the underlying predictive models' quality and engagement data availability. For instance, students with limited access to technology may not generate sufficient interaction logs, potentially limiting the tool's ability to provide accurate feedback. Addressing this challenge requires ongoing efforts to bridge the digital divide and ensure equitable access to educational resources (Afzal et al. 2023).

### 6.2.3 Synergistic Impact of CISE and ReflectMate

CISE and ReflectMate exemplify the transformative potential of combining predictive modeling with reflective analytics. While CISE provides educators with reliable insights into student performance, ReflectMate translates these insights into actionable student feedback, creating a virtuous cycle of engagement and improvement. This synergy addresses two critical dimensions of education: institutional decision-making and individual empowerment.

For example, integrating CISE's predictive insights with ReflectMate's self-reflection tools fosters cognitive self-regulation, enabling students to develop skills such as self-awareness, self-evaluation, and self-motivation (Zimmerman 2002). At the same time, ReflectMate's group-level overviews promote collaborative learning, positioning individual self-regulation and peer interaction as foundational pillars of a thriving educational ecosystem.

However, the success of this approach hinges on the seamless integration of predictive models with reflective tools. Challenges such as data privacy, interpretability of models, and the need for continuous refinement remain critical considerations. Future research should explore ways to enhance the transparency and fairness of predictive models while ensuring that reflective tools remain accessible and inclusive for all students (Khademi and Honavar 2020).

## 6.3 Insights from Entrance Exam Data at CADT

This section analyzes data from 4837 students' entrance exams administered by the **CADT** over three years (2020–2022) to explore further the digital divide's challenges and its impact on equitable access to education. These exams are a prerequisite for bachelor's students seeking admission to ICT-related programs. The analysis focuses on two critical dimensions: **gender disparities** and **location-based inequities**, as reflected in the application rates, success rates, and performance metrics. Personal data used in this analysis were anonymized to ensure confidentiality and compliance with data protection regulations.

### 6.3.1 Gender Disparities in ICT Education

The data reveals significant gender disparities in ICT education at CADT, particularly in primary selection and academic success. Table 6.1 summarizes key metrics that illustrate these gaps.

Table 6.1 – Statistics on Gender Disparities in ICT Education at CADT

| **Metric** | **Male (%)** | **Female (%)** | **Observations** |
|---|---|---|---|
| Major Selection in Computing and Technology | 70.55 | 29.45 | Male dominance in technical fields. |
| Major Selection in Digital Management and Business | 42.67 | 57.33 | Female preference for management-oriented fields. |
| Major Selection in Telecommunications and Networking | 61.69 | 38.31 | Gender disparity persists but is less pronounced. |
| Success Rates in Computing and Technology | 74.03 | 67.61 | Higher male success rates. |
| Success Rates in Digital Management and Business | 72.08 | 55.67 | Female students face challenges in management success. |
| Success Rates in Telecommunications and Networking | 66.72 | 57.03 | Female students face challenges in management success. |
| Success Rates from High Human Development Areas | 73.76 | 26.24 | Male students dominate regions with higher resources. |
| Success Rates from Low Human Development Areas | 61.95 | 38.05 | Female underrepresentation in disadvantaged areas. |

From Table 6.1, it is evident that male students dominate technical fields such as **Com-**

**puting, and Technology** and **Telecommunications and Networking**, accounting for **70.55%** and **61.69%** of applications, respectively. In contrast, female students show a stronger preference for **Digital Management and Business**, where they represent **57.33%** of applicants. However, even in this field, female success rates lag behind those of their male counterparts (**55.67% vs. 72.08%**). These disparities highlight female students' persistent barriers to accessing and succeeding in ICT education. Societal norms, gender biases, and limited exposure to STEM subjects during early education likely contribute to these inequities. Addressing these challenges requires targeted interventions, such as mentorship programs, scholarships, and initiatives to promote female participation in technical fields.

### 6.3.2 Location-Based Inequities in ICT Education

In addition to gender disparities, the data reveals significant location-based inequities in ICT education at CADT. Table 6.2 shows application and success rates by geographic location.

Table 6.2 underscores the disproportionate representation of students from **High Human Development Areas**, who account for **32.33%** of applications but achieve a success rate of **72.37%**. In contrast, students from **Low Human Development Areas** represent only **3.10%** of applicants and have a success rate of **61.95%**, reflecting underrepresentation and lower academic outcomes.

These findings highlight the profound impact of the digital divide on educational equity. Students from low-development areas often lack access to essential resources such as reliable internet connectivity, digital devices, and quality education. This lack of access limits their ability to participate in ICT education, affecting their performance in entrance exams and subsequent academic success. Bridging this gap requires concerted efforts to provide technological infrastructure, digital literacy training, and equitable access to educational opportunities for students from disadvantaged regions.

Table 6.2 – Statistics on Location-Based Disparities in ICT Education at CADT

| **Metric** | **Applications (%)** | **Success Rate (%)** | **Observations** |
|---|---|---|---|
| Representation from High Human Development Areas | 32.33 | 72.37 | Students dominate regions with higher resources. |
| Representation from Medium-high Human Development Areas | 41.29 | 68.15 | Students predominantly come from regions with relatively better resources and access to opportunities. |
| Representation from Medium-low Human Development Areas | 23.29 | 66.47 | Students predominantly come from regions with moderately better resources compared to lower development areas. |
| Representation from Low Human Development Areas | 3.10 | 61.95 | Students underrepresentation in disadvantaged areas. |

### 6.3.3 Implications for Predictive Modeling and ReflectMate

The insights from the entrance exam data have important implications for developing and implementing tools like CISE and ReflectMate. By integrating predictive modeling with reflective analytics, these tools can help identify and address the root causes of inequities in ICT education. For example:

— **CISE** can be used to predict the likelihood of success for students from different demographic groups, enabling targeted interventions to support underrepresented populations.

— **ReflectMate** can empower students by providing personalized feedback and actionable recommendations, helping them overcome barriers related to engagement, resource access, and self-regulation.

Moreover, including location-based and gender-specific attributes in predictive models ensures that these tools are accurate, equitable, and inclusive. By addressing the challenges posed by the digital divide, CISE and ReflectMate contribute to creating a more level playing field for all students, regardless of their background or circumstances.

The CADT entrance exam data analysis reveals significant disparities in ICT education based on gender and geographic location. These inequities underscore the urgent need for interventions that promote inclusivity and equity in education. Tools like **CISE** and **ReflectMate** offer promising solutions by leveraging predictive modeling and reflective analytics to identify at-risk students and provide them with the support they need to succeed. We can work towards a more equitable and inclusive educational ecosystem by addressing the challenges highlighted in this analysis.

Chapter 7

# Conclusion

## 7.1 Introduction

The culmination of this dissertation represents not only the resolution of the research questions posed at its inception but also the realization of a broader vision: to harness the power of predictive modeling and reflective learning analytics to address critical challenges in education, particularly in diverse and resource-constrained settings. This study was motivated by the pressing need to bridge the digital divide, enhance educational equity, and empower students through data-driven insights. Addressing these challenges, the research sought to answer our research questions.

Through developing and validating the **Cross-Institutional Stacking Ensemble (CISE)** model, this research has demonstrated that predictive models can achieve high levels of accuracy and generalizability when trained on diverse datasets from multiple institutions. The CISE model, which integrates base classifiers such as Decision Trees, Random Forests, Neural Networks, and Support Vector Machines, was rigorously tested on datasets from institutions like IDT, IDG, ITC, and ENSIM, achieving an F1 score of **78.25%** on the ENSIM validation dataset. These findings underscore the importance of cross-institutional validation in ensuring that predictive models are robust and applicable across varying educational contexts, thereby addressing one of the most significant limitations of current research in educational data mining.

Simultaneously, integrating **ReflectMate**, a reflective learning tool, with learning analytics has illuminated the transformative potential of empowering students through personalized feedback and actionable insights. ReflectMate leverages predictive analytics to give students real-time feedback on their engagement metrics—such as time spent on the LMS, quiz attempts, assignment submissions, and interaction logs—enabling them to reflect on their learning behaviors and make informed decisions about their academic strategies. The positive outcomes observed in student surveys and engagement data highlight the tool's efficacy in fostering self-awareness, self-regulation, and metacognition, which are essential for lifelong learning and academic success.

Furthermore, this research has shed light on the profound impact of the digital divide on educational equity, particularly in developing countries like Cambodia. The analysis of entrance exam data from CADT revealed significant gender and location-based disparities in ICT education, with male students dominating technical fields and students from high human development areas achieving higher success rates. These findings emphasize the urgent need for targeted interventions to address systemic inequities and ensure all students have equitable access to educational resources and opportunities.

Finally, the ethical implications of predictive modeling in education have been carefully considered throughout this study. Ensuring fairness, transparency, and the responsible use of student data has been a guiding principle in the design and implementation of both the CISE model and ReflectMate. By addressing issues such as bias, interpretability, and privacy, this research provides a framework for developing ethical and trustworthy predictive models that can be widely adopted in educational settings.

In this concluding chapter, we synthesize the contributions of this research in light of the findings presented in previous chapters. Each contribution is framed as a response to the research questions, illustrating how the results can be exploited to advance educational practices, inform policy decisions, and promote equity in education. By doing so, this chapter not only summarizes the key achievements of the study but also highlights its broader implications for the future of education in an increasingly digital world.

## 7.2 Limitations and Future Directions

While the development and implementation of **CISE** and **ReflectMate** represent significant advancements in educational data mining and reflective learning analytics, it is essential to acknowledge the limitations of this research and identify areas for future exploration. These limitations stem from challenges inherent in predictive modeling, reflective analytics, and the broader context of addressing the digital divide in education.

### 7.2.1 Limitations of Predictive Modeling

The **CISE model**, despite its robust performance across diverse datasets, is not without limitations. One key challenge lies in the variability of educational data, which often reflects differences in curricular structures, teaching practices, and student demographics. While the model demonstrated strong generalizability across institutions such as IDT, IDG, and ITC, its applicability to other contexts—particularly those with significantly different grading systems or engagement metrics—remains untested. Future iterations of the model could benefit from expanding the dataset to include more institutions and regions, ensuring that it captures a broader spectrum of educational environments (Baker et al. 2019).

Another limitation is the reliance on oversampling techniques to address class imbalance. While these methods improved the model's ability to predict underrepresented categories, such as top and bottom performers, they may introduce biases if not carefully calibrated. Future research should explore alternative approaches, such as transfer learning or domain adaptation, to enhance the model's robustness and fairness (S. J. Pan and

Yang 2009).

Additionally, the interpretability of the CISE model remains a concern. Although ensemble methods like stacking effectively improve predictive accuracy, their "black box" nature can make it challenging for educators and students to understand how predictions are made. Developing explainable AI techniques, such as feature importance measures or decision rules, could help bridge this gap and build trust in the model's outputs (Abdelqader et al. 2022).

### 7.2.2 Limitations of ReflectMate

ReflectMate, while empowering students through personalized feedback and actionable insights, also faces several limitations. One critical issue is the tool's dependence on high-quality engagement data from the LMS. Students from low human development areas or those with limited access to technology may generate incomplete or biased interaction logs, potentially limiting the tool's effectiveness in providing accurate feedback (Farooqi, Khalid, and A. Khan 2022). Addressing this challenge requires ongoing efforts to bridge the digital divide and ensure equitable access to educational resources.

Finally, the current version of ReflectMate primarily targets academic outcomes, such as quiz scores and assignment submissions. However, it does not fully account for non-academic factors influencing student success, such as socio-economic background, mental health, or extracurricular commitments. Incorporating these factors into the feedback engine could provide a more holistic view of student performance and well-being (Hirokawa 2018).

### 7.2.3 Broader Challenges in Addressing the Digital Divide

This research highlights the digital divide's profound impact on educational equity, particularly in developing countries like Cambodia. While tools like CISE and ReflectMate offer promising strategies for mitigating these disparities, their effectiveness is constrained by systemic issues such as unequal access to technology, digital literacy gaps, and socio-cultural norms that perpetuate gender and location-based inequities (Bon, Saa-Dittoh, and Akkermans 2024; Vassilakopoulou and Hustad 2023).

For instance, Table 6.2 reveals that only **3.10%** of applications come from low human development areas, where students face significant barriers to accessing ICT education. Similarly, Table 6.1 underscores the persistent gender disparities in technical fields, with male students dominating areas like Computing and Technology. Addressing these structural inequities requires concerted efforts beyond the scope of predictive modeling and reflective analytics, including policy interventions, infrastructure investments, and community engagement initiatives.

### 7.2.4 Future Directions

To address the limitations identified in this study and build on the successes of **CISE** and **ReflectMate**, several actionable avenues for future research are proposed:

1. **Scaling Research at CADT Research Lab:** The **CISE model** provides a strong foundation for expanding research initiatives within the CADT research lab. Future work could focus on adapting the model for broader applications, such as predicting outcomes in non-ICT disciplines or evaluating its effectiveness across diverse student populations. Collaborating with researchers across disciplines and applying advanced data-driven methodologies would enhance learning outcomes and promote scalability. Refining predictive algorithms to improve accuracy and

usability will also be critical for addressing generalizability challenges.

2. **Integrating Explainable AI (XAI) into Predictive Model:** Future iterations of **CISE** could incorporate **Explainable AI (XAI)** techniques to address concerns about transparency and trust in predictive modeling. For example, SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could provide clear, interpretable insights into why specific performance trends occur. This approach would empower students and educators to take informed actions, fostering fairness, bias detection, and trust in AI-driven education. Furthermore, studying explainability would contribute to broader research on ethical considerations in educational technologies.

3. **Addressing Class Imbalance and Dataset Heterogeneity:** While oversampling techniques effectively mitigate class imbalance, future research could explore advanced strategies such as cost-sensitive learning or hybrid ensemble methods to improve predictions for underrepresented categories further.

4. **Enhancing ReflectMate's Accessibility and Usability:** Ensuring that **ReflectMate** is accessible to all students, including those with disabilities or limited digital literacy, is crucial for promoting equity. Future research should investigate ways to optimize the tool's user interface, language support, and compatibility with assistive technologies. These efforts would ensure inclusive access and maximize the tool's impact on underserved populations.

5. **Exploring Emerging Technologies and Data Sources:** As educational technologies evolve, new data types—such as those generated by virtual reality (VR), gamified learning platforms, and AI-driven tutoring systems—will become increasingly available. Future research should explore how these novel data sources can be integrated into predictive models to capture a more comprehensive picture

of student learning behaviors. Additionally, leveraging emerging technologies like blockchain for secure data sharing and privacy-preserving analytics could enhance the ethical use of student data.

6. **Developing Policies to Reduce the Digital Divide:** Bridging the digital divide remains a critical priority for ensuring equitable access to educational resources. Policymakers and institutions should invest in digital infrastructure, provide scholarships and resources for students from low human development areas, and establish mentorship programs to support underrepresented groups. Evaluating the effectiveness of these policies through mixed-methods approaches would provide valuable insights into reducing systemic inequities in education.
7. **Conducting Longitudinal Studies on ReflectMate's Impact:** While this study highlights ReflectMate's short-term benefits in enhancing student engagement and academic performance, longitudinal studies could assess its long-term effects on metacognition, self-regulation, and career outcomes. Such studies would provide deeper insights into the sustained impact of reflective learning tools on lifelong learning skills.

In conclusion, while CISE and ReflectMate have demonstrated significant potential to improve educational outcomes and promote equity, their limitations underscore the need for continued innovation and refinement. By addressing these challenges and exploring new directions, future research can build on this study's successes to create more inclusive, data-driven educational systems that empower all students to succeed.

## 7.3 Contributions

This dissertation addresses critical educational challenges by leveraging predictive modeling and reflective learning analytics to enhance student outcomes and promote equity. The contributions of this research are presented as direct responses to the research questions outlined in Chapter 1, illustrating how the findings can be exploited to advance educational practices and policies. Each contribution is framed in the context of the tools developed—**CISE** and **ReflectMate**—and their implications for addressing the digital divide, fostering reflective learning and ensuring ethical and equitable use of data-driven technologies.

### 7.3.1 Addressing Research Question 1: How can a predictive model be designed to predict student performance across institutions with generalizability?

The first research question focuses on developing and validating predictive models that are robust and applicable across diverse educational settings. This study responds to this challenge by introducing the **CISE model**. This innovative ensemble approach integrates multiple base classifiers and meta-models to achieve high predictive accuracy and generalizability. Key contributions include:

1. **Development of a Cross-Institutional Predictive Model:** The CISE model was trained and validated on datasets from four distinct institutions—IDT, IDG, ITC, and ENSIM—spanning different curricular structures, grading systems, and student demographics. By addressing variability in feature distributions and class imbalances, the model achieved an F1 score of 78.25% on the ENSIM validation dataset, outperforming individual base models. It demonstrates the model's ability to generalize across institutional boundaries, making it a valuable tool for educators seeking reliable insights into student performance.
2. **Mitigating Class Imbalance Through Oversampling Techniques:** One of the significant challenges in predictive modeling is the uneven distribution of grades,

particularly for underrepresented categories such as top and bottom performers. The CISE model addressed this issue through oversampling techniques, ensuring that predictions were balanced and inclusive. This approach highlights the importance of fairness in predictive modeling and sets a precedent for future research in this area.

3. **Providing Actionable Insights for Early Interventions:** The CISE model's ability to predict student outcomes accurately enables educators to identify at-risk students early in the academic term. By providing timely interventions, institutions can improve retention rates, enhance academic performance, and reduce disparities among student groups. This contribution underscores the transformative potential of predictive analytics in creating proactive and equitable educational environments.

### 7.3.2 Addressing Research Question 2: How can insights from student performance predictive models be leveraged to design a reflective learning dashboard to enhance student engagement and academic performance?

The second research question explores the impact of integrating reflective learning tools with learning analytics to empower students and enhance their academic outcomes. This study introduces **ReflectMate**, a tool designed to translate predictive insights into actionable feedback, fostering self-awareness, self-regulation, and continuous improvement. Key contributions include:

1. **Empowering Students Through Personalized Feedback:** ReflectMate leverages engagement metrics from LMS platforms to give students real-time insights into their learning behaviors. Features such as progress tracking, personalized rec-

ommendations, and visualizations enable students to reflect on their strengths and areas for improvement. For example, students with low quiz engagement receive tailored suggestions, such as "Consider taking quizzes more frequently to reinforce your understanding." This personalized approach ensures that all students receive targeted support to improve their outcomes regardless of their starting point.

2. **Fostering Metacognition and Lifelong Learning Skills:** By promoting self-awareness, self-regulation, self-evaluation, and self-motivation, ReflectMate develops metacognitive skills essential for lifelong learning. Survey results from CADT students demonstrate the tool's effectiveness in enhancing metacognition, with participants reporting increased motivation and accountability. These findings highlight the role of reflective learning analytics in transforming passive learners into active participants in their educational journeys.

3. **Bridging the Gap Between Predictive Insights and Actionable Practices:** ReflectMate bridges the gap between predictive modeling and practical application by translating complex data into understandable and actionable feedback. This dual benefit empowers educators and students: educators gain insights to tailor interventions. In contrast, students take ownership of their learning strategies. The synergy between CISE and ReflectMate exemplifies how predictive analytics and reflective tools can work together to create holistic solutions for improving educational outcomes.

### 7.3.3 Addressing Research Question 3: How can predictive modeling and a reflective learning dashboard help reduce disparities based on gender and location by providing fair predictions and personalized support for all students?

The third research question examines the influence of the digital divide on the implementation and effectiveness of predictive models and reflective learning tools. This study provides valuable insights into gender and location-based disparities in ICT education at CADT, offering strategies to mitigate these inequities. Key contributions include:

1. **Highlighting Structural Inequities in ICT Education:** Statistical analyses reveal significant disparities in access to educational resources and opportunities. For instance, only 3.10% of applications come from low human development areas, where students face systemic barriers to success. Similarly, male students dominate technical fields like Computing and Technology, while female students are underrepresented in disadvantaged regions. These findings underscore the urgent need for targeted interventions to address structural inequities.

2. **Leveraging Data-Driven Approaches to Promote Equity:** The integration of CISE and ReflectMate offers promising strategies for mitigating the effects of the digital divide. ReflectMate's focus on individual progress indicators ensures that students from underrepresented groups receive tailored guidance. At the same time, the CISE model's cross-institutional insights enable institutions to identify and address systemic patterns of performance disparities. Together, these tools contribute to the broader goal of creating inclusive and equitable educational systems.

3. **Informing Policy and Practice on Digital Inclusion:** This study's findings have significant implications for policymakers and institutions seeking to bridge the

digital divide. Recommendations include providing scholarships and resources for students from low-human-development areas, establishing mentorship programs to support female students in technical fields, and investing in digital infrastructure to reduce regional disparities. These strategies align with the broader goals of digital inclusion and equity in education.

### 7.3.4 Addressing Research Question 4: What Are the Ethical Considerations Associated with the Use of Predictive Models in Education, and How Can They Be Addressed to Ensure Fairness and Transparency?

The fourth research question addresses the ethical implications of using predictive models in education, focusing on fairness, transparency, and the responsible use of student data. This study contributes to the growing body of literature on the ethics of educational data mining by emphasizing the importance of ethical considerations in designing and implementing predictive tools. Key contributions include:

1. **Ensuring Fairness and Transparency in Predictive Modeling:** The CISE model incorporates oversampling and feature engineering techniques to mitigate biases and ensure fairness. By addressing class imbalance issue, the model sets a benchmark for developing ethical and trustworthy predictive tools that can be widely adopted in educational settings.
2. **Protecting Student Privacy and Consent:** ReflectMate prioritizes student privacy by anonymizing data and ensuring that feedback is generated securely and transparently. This approach fosters trust among students and educators, encouraging the responsible use of technology in education.
3. **Promoting Ethical Guidelines for Future Research:** The ethical framework

developed in this study guides future research in educational data mining and learning analytics. By emphasizing fairness, transparency, and the responsible use of data, this framework ensures that technological advancements in education are aligned with students' rights and interests.

### 7.3.5 Synthesis of Contributions

In summary, this dissertation's contributions respond directly to the research questions, providing actionable solutions to key educational challenges. The CISE model and ReflectMate exemplify how predictive modeling and reflective learning analytics can synergistically enhance institutional decision-making and individual student empowerment. By addressing the digital divide, fostering reflective learning, and ensuring ethical and equitable use of data-driven technologies, this research lays the foundation for creating inclusive, data-driven educational systems that empower all students to succeed.

These contributions have far-reaching implications for educational practice, policy, and future research. Educators can leverage the tools and methodologies developed in this study to tailor interventions, promote equity, and foster lifelong learning skills. Policymakers can use the findings to inform targeted strategies for bridging the digital divide and ensuring equal opportunities in education. Researchers can build on this work to explore new avenues for integrating advanced analytics into educational systems, expanding datasets, and incorporating socio-economic factors to capture the full spectrum of influences on student performance.

## 7.4 Conclusion

This chapter has synthesized the key contributions of this dissertation, highlighting its potential to transform educational practices through predictive modeling, reflective learning analytics, and efforts to address the digital divide. By developing and validating the **CISE** model, this research has demonstrated the feasibility of creating generalizable predictive models that can be applied across diverse educational settings. Integrating **ReflectMate**, a reflective learning tool, has further underscored the importance of empowering students through data-driven insights and fostering self-awareness, self-regulation, and lifelong learning skills.

The future directions outlined in Section 7.2.4 present exciting opportunities to build on these contributions. Expanding datasets, enhancing reflective tools, addressing the digital divide, and ensuring ethical practices will be critical for advancing the field of educational data mining and learning analytics. These efforts align with creating inclusive, equitable, and effective educational systems that empower all students to succeed.

In conclusion, this research underscores the transformative potential of integrating predictive analytics and reflective learning tools in education. This study lays the foundation for a future where technology enhances learning outcomes while promoting fairness and inclusivity by addressing key challenges such as generalizability, interpretability, and equity. As we move forward, we must continue exploring innovative solutions that bridge gaps in access, support diverse learners, and inspire lifelong learning.

# Bibliography

Abdelqader, Husam et al. (2022), « Interpretable and Reliable Rule Classification Based on Conformal Prediction », *in*: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 385–401.

Afzal, Arfa et al. (2023), « Addressing the Digital Divide: Access and Use of Technology in Education », *in*: *Journal of Social Sciences Review* 3.*2*, pp. 883–895.

Ahmed, Allam (2007), « Open access towards bridging the digital divide–policies and strategies for developing countries », *in*: *Information Technology for Development* 13.*4*, pp. 337–361.

Alam, Ashraf (2023), « The Secret Sauce of Student Success: Cracking the Code by Navigating the Path to Personalized Learning with Educational Data Mining », *in*: *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, IEEE, pp. 1–8.

Ali, Haseeb et al. (2019), « Imbalance class problems in data mining: A review », *in*: *Indonesian Journal of Electrical Engineering and Computer Science* 14.*3*, pp. 1560–1571.

AlShammari, Iqbal, Mohammed Aldhafiri, and Zaid Al-Shammari (2013), « A meta-analysis of educational data mining on improvements in learning outcomes », *in*: *College Student Journal* 47.*2*, pp. 326–333.

Alshareef, Fatima et al. (2020), « Educational data mining applications and techniques », *in*: *International Journal of Advanced Computer Science and Applications* 11.*4*.

Alturki, Sarah and Nazik Alturki (2021), « Using educational data mining to predict students' academic performance for applying early interventions », *in*: *Journal of Information Technology Education: JITE. Innovations in Practice: IIP* 20, pp. 121–137.

Andrade, Heidi L (2019), « A critical review of research on student self-assessment », *in*: *Frontiers in Education*, vol. 4, Frontiers Media SA, p. 87.

Antonio, Amy and David Tuffley (2014), « The gender digital divide in developing countries », *in*: *Future Internet* 6.*4*, pp. 673–687.

Arizmendi, Cara J. et al. (2022), « Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work », *in*: *Behavior Research Methods* 55, pp. 3026–3054, URL: `https://api.semanticscholar.org/CorpusID:251842247`.

Arthars, Natasha et al. (2019), « Empowering Teachers to Personalize Learning Support: Case Studies of Teachers' Experiences Adopting a Student-and Teacher-Centered Learning Analytics Platform at Three Australian Universities », *in*: *Utilizing learning analytics to support study success*, pp. 223–248.

Ashfaq, Usman, PM Booma, and Raheem Mafas (2020), « Managing student performance: A predictive analytics using imbalanced data », *in*: *International Journal of Recent Technology and Engineering* 8.*6*, p. 6.

Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed (2020), « An intelligent prediction system for educational data mining based on ensemble and filtering approaches », *in*: *Procedia Computer Science* 167, pp. 1471–1483.

Asif, Raheela, Agathe Merceron, Syed Abbas Ali, et al. (2017), « Analyzing undergraduate students' performance using educational data mining », *in*: *Computers & education* 113, pp. 177–194.

Asif, Raheela, Agathe Merceron, and Mahmood K Pathan (2014), « Predicting student academic performance at degree level: a case study », *in*: *International Journal of Intelligent Systems and Applications* 7.*1*, pp. 49–61.

Assegie, Tsehay Admassu et al. (2024), « Evaluation of Random Forest and Support Vector Machine Models in Educational Data Mining », *in*: *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, IEEE, pp. 131–135.

Bai, Xiaomei et al. (2021), « Educational big data: Predictions, applications and challenges », *in*: *Big Data Research* 26, p. 100270.

Baker, Ryan S et al. (2019), « Challenges for the future of educational data mining: The Baker learning analytics prizes », *in*: *Journal of educational data mining* 11.*1*, pp. 1–17.

Banihashem, Seyyed Kazem et al. (2022), « A systematic review of the role of learning analytics in enhancing feedback practices in higher education », *in*: *Educational Research Review*, p. 100489.

Barakeh, Abdullah M, Mohammad A Mezher, and Banan A Alharbi (2024), « Literature Review for Educational Data Mining Systems—Fahad Bin Sultan University Case Study », *in*: *Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability*, Springer, pp. 435–453.

Barbosa, Pedro Luis Saraiva et al. (2024), « Adaptive learning in computer science education: A scoping review », *in*: *Education and Information Technologies* 29.*8*, pp. 9139–9188.

Black, Paul and Dylan Wiliam (1998), « Assessment and classroom learning », *in*: *Assessment in Education: principles, policy & practice* 5.*1*, pp. 7–74.

Blanchard, Emmanuel G (2015), « Socio-cultural imbalances in AIED research: Investigations, implications and opportunities », *in*: *International Journal of Artificial Intelligence in Education* 25, pp. 204–228.

Bon, Anna, Francis Saa-Dittoh, and Hans Akkermans (2024), « Bridging the digital divide », *in*: *Hannes Werthner · Carlo Ghezzi · Jeff Kramer · Julian Nida-Rümelin · Bashar Nuseibeh · Erich Prem ·*, p. 283.

Boud, David, Rosemary Keogh, and David Walker (2013), *Reflection: Turning experience into learning*, Routledge.

Brahim, Ghassen Ben (2022), « Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features », *in*: *Arabian Journal for Science and Engineering* 47, pp. 10225–10243, URL: `https://api.semanticscholar.org/CorpusID:246028835`.

Breiman, Leo (1996), « Bagging predictors », *in*: *Machine learning* 24, pp. 123–140.

— (2001), « Random forests », *in*: *Machine learning* 45, pp. 5–32.

Calderón Gómez, Daniel (2019), « Technological capital and digital divide among young people: an intersectional approach », *in*: *Journal of youth studies* 22.*7*, pp. 941–958.

Carlson, Andrew and Alyssa M Isaacs (2018), « Technological capital: an alternative to the digital divide », *in*: *Journal of Applied Communication Research* 46.*2*, pp. 243–265.

Chatti, Mohamed Amine et al. (2021), « Designing theory-driven analytics-enhanced self-regulated learning applications », *in*: *Visualizations and dashboards for learning analytics*, Springer, pp. 47–68.

Chawla, Nitesh V et al. (2002), « SMOTE: synthetic minority over-sampling technique », *in*: *Journal of artificial intelligence research* 16, pp. 321–357.

Cheshmehzangi, Ali et al. (2023), « The growing digital divide in education among primary and secondary children during the COVID-19 pandemic: An overview of social exclusion and education equality issues », *in*: *Journal of Human Behavior in the Social Environment* 33.*3*, pp. 434–449.

Chetty, Krish et al. (2018), « Bridging the digital divide: measuring digital literacy », *in*: *Economics* 12.*1*, p. 20180023.

Costaa, Carolina, Helena Alvelosa, and Leonor Teixeiraa (2015), « CENTERIS 2012- Conference on ENTERprise Information Systems The use of Moodle e-learning platform : a study in a Portuguese University », *in*: URL: `https://api.semanticscholar.org/CorpusID:110069063`.

Dawson, Shane et al. (2019), « Increasing the impact of learning analytics », *in*: *Proceedings of the 9th international conference on learning analytics & knowledge*, pp. 446–455.

Dawson, Shane P, Leah Macfadyen, and Lori Lockyer (2009), « Learning or performance: Predicting drivers of student motivation », *in*.

Demartini, Claudio Giovanni et al. (2024), « Artificial Intelligence Bringing Improvements to Adaptive Learning in Education: A Case Study », *in*: *Sustainability* 16.*3*, p. 1347.

Dewey, John (1933), « How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process », *in*: *Boston: DC HeathandCompany. pp*, pp. 9–107.

Dietterich, Thomas G (2000), « Ensemble methods in machine learning », *in*: *International workshop on multiple classifier systems*, Springer, pp. 1–15.

Duan, Lian and Wenjun Wang (2024), « Early detection of serious but struggling learners in MOOCs using ensemble deep learning », *in*: *Enterprise Information Systems* 18.*11*, p. 2415596.

Duch, Dynil, Madeth May, and Sébastien George (2024), « Empowering Students: A Reflective Learning Analytics Approach to Enhance Academic Performance », *in*: *16th International Conference on Computer Supported Education (CSEDU 2024)*, SCITEPRESS-Science and Technology Publications, pp. 385–396.

Economic Co-operation, Organisation for and Development (OECD) (2018), « Bridging the digital gender divide: Include, upskill, innovate », *in*: *OECD*.

Enughwure, Akpofure Avwerosuoghene and Mercy Ebitiminipre Ogbise (2020), « Application of machine learning methods to predict student performance: a systematic literature review », *in*: *Int. Res. J. Eng. Technol* 7.*05*, pp. 3405–3415.

Farooqi, A, U Khalid, and AM Khan (2022), « Understanding the Digital Divide in the Contemporary Digital World », *in*: *Global Political Review, VII*, pp. 7–14.

Fatma Gizem Karaoglan Yilmaz, Ramazan Yilmaz (2020), « Student Opinions About Personalized Recommendation and Feedback Based on Learning Analytics », *in.*

— (2022), « Learning Analytics Intervention Improves Students' Engagement in Online Learning », *in.*

Félix, Igor Moreira et al. (2018), « Data Mining for Student Outcome Prediction on Moodle: a systematic mapping », *in*: *Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*, URL: `https://api.semanticscholar.org/CorpusID:53612469`.

Feng, Guiyun and Muwei Fan (2024), « Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization », *in*: *Expert Systems with Applications* 237, p. 121555.

Flavell, John H (1979), « Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. », *in*: *American psychologist* 34.*10*, p. 906.

Freeman, Scott et al. (2014), « Active learning increases student performance in science, engineering, and mathematics », *in*: *Proceedings of the national academy of sciences* 111.*23*, pp. 8410–8415.

Freund, Yoav and Robert E Schapire (1997), « A decision-theoretic generalization of online learning and an application to boosting », *in*: *Journal of computer and system sciences* 55.*1*, pp. 119–139.

Friedman, Jerome H (2001), « Greedy function approximation: a gradient boosting machine », *in*: *Annals of statistics*, pp. 1189–1232.

Froehlich, Dominik E, Sara Van Waes, and Hannah Schäfer (2020), « Linking quantitative and qualitative network approaches: A review of mixed methods social network analysis in education research », *in*: *Review of research in education* 44.*1*, pp. 244–268.

Gardner, Joshua et al. (2023), « Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity », *in*: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1664–1684.

Gašević, Dragan, Shane Dawson, and George Siemens (2015), « Let's not forget: Learning analytics are about learning », *in*: *TechTrends* 59, pp. 64–71.

Gorbunova, Anna et al. (2024), « The Interplay of Self-Regulated Learning, Cognitive Load, and Performance in Learner-Controlled Environments », *in*: *Education Sciences* 14.*8*, p. 860.

Haleem, Abid et al. (2022), « Understanding the role of digital technologies in education: A review », *in*: *Sustainable operations and computers* 3, pp. 275–285.

Hattie, John (2009), « Visible learning. Hattie ranking: Influences and effect sizes related to student achievement », *in*: *Retrieved September* 28, p. 2019.

He, Haibo and Edwardo A Garcia (2009), « Learning from imbalanced data », *in*: *IEEE Transactions on knowledge and data engineering* 21.*9*, pp. 1263–1284.

Hegde, Vinayak, Anusha R Pai, and Ranjitha J Shastry (2022), « Personalized Formative Feedbacks and Recommendations Based on Learning Analytics to Enhance the Learning of Java Programming », *in*: *ICT Infrastructure and Computing: Proceedings of ICT4SD 2022*, Springer, pp. 655–666.

Hernández-de-Menéndez, Marcela et al. (2022), « Learning analytics: state of the art », *in*: *International Journal on Interactive Design and Manufacturing (IJIDeM)* 16.*3*, pp. 1209–1230.

Hirokawa, Sachio (2018), « Key attribute for predicting student academic performance », *in*: *Proceedings of the 10th International Conference on Education Technology and Computers*, pp. 308–313.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant (2013), *Applied logistic regression*, John Wiley & Sons.

Hu, Qian and Huzefa Rangwala (2020), « Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. », *in*: *International Educational Data Mining Society.*

Joksimović, Srećko, Vitomir Kovanović, and Shane Dawson (2019), « The journey of learning analytics », *in*: *HERDSA Review of Higher Education* 6, pp. 27–63.

Jones, Hazel (2016), « Ethical considerations in the use of student data: International perspectives and educators' perceptions », *in*: *ASCILITE Publications*, pp. 300–304.

Juhaňák, Libor, Jiří Zounek, and Lucie Rohlíková (2019), « Using process mining to analyze students' quiz-taking behavior patterns in a learning management system », *in*: *Computers in Human Behavior* 92, pp. 496–506.

Kamal, Preet and Sachin Ahuja (2019), « An ensemble-based model for prediction of academic performance of students in undergrad professional course », *in*: *Journal of Engineering, Design and Technology* 17.*4*, pp. 769–781.

Karaoglan Yilmaz, Fatma Gizem (2022), « The effect of learning analytics assisted recommendations and guidance feedback on students' metacognitive awareness and academic achievements », *in*: *Journal of Computing in Higher Education* 34.*2*, pp. 396–415.

Kernbach, Julius M and Victor E Staartjes (2022), « Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and overfitting », *in*: *Machine Learning in Clinical Neuroscience: Foundations and Applications*, pp. 15–21.

Kerras, Hayet et al. (2020), « The impact of the gender digital divide on sustainable development: comparative analysis between the European Union and the Maghreb », *in*: *Sustainability* 12.*8*, p. 3347.

Khademi, Aria and Vasant Honavar (2020), « Algorithmic bias in recidivism prediction: A causal perspective (student abstract) », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 10, pp. 13839–13840.

Khan, Majid et al. (2023), « Utilizing machine learning models to predict student performance from LMS activity logs », *in*: *IEEE Access*.

Koedinger, Kenneth R et al. (2013), « New potentials for data-driven intelligent tutoring system development and optimization », *in*: *AI Magazine* 34.*3*, pp. 27–41.

Kolb, David A (2014), *Experiential learning: Experience as the source of learning and development*, FT press.

Kotsiantis, Sotiris and Dimitris Kanellopoulos (2012), « Combining bagging, boosting and random subspace ensembles for regression problems », *in*: *International Journal of Innovative Computing, Information and Control* 8.*6*, pp. 3953–3961.

Krawczyk, Bartosz (2016), « Learning from imbalanced data: open challenges and future directions », *in*: *Progress in artificial intelligence* 5.*4*, pp. 221–232.

Kreber, Carolin (2012), « Critical reflection and transformative learning », *in*: *The handbook of transformative learning: Theory, research, and practice*, pp. 323–341.

Kwak, Chanyeong and Alan Clayton-Matthews (2002), « Multinomial logistic regression », *in*: *Nursing research* 51.*6*, pp. 404–410.

Larsson, Anthony and Yamit Viitaoja (2019), « Identifying the digital gender divide: How digitalization may affect the future working conditions for women », *in*: *The digital transformation of labor*, Routledge, pp. 235–253.

Le Quy, Tai (2024), « Fairness-aware Machine Learning in Educational Data Mining », *in*.

Li, Warren et al. (2022), « Disparities in students' propensity to consent to learning analytics », *in*: *International Journal of Artificial Intelligence in Education* 32.*3*, pp. 564–608.

Liaw, A (2002), « Classification and regression by randomForest », *in*: *R news*.

Liñán, Laura Calvet and Ángel Alejandro Juan Pérez (2015), « Educational Data Mining and Learning Analytics: differences, similarities, and time evolution », *in*: *RUSC. Universities and Knowledge Society Journal* 12.*3*, pp. 98–112.

Lukwaro, Elia Ahidi Elisante, Khamisi Kalegele, and Devotha G Nyambo (2024), « A Review on NLP Techniques and Associated Challenges in Extracting Features from Education Data », *in*: *Int. J. Com. Dig. Sys* 16.*1*.

Macfadyen, Leah P and Shane Dawson (2010), « Mining LMS data to develop an "early warning system" for educators: A proof of concept », *in*: *Computers & education* 54.*2*, pp. 588–599.

Madiah, Haziman and Rosmayati Mohemad (2023), « A review of learning management systems (LMS) framework towards the element of outcome based education (OBE) », *in*: *AIP Conference Proceedings*, vol. 2484, 1, AIP Publishing.

Manu, Guleria Pratiyushand Sood (2016), « Classifying educational data using support vector machines: A supervised data mining technique », *in*: *Indian Journal of Science and Technology* 9.*34*.

Mariscal, Judith et al. (2019), « Bridging the gender digital gap », *in*: *Economics* 13.*1*, p. 20190009.

Márquez-Vera, Carlos et al. (2016), « Early dropout prediction using data mining: a case study with high school students », *in*: *Expert Systems* 33.*1*, pp. 107–124.

Marzano, Gilberto and Velta Lubkina (2019), « The Digital Gender Divide: An Overview », *in*: *SOCIETY. INTEGRATION. EDUCATION. Proceedings of the International Scientific Conference*, vol. 5, pp. 413–421.

Masrom, Suraya et al. (2024), « Machine learning prediction for academic misconduct prediction: an analysis of binary classification metrics », *in*: *Bulletin of Electrical Engineering and Informatics* 13.*1*, pp. 388–395.

McMillan, James H and Jessica Hearn (2008), « Student self-assessment: The key to stronger student motivation and higher achievement », *in*: *Educational horizons* 87.*1*, pp. 40–49.

Mengash, Hanan Abdullah (2020), « Using data mining techniques to predict student performance to support decision making in university admission systems », *in*: *Ieee Access* 8, pp. 55462–55470.

Merceron, Agathe and Ange Tato (2023), « Introduction to Neural Networks and Uses in EDM. », *in*: *EDM.*

Moore, Raeal, Dan Vitale, and Nycole Stawinoga (2018), « The Digital Divide and Educational Equity: A Look at Students with Very Limited Access to Electronic Devices at Home. Insights in Education and Work. », *in*: *ACT, Inc.*

Mumporeze, Nadine and Michael Prieler (2017), « Gender digital divide in Rwanda: A qualitative analysis of socioeconomic factors », *in*: *Telematics and Informatics* 34.*7*, pp. 1285–1293.

Nafea, Ahmed Adil et al. (2023), « Enhancing Student's Performance Classification Using Ensemble Modeling », *in*: *Iraqi Journal For Computer Science and Mathematics* 4.*4*, pp. 204–214.

Nakayama, Luis Filipe et al. (2023), « The Digital Divide in Brazil and Barriers to Telehealth and Equal Digital Health Care: Analysis of Internet Access Using Publicly Available Data », *in*: *Journal of Medical Internet Research* 25, e42483.

Namoun, Abdallah and Abdullah Alshanqiti (2020), « Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review », *in*: *Applied Sciences*, URL: `https://api.semanticscholar.org/CorpusID:234387676`.

Ndukwe, Ifeanyi Glory and Ben Kei Daniel (2020), « Teaching analytics, value and tools for teacher data literacy: A systematic and tripartite approach », *in*: *International Journal of Educational Technology in Higher Education* 17.*1*, pp. 1–31.

Nguyen, NT, Larry C Allen, and Katherine Fraccastoro (2005), « Personality predicts academic performance: Exploring the moderating role of gender », *in*: *Journal of Higher Education Policy and Management* 27.*1*, pp. 105–117.

Niyogisubizo, Jovial et al. (2022), « Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization », *in*: *Computers and Education: Artificial Intelligence* 3, p. 100066.

Okewu, Emmanuel et al. (2021), « Artificial neural networks for educational data mining in higher education: A systematic literature review », *in*: *Applied Artificial Intelligence* 35.*13*, pp. 983–1021.

Okike, Ezekiel U and Merapelo Mogorosi (2020), « Educational data mining for monitoring and improving academic performance at university levels », *in*: *International Journal of Advanced Computer Science and Applications* 11.*11*.

Öz, Eda and H Şenay Şen (2021), « THE EFFECT OF SELF-REGULATED LEARNING ON STUDENTS'LIFELONG LEARNING AND CRITICAL THINKING TENDENCIES », *in*: *Elektronik Sosyal Bilimler Dergisi* 20.*78*, pp. 934–960.

Palacios, Carlos A et al. (2021), « Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile », *in*: *Entropy* 23.*4*, p. 485.

Pan, Sinno Jialin and Qiang Yang (2009), « A survey on transfer learning », *in*: *IEEE Transactions on knowledge and data engineering* 22.*10*, pp. 1345–1359.

Pan, Zilong et al. (2024), « A Systematic Review of Learning Analytics: Incorporated Instructional Interventions on Learning Management Systems », *in*: *Journal of Learning Analytics*, pp. 1–21.

Papamitsiou, Zacharoula and Anastasios A Economides (2014), « Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence », *in*: *Journal of Educational Technology & Society* 17.*4*, pp. 49–64.

Patidar, Preeti, Jitendra Dangra, and MK Rawar (2015), « Decision tree C4. 5 algorithm and its enhanced approach for educational data mining », *in*: *Engineering Universe for Scientific Research and Management* 7.*2*, pp. 1–14.

Pérez-Castro, Miguel Ángel, Miriem Mohamed-Maslouhi, and Miguel Ángel Montero-Alonso (2021), « The digital divide and its impact on the development of Mediterranean countries », *in*: *Technology in Society* 64, p. 101452.

Piaget, J (1952), « The origins of intelligence in children », *in*: *International University.*

Pong-inwong, Chakrit and Wararat Rungworawut (2012), « Teaching evaluation using data mining on moodle LMS forum », *in*: *2012 6th international conference on new trends in information science, service science and data mining (ISSDM2012)*, IEEE, pp. 550–555.

Porras, José Manuel et al. (2023), « A Case-Study Comparison of Machine Learning Approaches for Predicting Student's Dropout from Multiple Online Educational Entities », *in*: *Algorithms* 16.*12*, p. 554.

Purwaningsih, Jamila Nunik and Yuli Suwarno (2016), « Predicting students achievement based on motivation in vocational school using data mining approach », *in*: *2016 4th International Conference on Information and Communication Technology (ICoICT)*, IEEE, pp. 1–5.

QAZDAR, Aimad et al. (2022), « Learning Analytics for Tracking Student Progress in LMS », *in*.

Quinn, Rory Joseph and Geraldine Gray (2019), « Prediction of student academic performance using Moodle data from a Further Education setting. », *in*: *Irish Journal of Technology Enhanced Learning*, URL: `https://api.semanticscholar.org/CorpusID:209070178`.

Reich, Justin (2020), *Failure to disrupt: Why technology alone can't transform education*, Harvard University Press.

Romero, Cristobal and Sebastian Ventura (2020), « Educational data mining and learning analytics: An updated survey », *in*: *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 10.*3*, e1355.

Rosenthal, Robert and Lenore Jacobson (1968), « Pygmalion in the classroom », *in*: *The urban review* 3.*1*, pp. 16–20.

Santos, Rita, José Azevedo, and Luís Pedro (2013), « Digital divide in higher education students' digital literacy », *in*: *Worldwide Commonalities and Challenges in Information Literacy Research and Practice: European Conference on Information Literacy, ECIL 2013 Istanbul, Turkey, October 22-25, 2013 Revised Selected Papers 1*, Springer, pp. 178–183.

Sarker, Sazol et al. (2024), « Analyzing students' academic performance using educational data mining », *in*: *Computers and Education: Artificial Intelligence* 7, p. 100263.

Scholz, Felix, Thomas Elmar Kolb, and Julia Neidhardt (2024), « Classifying User Roles in Online News Forums: A Model for User Interaction and Behavior Analysis », *in*: *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 240–249.

Schon, Donald A (2008), *The reflective practitioner: How professionals think in action*, Basic books.

Segura-Morales, Marco and Edison Loza-Aguirre (2017), « Using decision trees for predicting academic performance based on socio-economic factors », *in*: *2017 International*

*Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, pp. 1132–1136.

Al-Shabandar, Raghad et al. (2019), « Detecting at-risk students with early interventions using machine learning techniques », *in*: *IEEE Access* 7, pp. 149464–149478.

Siemens, George and Phil Long (2011), « Penetrating the fog: Analytics in learning and education. », *in*: *EDUCAUSE review* 46.*5*, p. 30.

Silvola, Anni et al. (2021), « Expectations for supporting student engagement with learning analytics: An academic path perspective », *in*: *Computers & Education* 168, p. 104192.

Smirani, Lassaad K et al. (2022), « Using ensemble learning algorithms to predict student failure and enabling customized educational paths », *in*: *Scientific Programming* 2022.*1*, p. 3805235.

Smith, Stephen, David Cobham, and Kevin Jacques (2022), « The use of data mining and automated social networking tools in virtual learning environments to improve student engagement in higher education », *in*: *International Journal of Information and Education Technology* 12.*4*, pp. 263–271.

Susanto, Heru, Ching Kang Chen, and Mohammed Nabil Almunawar (2018), « Revealing big data emerging technology as enabler of LMS technologies transferability », *in*: *Internet of things and big data analytics toward next-generation intelligence*, pp. 123–145.

Trindade, Fernando Ribeiro and Deller James Ferreira (2021), « Student Performance Prediction Based on a Framework of Teacher's Features », *in*: *International Journal for Innovation Education and Research*, URL: `https://api.semanticscholar.org/CorpusID:234030418`.

Umer, Muhammad et al. (2023), « Analysing behavioural and academic attributes of students using educational data mining », *in*: *International Journal of Nanotechnology* 20.*5-10*, pp. 451–476.

Valverde-Berrocoso, Jesús et al. (2021), « The educational integration of digital technologies preCovid-19: Lessons for teacher education », *in*: *PloS one* 16.*8*, e0256283.

Vassilakopoulou, Polyxeni and Eli Hustad (2023), « Bridging digital divides: A literature review and research agenda for information systems research », *in*: *Information Systems Frontiers* 25.*3*, pp. 955–969.

Viberg, Olga et al. (2018), « The current landscape of learning analytics in higher education », *in*: *Computers in human behavior* 89, pp. 98–110.

Wang, Yinkai et al. (2021), « Graph-based Ensemble Machine Learning for Student Performance Prediction », *in*: *arXiv preprint arXiv:2112.07893*.

Warschauer, Mark (2004), *Technology and social inclusion: Rethinking the digital divide.*

West, Darrell M (2015), « Digital divide: Improving Internet access in the developing world through affordable services and diverse content », *in*: *Center for Technology Innovation at Brookings*, pp. 1–30.

Wolpert, David H (1992), « Stacked generalization », *in*: *Neural networks* 5.*2*, pp. 241–259.

Wong, Billy Tak-ming and Kam Cheong Li (2020), « A review of learning analytics intervention in higher education (2011–2018) », *in*: *Journal of Computers in Education* 7.*1*, pp. 7–28.

Wong, Billy Tak-Ming, Kam Cheong Li, and Samuel Ping-Man Choi (2018), « Trends in learning analytics practices: A review of higher education institutions », *in*: *Interactive Technology and Smart Education* 15.*2*, pp. 132–154.

Wongvorachan, Tarid, Surina He, and Okan Bulut (2023), « A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining », *in*: *Information* 14.*1*, p. 54.

Wu, Weijun (2025), « Comparing the Effectiveness of Ensemble Models and Single Algorithms in their Success in Predicting At-Risk Students », B.S. thesis, University of Twente.

Yağcı, Mustafa (2022a), « Educational data mining: prediction of students' academic performance using machine learning algorithms », *in*: *Smart Learning Environments* 9.*1*, p. 11.

— (2022b), « Educational data mining: prediction of students' academic performance using machine learning algorithms », *in*: *Smart Learning Environments* 9, URL: `https://api.semanticscholar.org/CorpusID:247233325`.

Young, Nicholas T and Marcos D Caballero (2021), « Predictive and explanatory models might miss informative features in educational data », *in*: *arXiv preprint arXiv:2103.14513*.

Yu, Rebecca P et al. (2016), « Mapping the two levels of digital divide: Internet access and social network site adoption among older adults in the USA », *in*: *Information, Communication & Society* 19.*10*, pp. 1445–1464.

Yürüm, Ozan Raşit, Tuğba Taşkaya-Temizel, and Soner Yıldırım (2023), « The use of video clickstream data to predict university students' test performance: A comprehensive educational data mining approach », *in*: *Education and Information Technologies* 28.*5*, pp. 5209–5240.

Zhang, Niu and Charles NR Henderson (2019), « Predicting stress and test anxiety among 1st-year chiropractic students », *in*: *Journal of Chiropractic Education* 33.*2*, pp. 133–139.

Zhen, Yang and Xiaoyan Zhu (2024), « An ensemble learning approach based on TabNet and machine learning models for cheating detection in educational tests », *in*: *Educational and Psychological Measurement* 84.*4*, pp. 780–809.

Zilka, Gila Cohen et al. (2021), « Implications of the digital divide for the learning process during the COVID-19 crisis », *in*: *Rev. Eur. Stud.* 13, p. 57.

Zimmerman, Barry J (2002), « Becoming a self-regulated learner: An overview », *in*: *Theory into practice* 41.*2*, pp. 64–70.

Appendix A

# Academic Grading System: Percentage to Letter Grade Conversion

Table A.1 – Grade Conversion Table with GPA, Descriptor, and Encoded Labels

| **Letter Grade** | **Percentage Range** | **GPA Equivalent** | **Descriptor** | **Encoded Label** |
|---|---|---|---|---|
| A | 85% – 100% | 4.00 | Excellent | 0 |
| B+ | 80% – 84% | 3.50 | Very Good | 1 |
| B | 70% – 79% | 3.00 | Good | 2 |
| C+ | 65% – 69% | 2.50 | Fairly Good | 3 |
| C | 50% – 64% | 2.00 | Fair | 4 |
| D | 45% – 49% | 1.50 | Poor | 5 |
| E | 40% – 44% | 1.00 | Very Poor | 6 |
| F | $<$ 40% | 0.00 | Failure | 7 |

Appendix B

# Complete List of Feedback Scenarios in ReflectMate

Table B.1: Comprehensive Feedback Scenarios in ReflectMate

| **Behavior** | **Feedback Provided** |
|---|---|
| Low Quiz Attempts | "Consider taking quizzes more frequently to test your knowledge and reinforce learning." |
| High Module Engagement with Low Quiz Scores | "You have spent a lot of time on Module 3 but scored below the average in quizzes. Try reviewing key concepts or discussing with a peer." |
| Infrequent LMS Access | "Your login frequency is below the average. Try to access the LMS regularly to stay up-to-date with new course materials." |
| Low Participation in Discussion Forums | "Engaging in discussions can enhance your understanding. Try contributing to forum threads related to challenging topics." |
| Incomplete Video Lecture Viewing | "You haven't completed watching Lecture 5. Finish it to ensure you don't miss important concepts." |

| | |
|---|---|
| Poor Performance on Specific Question Types | "You scored lower on multiple-choice questions. Review these question types to improve your test-taking strategies." |
| Irregular Attendance in Synchronous Sessions | "Attending live sessions improves engagement. Make it a priority to join scheduled classes." |
| Underutilization of Supplementary Resources | "Supplementary materials like practice exercises can help reinforce learning. Explore these resources today." |
| Consistent Late Submissions | "Submitting assignments late may impact your grades. Aim to complete tasks ahead of deadlines to avoid penalties." |
| Lack of Interaction with Peer Feedback | "Providing and receiving peer feedback enhances collaborative learning. Engage more actively in peer review activities." |
| Over-reliance on Instructor Support | "While seeking help is encouraged, try solving problems independently first. This will build your problem-solving skills." |
| Repetitive Mistakes in Assignments | "You seem to make similar errors across assignments. Focus on understanding the underlying concepts to avoid repeating these mistakes." |
| Skipping Optional Learning Activities | "Optional activities are designed to deepen your understanding. Consider completing them for extra practice." |
| Limited Use of Progress Tracking Tools | "Track your progress using the available tools to identify areas where you need improvement." |
| High Dropout Rate from Courses | "It seems you've stopped participating in some courses. Revisit your goals and consider restarting or focusing on one course at a time." |
| Neglecting Self-Assessment Opportunities | "Self-assessments help gauge your understanding. Take advantage of these opportunities to reflect on your learning." |
| Failing to Set Personal Learning Goals | "Setting clear learning goals can guide your efforts. Define what you want to achieve in each module and track your progress." |

| | |
|---|---|
| Excessive Time Spent on Easy Modules | "You’re spending too much time on easier modules. Allocate more time to challenging topics to maximize learning efficiency." |
| Avoiding Challenging Topics | "Avoiding difficult topics may hinder your overall progress. Tackle these areas systematically to strengthen your knowledge." |

Appendix C

# Survey Questions

This appendix contains the survey questions used to evaluate the ReflectMate tool.

— Q1: Please share your thoughts on the dashboard experience by rating the following points. (Accessibility, Data Indicators, Visualization, and Recommendation Tool)

— Q2: How do you rate the following usefulness indicators data in affecting your learning performance? (Attendance Data, Number of Interaction Logs to LMS, and Quizzes, Assignments, Tasks Submitted)

— Q3: How interested are you in utilizing the Student Performance Prediction Algorithm to enhance the Reflective Tools for the following aspects of your learning experience? (Self-awareness, Self-regulation, Self-evaluation, Self-motivation, Interaction with Course Materials and Peers, and Academic Production and Outcome)

— Q4: Do you have any additional comments or suggestions?

Appendix D

# Additional Statistical Analysis

This appendix includes additional statistical data in Chapter 4 and statistical data that supports the findings discussed in Chapter 6.

Table D.1 – Application and Success Rates by Major and Gender

| Major | Applications (rate) | | Success (rate) | |
|---|---|---|---|---|
| | Men | Women | Men | Women |
| Computing and Technology | 3639 (70.55%) | 1519 (29.45%) | 2694 (74.03%) | 1027 (67.61%) |
| Telecommunications and Networking | 607 (61.69%) | 377 (38.31%) | 405 (66.72%) | 215 (57.03%) |
| Digital Management and Business | 591 (42.67%) | 794 (57.33%) | 426 (72.08%) | 442 (55.67%) |

Table D.2 – Application and Success Rates by Geography Location

| Geography Location | Applications (rate) | | Success (rate) | |
|---|---|---|---|---|
| | Number | Percentage | Number | Percentage |
| High development | 1180 | 32.33% | 854 | 72.37% |
| Medium-high development | 1507 | 41.29% | 1027 | 68.15% |
| Medium-low development | 850 | 23.29% | 565 | 66.47% |
| Low development | 113 | 3.10% | 70 | 61.95% |

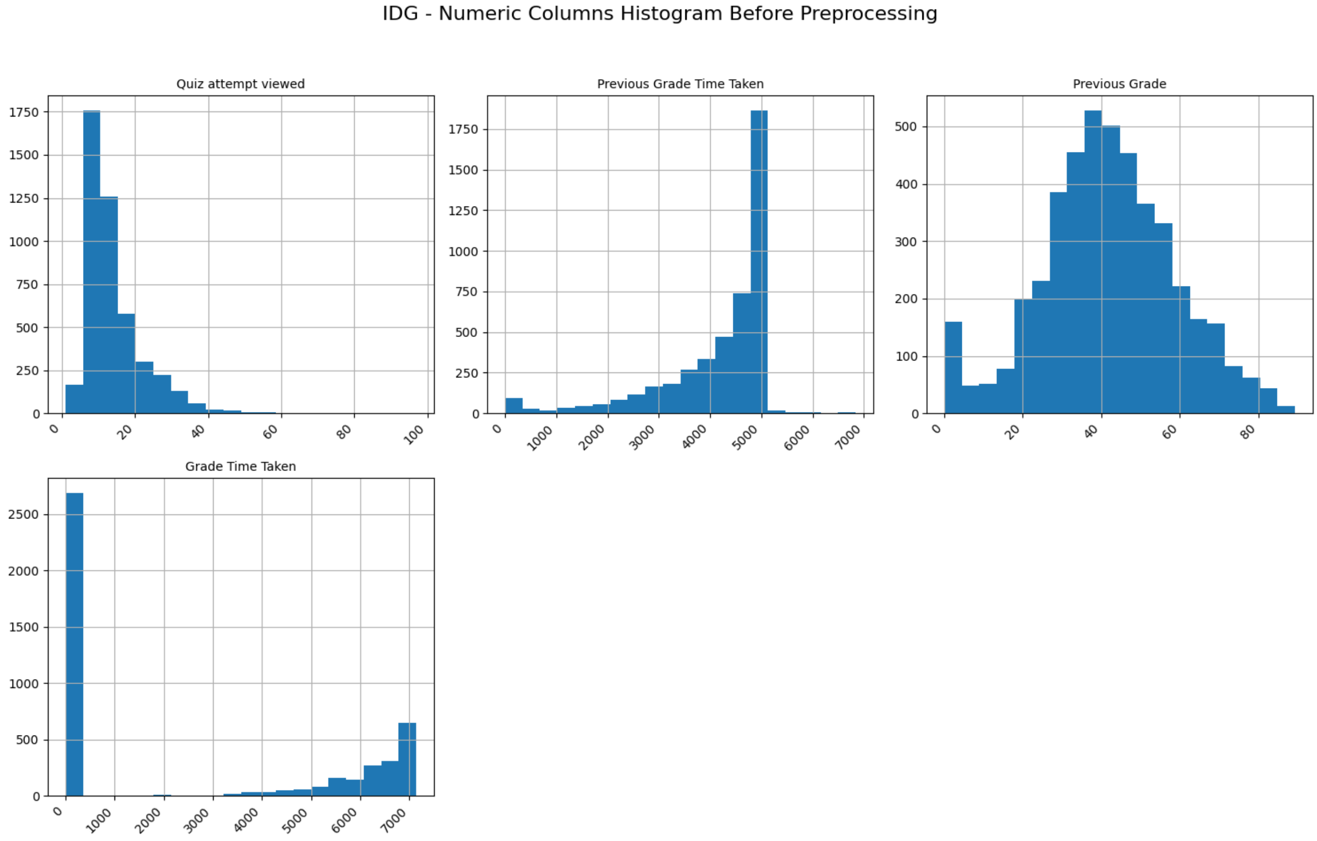


Figure D.1 – IDG - Numeric columns histogram before preprocessing

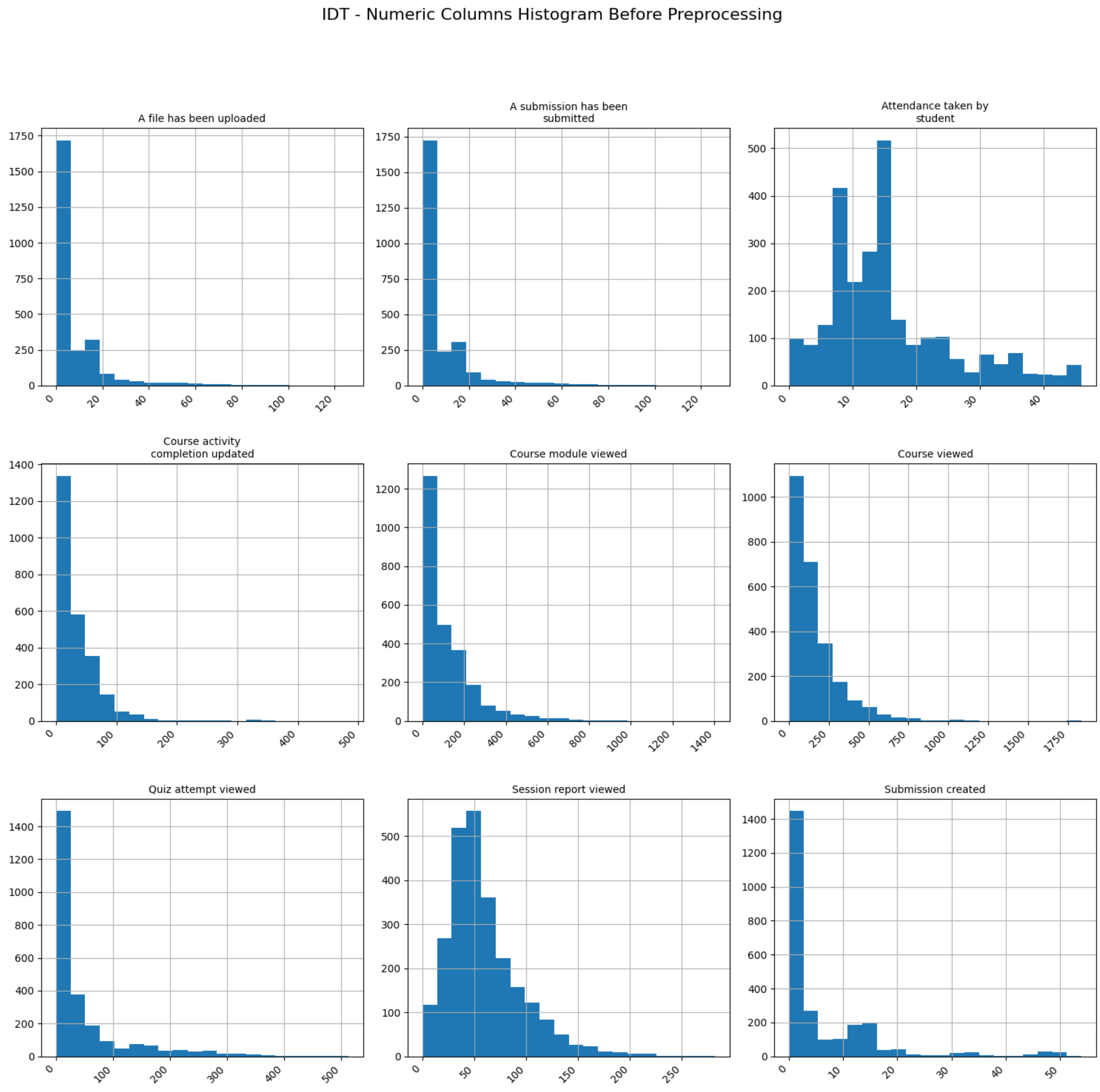


Figure D.2 – IDT - Numeric columns histogram before preprocessing (Part 1)

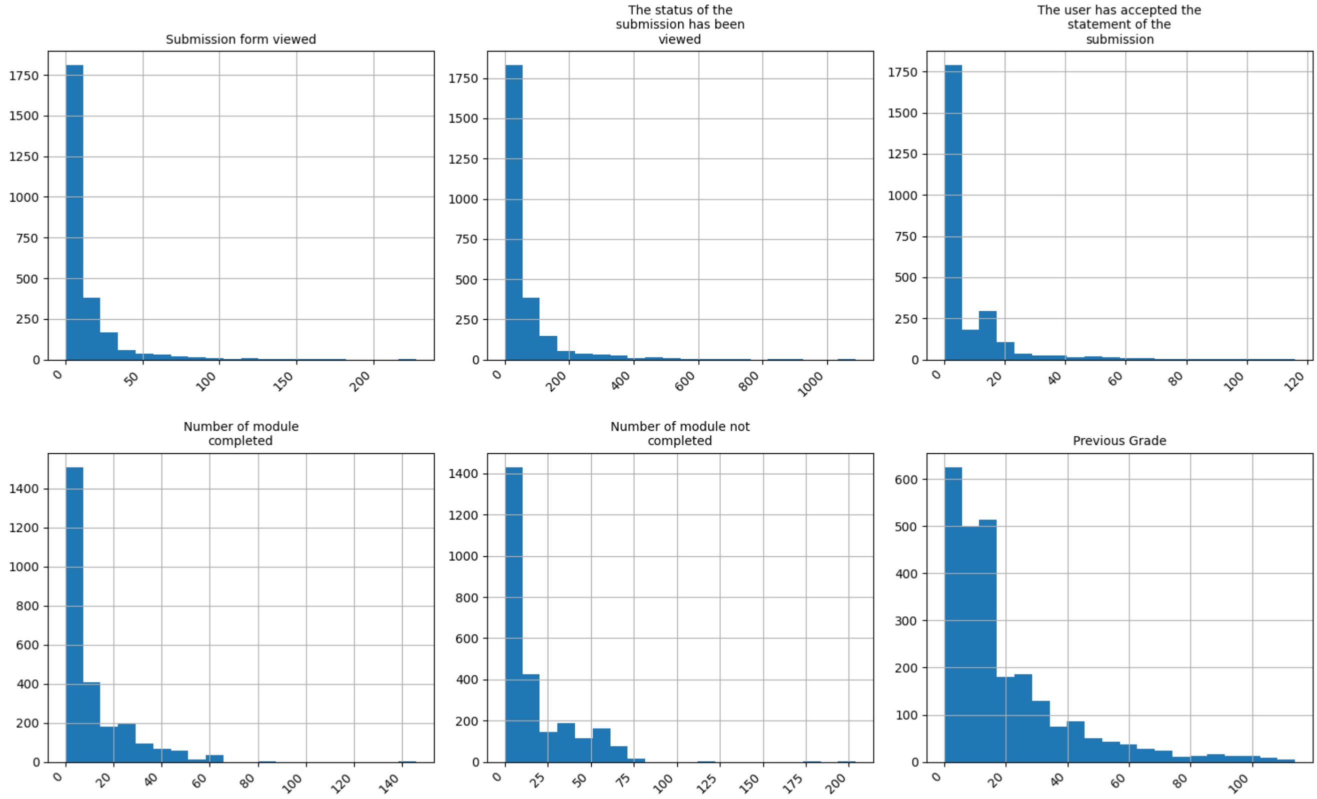


Figure D.3 – IDT - Numeric columns histogram before preprocessing (Part 2)

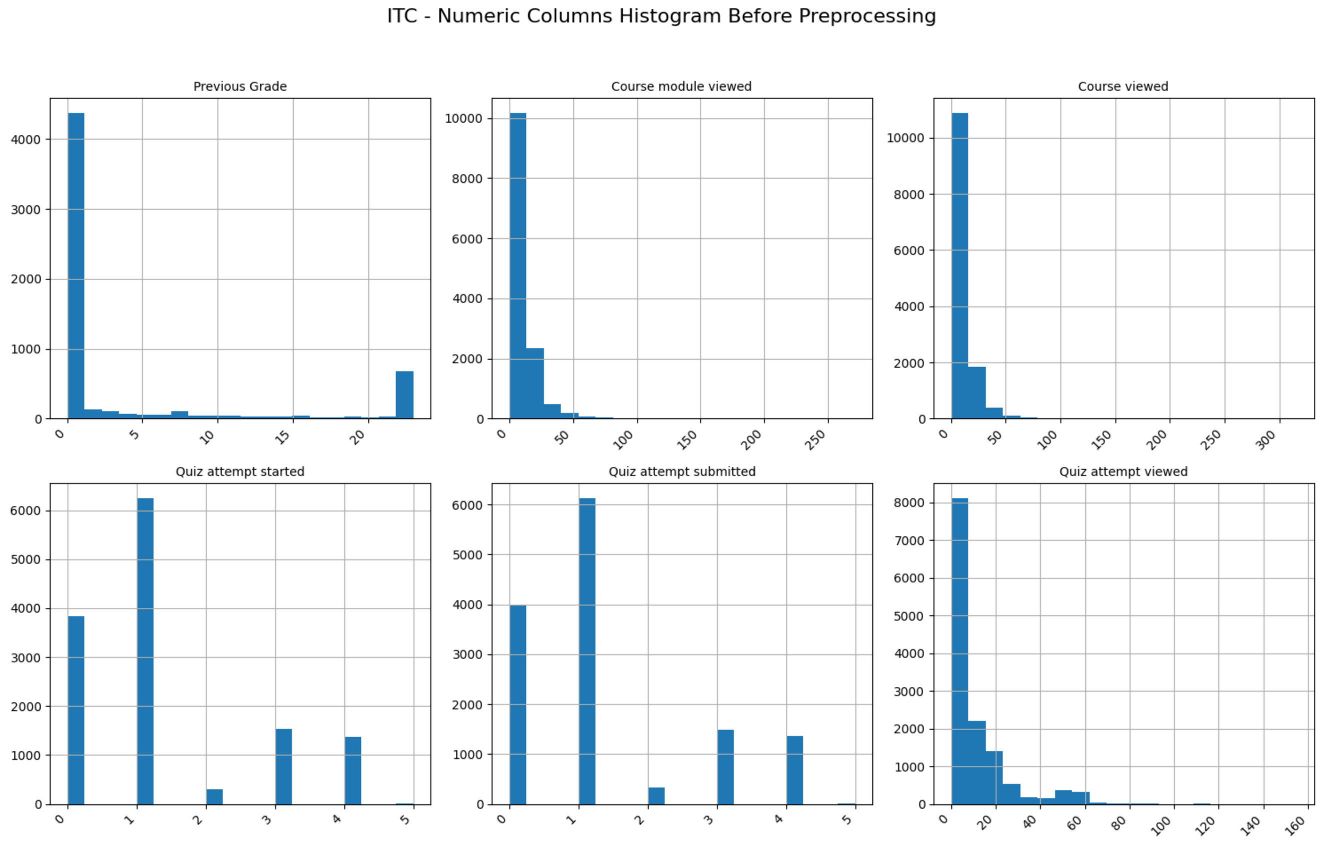


Figure D.4 – ITC - Numeric columns histogram before preprocessing

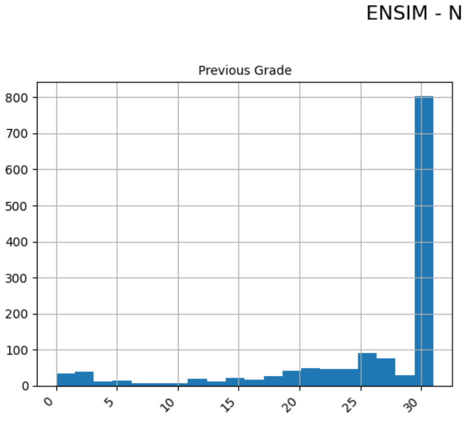


Figure D.5 – ENSIM - Numeric columns histogram before preprocessing

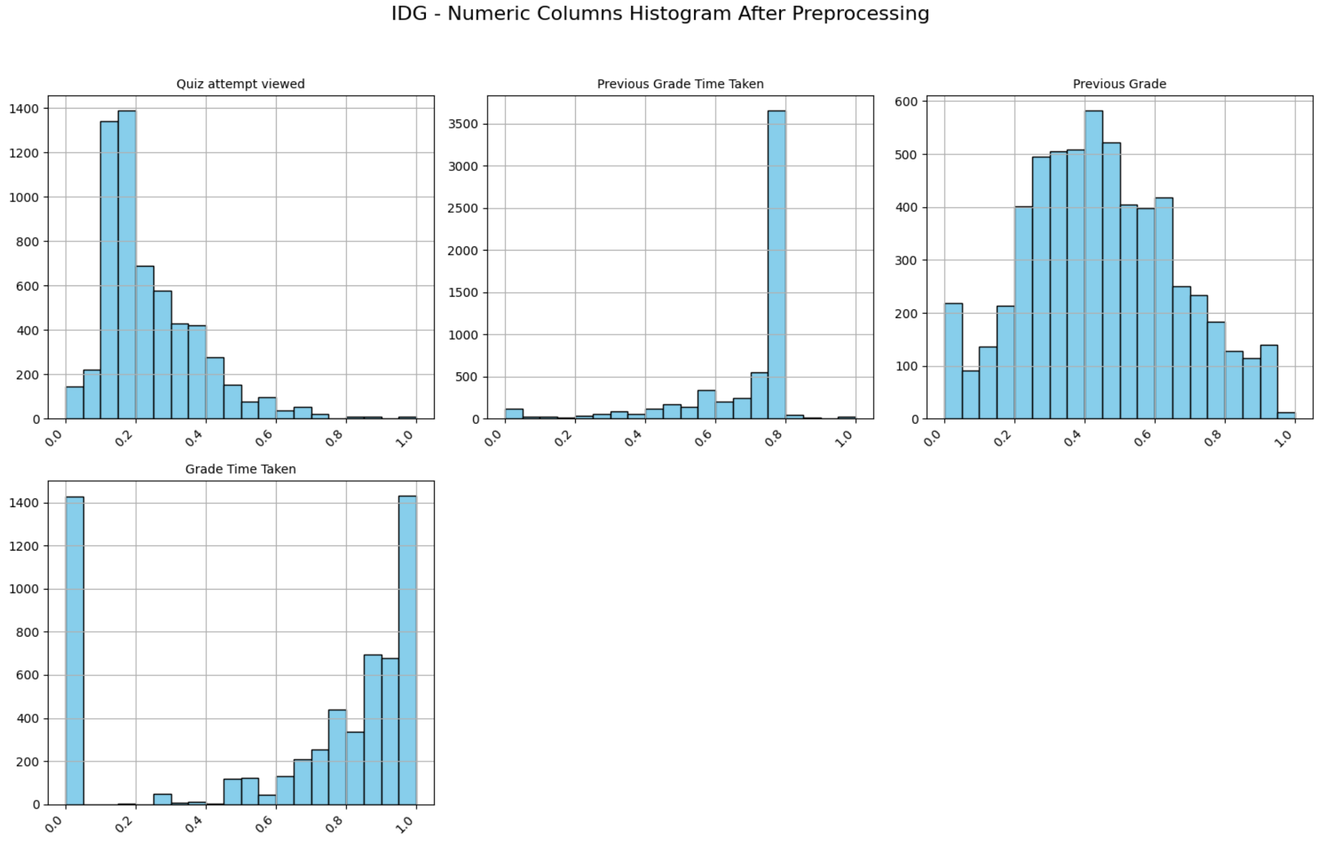


Figure D.6 – IDG - Numeric columns histogram after preprocessing

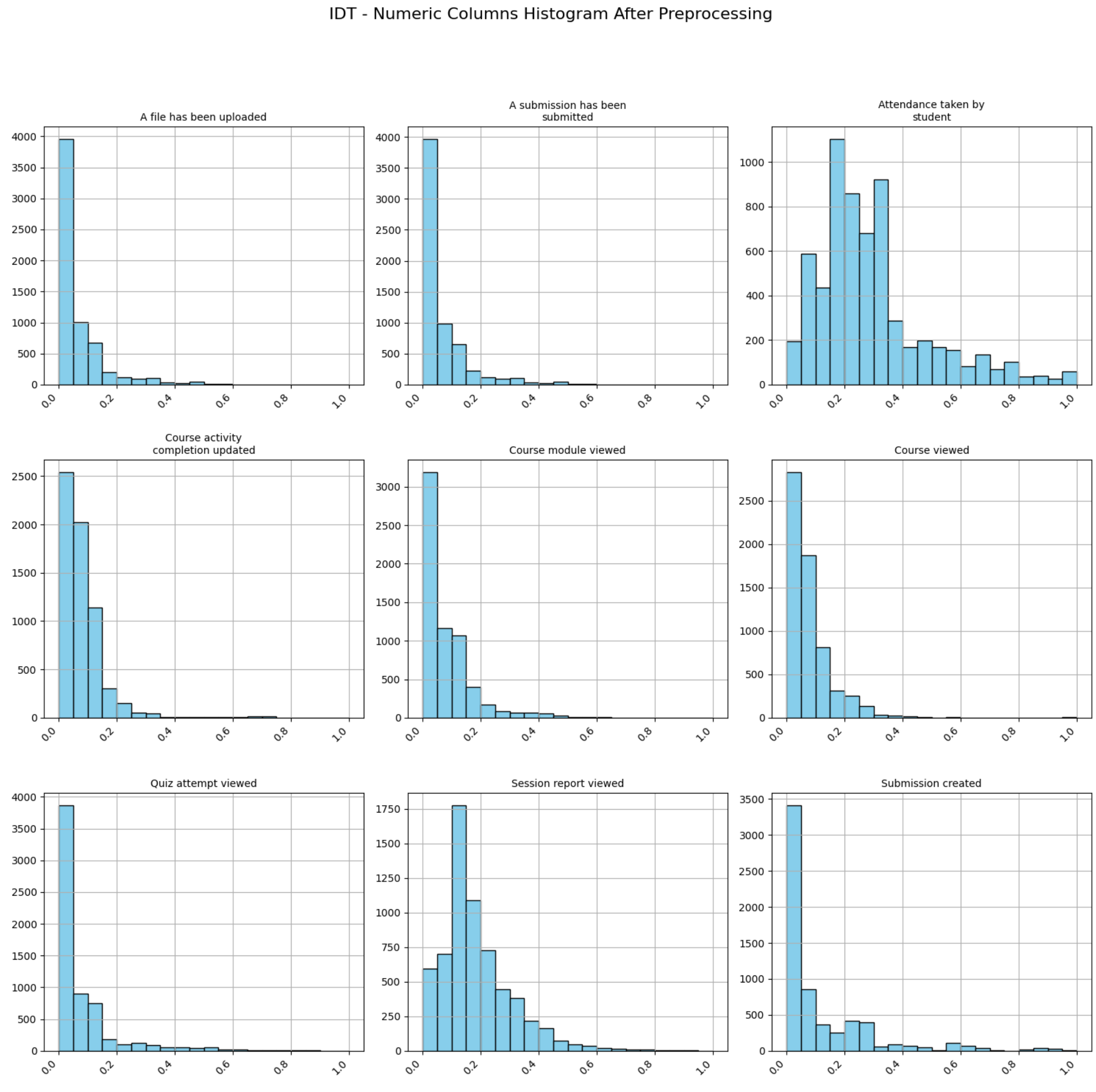


Figure D.7 – IDT - Numeric columns histogram after preprocessing (Part 1)

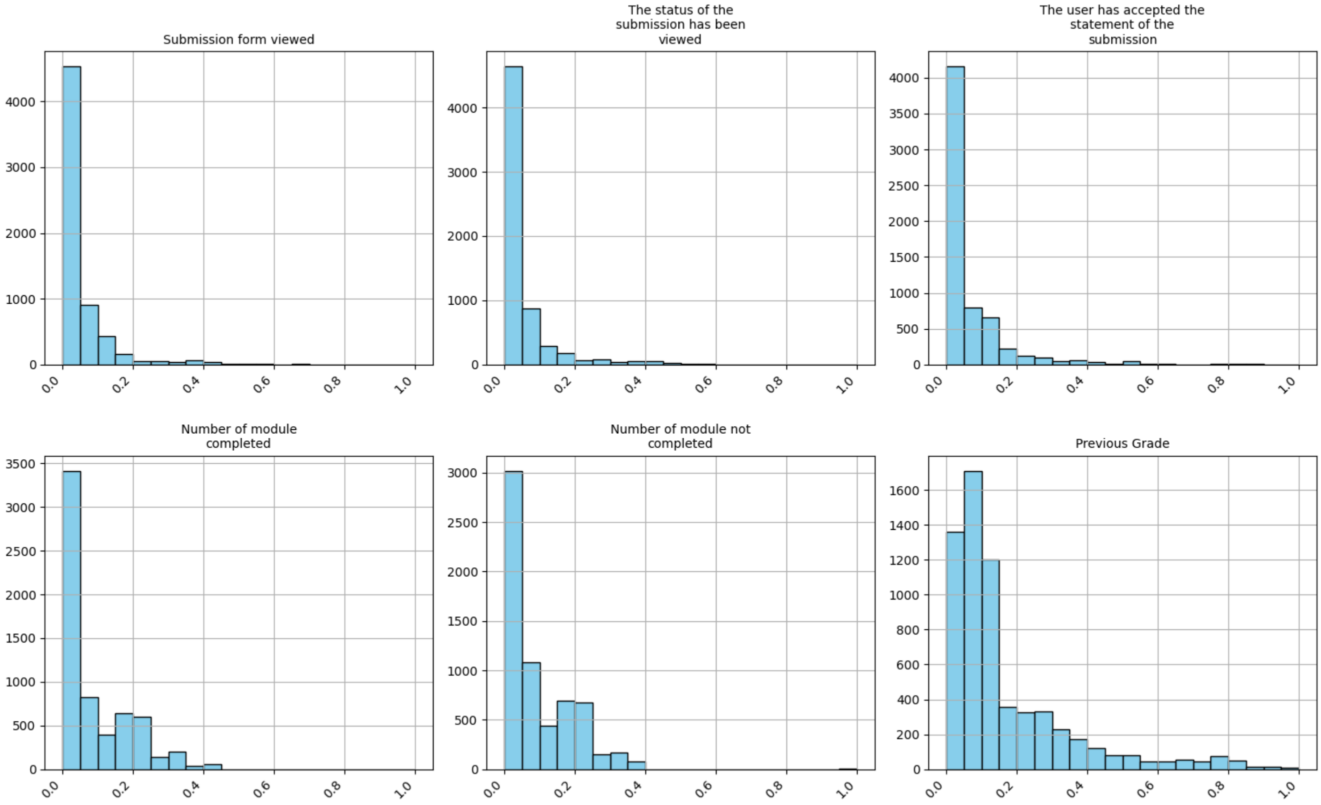


Figure D.8 – IDT - Numeric columns histogram after preprocessing (Part 2)

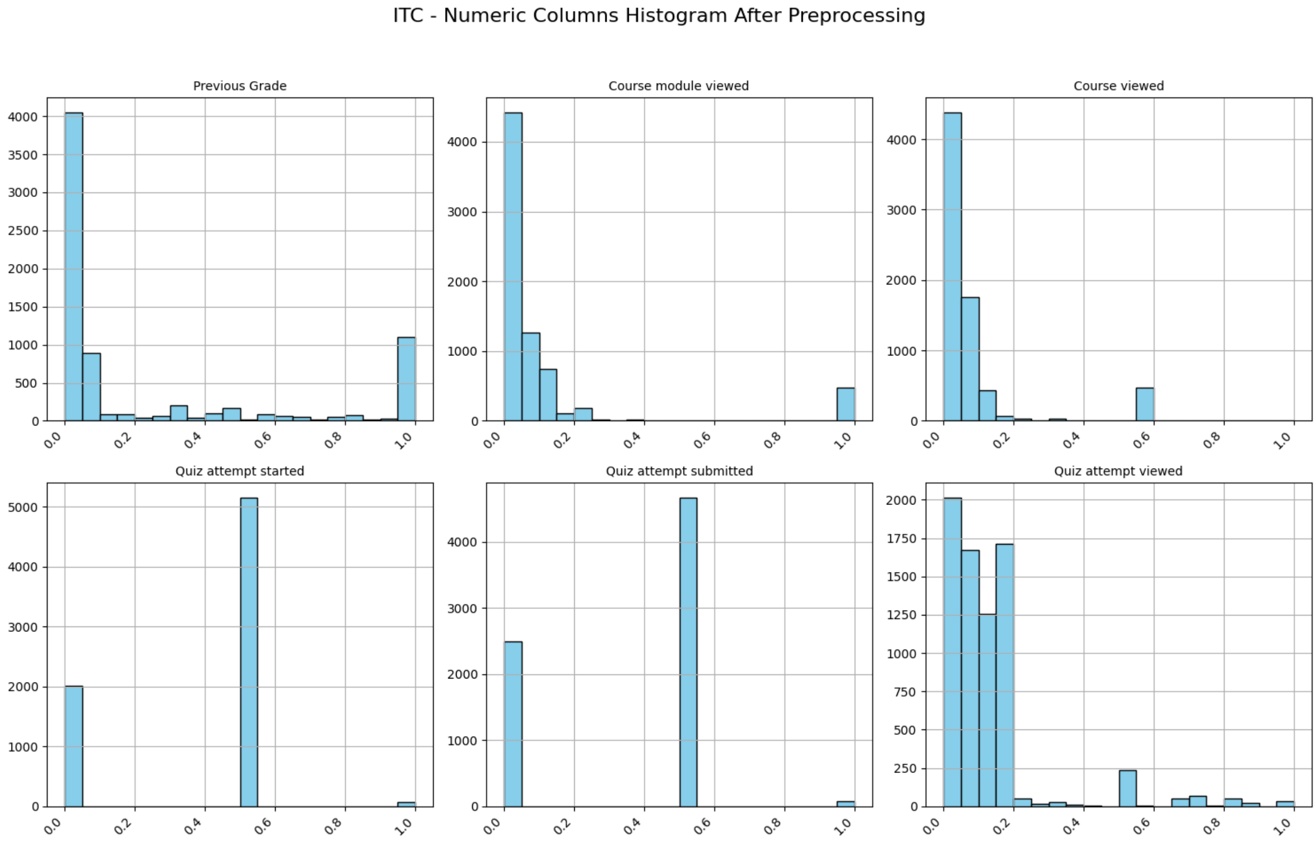


Figure D.9 – ITC - Numeric columns histogram after preprocessing

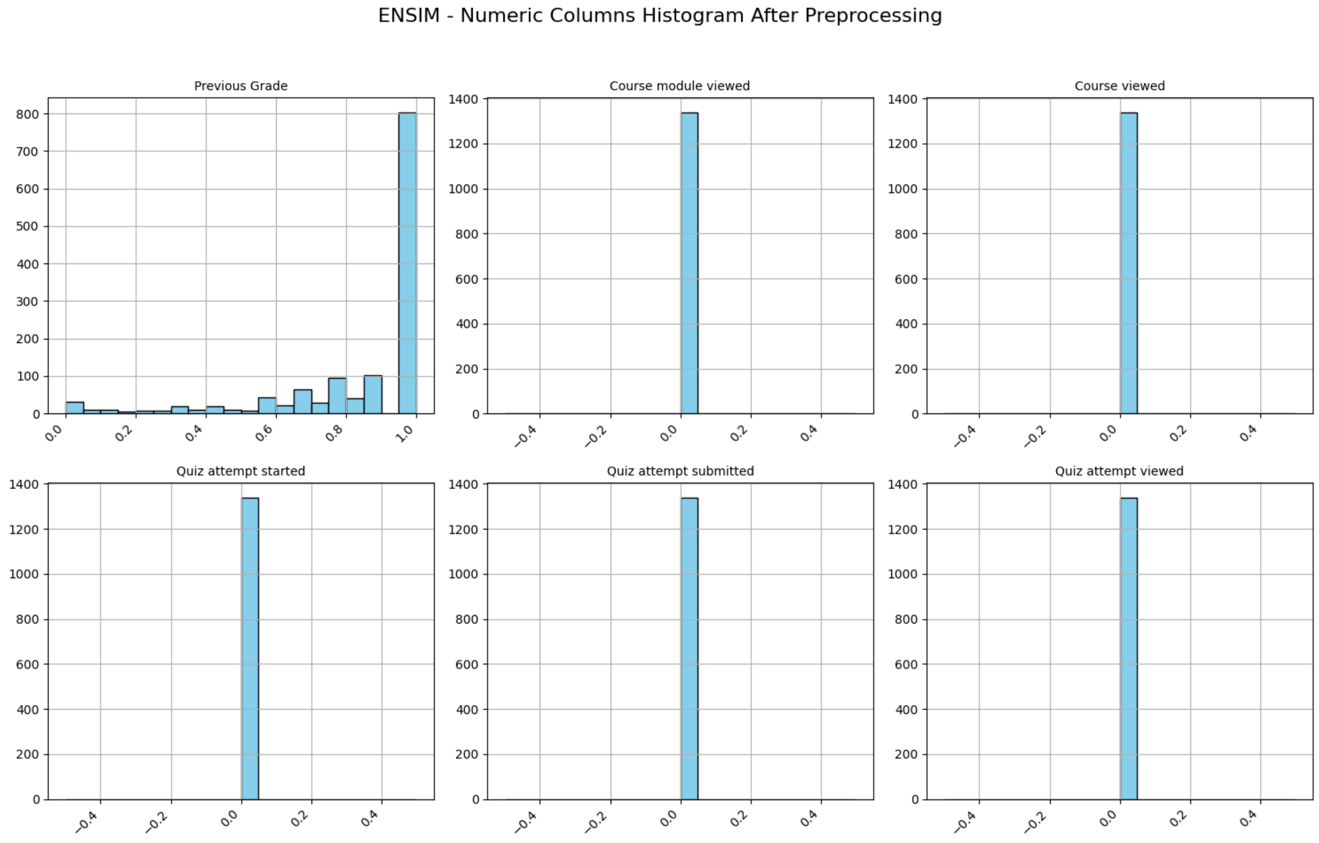


Figure D.10 – ENSIM - Numeric columns histogram after preprocessing

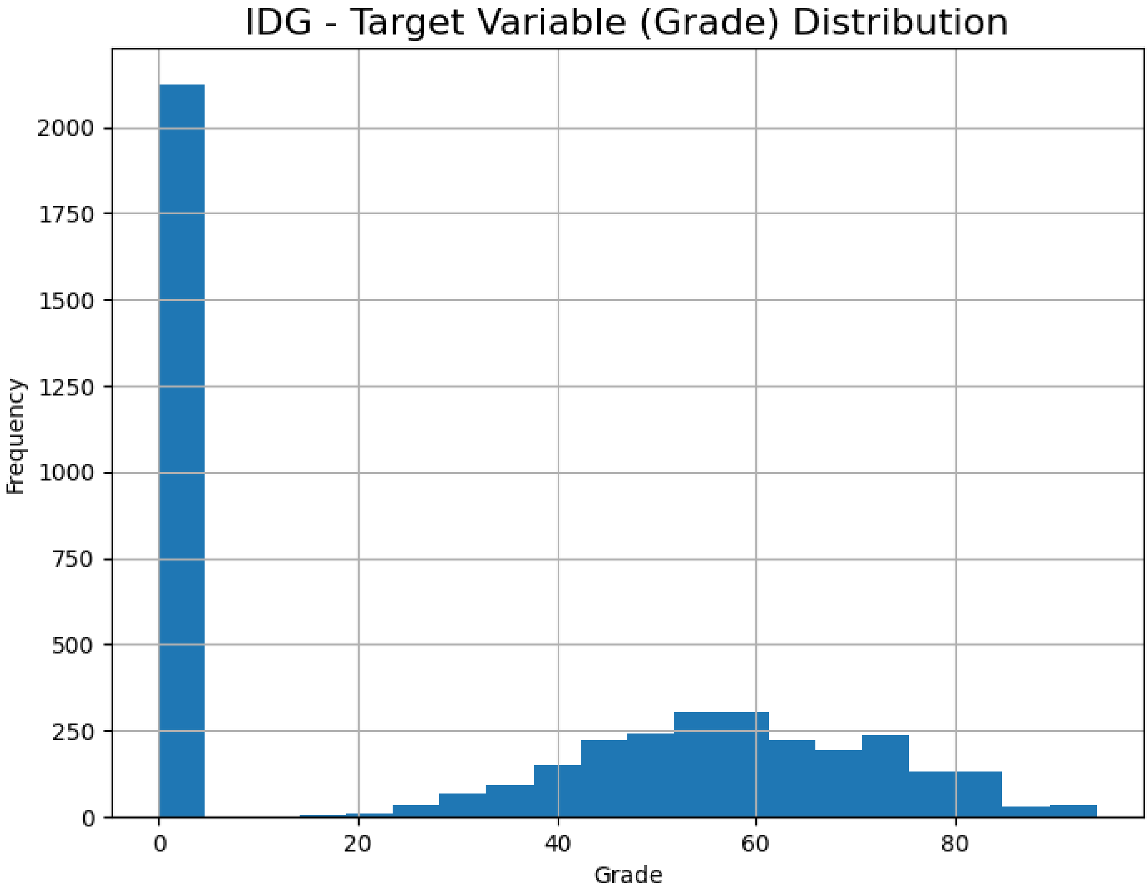


Figure D.11 – IDG - Target variable distribution

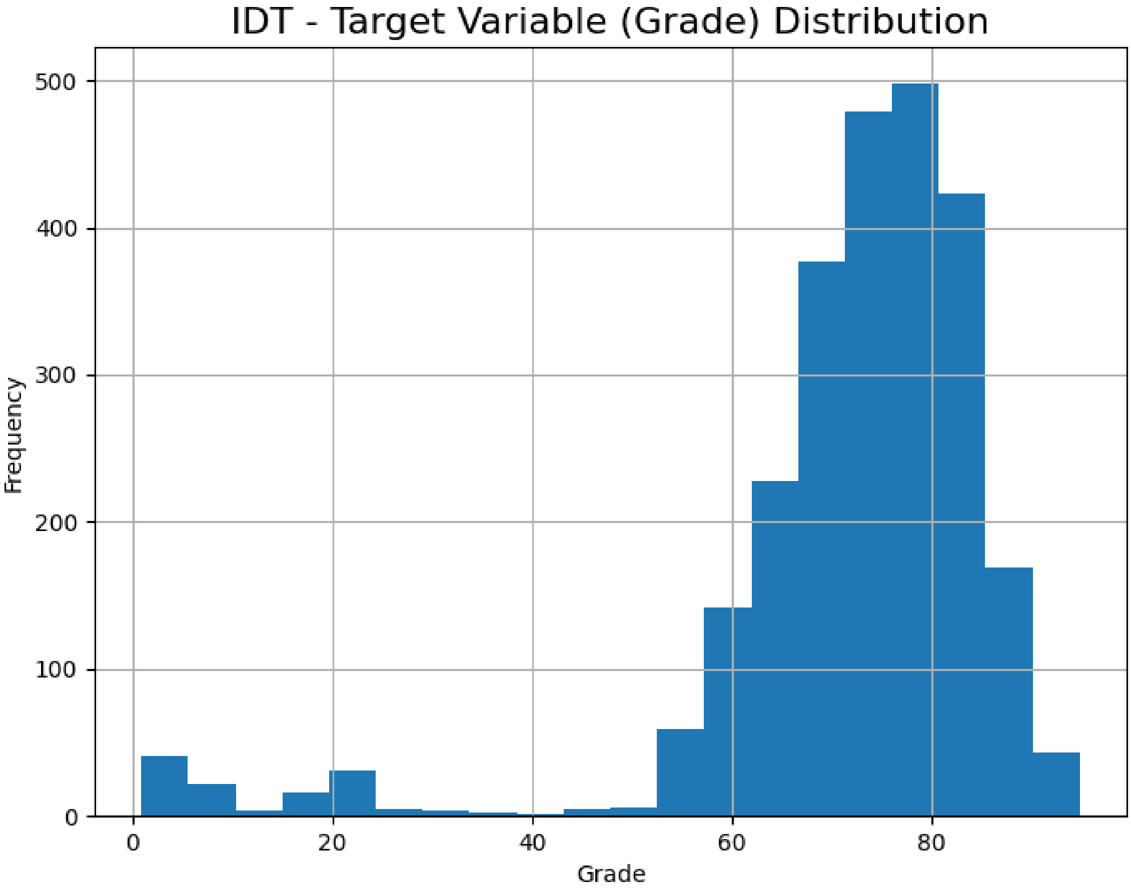


Figure D.12 – IDT - Target variable distribution

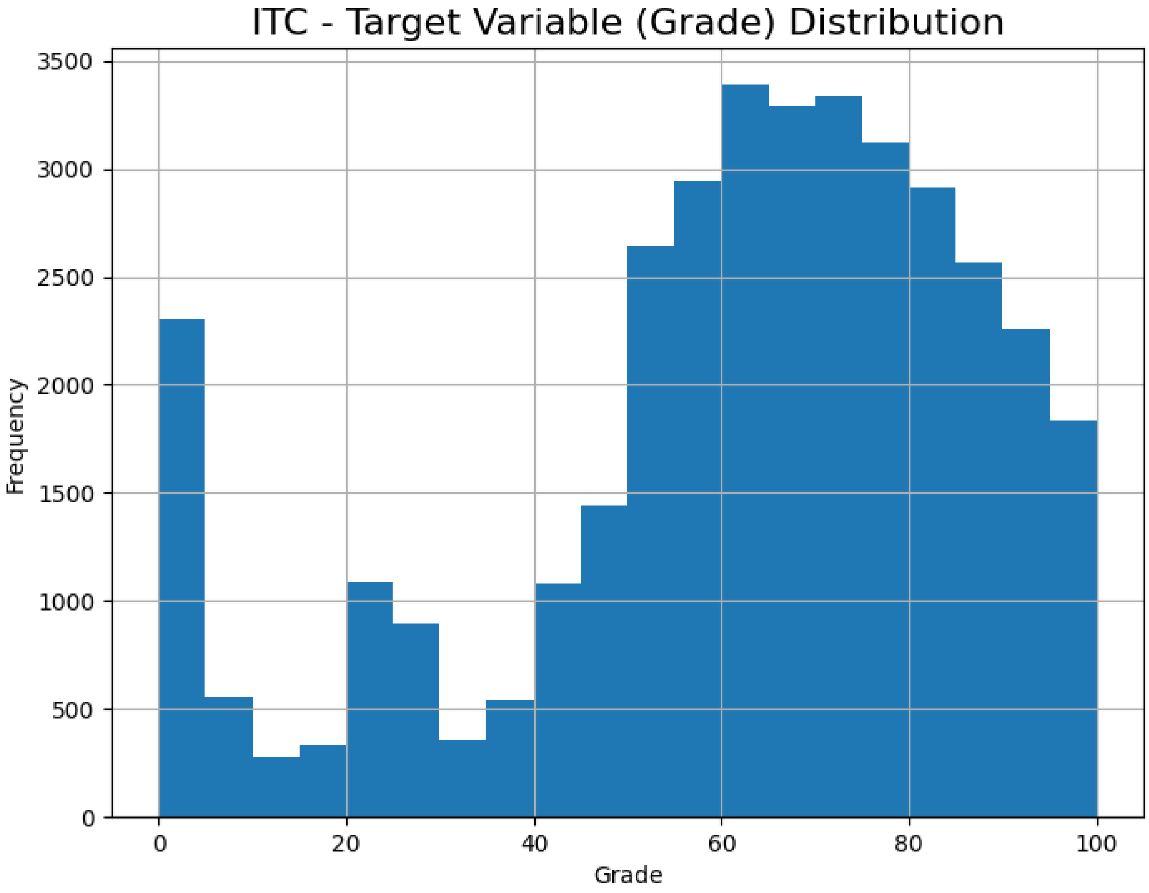


Figure D.13 – ITC - Target variable distribution

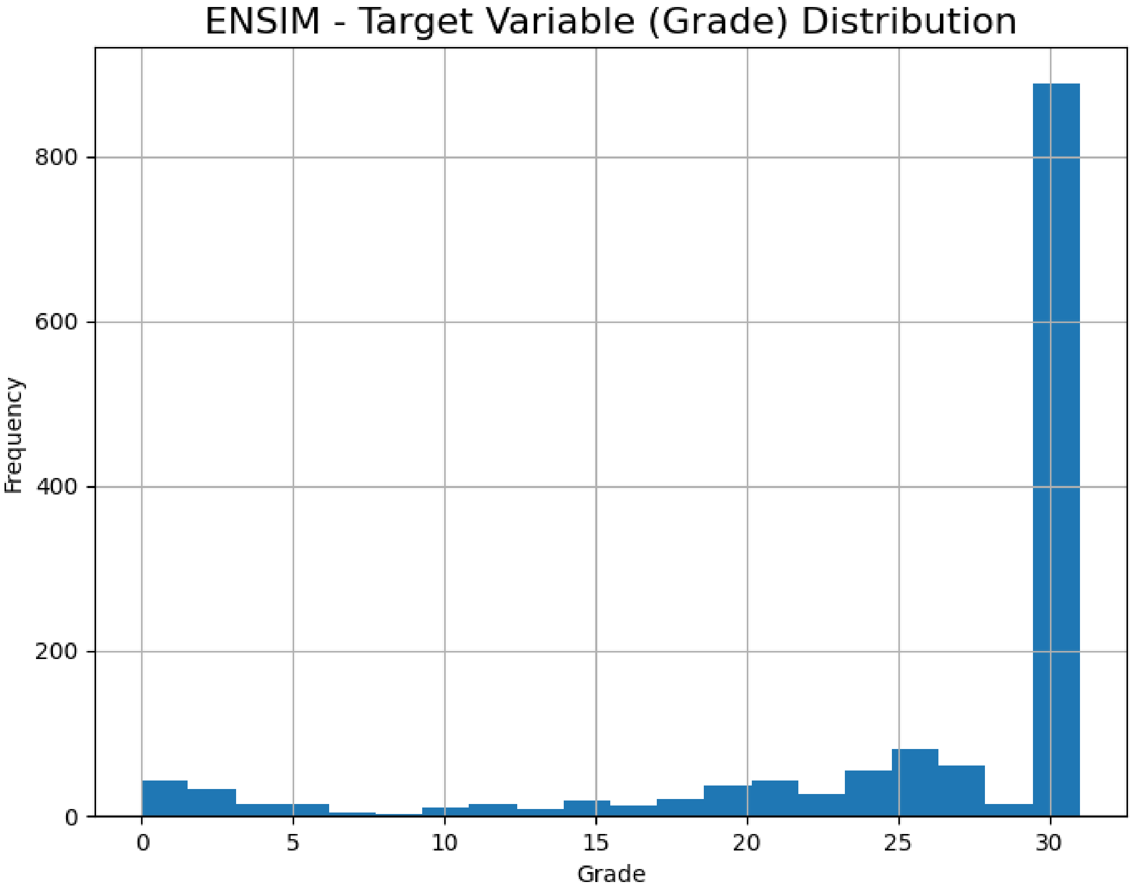


Figure D.14 – ENSIM - Target variable distribution

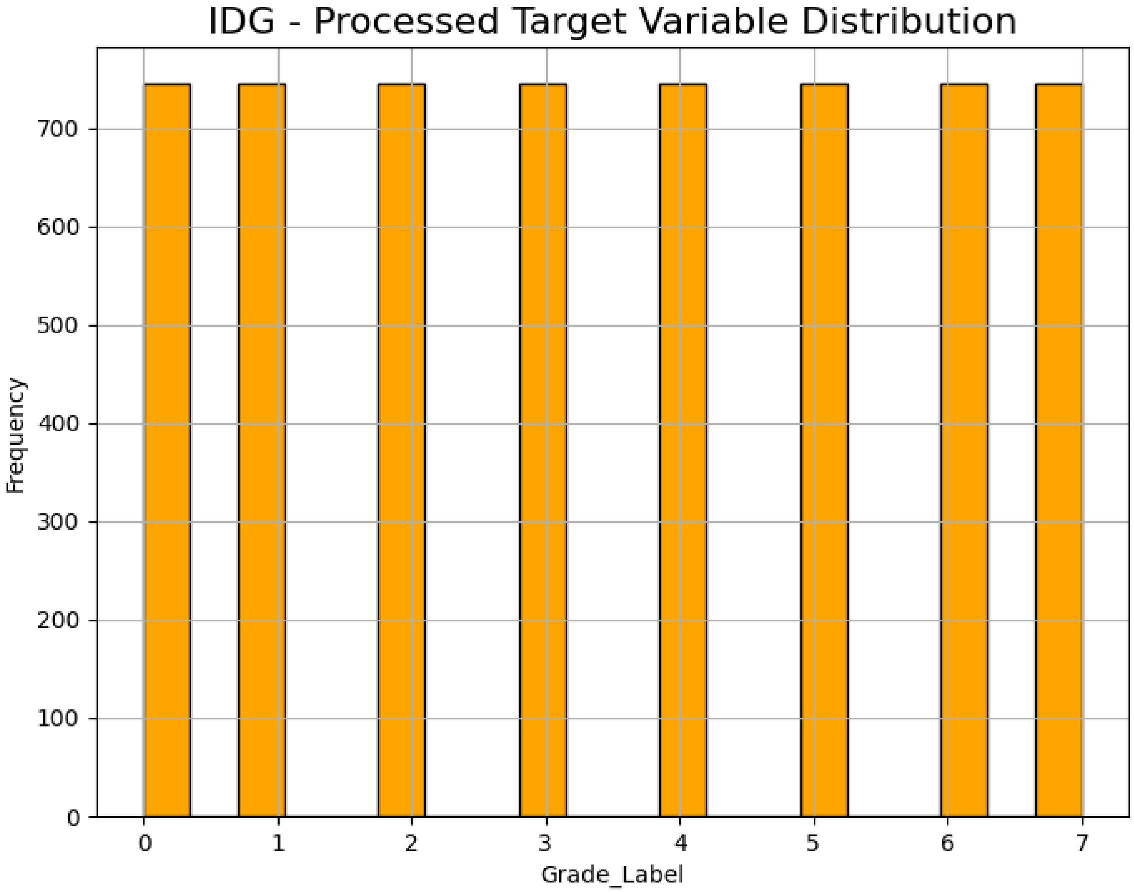


Figure D.15 – IDG - Processed Target variable distribution

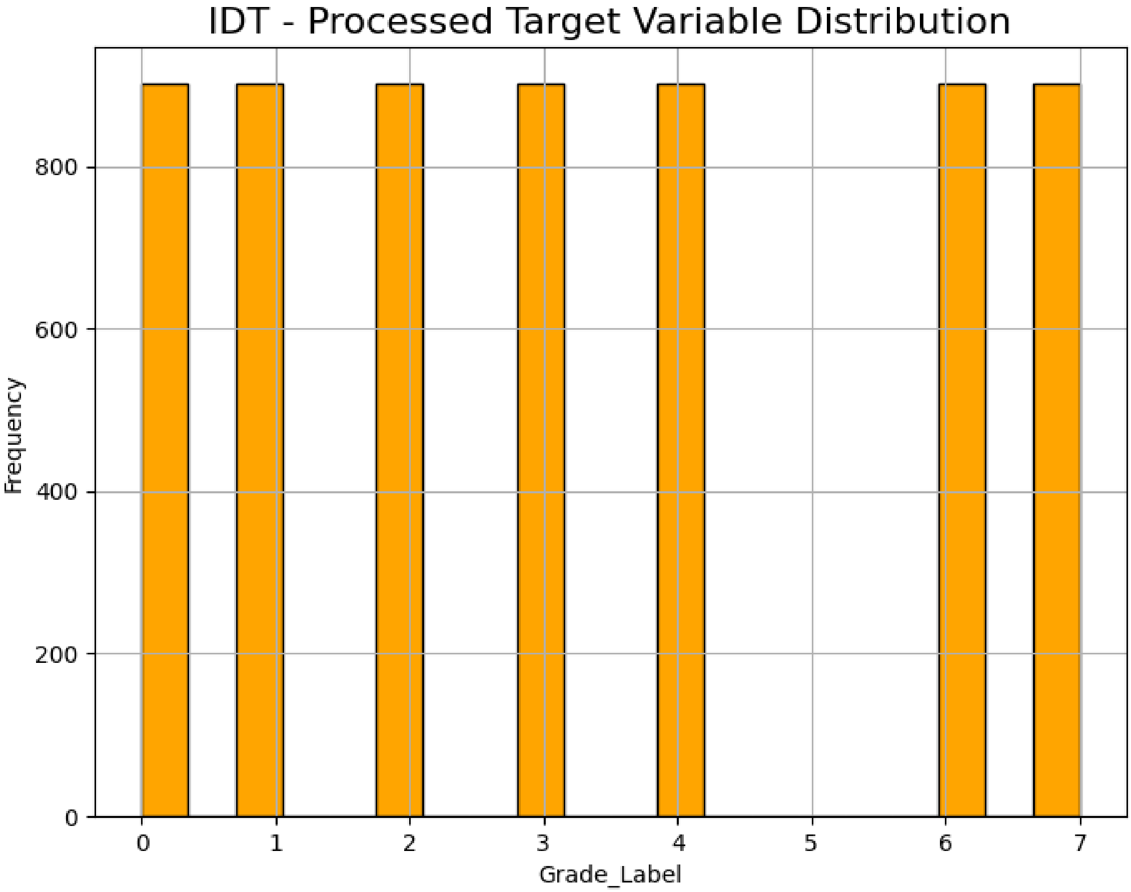


Figure D.16 – IDT - Processed Target variable distribution

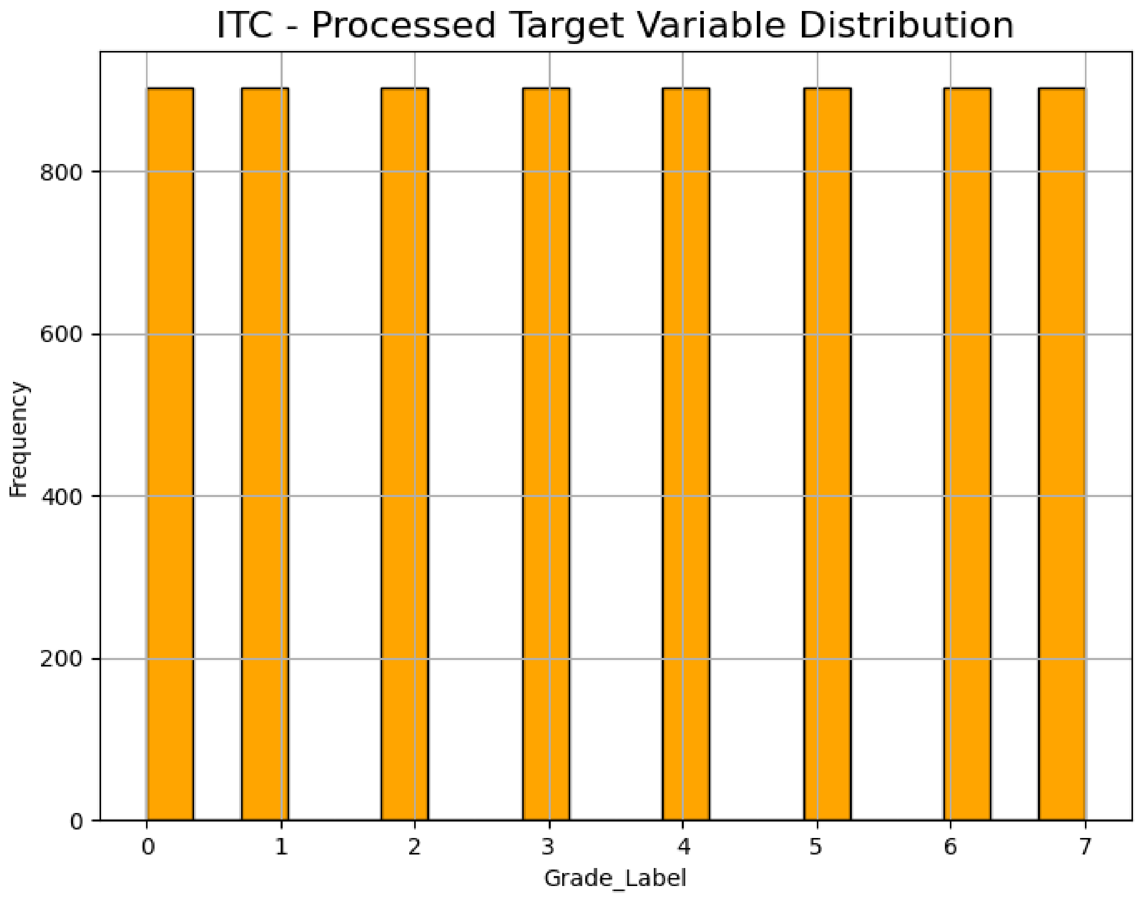


Figure D.17 – ITC - Processed Target variable distribution

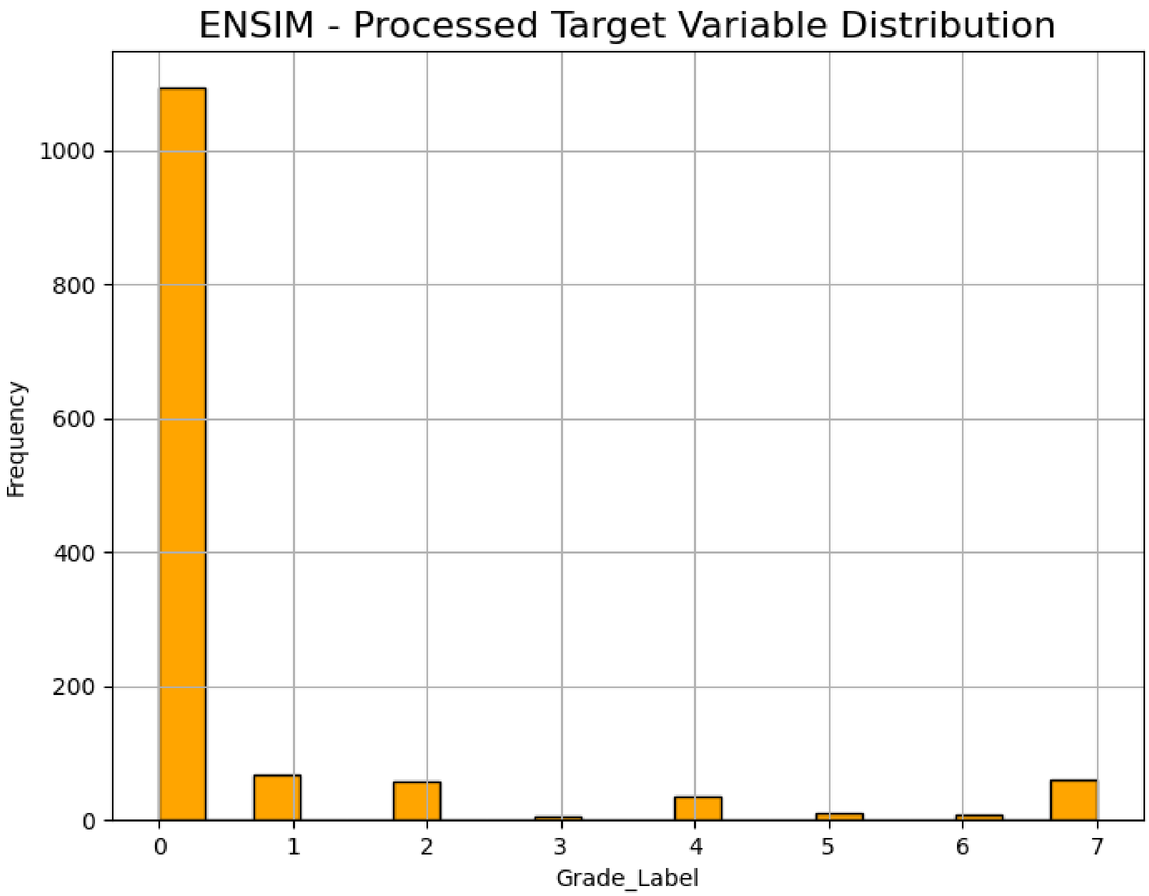
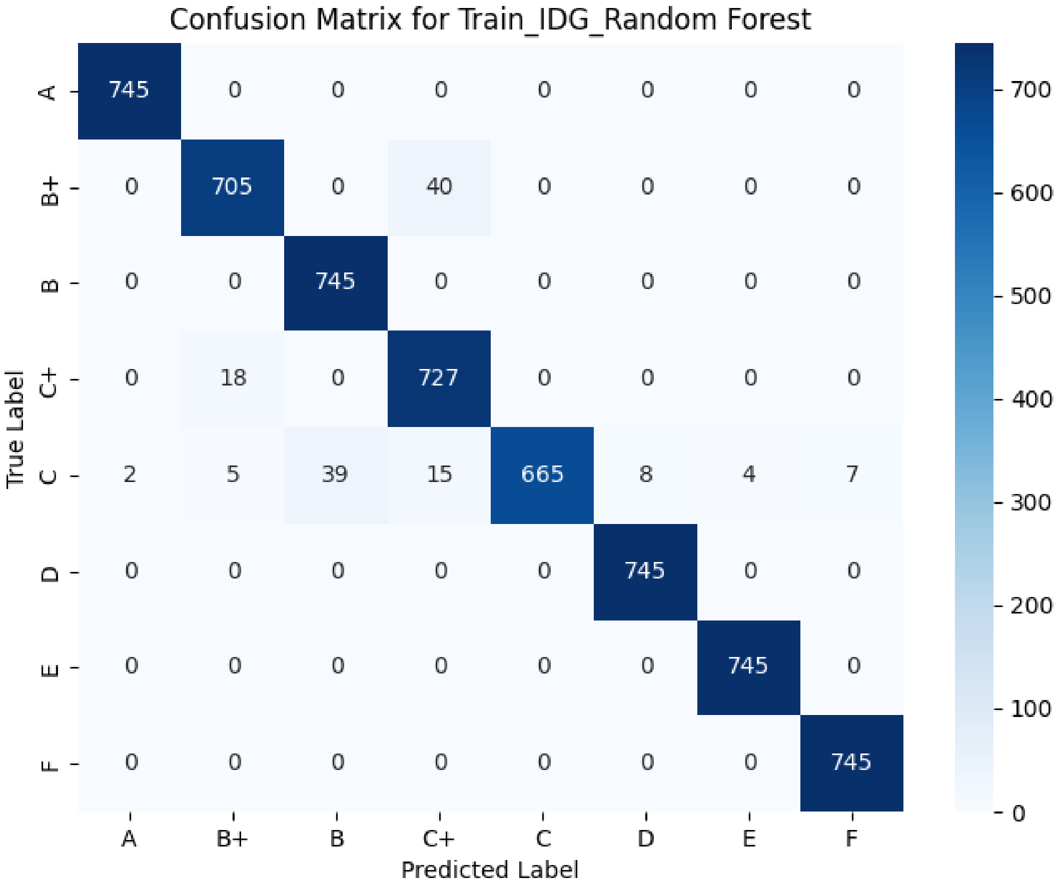
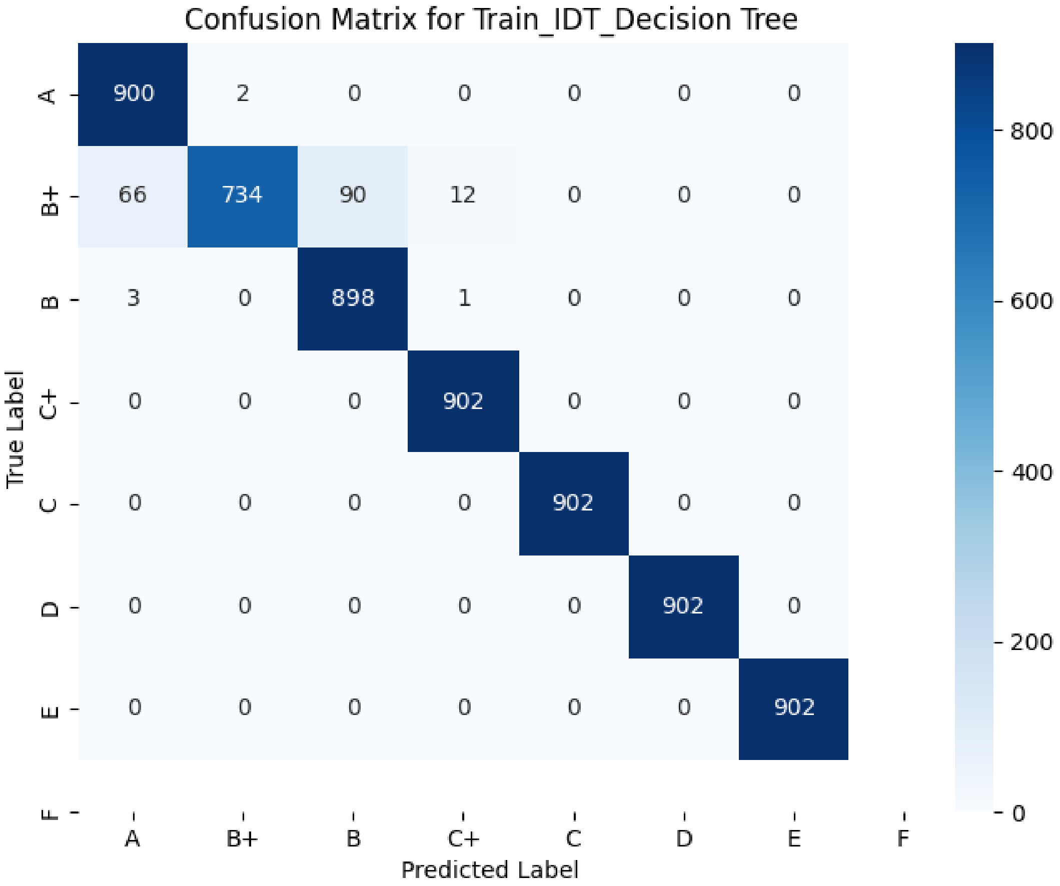


Figure D.18 – ENSIM - Processed Target variable distribution



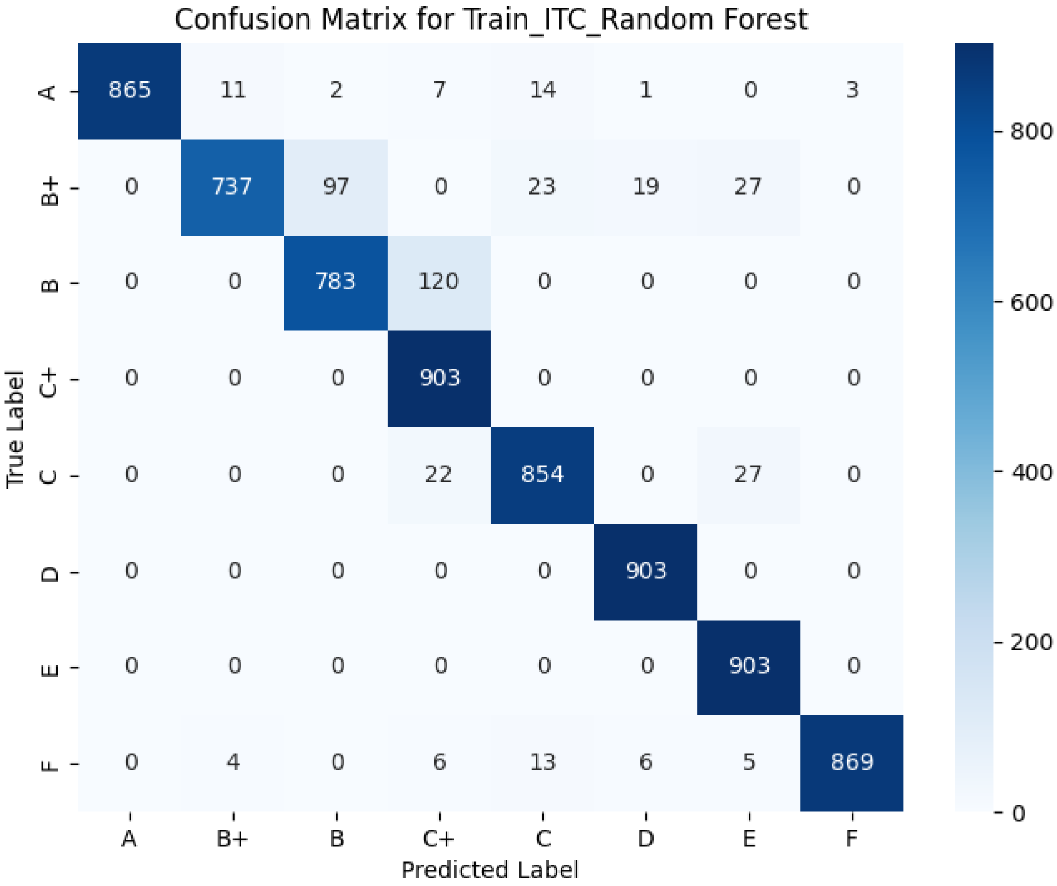
| | A | B+ | B | C+ | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| A | 745 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B+ | 0 | 705 | 0 | 40 | 0 | 0 | 0 | 0 |
| B | 0 | 0 | 745 | 0 | 0 | 0 | 0 | 0 |
| C+ | 0 | 18 | 0 | 727 | 0 | 0 | 0 | 0 |
| C | 2 | 5 | 39 | 15 | 665 | 8 | 4 | 7 |
| D | 0 | 0 | 0 | 0 | 0 | 745 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 745 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 745 |

Figure D.19 – Confusion Matrix for training the IDG dataset with Random Forest



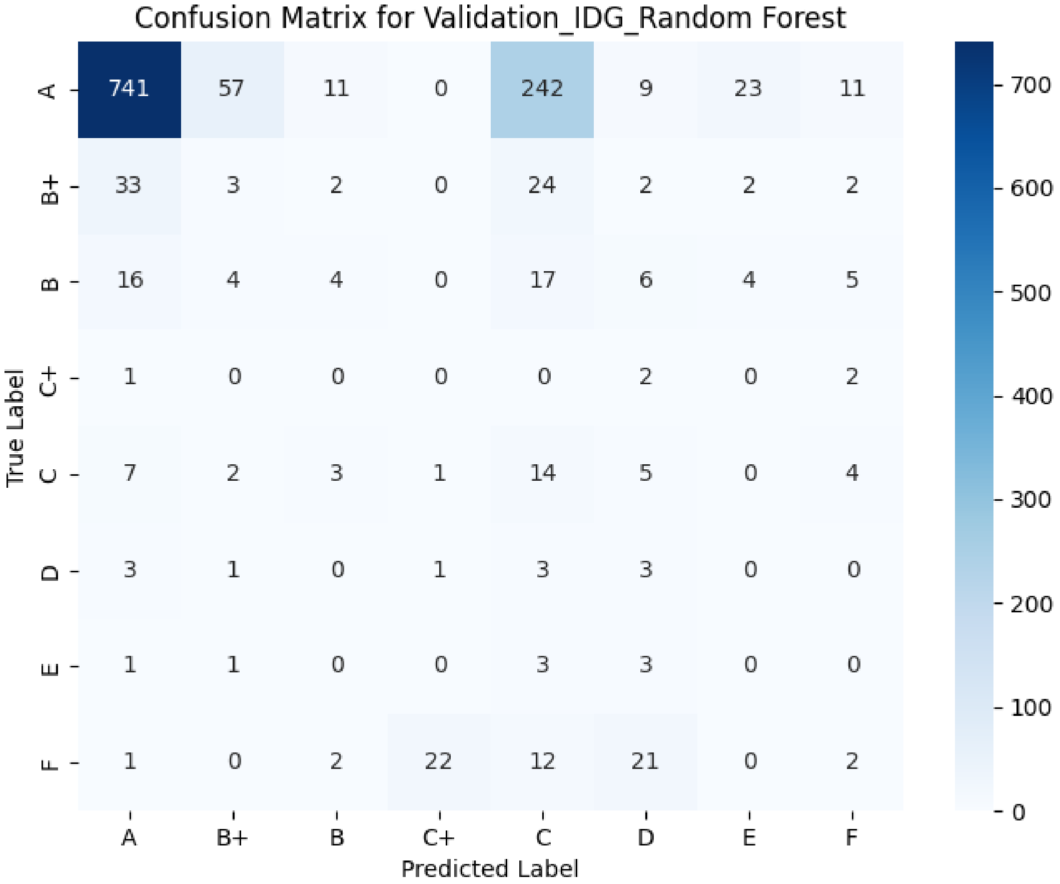
|  | A | B+ | B | C+ | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| A | 900 | 2 | 0 | 0 | 0 | 0 | 0 |  |
| B+ | 66 | 734 | 90 | 12 | 0 | 0 | 0 |  |
| B | 3 | 0 | 898 | 1 | 0 | 0 | 0 |  |
| C+ | 0 | 0 | 0 | 902 | 0 | 0 | 0 |  |
| C | 0 | 0 | 0 | 0 | 902 | 0 | 0 |  |
| D | 0 | 0 | 0 | 0 | 0 | 902 | 0 |  |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 902 |  |
| F |  |  |  |  |  |  |  |  |

Figure D.20 – Confusion Matrix for training the IDT dataset with Decision Tree



| True Label \ Predicted Label | A | B+ | B | C+ | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| A | 865 | 11 | 2 | 7 | 14 | 1 | 0 | 3 |
| B+ | 0 | 737 | 97 | 0 | 23 | 19 | 27 | 0 |
| B | 0 | 0 | 783 | 120 | 0 | 0 | 0 | 0 |
| C+ | 0 | 0 | 0 | 903 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 22 | 854 | 0 | 27 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 903 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 903 | 0 |
| F | 0 | 4 | 0 | 6 | 13 | 6 | 5 | 869 |

Figure D.21 – Confusion Matrix for training the ITC dataset with Random Forest



| | A | B+ | B | C+ | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| A | 741 | 57 | 11 | 0 | 242 | 9 | 23 | 11 |
| B+ | 33 | 3 | 2 | 0 | 24 | 2 | 2 | 2 |
| B | 16 | 4 | 4 | 0 | 17 | 6 | 4 | 5 |
| C+ | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 2 |
| C | 7 | 2 | 3 | 1 | 14 | 5 | 0 | 4 |
| D | 3 | 1 | 0 | 1 | 3 | 3 | 0 | 0 |
| E | 1 | 1 | 0 | 0 | 3 | 3 | 0 | 0 |
| F | 1 | 0 | 2 | 22 | 12 | 21 | 0 | 2 |

Figure D.22 – Confusion Matrix for the IDG dataset evaluated using a Random Forest model with the ENSIM validation dataset

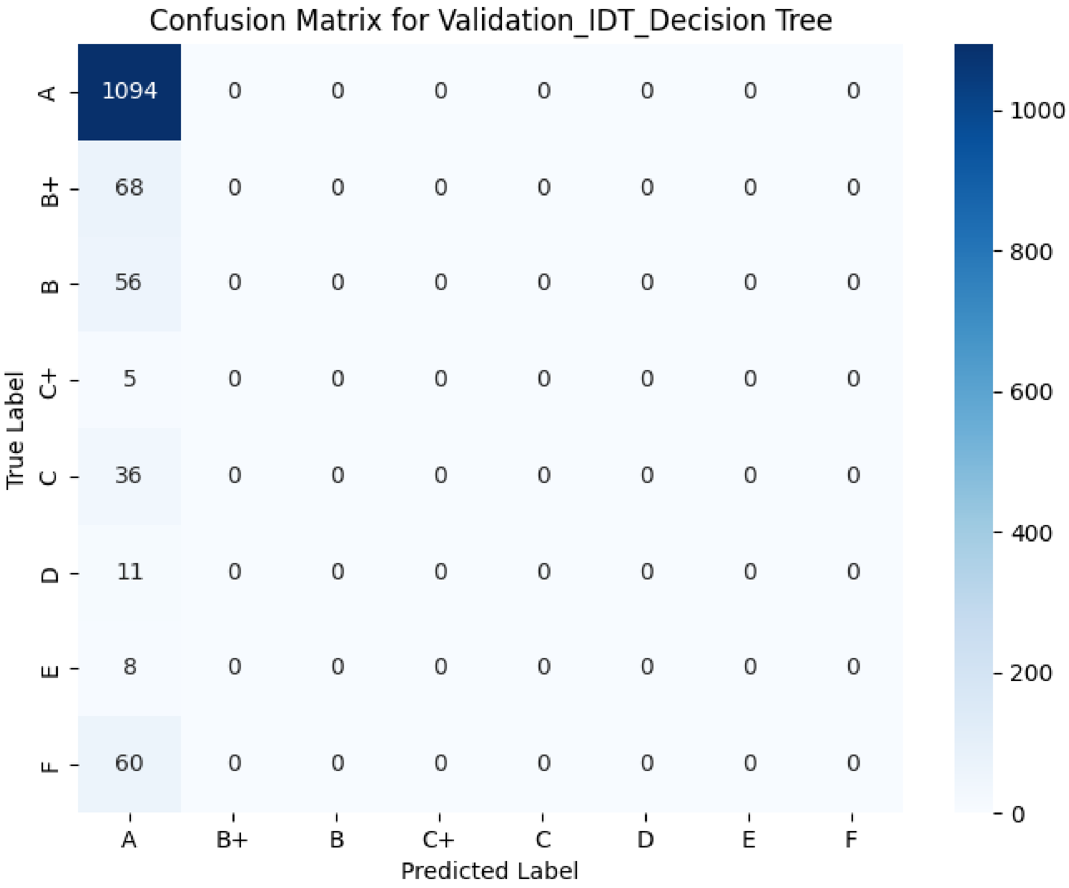


Figure D.23 – Confusion Matrix for the IDT dataset evaluated using a Decision Tree model with the ENSIM validation dataset

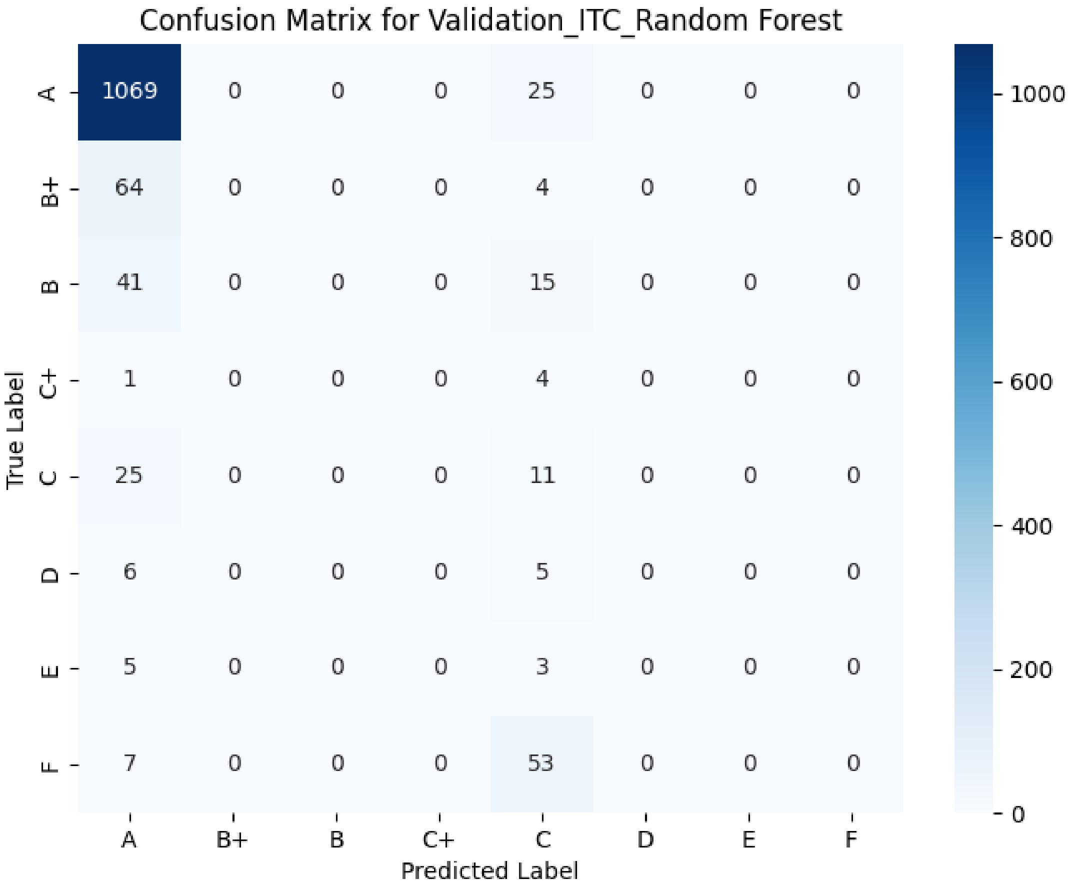
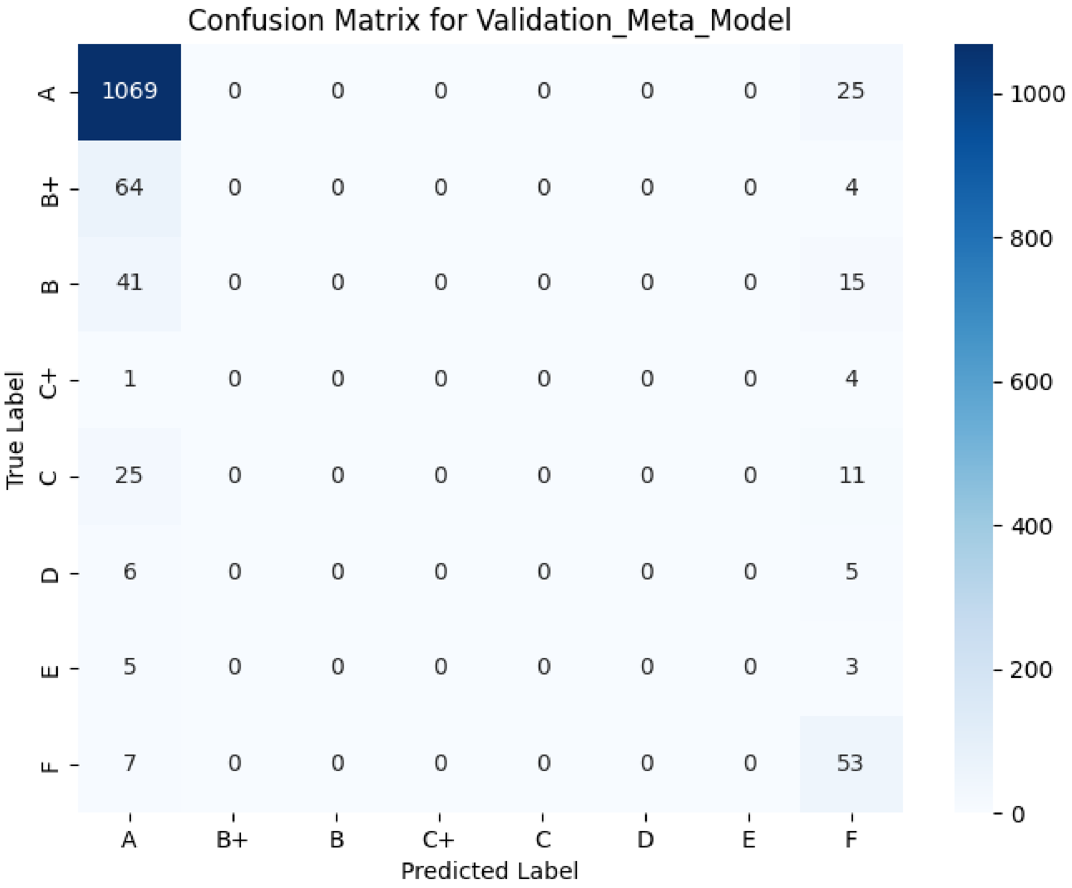


Figure D.24 – Confusion Matrix for the ITC dataset evaluated using a Random Forest model with the ENSIM validation dataset



| | A | B+ | B | C+ | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| A | 1069 | 0 | 0 | 0 | 0 | 0 | 0 | 25 |
| B+ | 64 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| B | 41 | 0 | 0 | 0 | 0 | 0 | 0 | 15 |
| C+ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| C | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| D | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| E | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| F | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 53 |

Figure D.25 – Confusion Matrix of the Meta Model evaluated on the ENSIM validation dataset

COLLEGE DOCTORAL PAYS DE LA LOIRE
ED MATHEMATIQUES STIC

Le Mans Université

**Titre :** Prédire les performances des étudiants grâce à la fouille de données trans-institutionnel : développement du modèle CISE et de l'outil ReflectMate

**Mot clés :** Prédiction des performances des étudiants, analyse de l'apprentissage, modélisation prédictive, apprentissage réflexif, validation interinstitutionnelle, engagement des étudiants

**Résumé :** L'évolution rapide des environnements d'apprentissage numériques a mis en évidence le besoin d'outils innovants pour améliorer les performances, l'engagement et l'équité des étudiants. Cette thèse présente deux cadres : le modèle prédictif **CISE (Cross-Institutional Stacking Ensemble)** et **ReflectMate**, un outil d'analyse réflexive de l'apprentissage. Elle aborde les défis de l'éducation en TIC, tels que l'hétérogénéité des ensembles de données, le déséquilibre entre les classes et les disparités causées par la fracture numérique, transformant ainsi les pratiques éducatives grâce à l'analyse prédictive et à l'apprentissage réflexif.

Le modèle **CISE** est un cadre d'apprentissage conçu pour prédire les performances académiques des étudiants dans divers contextes institutionnels. Il intègre cinq classificateurs d'apprentissage automatique : Decision Tree, Random Forest, Naive Bayes, Neural Network et Support Vector Machine, synthétisés via un méta-modèle basé sur la régression logistique. Cette approche d'ensemble empilé a démontré une précision prédictive et une généralisabilité supérieures, atteignant un taux de **78,25 % (score F1)** sur le jeu de données utilisé pour valider le méta-modèle.

En complément de CISE, **ReflectMate** a été conçu et développé pour permettre aux apprenants de s'engager activement dans leur parcours académique. Cet outil réflexif d'analyse des activités d'apprentissage, centré sur l'apprenant, fournit des informations en temps réel sur les interactions avec les plateformes LMS, notamment des indicateurs liés aux exercices, aux devoirs notés ainsi qu'au temps consacré aux activités d'apprentissage.

Une étude a également été menée sur la **fracture numérique**, en particulier dans le contexte de l'éducation en TIC au Cambodge. Des analyses statistiques ont révélé des disparités dans les résultats académiques en fonction du genre et de la localisation géographique, mettant en évidence des inégalités systémiques qui affectent les opportunités d'apprentissage.

Les contributions majeures de cette thèse portent sur l'exploration des données éducatives et l'analyse réflexive de l'apprentissage grâce à des solutions centrées sur les apprenants. En intégrant le modèle prédictif, ces travaux visent à améliorer les résultats académiques et à favoriser le développement de l'autonomie des apprenants.

COLLEGE DOCTORAL PAYS DE LA LOIRE
ED MATHEMATIQUES STIC

Le Mans Université

**Title:** Predicting Student Performance through Cross-Institutional Learning Analytics: Development of the CISE Model and ReflectMate Tool

**Keywords:** Student performance prediction, learning analytics, predictive modeling, reflective learning, cross-institutional Validation, student engagement

**Abstract:** The rapid evolution of digital learning environments has underscored the need for innovative tools to enhance student performance, engagement, and equity. This dissertation introduces two frameworks: the **Cross-Institutional Stacking Ensemble (CISE)** predictive model and **ReflectMate**, a reflective learning analytics tool. Together, they address challenges in ICT education, such as dataset heterogeneity, class imbalance, and disparities caused by the digital divide, transforming educational practices through predictive analytics and reflective learning.

The **CISE** model is an ensemble learning framework that forecasts student academic performance across diverse institutional contexts. It integrates five machine learning classifiers—Decision Tree, Random Forest, Naive Bayes, Neural Network, and Support Vector Machine—synthesized via a logistic regression-based meta-model. Achieving an **F1 score of 78.25%** on the ENSIM validation dataset, CISE demonstrates superior accuracy, generalizability, and reliability in identifying at-risk students, offering a robust solution for cross-institutional data mining.

Complementing CISE, **ReflectMate** empowers students by providing real-time insights into their learning behaviors, such as quiz attempts, assignment submissions, and LMS interactions. With features like progress tracking, tailored feedback, and an interactive dashboard, ReflectMate fosters self-reflection, self-regulation, and accountability. Surveys among CADT students highlight its effectiveness in promoting metacognition and academic engagement.

This research addresses the **digital divide** by revealing disparities in ICT education based on gender and location, emphasizing systemic inequities. The integration of CISE and ReflectMate bridges these gaps by delivering equitable access to resources and personalized support, empowering underserved populations. This dissertation advances educational data mining and learning analytics by combining predictive modeling with reflective practices, offering scalable, student-focused solutions to enhance outcomes and promote digital equity. Practical implications and future directions include expanding datasets, incorporating socio-economic factors, and evaluating long-term impacts on academic success and inclusion.