



**UNIVERSITY
OF OULU**

FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Fatemeh Mahjouyanmoghaddam

**PREDICTING GRADUATION TIME USING EDA
AND ML: A CASE STUDY AT THE UNIVERSITY
OF OULU**

Master's Thesis
Degree Programme in Business Analytics
June 2025

Mahjouyanmoghammad F. (2025) Predicting Graduation Time Using EDA and ML: A Case Study at the University of Oulu. University of Oulu, Degree Programme in Business Analytics, 72 p.

ABSTRACT

Predicting the time to graduation is a critical challenge for higher education institutions, as delays in degree completion can have significant financial and academic implications. This study analyzes academic data collected at the University of Oulu from students registered from 2004 to 2023 and develops a machine learning model to forecast the remaining time to graduate for still-studying students enrolled in a combined Bachelor's and Master's degree program. Using a dataset of student academic records, we apply data preprocessing techniques, feature engineering, and filtering criteria to extract key variables influencing graduation timelines. The selected features include study interruptions (Total Gap), credit accumulation (Total Credits, Average Credits per Term), and study pace (Rolling Average of Credits, Credit Growth Rate, Study Pace).

The predictive model is built using the Extreme Gradient Boosting (XGBoost) algorithm, optimized through hyperparameter tuning. The model achieves a high predictive accuracy, with an R-squared (R^2) score of 0.9596 on the test set. As a benchmark, a Random Forest model was also developed, which performed slightly lower but confirmed the robustness of the selected features across different tree-based methods. Feature importance analysis highlights that Total Gap and Average Credits per Term are the most influential predictors of graduation time, reinforcing the role of credit accumulation and study continuity in student progression.

Case study comparisons provide deeper insights into the behavioural patterns of still-studying students relative to graduated students. The findings reveal that students with slower study pace and lower credit accumulation rates tend to have prolonged graduation timelines, while those with consistent credit accumulation and fewer interruptions are more likely to complete their studies on time.

The results offer valuable insights for academic policymakers and advisors to identify students at risk of delayed graduation and implement data-driven interventions. Future research could explore additional demographic, behavioural factors, and external commitments, to further enhance prediction accuracy and policy recommendations.

Keywords: Educational Data Mining, Student Progression Analysis, XGBoost, Academic Performance

TABLE OF CONTENTS

ABSTRACT	
TABLE OF CONTENTS	
FOREWORD	
LIST OF ABBREVIATIONS AND SYMBOLS	
1. INTRODUCTION.....	7
1.1. Author’s Contributions and the Role of Artificial Intelligence	8
2. LITERATURE REVIEW.....	9
2.1. Introduction.....	9
2.2. BI Tools and DAX in Analysis of Educational Data	9
2.3. Predictive Modeling for Graduation Time.....	11
2.3.1. Core Predictive Studies.....	11
2.3.2. Advanced Implementations and Interventions	12
2.3.3. Contextual and Institutional Factors.....	13
2.3.4. Comparative Model Performances.....	13
2.3.5. Challenges and Limitations	13
2.4. Focusing on the Use of XGBoost and Random Forest in Education.....	14
2.4.1. Applications of Random Forest and XGBoost in Education.....	14
2.4.2. Enhancing Model Interpretability and Feature Selection	15
2.4.3. Evaluating Random Forest in Performance Prediction.....	15
2.4.4. Challenges and Limitations	16
3. METHOD AND TOOLS.....	17
3.1. Overview of Analytical Tools and Environment	17
3.2. Selection of Machine Learning Models	18
3.2.1. Random Forest.....	19
3.2.2. XGBoost (Extreme Gradient Boosting).....	20
3.2.3. Comparison Between Random Forest and XGBoost	21
3.3. Model Optimization and Evaluation	22
3.3.1. Evaluation Metrics	22
3.3.2. Validation Techniques and Residual Analysis.....	23
3.4. Feature Importance Techniques.....	24
3.5. Model Interpretation with SHAP.....	25
3.6. Python Libraries	25
3.7. Data Preprocessing and Additional Calculations	26
3.8. Tools for Visualization	27
3.9. Tools for Data Export and Machine Learning Integration.....	27
3.10. Summary: General Machine Learning Workflow	28
4. DATA AND INSIGHTS	30
4.1. Dataset Overview.....	30
4.1.1. Initial State of the Dataset.....	30
4.1.2. Statistical Overview.....	32
4.2. Data Preprocessing	32
4.3. Calculated Fields and Explanatory Analysis	34
5. MACHINE LEARNING IMPLEMENTATION	49

5.1.	Data Preparation and Filtering	49
5.2.	Feature Engineering and Selection	49
5.3.	Correlation Analysis of Key Features	50
5.4.	Train-Test Split and Data Preparation for ML	51
5.5.	Machine Learning Model Development and Selection	52
5.5.1.	Machine Learning Workflow	52
5.5.2.	Training and Evaluation Process.....	52
5.5.3.	Feature Importance Analysis	54
6.	RESULTS.....	56
6.1.	Prediction and Interpretation.....	56
6.2.	Evaluation of Prediction Quality	57
6.3.	SHAP Analysis.....	58
6.4.	Case Studies Using A Dashboard.....	59
6.4.1.	Case Study 1: Fast Track Group	59
6.4.2.	Case Study 2: Moderate Progress Group.....	61
6.4.3.	Case Study 3: Extended Study Group	62
6.5.	Summary of Findings	64
7.	DISCUSSION	65
7.1.	RQ1: What Academic Behavioural Features Most Significantly Influence a Student's Time to Graduate?.....	65
7.2.	RQ2: Which Machine Learning Model Provides the Most Accurate Prediction of Graduation Timelines for Still-Studying Students?.....	65
7.3.	RQ3: How Can the Model's Predictions Be Used to Identify Patterns and Support Students at Risk of Delayed Graduation?.....	66
7.4.	Limitations and Future Work	66
7.5.	Conclusion	67
8.	SUMMARY	68
9.	REFERENCES	69

FOREWORD

I would like to express my deepest gratitude to Prof. Olli Silvén for his invaluable guidance, support, and inspiration throughout this research. His ideas laid the foundation for this study, and his encouragement motivated me to explore the topic in depth.

I am sincerely thankful to Dr. Lauri Lovén for his dedicated supervision, insightful feedback, and patience during the thesis process. His support was instrumental in shaping both the direction and execution of this work.

My heartfelt appreciation also goes to Adj. Prof. Susanna Pirttikangas, whose support during the early stages of my studies and internship greatly contributed to my development as a researcher.

I would also like to thank the University of Oulu for providing the opportunity and academic environment to carry out this research. I am honoured to have been part of this community.

Finally, I wish to thank my family for their unwavering emotional support and my dear friend, Jani Launonen, for his continued encouragement throughout this journey.

Oulu, June 2nd, 2025

Fatemeh Mahjouyanmoghaddam

LIST OF ABBREVIATIONS AND SYMBOLS

AI	Artificial Intelligence
BI	Business Intelligence
CV	Cross-Validation
DAX	Data Analysis Expressions
EDA	Exploratory Data Analysis
ETL	Extract, Transform, Load
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
R^2	Coefficient of Determination
RF	Random Forest
SHAP	SHapley Additive exPlanations
SQL	Structured Query Language
SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting
AvgCreditPerTerm	Average credits taken per term
EducationLength	Total study duration from start to graduation
IntakeDelay	Time between registration and the start of studies
RunningTotal	Accumulated credits up to a given term
TermSerial	Sequential number of terms from start
TotalCredit	Total credits earned by the student
TotalGap	Sum of all inactive study periods, including final gap
StudyPace	Credit growth relative to time passed
CreditGrowthRate	Acceleration in credit accumulation over time
RollingAvgCredits	Rolling average of credits over past terms

1. INTRODUCTION

Higher education institutions are increasingly focused on understanding student progression and graduation timelines, as timely degree completion is crucial for both students and universities. Delayed graduation can lead to financial burdens for students, increased resource allocation by institutions, and inefficiencies in academic planning [1, 2]. Predicting students' time to graduate is, therefore, an essential aspect of academic decision-making, as it enables universities to identify at-risk students, optimize resource distribution, and improve retention rates [3].

Machine learning techniques have emerged as powerful tools for analyzing large educational datasets, uncovering patterns in student behaviour, and providing data-driven insights [4, 5, 6]. In this study, extensive data analysis was first conducted using Power BI¹ and DAX (Data Analysis Expressions)² to explore academic trends, visualize student progress, and generate interpretable features. These analyses offered a foundation for understanding behavioural patterns and identifying key indicators of academic delay or success. Building on these insights, machine learning models were then employed to predict the remaining time to graduation. Unlike traditional statistical methods, machine learning approaches can adapt to complex, nonlinear relationships between academic factors and graduation outcomes, offering more precise predictions. While previous research has explored student performance prediction [7], dropout risk analysis [8, 9], and graduation likelihood estimation using various machine learning approaches [10, 11], limited studies have focused on estimating the remaining time to graduation for students who are still enrolled. This study aims to fill that gap by integrating visual data analytics and machine learning to deliver actionable predictions grounded in real academic behaviour.

This research specifically focuses on combined degree students, who pursue both Bachelor's and Master's degrees within a single academic track. The dataset used in this study includes educational records such as registration and graduation dates, and credits earned during each academic period. While the original dataset contained limited demographic or personal information, critical behavioural features were derived through a detailed preprocessing and feature engineering process. These included metrics such as total credits earned, cumulative study gaps, average credit load per term, and credit accumulation patterns. One of the key challenges in predicting graduation time lies in capturing variations in academic trajectories, some students progress steadily, while others experience study delays or irregular credit accumulation. By leveraging these engineered features and applying advanced machine learning techniques, this study aims to provide a robust and interpretable prediction model that supports academic advisors and institutional decision-makers in identifying students at risk of prolonged study durations.

The primary objective of this study is to develop a predictive model that estimates the remaining time to graduation for still-studying students in a combined degree program. The methodology involves detailed feature engineering, machine learning model development, and evaluation based on predictive accuracy. Two models, XGBoost and Random Forest, were trained and validated using historical data from graduated students. The model demonstrating superior performance was then applied

¹<https://powerbi.microsoft.com>

²<https://learn.microsoft.com/en-us/dax/>

to predict outcomes for still-enrolled students. Additional analyses, including feature importance interpretation and case study comparisons, were conducted to gain a deeper understanding of the factors influencing study duration.

To guide this investigation, the study addresses the following research questions:

1. What academic behavioural features most significantly influence a student's time to graduate?
2. Which machine learning model provides the most accurate prediction of graduation timelines for still-studying students?
3. How can the model's predictions be used to identify patterns and support students at risk of delayed graduation?

The findings contribute to both academic research and institutional planning by offering a data-driven framework for understanding student progress. Beyond prediction, the results support proactive interventions, targeted advising, and strategic policy development to improve student retention and timely degree completion.

1.1. Author's Contributions and the Role of Artificial Intelligence

The author of this thesis conducted all data preprocessing, feature engineering, model development, evaluation, and interpretation independently. The predictive models and visualizations were designed, implemented, and analyzed by the author using Power BI, Python, and related libraries.

In the preparation of this document, ChatGPT (GPT-4) by OpenAI was used to support language refinement. Specifically, the tool was employed to correct grammar, improve sentence clarity, and restructure paragraphs for better readability in select parts of the thesis. All technical content, methodological design, and interpretation were authored and validated solely by the researcher.

2. LITERATURE REVIEW

2.1. Introduction

The objective of this literature review is to provide a structured understanding of prior research related to predicting graduation timelines using data-driven methods. Specifically, the review focuses on three key areas: (1) the role of Business Intelligence (BI) tools and data transformation techniques (such as Power BI and DAX) in educational analytics, (2) the application of machine learning models to predict academic outcomes like performance, dropout risk, or graduation likelihood, and (3) the use and interpretability of tree-based ensemble algorithms, particularly Random Forest and XGBoost, for structured educational data. These themes were selected to align with the tools and methods used in this thesis and to ensure theoretical grounding for both the technical approach and the analysis of results.

The initial phase of the literature review focused on identifying high-quality studies published between 2014 and 2024. Using bibliographic databases and academic platforms, the search was guided by keywords such as “graduation prediction,” “academic performance,” “student outcomes,” “machine learning in education,” and “educational data mining.” This process yielded a large number of publications covering themes like graduation time estimation, academic performance forecasting, and the use of machine learning in educational settings.

To ensure alignment with the thesis objectives, the findings were filtered in a single structured step. Studies were selected based on criteria such as methodological soundness, practical implementation of machine learning techniques, empirical data usage, peer-reviewed status, and contextual relevance. Additionally, studies that lacked full-text access or did not contribute meaningfully to the research scope were excluded.

Following this screening process, 44 studies were retained to support the development of the theoretical and methodological foundation of this thesis. These selected works informed the framing of research questions, the choice of modelling techniques, and the identification of challenges and opportunities within the educational data science landscape.

2.2. BI Tools and DAX in Analysis of Educational Data

This section focuses on Business Intelligence (BI) tools and DAX because these technologies form the foundation of the visual analytics and data preprocessing stages in this thesis. The inclusion of this topic was guided by its methodological relevance to the dashboard design and data modeling processes used in the empirical phase.

Studies were selected based on thematic relevance, specifically their focus on real-time data visualization, dashboard usability, interactive reporting, and BI-supported decision-making. While some studies target general business domains, they were included due to the transferability of their approaches to educational contexts. This thematic selection ensures a comprehensive understanding of how BI tools such as Power BI, Tableau, and QlikView, along with DAX expressions, can support academic data workflows.

The integration of Business Intelligence tools and Data Analysis Expressions (DAX) has significantly advanced data-driven decision-making in educational settings, offering stakeholders the ability to interpret and act on complex datasets. This section reviews studies exploring BI tools in educational analytics, highlighting their methodologies, findings, and limitations.

Studies have explored the integration of Business Intelligence (BI) tools and Data Analysis Expressions (DAX) in educational analytics, emphasizing their value in enabling stakeholders to interpret and act on complex datasets. One such study examined the usability and user satisfaction of different visualization designs in Big Data analytics, concluding that Cartesian coordinate visualizations outperformed polar-coordinate alternatives in terms of usability [12]. Although the study did not focus directly on educational data, its findings are highly relevant for improving dashboard design in academic contexts. The authors used interactive visualization tools and MANCOVA on a simulated dataset of 9,961 records and 14 dimensions. However, the limited visualization range and user interpretation challenges may reduce its applicability in broader educational settings.

Microsoft Power BI has been shown to effectively analyze and visualize educational data in practice. One study presented the usage of Power BI and DAX for the preparation of an online, web-based, interactive dashboard regarding the progression of Placement and Student's Marks for real-time decisions [13]. However, given its merits, detailed discussion of comprehensiveness was beyond the scope of this study; therefore, it was not fully able to address such nuances for a more complex type of educational data. Nevertheless, it gave some hints on how institutional decisions may be supported by BI tools.

A BI framework has been proposed that integrates data mining techniques, such as J48 classification, to analyze student dropout in distance learning programs [14]. The study identified critical factors contributing to dropout rates, enabling targeted interventions. The framework utilized WEKA for data mining and Microsoft SQL for Extract, Transform, and Load (ETL) processes, analyzing a dataset of 3,207 students. However, its reliance on academic and financial data limited its ability to capture broader influencing factors, emphasizing the need for more diverse datasets. .

The efficacy of tools such as Dash and Tableau in the formulation of interactive dashboards for predictive analytics purposes has been examined [15]. Although the investigation did not specifically concentrate on educational environments, the methodologies articulated, such as exploratory data analysis and predictive modelling, are amenable to application within educational frameworks. These strategies possess considerable ramifications for the development of intuitive and engaging platforms tailored for educational stakeholders.

QlikView BI software has been used to analyze trends in attendance and performance among university students [8]. The study identified that female students outperformed their male counterparts and highlighted a positive correlation between attendance and academic performance. However, the dataset's small size, consisting of only 575 records over three years, limited the generalizability of its findings. Despite these limitations, the research demonstrated the value of BI tools in uncovering actionable insights to improve student outcomes.

The reviewed studies collectively illustrate the transformative potential of BI tools in educational analytics. Tools like Power BI, Tableau, and QlikView provide robust

solutions for data visualization and decision-making. Some studies contributed general insights into visualization design and predictive modelling [12], [15], while others highlighted the direct application of BI tools in educational settings [13], [14]. Despite these contributions, challenges such as small datasets, narrow visualization scopes, and incomplete data persist. Future research should aim to integrate more comprehensive datasets, explore advanced visualization techniques, and address these limitations. By doing so, BI tools can better support data-driven educational practices and enhance institutional outcomes.

2.3. Predictive Modeling for Graduation Time

The utilization of machine learning (ML) for forecasting graduation timelines has gained substantial scholarly interest in recent years, offering new pathways for improving educational planning and student outcomes. This section synthesizes key research efforts that apply predictive models to structured educational data, focusing on graduation timing, dropout risks, and academic success metrics. The studies discussed vary in methodological complexity and in the types of student-related factors they examine.

The inclusion of this topic in the literature review is based on its direct thematic relevance to the central objective of this thesis, predicting students' time to graduate. Selected studies apply machine learning techniques to model graduation-related outcomes using real or simulated institutional data, providing a foundation for understanding how such predictions are constructed and validated in academic research.

To organize the findings, this section is divided into four thematic subsections: **Core Predictive Studies**, which examine foundational predictors such as GPA and study duration; **Advanced Implementations and Interventions**, focusing on studies that employ state-of-the-art techniques like SMOTE, LSTM, and XGBoost; **Contextual and Institutional Factors**, which highlight the impact of pre-admission and enrollment characteristics on graduation outcomes; and **Comparative Model Performances**, which analyze the effectiveness of various algorithms across different educational contexts.

The grouping was based on the complexity of the applied methods and the focus of each study, whether it emphasized general predictors, cutting-edge techniques, institutional background, or algorithmic comparisons. This categorization supports a clearer understanding of methodological trade-offs, data dependencies, and prediction accuracies within the domain of graduation time forecasting.

2.3.1. Core Predictive Studies

Studies in this category focus directly on predicting graduation timelines or academic performance using machine learning models. They emphasize core variables such as GPA, academic probation status, and study duration, providing foundational insights into student outcomes.

Several studies illustrate the effectiveness of ML algorithms in predicting graduation timelines, emphasizing the importance of data quality, algorithm selection, and stakeholder collaboration. These works use classification, regression, and deep learning models, and also highlight ethical challenges such as GDPR compliance. Metrics such as GPA and employment success rates are often used to inform graduation predictions [16].

One study focuses on students on academic probation, identifying study duration and secondary school performance as significant predictors of academic success. Using supervised ML algorithms, including J48 and Random Forest, it achieves an accuracy of 82.4%, showcasing the potential of these methods to identify at-risk students. Although the study centers on academic performance, its findings contribute to understanding factors affecting graduation timelines [17].

A targeted review identifies GPA as the most influential factor in predicting graduation time. The application of Neural Networks and Support Vector Machines achieves accuracy rates of 95% and 93.95%, respectively, directly addressing graduation predictions and emphasizing the importance of early interventions for at-risk students [6].

One study explores ensemble methods, combining Logistic Regression with Decision Tree models, to achieve an accuracy of 88.3% in predicting graduation. While the methods are robust, limitations such as data collection biases and a lack of longitudinal data restrict their ability to analyze long-term trends [18].

A broader perspective is provided in a systematic review of predictive learning analytics over a decade [5]. The study highlights advancements in ML techniques, such as Gradient Boosting and Neural Networks, which achieve accuracies of 84–93% in predicting graduation rates. However, small sample sizes and privacy concerns remain significant barriers to generalizability.

One study explores algorithms like K-Nearest Neighbors (KNN) and Artificial Neural Networks to forecast academic outcomes [4]. Although the study does not explicitly focus on graduation timelines, its emphasis on data acquisition and accuracy challenges offers valuable insights for related applications.

2.3.2. Advanced Implementations and Interventions

This section includes studies that adopt advanced modelling approaches or propose specific interventions based on predictive outcomes. These studies employ techniques such as SMOTE, LSTM, and XGBoost to address challenges like data imbalance or precision in at-risk student identification.

Advanced ML techniques have been employed to enhance prediction accuracy. One study applies Long Short-Term Memory (LSTM) models alongside the Synthetic Minority Oversampling Technique (SMOTE) to address the class imbalance in predicting late graduates [19]. Achieving an accuracy of 85%, the study highlights the potential of these techniques in improving retention strategies and graduation outcomes.

Similarly, a study demonstrates the superiority of XGBoost in predicting academic performance among slow learners, achieving a precision rate of 75%, with implications for understanding delayed graduation [7].

One study employs a comparative approach to evaluating Decision Trees, Neural Networks, and Support Vector Machines (SVM) for predicting graduation time. SVM, achieving the highest accuracy at 85.18%, underscores the importance of selecting appropriate models for early intervention strategies to enhance timely graduation rates [20]. Another study uses Artificial Neural Networks with backpropagation to distinguish between on-time and late graduates, achieving a precision rate of 97%. The study emphasizes the role of targeted interventions in improving graduation outcomes, showcasing the practical application of ML in academic decision-making [10].

2.3.3. Contextual and Institutional Factors

This subsection addresses studies that investigate external and institutional variables influencing graduation time. These factors, often overlooked in core predictive models, include pre-university academic background, enrollment behaviour, and institutional policies, which can substantially impact students' academic trajectories.

Pre-institutional factors are critical in understanding graduation timelines. One study highlights high school GPA, ACT scores, and financial aid as significant predictors of graduation delays. Deep Boltzmann Machines (DBMs) outperform other models in computational efficiency and class recall, offering robust solutions for predicting graduation delays [21]. Another study focuses on enrollment factors, such as major changes, as significant predictors of graduation timing. Using XGBoost, the study outperforms logistic regression, particularly for early-semester predictions. However, its reliance on proxy measures for social integration indicates the need for more comprehensive interaction data [22].

2.3.4. Comparative Model Performances

This subsection reviews studies that compare the performance of various machine learning models in predicting graduation outcomes. These works emphasize how different algorithmic choices affect prediction accuracy, particularly when applied to diverse educational and contextual data.

Random Forests and genetic algorithms have been shown to be effective in predicting graduation success [23]. The C4.5 decision tree algorithm was used in another study, achieving a prediction accuracy of 90% [11]. Additional studies have incorporated socioeconomic and behavioural data into their analyses to improve prediction performance. One study focused on dropout and success rates [9], while another used Artificial Neural Networks to predict degree completion with 100% accuracy in specific classifications [1]. These findings underscore the value of selecting appropriate models and integrating diverse data sources to enhance predictive accuracy.

2.3.5. Challenges and Limitations

Despite significant advancements, challenges persist in applying machine learning (ML) to educational data. Data quality issues, algorithmic biases, and limited

generalizability are recurring themes. One study used AutoML functionalities to predict course completion times, achieving over 90% accuracy. However, imbalanced datasets and the exclusion of dropout records limited the applicability of the findings [2]. Another emphasized rigorous data preprocessing and feature selection, attaining a top accuracy of 96.96% with ensemble methods [3]. A third identified cumulative credits and average grades as key predictors of on-time graduation, achieving 85% accuracy [24].

These studies collectively illustrate the transformative potential of predictive modelling in improving graduation outcomes. They highlight that critical academic, demographic, and behavioural features, such as credit accumulation and prior performance play pivotal roles in shaping graduation timelines. At the same time, the challenges they encountered including dataset imbalance [2], the importance of refined features [3], and small sample limitations [24] underscore the need for future research to address data quality, generalizability, and ethical considerations in educational machine learning applications.

2.4. Focusing on the Use of XGBoost and Random Forest in Education

The utilization of machine learning (ML) has evolved into an indispensable instrument within the domain of educational data mining, enabling institutions to enhance decision-making through predictive analytics. Among the wide range of ML algorithms, XGBoost and Random Forest have consistently demonstrated high predictive performance, scalability, and adaptability across various structured data applications. Their repeated use and superior accuracy in numerous educational prediction tasks, such as student success, dropout detection, and graduation forecasting make them particularly relevant to this thesis.

The inclusion of this subsection is thus thematically motivated by their prominence and proven effectiveness in the literature, as well as by their direct relevance to the modelling approach adopted in this thesis. This section synthesizes studies that have specifically implemented XGBoost or Random Forest in educational contexts, highlighting how these algorithms have been applied, validated, and interpreted across different academic datasets.

2.4.1. Applications of Random Forest and XGBoost in Education

Studies in this subsection apply Random Forest and XGBoost algorithms specifically to educational prediction tasks, including academic performance, early intervention, and identification of at-risk students. While not all directly predict graduation timelines, the outcomes contribute to understanding key precursors that influence graduation success.

Random Forest has also proved to be versatile in a number of studies that have tried to predict academic performance for betterment in educational outcomes. One study developed a classifier model in which Random Forest outperformed other algorithms, achieving an accuracy of 76.59% [25]. By identifying prior qualifications and gender as key predictors, the model indirectly related academic performance to potential

graduation outcomes. However, the exclusion of biased attributes such as school type raised concerns about the completeness and generalizability of the findings [25].

In a separate study, Random Forest was applied within the context of higher education and achieved an accuracy of 97.03% in predicting the academic performance of first-year IT students [26]. The study emphasized the role of early intervention in supporting at-risk students, thereby indirectly contributing to improved graduation rates. Nonetheless, challenges such as class imbalance and the omission of critical performance indicators highlighted methodological limitations that warrant further refinement [26].

The potential of XGBoost has been demonstrated in evaluations of machine learning techniques for identifying at-risk students, where it achieved the highest AUC score (0.92), indicating strong suitability for retention strategy enhancements [27]. Although the focus was not specifically on graduation timelines, the applied methodologies provided relevant insights into broader educational applications of machine learning. However, the study's reliance on condition-specific features introduced computational challenges that limited its scalability [27].

Another study addressed socio-economic disparities by employing both Random Forest and Extreme Gradient Boosting to predict academic performance among disadvantaged students [28]. The identification of prior academic performance and admission test scores as significant predictors indirectly contributed to understanding graduation outcomes [28].

2.4.2. Enhancing Model Interpretability and Feature Selection

Recent research has increasingly focused on improving the interpretability and reliability of machine learning models in educational settings. One study employed Gradient Boosting in combination with Biogeography-Based Optimization (BBO) for feature selection, achieving an accuracy of 73.88% and utilizing SHAP and LIME to enhance model transparency [29]. Although not directly aimed at predicting graduation timelines, the applied techniques offered valuable insights for refining ML models applicable to academic data. Limitations such as reliance on self-reported data and missing attributes highlighted the importance of comprehensive datasets [29].

Another study addressed class imbalance challenges by combining Random Forest with the Synthetic Minority Oversampling Technique (SMOTE), reaching a precision and recall rate of 0.83 despite using a relatively small dataset [30]. While the primary objective was not graduation prediction, the methodology emphasized the critical role of dataset balance in improving predictive accuracy, which holds relevance for similar educational applications [30].

2.4.3. Evaluating Random Forest in Performance Prediction

The robustness and reliability of Random Forest have been further validated in various educational contexts. One study achieved a perfect accuracy score of 100% in predicting academic performance using Random Forest, focusing on demographic and test preparation factors [31]. The findings emphasized the algorithm's strength

in academic performance analysis. However, the absence of detailed constraints and challenges within the dataset raised questions about the generalizability of the results to broader contexts [31].

2.4.4. Challenges and Limitations

Many of the reviewed studies report common challenges such as class imbalance, biased data, and incomplete datasets. For example, one study achieved very high accuracy, but the risk of prediction bias remained due to a strong imbalance between positive and negative samples [26]. Another study emphasized that dataset balance is particularly crucial in small-scale studies, as imbalance may significantly distort predictive accuracy [30]. The improvement of interpretability in ML models remains another notable challenge. While one study attempted to address this using SHAP and LIME, the methods still require further refinement to enhance transparency without compromising predictive power [29]. These challenges indicate the need for improved algorithmic design and data collection practices to maximize the effectiveness of ML in educational settings.

3. METHOD AND TOOLS

This section provides a theoretical overview of the machine learning methods, evaluation metrics, and tools used in this study. The goal is to predict the remaining time to graduation for students enrolled in a combined Bachelor's and Master's degree program. The methods selected are based on their suitability for structured educational data, their interpretability, and their proven effectiveness in regression-based prediction tasks.

Multiple environments and technologies were employed, including Power BI for data visualization and feature creation, Python for machine learning model development, and R for custom visualizations. These tools were chosen to support the full lifecycle of the machine learning pipeline: from preprocessing to prediction and interpretation. Their individual roles in the data-specific workflow are discussed later in Section 5.

While data characteristics often inform method selection, this chapter presents the analytical tools and modelling techniques in advance to provide a comprehensive theoretical foundation. The structure reflects an emphasis on methodological transparency, with dataset-specific preprocessing and insights detailed subsequently in Section 4. This separation allows for a clear distinction between generalizable machine learning principles and context-specific data handling procedures.

3.1. Overview of Analytical Tools and Environment

This study adopts a modular, tool-driven architecture for predictive modelling, with each platform selected based on its specific analytical strengths. The analytical environment comprises business intelligence tools, programming languages, and machine learning libraries, each contributing to different stages of the workflow, ranging from data preprocessing and modelling to evaluation and interpretability.

- **Power BI:** A leading business intelligence tool employed for initial data exploration, preprocessing, and transformation. Power BI supports the DAX (Data Analysis Expressions) language, which enables the construction of calculated columns and dynamic aggregations to support structured educational data analysis. Prior studies have demonstrated the effectiveness of Power BI in academic analytics and predictive modelling scenarios [13, 14, 10, 12].
- **DAX Studio:** This external tool interfaces with Power BI to enable advanced querying and extraction of tabular model data. It enhances reproducibility and precision when exporting feature-engineered datasets for use in external machine learning environments, complementing the Power BI ecosystem [13].
- **Python:** A versatile programming language widely adopted for data analysis and machine learning tasks. It provides access to comprehensive libraries such as `scikit-learn` for classical models [32, 33] and `xgboost` for optimized gradient boosting methods [34]. Python served as the primary environment for model training, evaluation, and interpretation.

- **Google Colab:** A cloud-based Jupyter Notebook service that enables code execution without local configuration. It supports GPU acceleration and fosters collaborative, reproducible machine learning research. Its suitability for rapid experimentation makes it ideal for academic studies involving large datasets [15, 16].
- **SHAP (SHapley Additive exPlanations):** A model-agnostic interpretability framework grounded in cooperative game theory. SHAP attributes individual prediction outcomes to feature contributions, offering both global and local interpretability [35, 36, 37]. It is especially effective in education-related predictive studies for revealing which features most strongly influence graduation time predictions [17, 19].
- **R (via Power BI integration):** Within Power BI, R was integrated to enable custom visualizations for presenting model outputs. R scripting facilitated enhanced interactivity and interpretability of dashboard visuals, in line with educational analytics practices demonstrated in earlier research [6, 11].

This analytical environment supports a flexible and modular machine learning workflow. Each tool in the pipeline plays a distinct role, from preparing structured educational datasets and building predictive models to interpreting results and communicating insights effectively.

3.2. Selection of Machine Learning Models

Selecting the appropriate machine learning model is a foundational step in any supervised learning task, as it determines how well the algorithm can capture patterns in the data and make accurate predictions. In the context of educational data, where relationships between variables can be complex and non-linear, it is especially important to choose models that offer both high predictive accuracy and interpretability.

Several theoretical and practical criteria typically guide model selection:

- **Ability to Model Non-linear Relationships:** Many student behaviors and academic patterns do not follow linear trends. Effective models must therefore be able to capture complex interactions between features.
- **Interpretability and Explainability:** Educational applications often require clear explanations of predictions to support decision-making by non-technical stakeholders such as educators and administrators.
- **Resistance to Overfitting and Noise:** Real-world educational datasets often contain noise, missing values, and irregular study patterns. Robust models should generalize well to unseen data without being overly sensitive to outliers.
- **Scalability and Computational Efficiency:** As datasets grow in size and complexity, models must remain computationally efficient and scalable.

- **Compatibility with Feature Importance Techniques:** Since identifying key drivers of academic outcomes is a major objective of this study, models that support feature importance analysis and interpretation (e.g., SHAP values) are preferred.

Based on these criteria, this study selected two tree-based ensemble learning algorithms, **Random Forest** and **XGBoost**, which are known for their robustness, flexibility, and strong performance on structured tabular data [38, 32, 33]. Both models have been widely adopted in educational data mining and have demonstrated effectiveness in predicting academic performance and graduation outcomes [16, 17, 19].

3.2.1. *Random Forest*

Random Forest is a supervised machine learning algorithm that belongs to the family of ensemble methods, techniques that combine multiple models to improve predictive performance and robustness [38]. It is widely used for both classification and regression tasks and is especially valued for its ability to handle high-dimensional, structured data [32, 33].

Foundation: Decision Trees At the core of the Random Forest algorithm are decision trees. A decision tree is a model that recursively splits the input data based on feature values to arrive at a prediction. Each internal node of the tree corresponds to a decision rule on a feature, and each leaf node corresponds to a predicted outcome. While decision trees are simple to interpret, they are prone to overfitting when trained deeply on small or noisy datasets [33].

Ensemble Concept: Bagging (Bootstrap Aggregation) To reduce overfitting and increase prediction stability, Random Forest uses a technique called bootstrap aggregation, or bagging. In this approach, multiple bootstrap samples are drawn from the original dataset by random sampling with replacement. A separate decision tree is then trained on each bootstrap sample. The final prediction is obtained by aggregating the results of all individual trees: averaging the outputs in regression tasks or applying majority voting in classification tasks. This ensemble method reduces the model's variance, its sensitivity to fluctuations in the training data, without significantly increasing bias [38, 33].

Randomness in Feature Selection Random Forest further introduces randomness in the feature selection process. At each node split, instead of evaluating all features, it considers only a randomly selected subset. This decorrelates the trees and makes the overall ensemble more diverse, which is critical for improving generalization [32].

Advantages of Random Forest Random Forest offers several advantages that make it particularly suitable for educational prediction tasks. It exhibits robustness against overfitting by averaging across multiple trees, leading to better generalization compared to individual decision trees. The algorithm also naturally provides feature importance estimates, ranking features based on how often and how effectively they are used to split the data [33]. Furthermore, Random Forest is non-parametric, meaning that it does not assume any specific form for the underlying data distribution. Its ability

to handle high-dimensional data makes it especially effective in scenarios involving many features or mixed data types [16, 7].

Key Concepts

Three fundamental concepts are critical to understanding Random Forest. Overfitting refers to a situation where a model learns noise from the training data, resulting in poor performance on new, unseen data. Variance measures the extent to which a model's predictions would vary if it were trained on different subsets of the data. Bias, in contrast, represents the error introduced by overly simplistic assumptions in the learning algorithm. Random Forest strikes an effective balance between bias and variance by combining deep decision trees trained on different samples of data, leading to better overall generalization.

Limitations

Despite its strengths, Random Forest has some limitations. Although individual decision trees are easy to interpret, an ensemble of hundreds of trees can become complex and difficult to explain to non-technical stakeholders. Additionally, training and predicting with a Random Forest can be computationally intensive, especially with large datasets.

Random Forest offers a powerful yet relatively user-friendly approach to predictive modelling. By combining the predictions of multiple decision trees trained on bootstrapped datasets and randomized features, it reduces overfitting and improves generalization. These characteristics make it a preferred algorithm in a wide range of applications, including education-related predictive modelling [7, 16].

3.2.2. XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm designed for structured (tabular) data. It belongs to the family of ensemble learning methods and is particularly known for its predictive accuracy, speed, and scalability [34, 33].

At its core, XGBoost builds on decision trees, but instead of creating many independent trees and aggregating them like in Random Forest (bagging), it constructs trees sequentially in a process known as *boosting*. In boosting, each new tree is trained to correct the mistakes, or residual errors, made by the previous trees. A decision tree recursively splits the dataset based on feature values to minimize error at each step. In the boosting framework, trees are added one at a time, with each tree focusing on reducing the overall loss of the model. The final prediction is then made by combining the outputs of all trees.

During training, XGBoost works iteratively: it first computes the residuals, which are the differences between the actual target values and the model's current predictions. A new decision tree is then fitted to these residuals. The predictions from the new tree are added to the existing model in order to update it and reduce the total error. This process is repeated multiple times, progressively refining the model's accuracy with each iteration.

What sets XGBoost apart from traditional boosting algorithms is its use of regularization and optimized training strategies. Regularization techniques such as L1 (Lasso) and L2 (Ridge) penalties are incorporated directly into the objective function to

prevent overfitting and control model complexity [38, 33, 32]. L1 regularization adds the absolute values of the model's coefficients to the loss function, encouraging sparsity by shrinking some coefficients to zero and effectively performing feature selection [33]. L2 regularization adds the squared values of the coefficients, discouraging overly large weights and leading to smoother and more generalizable models [32].

In addition to regularization, XGBoost includes several other optimization strategies. Shrinkage, controlled by the learning rate, ensures that each new tree makes only a small contribution to the overall model, allowing gradual improvements and reducing the risk of overfitting [38]. Tree pruning is also employed to remove branches that do not significantly enhance predictive performance. Furthermore, XGBoost supports parallel computation during tree construction, considerably speeding up the training process [34].

Regarding model evaluation, XGBoost optimizes a specified loss function, such as mean squared error in regression tasks, using the gradient descent method. At each step, the model calculates the gradient (indicating how much the prediction needs to change), fits a new tree to predict these gradients (residuals), and updates its predictions accordingly. This iterative procedure explains why the method is termed *gradient boosting*.

Despite its complexity, XGBoost integrates well with interpretability tools. It provides built-in feature importance scores, helping users understand which features most influence predictions. Moreover, it is fully compatible with SHAP (SHapley Additive exPlanations) [35, 37], allowing both global and local model interpretability to be explored.

XGBoost has gained popularity not only for its predictive power but also for its practical advantages. It consistently delivers top-tier performance in both academic competitions and real-world tasks, offers fast computation optimized for large datasets, and provides flexibility for both classification and regression problems.

Overall, XGBoost is considered a state-of-the-art algorithm, particularly well-suited for problems involving structured data where high predictive accuracy, speed, and interpretability are required. Its robust optimization techniques and compatibility with explainability frameworks make it an ideal choice for educational data mining tasks such as predicting time to graduation.

3.2.3. Comparison Between Random Forest and XGBoost

The theoretical and practical strengths of both Random Forest and XGBoost make them highly suitable choices for predicting educational outcomes. Random Forest offers stability, ease of interpretation, and robustness against noise through its use of bagging and decision tree ensembling [38, 32]. Its ability to manage high-dimensional datasets and naturally estimate feature importance makes it particularly effective for structured educational data analysis.

In contrast, XGBoost leverages gradient-based optimization and advanced regularization techniques to achieve exceptional predictive accuracy while maintaining strong generalization capabilities [34, 33]. Its incorporation of shrinkage, tree pruning, and parallel computation further enhances both the efficiency and reliability of model training. Moreover, both algorithms support model interpretability, with

built-in feature importance measures and compatibility with SHAP (SHapley Additive exPlanations) [37, 35, 36], enabling detailed insights into the driving factors behind model predictions.

The inclusion of both Random Forest and XGBoost in this study enables a comprehensive evaluation of ensemble learning strategies, comparing bagging-based and boosting-based approaches. This comparative framework provides a robust basis for assessing prediction performance, interpretability, and practical applicability in the context of structured educational datasets.

3.3. Model Optimization and Evaluation

Hyperparameter optimization plays a critical role in improving the predictive performance and generalization ability of machine learning models. Several techniques exist for this purpose, including manual tuning, grid search, randomized search, Bayesian optimization, and evolutionary algorithms. Each method offers trade-offs between computational efficiency, thoroughness of search, and ease of implementation.

- **Random Forest:** Although grid search provides an exhaustive approach to hyperparameter tuning, it becomes computationally expensive in high-dimensional search spaces. In contrast, `RandomizedSearchCV` samples a fixed number of parameter combinations from the defined distribution, offering a more computationally efficient yet effective alternative. It has been shown to perform comparably to grid search with significantly fewer iterations, especially for tree-based models [39]. More advanced methods such as Bayesian optimization or evolutionary algorithms could also be applied, but their added complexity may not yield proportional benefits for Random Forest in most regression tasks.
- **XGBoost:** For this model, `BayesianOptimization` was selected due to its ability to efficiently navigate large and complex hyperparameter spaces. Unlike random or grid search, Bayesian optimization builds a probabilistic model of the objective function and selects hyperparameters that are likely to yield better results, balancing exploration and exploitation [40]. While alternatives such as random search or genetic algorithms are also applicable, Bayesian methods are particularly well-suited for boosting frameworks, where optimal tuning can substantially improve model performance and control overfitting.

In the following subsections, the evaluation metrics and validation techniques used for assessing machine learning model performance, such as MAE, MSE, R^2 , cross-validation, and residual analysis are explained in detail.

3.3.1. Evaluation Metrics

To evaluate the trained models, standard regression metrics were selected: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of

Determination (R^2). These metrics provide complementary perspectives on model accuracy and error characteristics [41, 42]. Other metrics such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), or Adjusted R^2 are also commonly used in regression analysis. However, MAE, MSE, and R^2 were chosen due to their interpretability, widespread use, and alignment with previous research in educational predictive modelling [4, 19, 16].

- **Mean Absolute Error (MAE)** represents the average magnitude of prediction errors, offering a clear and intuitive measure of model accuracy. It is calculated as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (1)$$

- **Mean Squared Error (MSE)** penalizes larger errors more severely, making it sensitive to high-variance predictions and thus suitable for detecting outlier influence:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (2)$$

- **R-squared (R^2)** indicates the proportion of variance in the dependent variable that is explained by the model, serving as a normalized indicator of goodness-of-fit:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3)$$

Together, these metrics allow for a holistic understanding of model performance, balancing simplicity, sensitivity to error magnitude, and explanatory power.

3.3.2. Validation Techniques and Residual Analysis

Cross-validation is a widely used model evaluation technique that helps estimate the generalizability of machine learning models to unseen data. It mitigates overfitting by testing the model on multiple train–validation splits, ensuring that performance estimates are not biased by any particular partition of the dataset.

In this study, five-fold cross-validation was employed. This technique partitions the training data into five equally sized subsets (folds). In each iteration, one fold is held out as a validation set while the remaining four are used to train the model. This process is repeated five times, with each fold serving once as the validation set. The resulting performance metrics are then averaged to produce a stable and unbiased estimate of model accuracy [43].

Several cross-validation strategies exist, including leave-one-out cross-validation, repeated k-fold cross-validation, and stratified sampling. Five-fold cross-validation was selected due to its favorable balance between computational efficiency and performance stability. Compared to leave-one-out approaches, it reduces variance without significantly increasing computational cost, making it a commonly recommended practice in regression modelling [43, 33].

Beyond model evaluation, cross-validation also plays a critical role in hyperparameter tuning, enabling consistent model selection and comparison while minimizing the risk of overfitting to a specific data split.

Residual analysis is a diagnostic tool used in regression modelling to assess the quality of predictions and detect potential model misspecifications. Residuals are defined as the differences between observed values and their corresponding predicted values. Mathematically, the residual for the i th observation is computed as:

$$e_i = y_i - \hat{y}_i \quad (4)$$

where y_i is the actual target value, \hat{y}_i is the predicted value, and e_i is the residual. In a well-fitted model, residuals should be randomly distributed around zero, indicating that the model has captured the underlying patterns in the data without systematic bias.

Visual inspection of residuals, typically through scatter plots of residuals versus predicted values, helps detect non-random patterns that may suggest problems such as heteroscedasticity, autocorrelation, omitted variables, or nonlinearity. An ideal residual plot exhibits homoscedasticity (constant variance) and no apparent structure [44].

Residual analysis complements quantitative evaluation metrics like MAE, MSE, and R^2 by providing qualitative insights into prediction errors. While performance metrics summarize the average magnitude of error, residual plots allow for examination of error distribution and variance across the prediction space, supporting model validation and improvement.

3.4. Feature Importance Techniques

Feature importance techniques are essential for interpreting machine learning models, particularly in applications where understanding the contribution of each input variable is critical for decision-making. This study considers two widely used methods for assessing feature relevance: model-based importance and permutation importance.

Model-Based Importance: In tree-based models such as XGBoost, feature importance can be derived from the internal structure of the model. These importance scores are typically computed based on metrics such as the frequency of splits, the average gain in information, or the reduction in impurity associated with a feature. Features that frequently and effectively split nodes contribute more to reducing prediction error and are assigned higher importance scores [38, 34]. This method is efficient and well-integrated into the training process but may be biased in the presence of correlated or high-cardinality features.

Permutation Importance: Permutation importance offers a model-agnostic approach by assessing how shuffling each feature's values affects model performance. After permuting a feature, the model's predictions are recomputed, and the resulting drop in accuracy or increase in error is measured. A larger decline in performance implies a greater dependence on that feature. This method is particularly robust in revealing true feature influence, especially when dealing with correlated inputs, and it allows for consistent comparison across different model types [37].

By combining both approaches, model-based and permutation importance, it is possible to obtain a more nuanced understanding of feature relevance. While model-based scores provide insight into internal model mechanics, permutation-based estimates offer a broader and less biased view of feature impact.

3.5. Model Interpretation with SHAP

Interpretability is a critical component in machine learning, especially in high-stakes domains such as education, where stakeholders must understand and trust the model's decisions. SHAP (SHapley Additive exPlanations) is a unified interpretability framework that attributes the contribution of each input feature to a model's prediction based on principles from cooperative game theory [35].

SHAP values are grounded in Shapley values, originally developed by Lloyd Shapley to fairly allocate payouts among players in a cooperative game. In the context of machine learning, the "players" are the input features, and the "payout" is the model's predicted output. SHAP calculates the average marginal contribution of each feature across all possible feature coalitions, providing a theoretically sound and consistent explanation of model behavior.

One of SHAP's key advantages is its ability to provide both local and global interpretability. Local explanations reveal how each feature contributes to individual predictions, while global explanations summarize overall feature importance across the dataset. Furthermore, SHAP satisfies desirable properties such as local accuracy, consistency, and missingness, which distinguish it from other explanation techniques [37].

For tree-based models such as XGBoost, the TreeSHAP algorithm enables fast and exact computation of SHAP values, making the method computationally feasible even on large datasets [36]. This efficiency, combined with its theoretical rigor, makes SHAP particularly suitable for interpreting complex models while maintaining transparency and accountability.

3.6. Python Libraries

The Python programming environment offers a rich ecosystem of libraries that support the entire machine learning lifecycle, from data preprocessing and modelling to evaluation and interpretability. The following libraries were used for implementing the techniques described in this study:

- **scikit-learn (sklearn):** A widely used machine learning library that provides tools for data splitting, preprocessing, classical algorithms, and model evaluation. Its consistent API design and modular architecture make it a foundation for supervised learning workflows [32].
- **XGBoost:** A high-performance gradient boosting library optimized for speed and accuracy. It supports parallel processing, regularization, and advanced tree pruning techniques, making it a preferred choice for structured prediction tasks [34].

- **pandas:** A powerful library for data manipulation and analysis, offering high-level data structures such as DataFrames. It enables efficient handling of tabular data and supports integration with other libraries in the Python data science stack.
- **NumPy:** A core numerical computing library that provides array structures and vectorized operations, serving as the computational backbone for many higher-level tools, including pandas and scikit-learn.
- **matplotlib and seaborn:** These libraries support data visualization and exploratory data analysis. Matplotlib offers low-level control over plots, while seaborn builds on it to provide aesthetically pleasing statistical visualizations.
- **SHAP:** A specialized interpretability library that implements SHapley Additive exPlanations (SHAP) to explain the output of machine learning models. It supports both local and global interpretability and integrates efficiently with tree-based models such as XGBoost [35, 36].

These libraries collectively provide a comprehensive and flexible environment for implementing, evaluating, and interpreting predictive models in educational data science.

3.7. Data Preprocessing and Additional Calculations

Effective data preprocessing is a foundational step in any machine learning project, as the quality of input data directly influences model performance and interpretability. This process involves cleaning, transforming, and enriching raw data to make it suitable for analysis and modelling [33, 32].

In this study, data preprocessing and feature engineering were primarily conducted using Power BI and DAX (Data Analysis Expressions). Power BI offers an interactive environment for handling structured educational data, while DAX allows for the creation of calculated columns, metrics, and aggregations, making it well-suited for large institutional datasets [13, 14, 12].

Key preprocessing concepts include:

- **Data Cleaning and Transformation:** Addressing inconsistencies in the data, handling missing values, and standardizing data formats. These steps ensure that models are trained on reliable and interpretable inputs [4].
- **Feature Engineering:** Constructing new variables from raw data to better capture meaningful patterns and relationships. This may include calculations such as averages, running totals, time gaps, or cumulative statistics, which are especially useful in longitudinal educational data [16, 6].
- **Outlier Detection and Handling:** Identifying anomalous values that could distort model training and evaluation. Common techniques include z-score analysis, IQR filtering, or domain-specific thresholds [30].

- **Aggregation and Structuring:** Summarizing and reshaping data, often into tabular or matrix form, to align with machine learning input requirements. This step improves model efficiency and ensures that temporal or grouped data is represented accurately [8].

Preprocessing not only improves data quality but also supports the interpretability and stability of machine learning models, especially in educational analytics, where time-based patterns and institutional factors are often central to prediction tasks.

3.8. Tools for Visualization

Visualization plays a critical role in both exploratory data analysis and the communication of machine learning results. In this study, Power BI served as the primary platform for building interactive dashboards and visual reports. Its user-friendly interface and integration with Data Analysis Expressions (DAX) allow for dynamic filtering, aggregation, and custom metric visualization [13, 14].

Power BI is particularly well-suited for working with structured educational data, enabling end-users to explore complex datasets without requiring programming expertise. It supports seamless integration with external tools and scripting environments, such as R and Python, enhancing its flexibility for advanced visualization tasks [12].

To extend its native capabilities, R scripts were embedded within Power BI to enable custom charting and interactive plots. This integration allowed for the creation of visual elements beyond Power BI's built-in options, supporting richer analytical narratives and tailored communication of results [6].

Together, Power BI and R offer a robust visualization framework that balances accessibility with analytical depth, making them effective tools for communicating insights in data-driven educational research.

3.9. Tools for Data Export and Machine Learning Integration

The integration of data transformation tools with external machine learning environments is a crucial step in modern analytical workflows. This study employed a seamless export mechanism from Power BI to Python-based modelling environments to maintain consistency across visualization and prediction tasks.

As introduced in Section 3.1, **DAX Studio** was used to extract feature-engineered, tabular datasets from Power BI. By enabling direct querying of Power BI models using DAX, DAX Studio ensures reproducibility and fidelity in data export [13].

The exported datasets were subsequently processed and modelled in **Google Colab**, a cloud-based development platform described earlier in Section 3.1. Colab supports Python-based machine learning workflows and provides an efficient environment for collaborative experimentation, training, and evaluation [15].

This toolchain supports the modular structure of the analytical workflow: preprocessing and feature calculation in Power BI, data extraction via DAX Studio, and predictive modelling in Colab using Python. Such integration enables consistent

data logic and traceability between the business intelligence layer and the machine learning models.

3.10. Summary: General Machine Learning Workflow

This section concludes the Method and Tools chapter by summarizing the overall workflow applied in this thesis. It brings together the model selection, evaluation techniques, interpretability tools, and implementation platforms described in previous sections. Figure 1 presents a visual overview of this machine learning pipeline. This modular framework ensures that models are both generalizable and interpretable, and it supports iterative improvements throughout the modelling process.

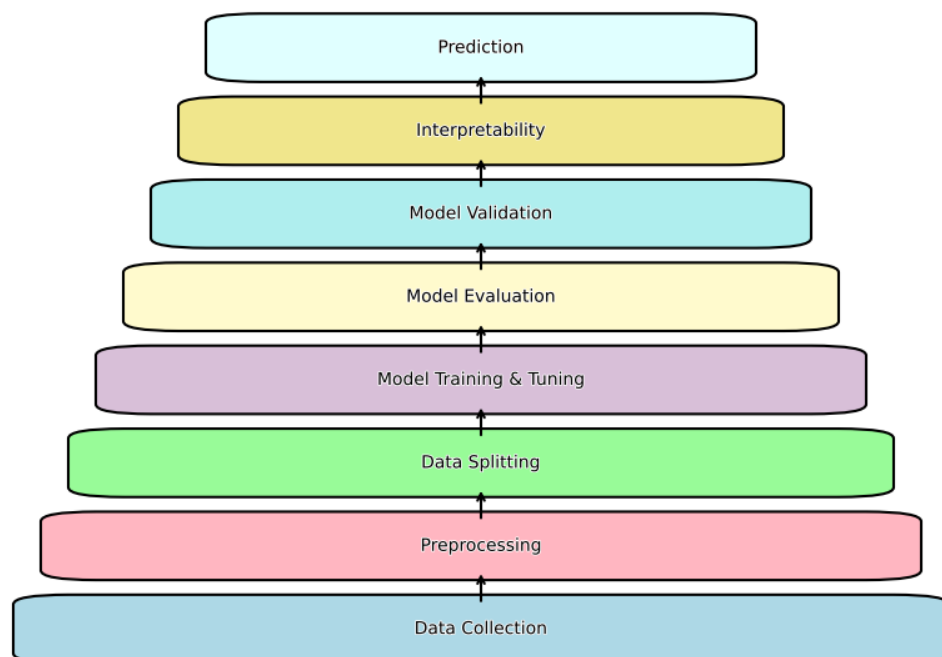


Figure 1. Overall machine learning workflow implemented in this thesis.

The core stages of a typical supervised regression workflow are summarized below:

- **Data Splitting:** The dataset is partitioned into separate training and testing subsets to evaluate model performance on unseen data. Stratified sampling may be used to preserve class distributions when appropriate. This step supports fair evaluation and guards against overfitting [33, 43].
- **Model Training and Hyperparameter Optimization:** Models are trained on the training data and optimized using techniques such as Randomized Search or Bayesian Optimization. These strategies aim to identify hyperparameter combinations that maximize predictive performance without incurring excessive computational costs [39, 40].

- **Model Evaluation:** Performance is assessed using standard regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). These metrics offer complementary insights into prediction accuracy, robustness, and variance explanation [41, 42, 19].
- **Residual Analysis:** Residual plots are used to diagnose model fit and identify potential issues such as heteroscedasticity, nonlinearity, or systematic bias in predictions. This step complements numerical metrics with qualitative visual diagnostics [44].
- **Feature Importance and Interpretability:** The influence of input variables is examined using model-based importance scores, permutation importance, and SHAP values. These tools provide transparency and support accountability in educational modelling scenarios [37, 35, 36, 17, 16].

This structured workflow underpins the modelling approach presented in later sections, ensuring that the techniques applied are both theoretically grounded and consistent with best practices in machine learning research.

4. DATA AND INSIGHTS

The dataset used in this study pertains to students from the University of Oulu across several faculties, from 2004 to 2023. While not publicly available, this dataset offers a comprehensive view of academic progress for 27,754 students over nearly two decades. It includes detailed academic information for each student from their registration date to graduation, making it a valuable resource for analyzing academic trajectories and predicting the time to graduation.

4.1. Dataset Overview

The dataset consists of 370,144 rows and 7 columns, providing longitudinal information to students in various degree programs, including bachelor's, master's, and combination degrees. Each row represents an educational period for a student, capturing key details about their academic progress. An overview of the dataset structure, including descriptions, types, and notes for each field, is presented in Table 1.

At the University of Oulu, the academic year is divided into two terms, and each term is further split into two periods, resulting in four instructional periods per academic year. Each “period” in the dataset corresponds to approximately 7–8 weeks of study and is used as a temporal unit for tracking credit accumulation and study progression.

The theoretical duration for completing a combined bachelor's and master's degree is five academic years (three years for the bachelor's degree and two for the master's degree), comprising a total of 300³ ECTS credits⁴. However, actual study durations may vary due to factors such as part-time enrollment or study gaps.

4.1.1. Initial State of the Dataset

All columns were initially in a general format, which required conversion into more usable data types. For example, dates were converted into standard date formats, and textual representations were parsed into numerical or categorical variables where necessary. This transformation was crucial for ensuring compatibility with machine learning models and analytical tools.

Challenging Fields: The Educational Year and Period field presents significant preprocessing challenges. Originally stored as “op-year-period,” this field required parsing and feature transformation operations to derive new columns (see Table 2).

One of the derived variables, *Intake Delay*, captures the time gap between a student's official registration and the actual start of their educational activity. While this may be uncommon in other educational systems, such delays at the University of Oulu

³For context, the official degree structure at the University of Oulu requires 180 ECTS for the Bachelor's degree and 120 ECTS for the Master's.

⁴Credits in this thesis refer to ECTS (European Credit Transfer and Accumulation System) credits, which are the standard measure of workload in European higher education. One ECTS credit typically corresponds to 25–30 hours of student work, and a full-time academic year usually comprises 60 ECTS credits.

Table 1. Overview of Primary Fields in the Dataset

Field Name	Description	Type	Notes
Student	Unique identifier for each student	Text	Used to uniquely identify and track each student across all records.
Study Right	Numeric representation of the student's study program	Text	Excluded from analysis due to irrelevance to predictive modelling objectives.
Registration Date	Date when the student registered for a study program	Date	Used to calculate the starting point of each student's academic timeline.
Bachelor Graduation Date	Date of bachelor's degree completion	Date	Applicable to bachelor's and combined degree students.
Master Graduation Date	Date of master's degree completion	Date	Applicable to master's and combined degree students.
Educational Year and Period	Academic year and period (e.g., "2-1,..., 5-4")	Text	Transformed during preprocessing to create standardized temporal identifiers.
Total Credits	Number of credits earned during a specific period	Decimal Number	Key metric for tracking academic progress per period.

can occur due to postponed enrollment, flexible starting periods, or administrative processes. This variable is measured in educational periods to maintain consistency with other time-based fields.

Table 2. Derived Fields from Educational Year and Period

Field Name	Description
Educational Year	Represents the academic year extracted from the original field.
Educational Period	Indicates the specific term or period within the academic year.
Educational Date	Standardized date format used for chronological alignment and analysis.
Education Length	Derived feature quantifying the cumulative educational duration for each student.
Educational Gap	The gap between two consecutive educational dates, equal to or longer than one period, measured in educational length units.
Intake Delay	Time difference between registration and the actual start of studies.

These transformations were essential for creating a data set that aligned with the study objectives and facilitated accurate modelling. This comprehensive data set, with its detailed academic records and longitudinal structure, serves as the foundation for predictive modelling and exploratory analysis in this thesis.

4.1.2. Statistical Overview

The credit overview before preprocessing is shown in Table 3.

Table 3. Descriptive Statistics of Credits

Min	Max	Average	Standard deviation	Variance	Median
0	525	16	13	173	13

In addition, the date ranges in the dataset, summarized in Table 4, provide context for the temporal coverage of academic activity.

Table 4. Date Ranges in the Dataset

Event	Date
First Register Date	08/01/2004
Last Register Date	08/01/2022
First Bachelor Graduation Date	09/22/2005
Last Bachelor Graduation Date	08/31/2023
First Master Graduation Date	01/25/2006
Last Master Graduation Date	08/31/2023

An analysis of registration trends revealed variations in the number of students registered per year, as shown in Figure 2. The highest registration year was 2005 with 1869 registrations, while the lowest was 2016 with 1392 registrations.

4.2. Data Preprocessing

The preprocessing phase involved converting the raw dataset into a structured and analyzable format while addressing data inconsistencies and calculating additional fields for the analysis. The steps included transforming data formats, handling anomalies, and generating calculated fields.

Data Transformation and Parsing

Initially, the dataset contained general formats for certain fields, which were converted to specific formats to facilitate analysis. For example, the "op-year-period" field, which combined year and period information, was parsed into two separate numerical fields: Educational Year and Educational Period. This transformation enabled better tracking and analysis of student academic progression.



Figure 2. Yearly Registration before preprocessing

Handling Duplicate and Ambiguous Records

A subset of 1,843 students had multiple registration dates under the same study right. Due to the lack of clarity on which records corresponded to specific majors, these records were excluded from the analysis. The ambiguity in these entries made it challenging to ensure the reliability of the calculated features.

Outlier Detection and Removal

To address anomalies, the dataset was visualized to examine the distribution of maximum credits taken per period by the students. Based on Figure 3 the analysis revealed that 97.15% of students had a maximum of 100 credits or fewer per period, as shown in the left chart. Students with more than 100 credits per period were considered outliers and removed from the dataset, the right chart breaks down the distribution for remaining values. This approach ensured that the analysis focused on records that aligned with realistic academic progression patterns.

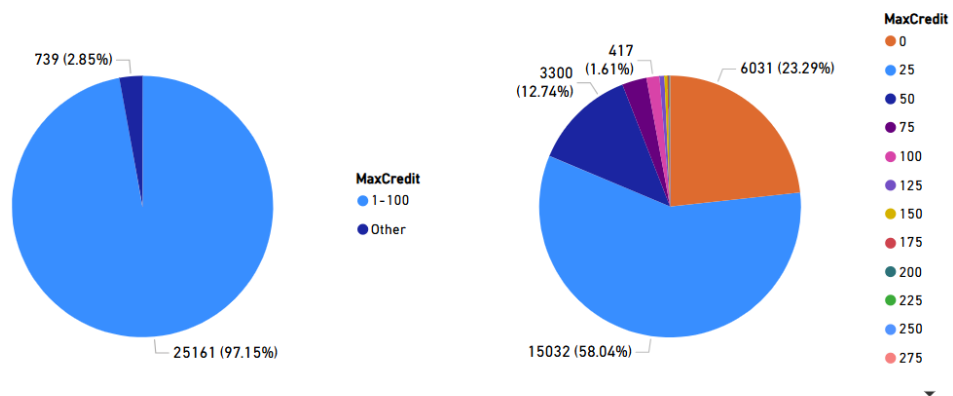


Figure 3. Number of students with more than 100 credits per period

Retention of Students with Special Cases

The dataset included 2,447 students with an Educational Year of 0. These records were retained, as they likely represented students who transferred credits from other institutions or obtained credits through open university education. Their inclusion was essential to maintain a comprehensive analysis of the student population.

Minimal Cleaning, Focus on Calculations

While the dataset required minimal cleaning, the preprocessing phase primarily emphasized generating calculated fields for subsequent analysis. These fields included cumulative credits, education length, intake delays, and other derived metrics that provided deeper insights into students' academic progress and graduation timelines.

Following two cleaning steps, the dataset now includes 325,600 records representing 25,104 students, ensuring a more streamlined and interpretable dataset for further analysis.

4.3. Calculated Fields and Explanatory Analysis

Degree The Degree field was calculated based on the presence of graduation dates in the dataset:

Bachelor's Degree Students: Students with only a Bachelor's graduation date.

Master's Degree Students: Students with only a Master's graduation date.

Combined Degree Students: Students with both Bachelor's and Master's graduation dates.

As this study focuses on students enrolled in combined-degree programs, only **graduated students**, those with both Bachelor's and Master's graduation dates, were selected for further analysis. The remaining students were excluded from subsequent steps. To illustrate the distribution of students across these categories, the following visualization is provided:

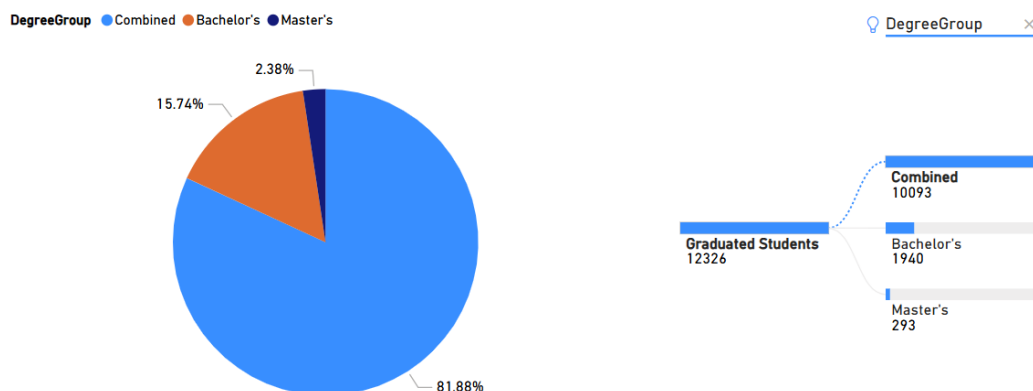


Figure 4. Distribution of Graduated Students Based on Degree Type

As shown in Figure 4, the pie chart on the left provides an overview of the proportions of graduated students across three-degree categories: the majority of graduated students (81.88%) are part of combined degree programs, while 15.74% have only Bachelor’s degrees and 2.38% have only Master’s degrees. The diagram on the right further breaks down these groups, showing the numerical distribution: Combined Degree with 10,093 Students, Bachelor’s Degree with 1940 Students, and Master’s Degree with 293 Students. This visualization highlights the focus of this study, students from the combined degree program, who make up the largest group of graduates. These students were selected as the target group for further analysis.

Education Date The Education Date is foundational for analyzing gaps between educational periods and calculating other fields.

Education Length The Education Length is calculated based on the difference between the first education date (start of studies) and the last education date for each student, and converting it into years and periods using a predefined unit system. It represents the total time a student spent in their educational program, which helps capture the duration of study, including gaps, for combined degree students.

The Education Length field represents the duration of a student’s education, measured in standardized units. One unit equals one academic year, with each period equivalent to 0.25 units (e.g., 3.75 represents three years and three periods). While the theoretical duration of combined Bachelor’s and Master’s degrees at the University of Oulu is five years (three for Bachelor’s and two for Master’s), actual study times may vary due to factors such as study gaps, part-time enrollment, or delays in course completion.

The following analysis demonstrates the typical graduation timelines for combined degree students while highlighting the impact of incomplete data for more recent cohorts on interpreting trends.

The pie chart in Figure 5 on the left provides an overview of the distribution of

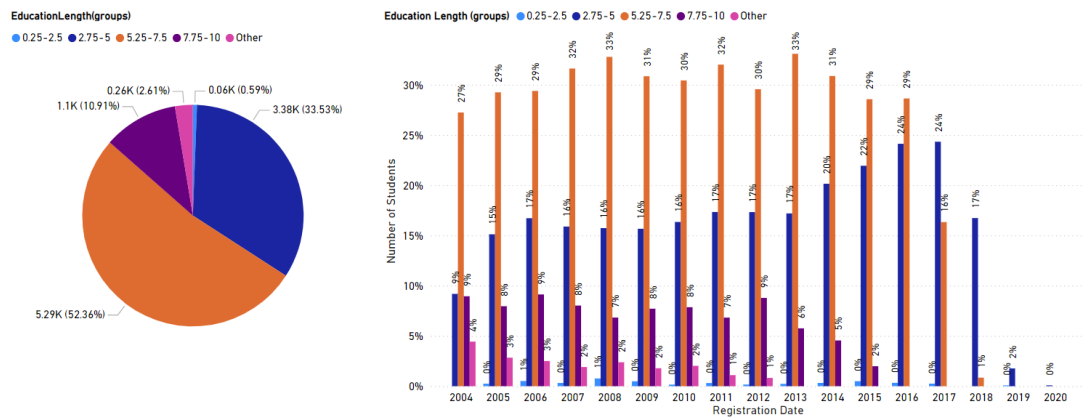


Figure 5. Distribution of Education Length and Graduation Rate by Registration Year and Education Length

Education Length for combined degree students. The majority of them (52.36%) complete their programs within 5.25-7.5 years, followed by 2.75-5 years (33.53%). Only 0.59% of students graduated in 1-2.5 years, likely representing students who transfer significant credits or accelerate their studies, while 10.91% fall into the 7.75-10 years bin, and 2.61% are categorized as "Other".

The bar chart, on the right in Figure 5 presents the distribution of Graduation Rate by Education Length across Registration Years. The largest group in most years is consistently the 5.25-7.5 years bin, underscoring its status as the standard timeframe for graduation.

For registration years after 2017, observed decreases in longer education durations (e.g., 7.75-10 years) reflect incomplete data coverage, as students from these cohorts are still in progress at the time of this study.

Comparison of Education Length by Graduation Year vs. Registration Year

It is crucial to examine how Education Length trends evolve over time among graduated students. The choice of reference year, whether based on graduation year or registration year, significantly influences the observed patterns and potential biases.

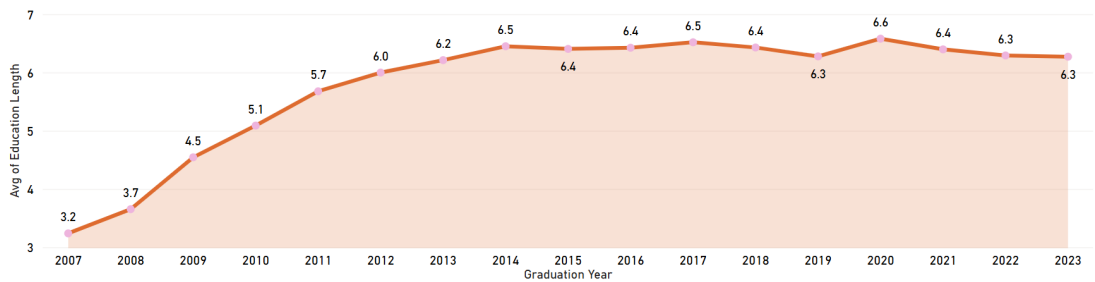


Figure 6. Average Education Length trend over Graduation Years among graduated students

The line chart in Figure 6 presents the average Education Length by graduation year. It shows a steady increase from 2007 to 2014, followed by a plateau of around 6.3 to 6.6 years, meaning that students who graduate after this period follow a consistent study duration. However, this graduation-year-based analysis tends to underestimate study length in the early years (2007–2010) because only students who registered in 2004 or later are included. Therefore, the dataset in the early years does not include those who graduated for a longer duration. Despite this limitation, from 2014 onward, the trend stabilizes between 6.3 and 6.6 years, even in 2023, the results remain stable, which suggests that this method captures a complete picture of graduation durations for most groups.

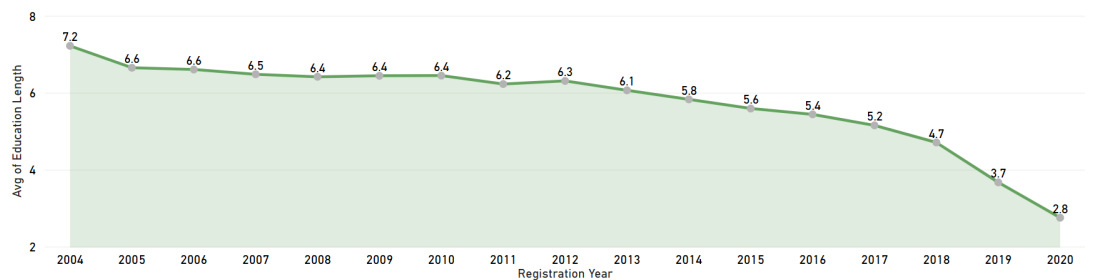


Figure 7. Average Education Length trend over Registration Years among graduated students

The registration year-based analysis follows a clear downward trend, but this pattern is influenced by data limitations.

On the one hand, older graduations (who registered between 2004–2012) have

complete graduation data, so we see higher education lengths (e.g., 7.2 years for 2004 registrants). On the other hand, more recent registrants (2016 onward) show shorter education lengths, but this does not necessarily mean that students are graduating faster. Instead, it reflects incomplete graduation data for these groups. For example, many students from the 2018 cohort are still studying. Since the analysis only includes combined degree graduates, those who take longer to finish are not yet counted, artificially lowering the average study length. This means that the registration-year approach systematically underestimates education length for recent years.

Intake Delay The Intake Delay field represents the gap between the Registration Date and the first Educational Date for each student. This field was calculated to identify how long students took to begin their studies after registering. The Intake Delay is expressed using the same Educational Length units as the Education Length field (e.g., 1 = 1 year, 0.25 = 1 period, 0.5 = 1 term).

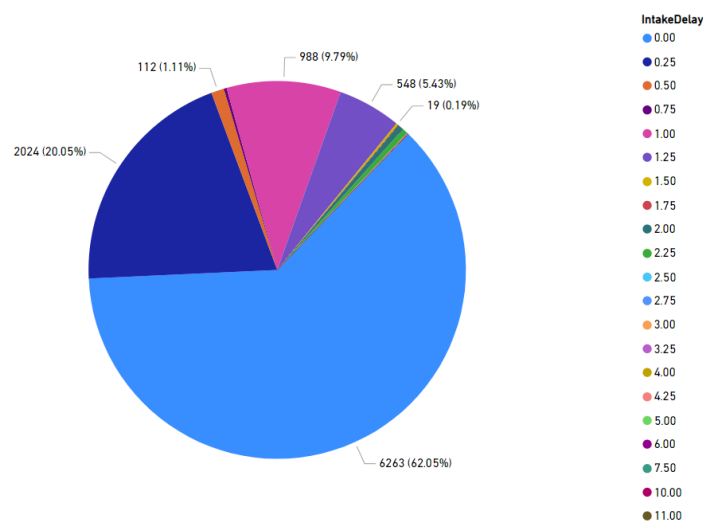


Figure 8. Distribution of Intake Delay among Combined Degree Graduates

The pie chart in Figure 8 presents the distribution of Intake Delay values across all combined degree students in the dataset. The delay is measured in Educational Length units, where 0.25 represents one period, 0.5 represents one term, and so on. Here are the key observations from the chart:

Most of the students, 62.05% (6263 students), started their studies immediately after registration, with no intake delay. About 20% (2024) of students started their studies within one period (0.25 units) after registration. Around 10% (988) of students had a delay of 1 year, and about 5% (548 students) started after 1 year and 1 period (1.25 units). The sharp drop-off after 1.25 units shows that extended delays are uncommon. Most students either begin their education immediately or within a relatively short time frame after registration.

Gap Between The Gap Between field represents the time interval between two consecutive educational dates (i.e., dates when the student earned credits) for each student, measured in units consistent with the education length (e.g., 0.25 for one period, 0.5 for one term, 1 for one year). This field provides insight into study patterns, such as periods of inactivity. A Gap Between value of 1.0 or higher indicates that the

student did not earn any credits for at least one full academic year between those two dates. This does not necessarily mean they were unregistered, but it does indicate a period of inactivity in credit accumulation, which may reflect time off, leave of absence, or disengagement from studies.

Total Gap The Total Gap field aggregates the individual gaps, with a key distinction between graduated and still-studying students:

Graduated Students: The 'Total Gap' is the sum of all 'Gap Between' intervals throughout their education. **Still-Studying Students:** The 'Total Gap' includes the sum of all 'Gap Between' intervals plus an additional 'Last Leave,' representing the time from the last recorded educational date to the current dataset update time.

This conditional calculation ensures that the 'Total Gap' accurately reflects the cumulative inactivity for both graduated and still-studying students, accounting for ongoing delays in the latter group. As shown in the donut chart on the left side of

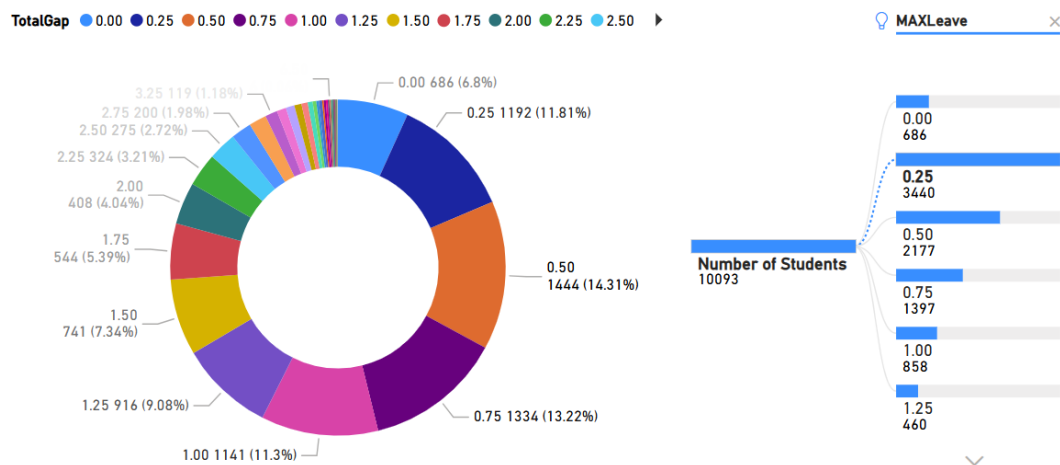


Figure 9. Distribution of Total Gap and Maximum Leave for Combined Degree Graduated Students

Figure 9, the most common Total Gap is 0.5 years (1 term or two consecutive periods), accounting for 14.31% (1,444 students). Other significant total gaps include 0.75 years (13.22%, 1,334 students), 0.25 years (11.81%, 1,192 students), and 1 year (11.3%, 1,141 students).

A zero-gap (0.00 years) accounts for only 6.8% (686 students), indicating that a small minority of students completed their education without any breaks. However, the overall distribution highlights that the majority of combined degree students experience short to medium gaps, with longer gaps becoming increasingly uncommon.

The bar chart on the right side of Figure 9 illustrates the Maximum Leave (i.e., the longest single gap) experienced by these students:

The largest group of students (34.08%, 3,440 students) experienced a maximum leave of 0.25 years (1 period). 21.56% (2,177 students) had a maximum leave of 0.5 years (two consecutive periods). 13.84% (1,397 students) had a maximum leave of 0.75 years (three consecutive periods). Longer gaps (e.g., 1 year, 1.25 years, or more) are progressively less frequent, with only 4.55% (460 students) experiencing a maximum leave of 1.25 years.

The relatively small number of students with no gaps underscores the frequent

occurrence of academic gaps in combined degree programs. Shorter gaps (e.g., 0.25–1.25 years) are common and may reflect planned academic breaks or minor interruptions due to parental leave, military service, or illness.

Comparison of Total Gap by Graduation Year vs. Registration Year To understand how study gaps have evolved, we examine two perspectives: Graduation Year-based Analysis and Registration Year-based Analysis. Both approaches provide valuable insights but also introduce unique biases due to data limitations, which will be discussed in the following analysis.

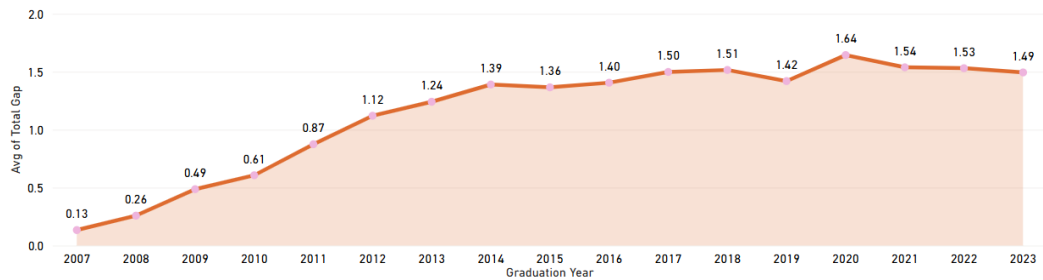


Figure 10. Total Gap trend over Graduation Years among Combined Degree Graduates

The line chart in Figure 10 visualizes the average total gap among combined degree graduates grouped by their graduation year. From 2007 to 2014, there is a steady increase in the total gap, starting from 0.13 years in 2007 and reaching 1.39 years in 2014. After this period, the total gap stabilizes between 1.39 and 1.64 years, with minor fluctuations. This trend suggests that students who graduated in earlier years had shorter gaps on average, possibly due to data limitations, as the dataset only includes students who registered from 2004 onward and started graduating in 2007. However, from 2014 onward, the results become more reliable and stable, since by this period, the dataset covers a full range of study behaviours, and no significant data are missing. The peak in 2020 (1.64 years) might be attributed to external factors such as policy changes or disruptions (e.g., COVID-19). Afterwards, the total gap stabilizes again at around 1.49 years in 2023, indicating a well-established trend in study behaviour for recent graduates. Graduation Year-based analysis is more stable for recent years (after 2014) because it includes students who completed their studies. This provides a more complete picture of actual study behaviours.

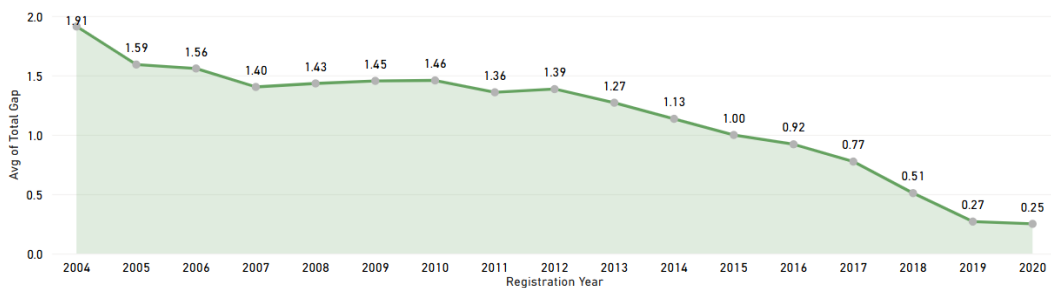


Figure 11. Total Gap trend over Registration Years among Combined Degree Graduates

The line chart in Figure 11 represents the average total gap by registration year, which follows a downward trend. The earliest cohorts (2004–2008) exhibit the highest

total gaps (1.91 years in 2004), reflecting longer study interruptions. After 2009, the total gap stabilizes around 1.4 to 1.5 years, aligning with the findings from the graduation-year-based analysis. However, a sharp decline occurs after 2014, with the most recent cohorts (e.g., 2018–2020) showing significantly shorter total gaps. This decline does not indicate that students are taking fewer breaks; rather, it is an effect of incomplete data, as many students from these cohorts are still studying. For instance, students who registered in 2018 may still be completing their degrees, and their full study gaps are not yet accounted for in the dataset. The registration Year-based analysis underestimates the total gaps for recent years (post-2014), as it excludes students who are still studying and have not yet contributed to the dataset.

Bachelor Graduation Gap The Bachelor Graduation Gap is a calculated field that represents the time elapsed between a student’s Bachelor’s graduation date and the dataset’s update time. It applies to students who graduated with a Bachelor’s degree and have not recorded any further education dates since their graduation. This gap is measured in education length units, where 1 unit corresponds to one academic year and 0.25 units represent one academic period. This metric is used to assess the time interval between Bachelor’s completion and potential continuation into a Master’s program. In the preprocessing stage, graduates with bachelor’s or master’s degrees are excluded from the analysis. In addition, students who have a recorded Bachelor’s graduation date but are still studying are excluded from the further analysis. The primary reason for this exclusion is that the dataset does not provide sufficient information on the expected time gap between Bachelor’s completion and continuation into a Master’s program.

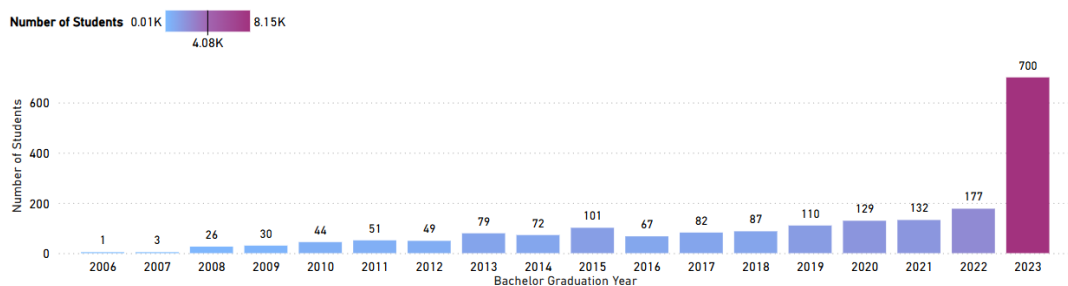


Figure 12. Distribution of Bachelor graduations by Graduation Year

As shown in Figure 12, the distribution of Bachelor’s graduation dates indicates a significant increase in recent years, making it unclear how many of these students will eventually continue their studies versus those who will not.

Furthermore, the time between Bachelor’s graduation and the dataset update time was calculated for Bachelor’s graduations to observe the gap distribution (Figure 13).

The variation in these gaps suggests that there is no clear threshold to determine whether a student will return for further studies. Since the focus of this study is on combined degree students those pursuing both Bachelor’s and Master’s degrees within the same timeframe students with a Bachelor’s degree who are still studying introduce uncertainty in modelling educational progression. For this reason, in the analysis of still-studying students, we will exclude students with a recorded Bachelor’s graduation date who are continuing their studies.

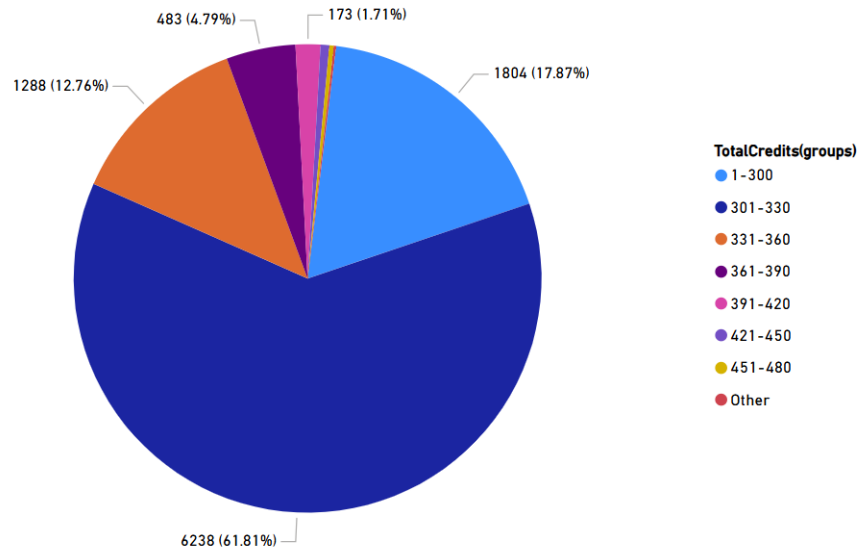


Figure 15. Distribution of total credits earned by Combined Degree Graduates

The pie chart in Figure 15 presents the distribution of total credits earned by combined degree graduates. Most students (61.81%) accumulate between 301 and 330 total credits, which aligns with the typical credit requirement for a combined Bachelor's and Master's degree. This suggests that the majority of students complete their studies within the expected credit range. A significant portion (17.87%) of students fall in the 1 to 300 credit range. This group may include students who either transferred out or had prior learning credits that were not counted in the dataset. Students earning between 331 and 360 credits make up 12.76% of the cohort. These students may have taken extra courses, electives, or specialization subjects beyond the standard credit requirement. A smaller percentage (4.79%) of students have total credits in the 361 and 390 range, and even fewer students fall into the higher credit brackets (above 391). These outliers likely represent students who either pursued additional coursework, dual-degree programs or extended their studies.

The vast majority of students complete their education within the expected standard credit range. There is a gradual decline in student count as the total credit range increases, indicating that fewer students exceed the necessary coursework. Exploratory analysis of students with exceptionally high or low total credits did not reveal noteworthy trends in study length, education gaps, or intake delays. This suggests that students accumulating unusually low or high credits do not follow distinct study patterns and might represent individual cases due to personal study choices, transfers, or curriculum variations.

Comparison of Total Credits by Graduation Year vs. Registration Year To examine how students' accumulated total credits have changed over time, two different perspectives are analyzed: Graduation Year and Registration Year.

The line chart in Figure 16 illustrates the average total credits of combined degree students who graduated each year. The trend shows an increase from 2007 to 2009, peaking at 330 credits, followed by a gradual decline and stabilization of around 317 to 320 credits between 2013 and 2020. A slight decrease is observed after 2020, reaching

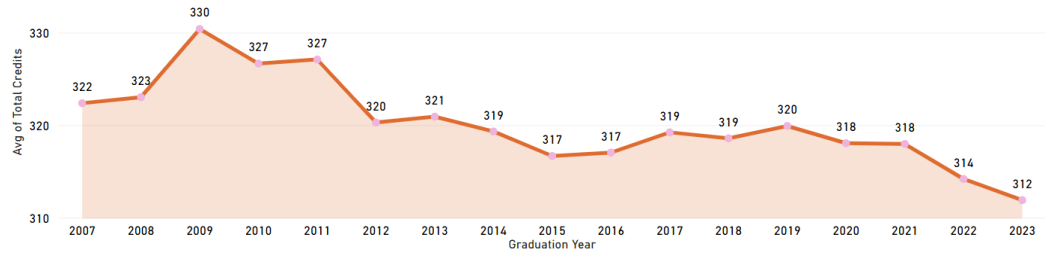


Figure 16. Average of Total Credits trend over Graduation Years

312 credits in 2023. This stabilization suggests that degree requirements became more standardized over time.

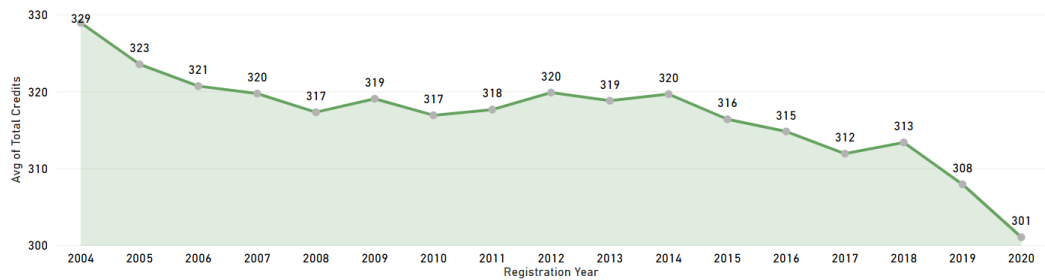


Figure 17. Average of Total Credits trend over Registration Years

As shown in Figure 17 the line chart presents the average total credits based on students' registration years. Unlike the graduation-year trend, this one shows a consistent downward trajectory. Combined degree graduates who registered in 2004 accumulated an average of 329 credits, whereas those who registered in 2020 had 301 credits. This decline suggests changes in curriculum structures, stricter academic policies, or more efficient study pathways.

While the graduation-year trend shows a peak in 2009, the registration-year trend exhibits a steady decline. Both trends stabilize around 2013–2020, indicating that credit accumulation became more predictable in these years. The registration-year trend declines more sharply due to missing data for students who are still studying, whereas the graduation-year trend remains stable. The decreasing trend in total credits over time suggests a shift toward more structured degree requirements, potentially limiting excessive credit accumulation. However, recent registration years (2018 onward) may be underestimated as many students from these cohorts have not yet graduated.

Educational Term The Educational Term is derived from the combination of the Educational Year and Educational Period. It includes values 1 and 2 for each Educational Year, representing the two academic terms within a year. This field plays a fundamental role in tracking the academic progression of students and is used to calculate the Term Serial Number, an essential metric for analyzing study paths.

Term Serial The Term Serial Number is an ordinal variable that assigns sequential numbers to all educational terms throughout a student's academic journey, starting from 1 for their first term. This numbering system enables a structured comparison of students across different cohorts, aligning them by their study progress rather than

calendar years. For example, students in their first term (Term Serial = 1) can be compared regardless of whether they started in different academic years. Similarly, students in Term Serial = 4 are in their fourth term, even though they may have started in different years. This approach allows for a more meaningful examination of study patterns and retention trends across time, independent of specific enrollment periods.

The Term Serial is particularly valuable for analyzing study pace, dropout rates, and progression consistency. Aligning students based on their term count rather than their registration date facilitates a more standardized comparison of study behaviours at equivalent academic stages.

As shown in Figure 5 more than 97% of graduated combined degree students completed their education within 10 years or less, analyzing the distribution of the students by Term Serial, filtered for this group, reveals key trends in study progression.

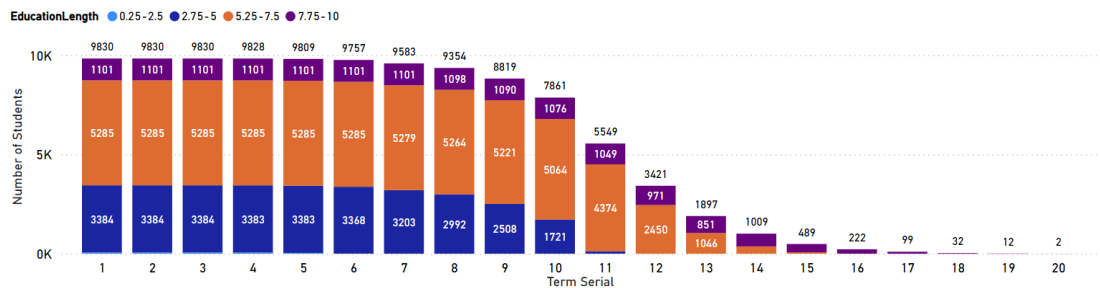


Figure 18. Distribution of graduated combined degree students by Term Serial, categorized by Education Length ≤ 10 years

The visualization in Figure 18 illustrates the progression of structured study in 10 years. Each bar represents the number of graduated students active in a given term (Term Serial), and the colour segments indicate their total Education Length at the time of graduation. Most students actively progress through the first six terms (three years), indicating a structured study pattern. After term 7th, an obvious decline begins as students approach expected graduation timelines, aligning with the standard duration for combined degrees. The sharp drop after term serial 9 suggests that around 43% of the students graduate before term 11, reinforcing this as a key threshold for the completion of the study. Beyond term 12 a small proportion of students extend their studies, potentially due to study gaps or interruptions.

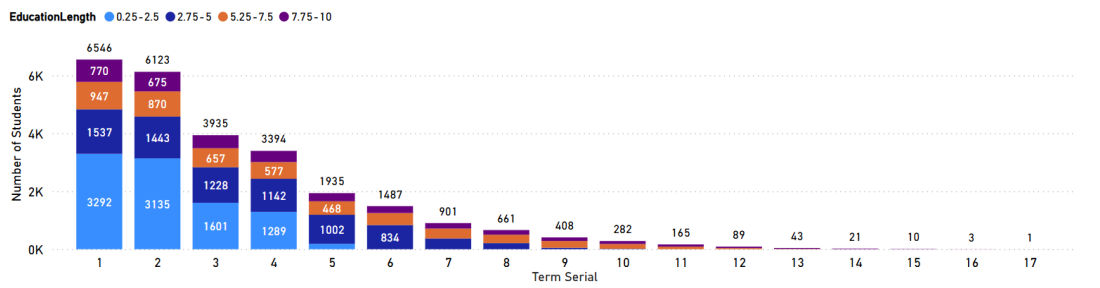


Figure 19. Distribution of Still-Studying Students by Term Serial, categorized by Education Length ≤ 10 years

A similar distribution pattern is observed for still-studying students with filtered Education Length ≤ 10 years (Figure 19). Each bar shows the number of still-studying

students active in a given term (Term Serial), with colour segments representing their accumulated Education Length at the current point in their studies. The visualization shows most students are concentrated in the first few terms (1,2), meaning they are still in the early stages, with a steady decline as the Term Serial increases. Notable drops occur after two Term Serials of 2 and 4, indicating that many students discontinue their studies around this point, indicating that many students either slow down their progress or drop out. Beyond Term Serial 6, the numbers decrease drastically, with very few students progressing past Term 10, mirroring the expected graduation timelines. After Term Serial 10, only a small number continue without completing their degrees, which suggests that they may still complete their degrees, but delays or interruptions are common.

Average credit per Term Serial Average Credits per Term Serial is a calculated metric that represents the average of each student's individual average credits per term serial throughout their education. The expected number of ECTS credits per term serial is 30, based on a standard academic load of 60 ECTS per year.

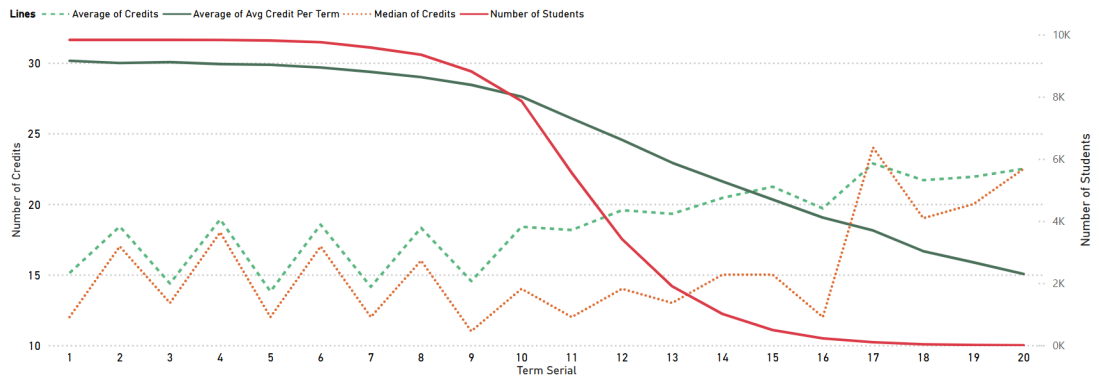


Figure 20. The overall average of students' average credits per Term Serial, alongside the actual average, median credits and the number of students, for graduated combined degree students filtered by Education Length ≤ 10 years.

In Figure 20 the average of these individual student averages, provides an overall measure of credit accumulation trends among all students. This metric highlights that students who remain on the chart at later term serials generally had lower average credits throughout their education, indicating a slower study pace. These students tend to extend their studies over a longer period, often accumulating credits at a lower rate per term compared to those who graduate earlier.

This visualization highlights key trends in student progression by Term Serial. The number of students declines sharply after Term Serial 10, reflecting expected graduation trends. Overall average credits per student remain relatively stable but show a gradual decline, indicating reduced study intensity over time. The actual average credits per term serial fluctuate, suggesting variations in study load across different terms. The median credits also exhibit fluctuations, reinforcing the impact of individual study patterns. The increase in average credits in later terms is influenced by the decreasing number of students rather than an actual rise in study activity.

The red line in Figure 21 represents the declining number of students as the Term Serial increases, which clearly shows that fewer students persist in later terms, a sharp decline after Term Serial 4, showing that a significant number of students discontinue

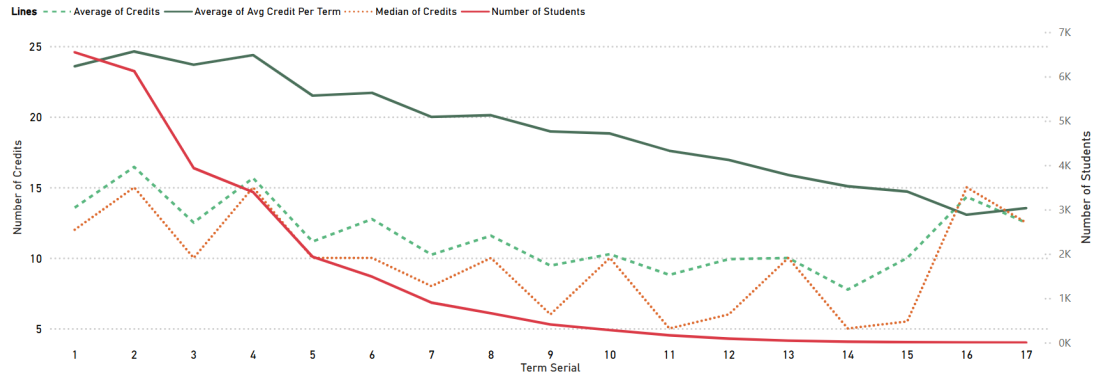


Figure 21. The overall average of students' average credits per Term Serial, alongside the actual average and median credits per Term Serial and the number of students, for still-studying students filtered by Education Length ≤ 10 years.

their studies early. The solid green line (Average of Average Credits Per Term Serial) shows that students who remain in the later terms generally have lower average credits over their study duration, indicating a slower study pace, this suggests that extended study durations are strongly correlated with lower credit accumulation rates. However, this supports the hypothesis that students with slower study progress tend to increase their education rather than drop out altogether. The dashed green line (Average credits load) and dotted orange line (Median Credits load) display fluctuating credit loads at each term serial for all students, reflecting variations in student enrollment patterns and study intensity. Moreover, this visualization shows that the apparent increase in average credits is actually due to fewer students remaining, not because they take more credits.

Running Total Credits Running Total Credits is a cumulative measure representing the total number of credits a student has earned up to a given Term Serial. It is calculated by summing all previously earned credits at each term, providing insight into students' academic progression over time. This metric helps analyze study pace, identify patterns in credit accumulation, and distinguish between students who follow a structured path and those with delays or irregular study patterns.

The visualization in Figure 22 illustrates the cumulative accumulation of credits of individual students over time, highlighting structured study progressions, common graduation timelines, and variations in study pace beyond the expected completion periods. Each line represents a student's cumulative credit accumulation over time⁵. The majority of students follow a structured credit accumulation path, with a steady increase in Running Total Credits as their Term Serial progresses. The density of overlapping lines highlights common study paths, while outliers and deviations indicate students who may have taken nontraditional routes. The visible jumps at specific term serials suggest common graduation timelines or significant credit accumulation points (e.g., final coursework completion). The spread in later terms

⁵Some students show unusually high credit accumulation in early terms due to prior studies in Open University or transferred credits from other institutions. These credits are typically recorded under Education Year 0, which is not displayed in this chart, resulting in elevated totals already at Term Serial 1.

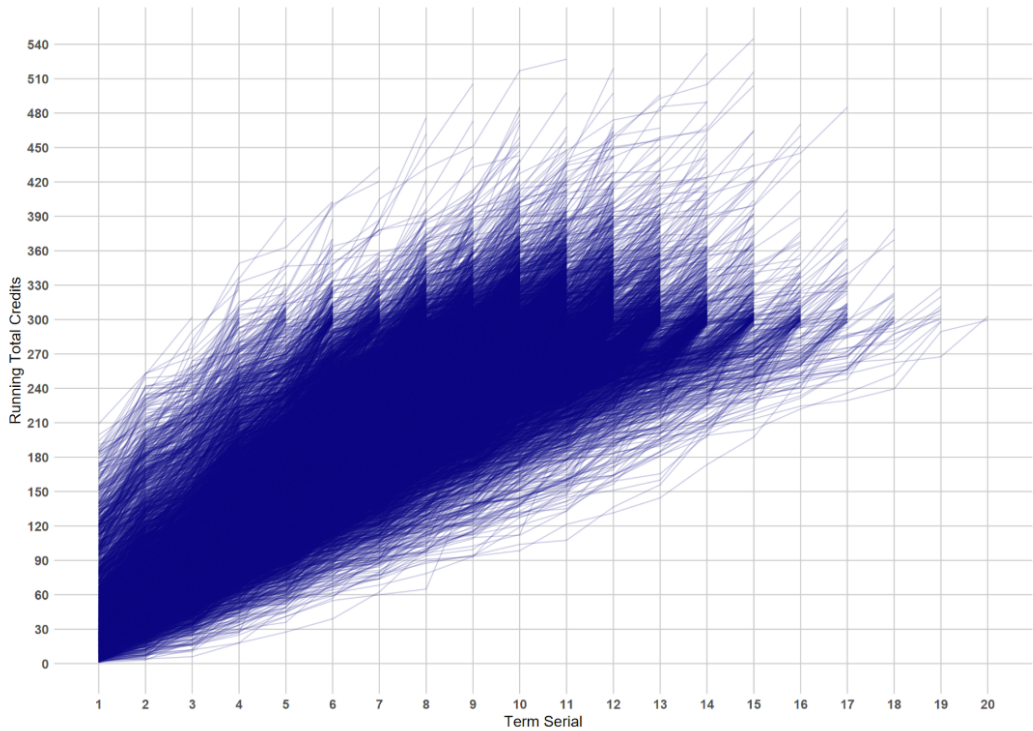


Figure 22. Running Total Credits per Term Serial for Combined Degree Graduates with Education Length ≤ 10 years.

suggests a minority of students take longer to graduate, potentially due to individual academic circumstances or study interruptions.

Graduation Probability by Credit Accumulation and Term Serial The visualization in Figure 23 presents the graduation probability of combined degree students who completed their studies within 10 years, analyzed based on Running Total Credits and Term Serial. Each cell represents the number of students who eventually graduated in 10 years, given their credit accumulated at each term.

RunningTotalCreditsgroups	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20			
Other								1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50					
421-450								1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00						
391-420								1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.92	0.33	1.00					
361-390								1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.91	0.76	0.86	0.71	0.50			
331-360								1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.97	0.92	0.85	0.72	0.74	0.17			
301-330								1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.98	0.97	0.94	0.84	0.78	0.54	0.33	0.32	0.08
271-300								1.00	1.00	1.00	0.99	0.99	0.99	0.96	0.92	0.87	0.79	0.64	0.51	0.43	0.29	0.20	
241-270								1.00	1.00	0.99	0.99	0.98	0.96	0.93	0.86	0.74	0.58	0.48	0.33	0.19	0.09		
211-240								0.97	0.98	0.99	1.00	0.99	0.97	0.95	0.90	0.80	0.70	0.55	0.46	0.29	0.22	0.10	
181-210								0.94	0.99	0.99	0.99	0.99	0.99	0.96	0.93	0.84	0.76	0.65	0.53	0.40	0.15		
151-180								1.00	0.99	0.98	0.98	0.99	0.98	0.96	0.90	0.83	0.73	0.66	0.45	0.27	0.11		
121-150								0.99	0.98	0.98	0.99	0.98	0.94	0.90	0.82	0.74	0.58	0.34	0.22	0.14			
91-120								0.98	0.98	0.99	0.98	0.94	0.90	0.81	0.69	0.53	0.33	0.14					
61-90								0.96	0.98	0.98	0.92	0.85	0.68	0.58	0.27								
31-60								0.98	0.97	0.93	0.81	0.67	0.36	0.11									
1-30								0.97	0.88	0.75	0.60	0.25											

Figure 23. Graduation probability of combined degree students who completed their studies within 10 years, analyzed by Running Total Credits and Term Serial.

Students who accumulate higher credits earlier (e.g., above 240 credits by Term Serial 8-10) show near-certain graduation probabilities (values close to 1.00). In contrast, students with low accumulated credits (e.g., below 90 credits by Term Serial 10) have significantly lower graduation probabilities. Students with steady credit accumulation maintain high graduation probabilities across terms. Those accumulating fewer credits per term (e.g., below 20 per term) have an uncertain outcome, often leading to longer study durations. Students in lower credit ranges (e.g., 1-90 credits) show decreasing graduation probabilities as they move toward higher Term Serials.

This analysis shows that Running Total Credits is a key predictor of graduation likelihood. It validates the need to incorporate credit accumulation trends into ML models. The declining probabilities for lower-credit students highlight early warning indicators for at-risk students, which can be leveraged in predictive models. Based on these insights, it is evident that credit accumulation plays a significant role in determining graduation outcomes. To further quantify and predict these outcomes, we now develop Machine Learning models that utilize these trends to forecast the likelihood of graduation for still-studying students.

5. MACHINE LEARNING IMPLEMENTATION

This study employs a data-driven approach to predict the remaining time to graduation for combined degree students using machine learning techniques. This section details the practical implementation of the methods and tools introduced earlier. It outlines the full machine learning pipeline, including data preparation, feature engineering, filtering criteria, model training, and performance evaluation using predictive analytics. These implementation steps form the technical foundation for generating reliable predictions and support the subsequent analysis and interpretation of model outcomes.

5.1. Data Preparation and Filtering

Before training the machine learning model, the dataset was preprocessed and filtered to ensure accurate and meaningful predictions.

Data Cleaning: Duplicate records were removed to ensure each student's progression was represented uniquely.

Degree Filtering: The focus of this study is on combined degree students, meaning that students who only completed a Bachelor's or Master's degree were excluded.

Still-Studying Student Selection: To identify students who are likely to graduate, we applied conditional filtering:

- Only students without a recorded graduation date were included.
- Those with a Bachelor's degree but no Master's enrollment were excluded, as their future study intentions remain uncertain.
- Students whose last leave (gap since their last recorded study date) was shorter than the maximum observed leave among graduates were retained, assuming they are likely to continue and graduate.

5.2. Feature Engineering and Selection

To develop a robust and explainable model, a set of features was engineered based on academic progression patterns and student study behaviours. These features build upon the theoretical foundations of feature engineering and time-based indicators discussed in Section 3.7.

Justification of Feature Selection: The selected features were designed to balance study pace, credit accumulation, and interruptions, providing a holistic view of student progress. Variables such as Rolling Average of Credits, Credit Growth Rate, and Study Pace reflect active study behaviours, while Intake Delay and Total Gap capture potential delays. Total Credits Earned and Average Credit Per Term ensure alignment with graduation requirements (see Table 5).

The importance levels listed in Table 5 (High vs. Medium) were based on the researcher's observations during the data exploration and feature engineering process. This classification reflects the consistency of patterns identified in exploratory

visualizations and the conceptual alignment of each feature with graduation-related outcomes. Features marked as High showed strong and stable trends across students (e.g., Total Credits, Intake Delay), and were directly tied to degree completion requirements. Features labeled as Medium also contributed meaningful insights, but exhibited more indirect effects or greater variability in patterns (e.g., Study Pace, Credit Growth Rate).

Table 5. Selected Features and Their Importance

Feature	Description	Importance
IntakeDelay	Time between registration and first education date, helping to assess delays affecting graduation time.	High
TotalGap	Sum of all study gaps, including the last gap for still-studying students.	High
RollingAvgCredits	Moving average of credits earned over the last 3 terms to track study pace.	Medium
CreditGrowthRate	Change in running total credits between consecutive terms, showing study progress.	Medium
StudyPace	Measures frequency of consecutive active terms to evaluate study consistency.	Medium
AvgCreditPerTerm	Student's average credit load per term over their entire study period.	High
TotalCredit	Total credits earned by a student, capturing overall progress.	High

5.3. Correlation Analysis of Key Features

To further validate the selected features, we conducted a correlation analysis to examine relationships between predictors and the target variable (Education Length). Figure 24 presents the correlation matrix, highlighting the strength and direction of associations.

From the matrix:

- **Total Gap(0.88 correlation with Education Length):** Strongly associated with longer study durations, reinforcing its importance in modelling delays.
- **Average Credit Per Term (-0.71 correlation with Education Length):** Negatively correlated, indicating that students with a high study pace tend to graduate faster.
- **Rolling Average of Credits (weak correlation):** Provides insights into credit-taking consistency but does not show a direct strong correlation with the target.
- **Intake Delay (0.03 correlation with Education Length):** Weak relationship, but still considered to assess initial enrollment delays.

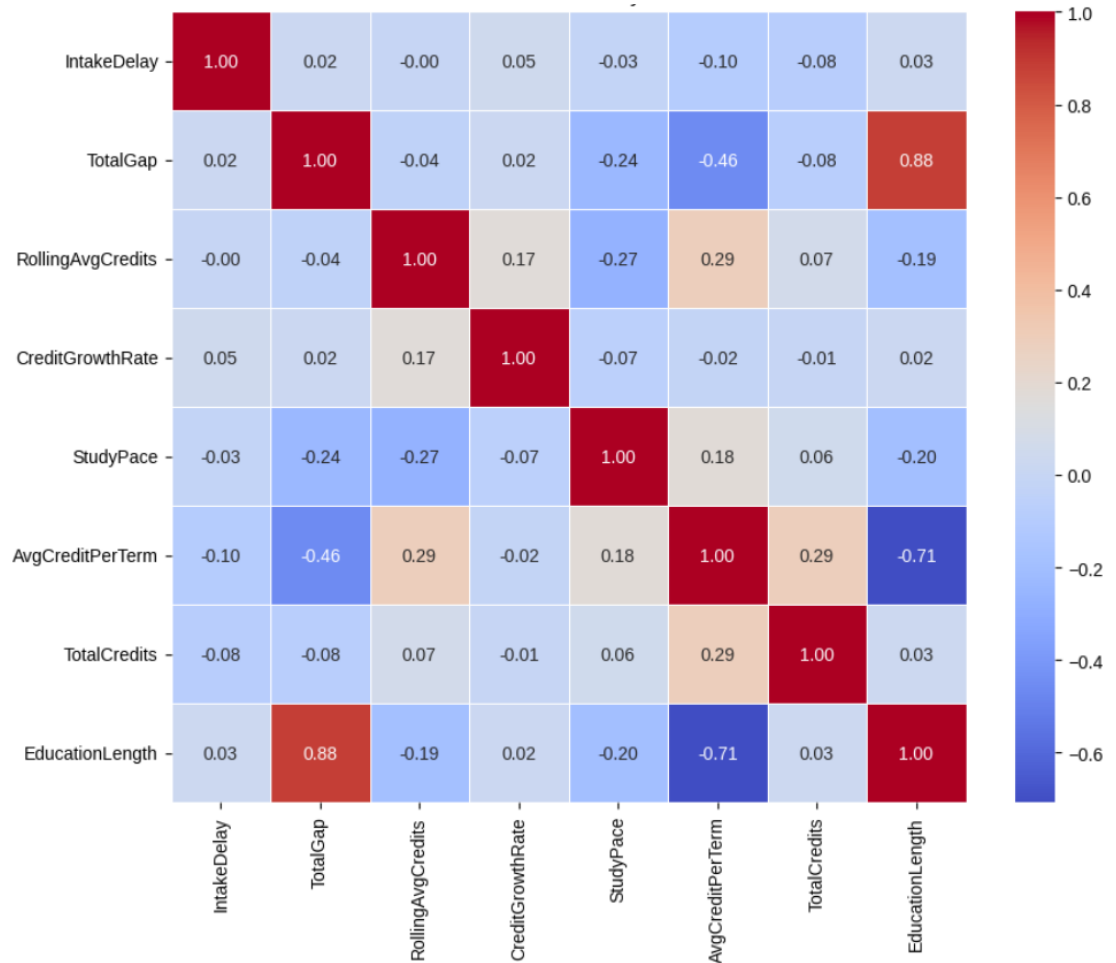


Figure 24. Correlation matrix of selected features, highlighting relationships between key predictors and the target variable (Education Length).

This analysis confirms that selected features align with expected academic progression trends, ensuring model interpretability. Although Intake Delay shows a relatively low correlation (0.03) with Education Length, it was retained due to its conceptual relevance: it captures delays between official registration and the actual start of academic activity. These delays are often institutionally or administratively driven and not directly reflected in credit accumulation patterns, making their correlation weaker. However, from a modelling standpoint, Intake Delay still offers explanatory value in cases where early inactivity contributes to extended study durations. By contrast, features like Study Pace and Credit Growth Rate, although showing similar or slightly stronger correlations, reflect more dynamic progress-based behaviours. The final feature set was therefore chosen not solely based on statistical correlation, but by balancing interpretability, conceptual significance, and insights from early exploratory analysis.

5.4. Train-Test Split and Data Preparation for ML

To maintain data integrity and prevent data leakage, the dataset was split by unique students rather than individual study records.

- **Training Data:** Students who had already graduated with a combined degree.
- **Testing Data:** A 25% sample of graduated students to validate model performance.
- **Still-Studying Data:** Students meeting the eligibility criteria for prediction but without graduation records.
- **Feature Selection for Model Training:** Features include: Intake Delay, Total Gap, Rolling Average Credits, Credit Growth Rate, Study Pace, Average Credit Per Term, Total Credits Earned, and the target variable is Education Length (Total Study Duration).

5.5. Machine Learning Model Development and Selection

This section outlines the implementation of both XGBoost and Random Forest regression models, focusing on their training, evaluation, and selection process.

5.5.1. Machine Learning Workflow

The following steps align with the general workflow described in Section 3.10.

- **Data Splitting:** The dataset was partitioned into training (75%) and testing (25%) subsets. Stratified sampling was applied where applicable to preserve the distribution of relevant outcome classes or student groups.
- **Cross-Validation:** A five-fold cross-validation strategy was implemented on the training set to assess model generalizability and guide hyperparameter optimization.
- **Feature Analysis:** Correlation analysis was conducted to evaluate inter-feature relationships and to identify potential multicollinearity among predictors.
- **Residual Analysis:** Residual plots were used to visualize the distribution of prediction errors and assess model fit beyond standard evaluation metrics.

5.5.2. Training and Evaluation Process

The dataset was split into training and testing sets using a student-wise split to ensure that individual students' data did not appear in both sets. Only graduated combined degree students were included in this phase to ensure the target variable (Education Length) was fully observed. An XGBoost and a Random Forest regression model were trained using optimized hyperparameters. Random Forest was tuned using `RandomizedSearchCV`, while XGBoost was optimized through `Bayesianoptimization`.

Model training and hyperparameter tuning were performed in Python using Google Colab. The environment ran on a standard cloud-backed kernel. Model training and prediction took approximately one minute per run. All key libraries used in the implementation included `matplotlib`, `numpy`, `xgboost`, `scikit-learn`, and `pandas`. The models' performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) on both training and testing sets.

Model Evaluation and Justification for Model Selection

To validate model performance and support the selection of the final model, XGBoost and Random Forest were compared using key evaluation metrics on both training and test datasets. The performance of XGBoost and Random Forest was evaluated based on the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) across training, testing, and cross-validation. Tables 6 and 7 summarize their performance:

Table 6. Comparison of XGBoost and Random Forest on Test Data

Model	MAE (Test)	MSE (Test)	R ² (Test)
XGBoost	0.2419	0.1115	0.9596
Random Forest	0.2531	0.1359	0.9508

Table 7. Comparison of XGBoost and Random Forest on Training Data

Model	MAE (Train)	MSE (Train)	R ² (Train)
XGBoost	0.2310	0.0991	0.9706
Random Forest	0.1243	0.0302	0.9910

XGBoost exhibited lower error rates on the test set (MAE = 0.2419, MSE = 0.1115) compared to Random Forest, indicating that it provided more accurate predictions. Although Random Forest achieved a higher R^2 in training (0.9910), reflecting an almost perfect fit, its R^2 dropped more significantly in the test set (0.9508), suggesting a higher degree of overfitting.

Further validation through 5-fold cross-validation showed that Random Forest had a slightly higher average R^2 (0.9816) compared to XGBoost (0.9643). However, this does not necessarily indicate better generalization, as the larger gap between training and test performance suggests that Random Forest overfitted the training data, while XGBoost maintained a more balanced performance across training, testing, and validation.

XGBoost was ultimately chosen due to its balanced performance across training, testing, and cross-validation. Its lower test error rates make it more reliable for real-world predictions, and its better generalization capability ensures robustness in handling unseen data. In contrast, Random Forest demonstrated signs of overfitting, as evidenced by its significantly higher training performance compared to its test results, making XGBoost the more stable and reliable choice for predicting students' time to graduation.

5.5.3. Feature Importance Analysis

To better understand the impact of each predictor on the model's output, we analyzed feature importance using model-based importance and permutation importance as described in Section 3.4.

Figure 25 presents the feature importance scores derived from the trained model.

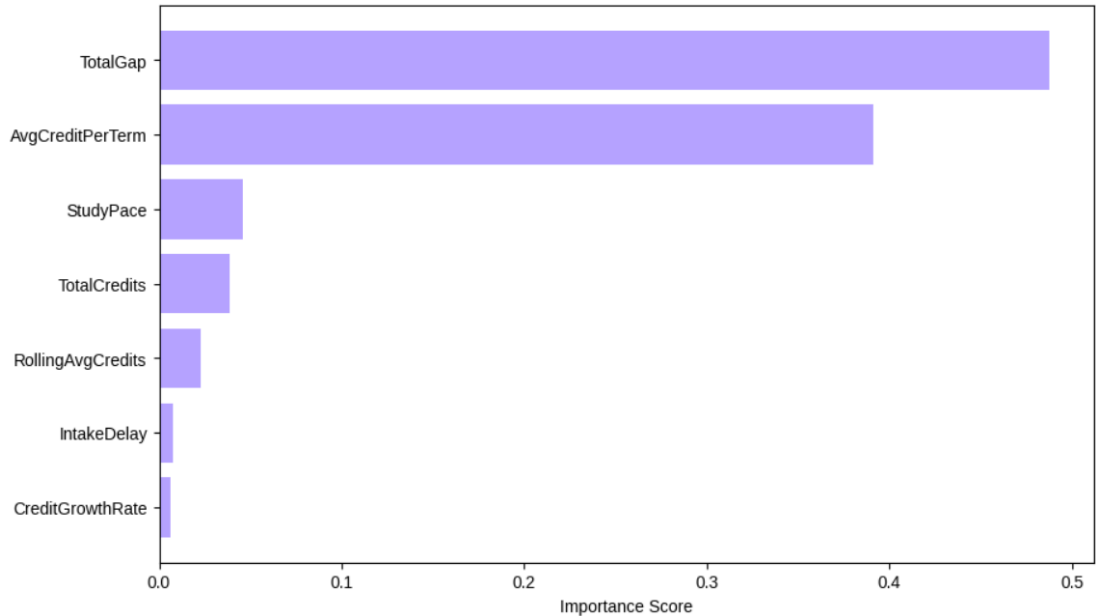


Figure 25. Feature Importance for Predicting Education Length using XGBoost.

The results highlight that the Total Gap and Average Credit Per Term are the most influential factors in predicting Education Length. This aligns with expectations, as study gaps can significantly extend the time required to graduate, while higher credit loads per term generally lead to faster completion.

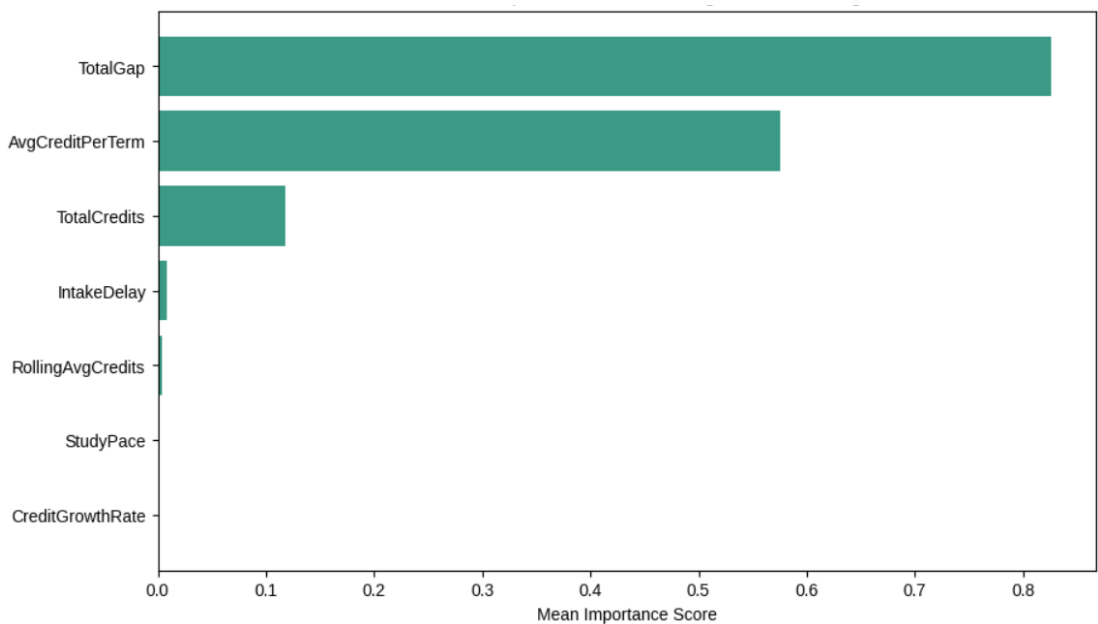


Figure 26. Permutation Importance for Predicting Education Length.

Figure 26 illustrates the results of permutation importance. The results confirm the dominance of Total Gap and Average Credit Per Term, reinforcing their strong predictive power.

In particular, Total Credits Earned exhibit a higher importance in permutation analysis compared to model-based importance, suggesting that cumulative credit accumulation plays a meaningful role in defining students' academic pathways.

Together, these analyses validate the relevance of the selected features and provide interpretability of the factors that influence students' time to graduation.

6. RESULTS

This section presents a comprehensive analysis of the machine learning models designed to predict the remaining time to graduation for still-studying students. The analysis begins with a visualization of the predicted graduation timelines through a histogram, providing an overview of the distribution of expected completion times. Next, an evaluation of the model's predictive performance is conducted using residual analysis, a residuals-versus-predicted scatter plot, and a comparative assessment of XGBoost and Random Forest models. The evaluation highlights XGBoost's superior predictive accuracy, lower error rates, and better generalization, leading to its selection as the final model.

Further, SHAP analysis is performed to interpret the contribution of key features, revealing that study interruptions (Total Gap), credit accumulation rate (Average Credits per Term), and overall credit completion (Total Credits Earned) have the most significant impact on graduation timelines. Finally, to contextualize the model's predictions, several case studies are examined, comparing the characteristics and study patterns of different groups of students, including both still-studying and graduated.

6.1. Prediction and Interpretation

The trained model was applied to still-studying students, with an Education Length of less than or equal to 5 years, with predictions based on their most recent term. The results were adjusted to reflect real-world education periods, ensuring the predicted remaining time followed the observed academic calendar. The histogram in Figure 27 illustrates the distribution of predicted remaining time to graduate for still-studying students. The x-axis represents the remaining years required for completion, while the y-axis indicates the number of students within each predicted duration range.

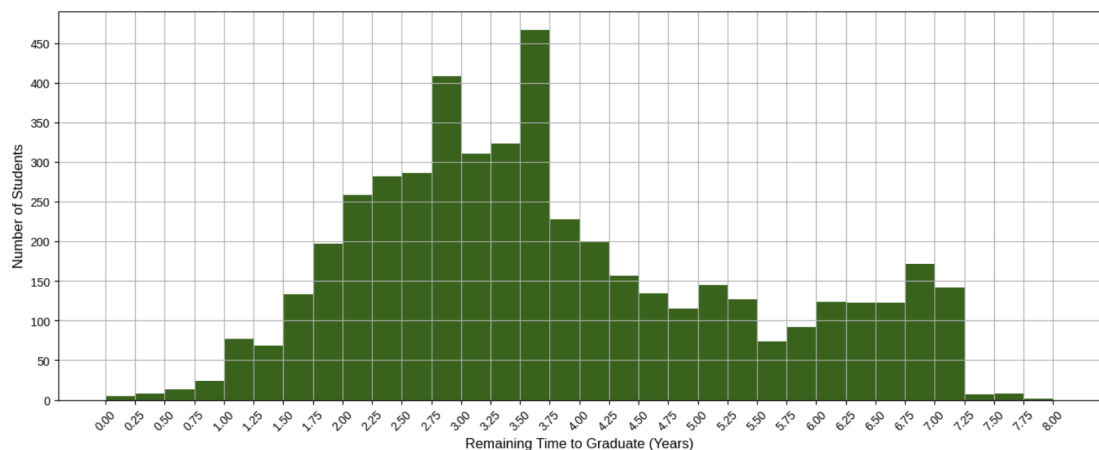


Figure 27. Distribution of Predicted Remaining Time to Graduate for Still-Studying Students.

The highest concentration of students is observed around 2 to 3.5 years, suggesting that the majority of still-studying students are expected to graduate within this time frame. This indicates a normal progression trend that aligns with the structure of the

academic system. The predictions vary widely, ranging from less than one year to over seven years. A smaller group of students is expected to complete their degrees in less than 1.5 years, likely representing those who have already accumulated a significant number of credits. In contrast, a portion of the students are projected to take more than 6 years, which could indicate those with a low study pace, extended study gaps, or irregular credit accumulation.

The distribution shows multiple peaks, particularly around 2.75 and 3.5 years, suggesting common graduation durations among students with similar study behaviours. Other peaks appear around 5 and 6.75 years, possibly reflecting a group of students with significant study delays or unique academic pathways. Although a majority of students are expected to graduate within 1.75 to 4 years, a tail extending towards 7+ years highlights students who may face academic challenges, interruptions, or lower course loads per term. These cases may require targeted educational interventions to support timely completion. These predictions provide valuable insight into student progression and can be used to support academic advising and institutional decision-making.

6.2. Evaluation of Prediction Quality

To assess prediction accuracy, residuals were analyzed using two key visualizations:

Residual Analysis: Based on Figure 28 Residual represents the distribution of prediction errors (difference between the actual and predicted values). The normal distribution, centred around zero, suggests that the model does not exhibit significant bias in underestimating or overestimating graduation time. The presence of minor outliers suggests that a small subset of predictions deviates from the expected range, but, overall, the residuals are well distributed, reinforcing the model's reliability.

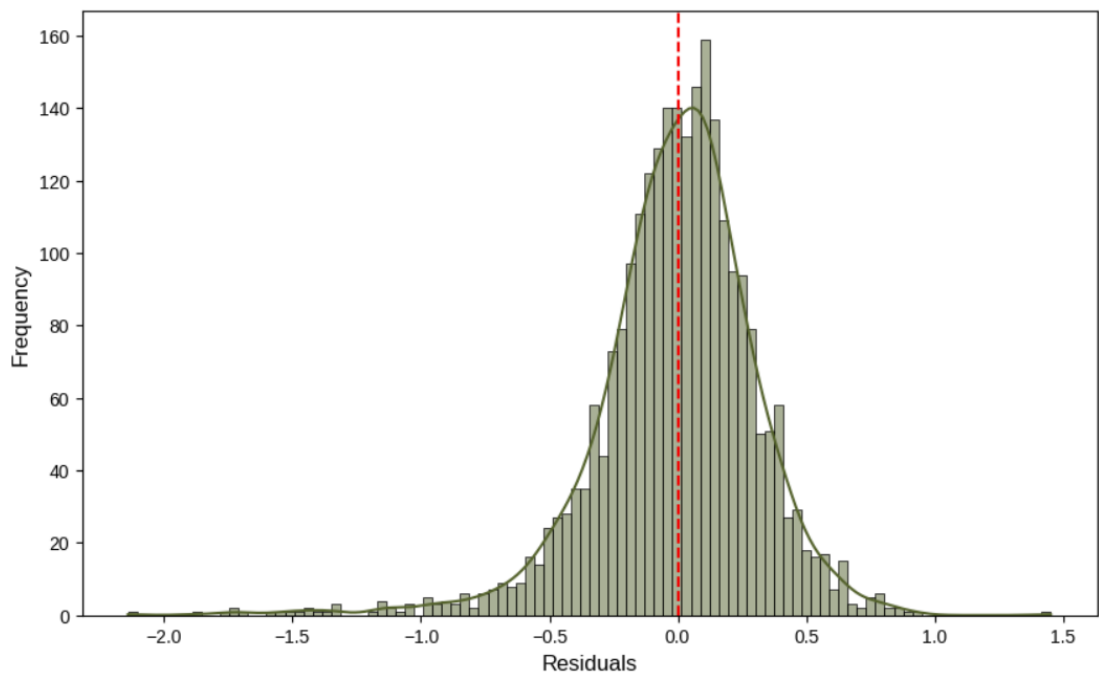


Figure 28. Distribution of residuals, indicating the model's prediction errors.

Residuals vs. Predicted Values: The scatter plot of residuals against predicted values in Figure 29 does not have visible patterns in the residual scatter plot, indicating homoscedasticity and good model fit. The majority of residuals are evenly spread around zero, without a clear pattern, which supports the assumption that errors are randomly distributed. However, a slight widening of the residuals at higher predicted values suggests that the model predictions might be less precise for students with longer expected graduation times. This could be an area for further refinement, possibly through feature adjustments or additional regularization.

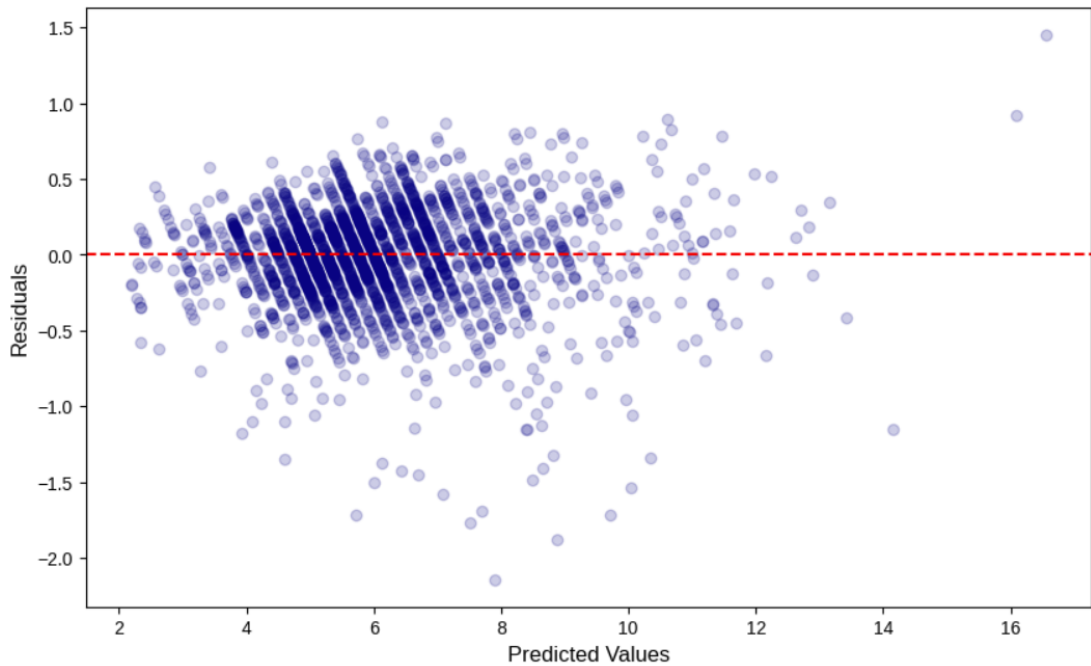


Figure 29. Scatter plot of residuals against predicted values, showing random distribution around zero.

6.3. SHAP Analysis

To interpret the XGBoost model's predictions, SHAP (SHapley Additive ExPlanations) analysis was conducted. SHAP values indicate how each feature contributes to the predicted remaining time to graduate. Figure 30 presents the SHAP summary plot, ranking features by their impact. Each point represents a prediction, with red indicating high feature values and blue indicating low values.

Total Gap has the highest impact, with larger gaps leading to longer predicted graduation times. Average Credit Per Term negatively correlates with Education Length, as students taking more credits per term tend to graduate faster. Total Credits Earned influences predictions, but less than the top two features. Intake Delay and Credit Growth Rate have minimal effects, suggesting that initial delays matter less than ongoing study habits.

This analysis confirms that study gaps and credit accumulation strongly affect graduation timelines. These insights can help universities identify at-risk students and optimize academic support strategies.

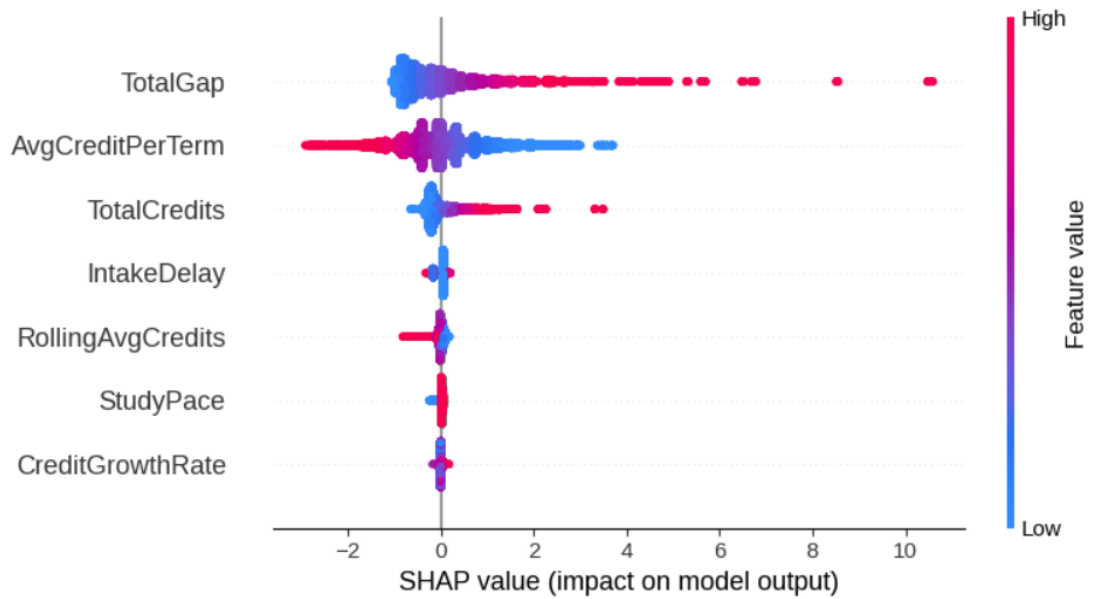


Figure 30. SHAP Summary Plot: Feature Contributions to Predicted Education Length

6.4. Case Studies Using A Dashboard

To further analyze the predictive insights generated by the model, we conduct a series of case study comparisons between still-studying students and graduated students. These comparisons aim to validate the model predictions by examining whether the academic behaviour of still-studying students aligns with those who have completed their degrees within similar time frames. Each case study focuses on students at different stages of their education. For each case, we compare study pace (average credits per term), study interruptions (total gap), and total accumulated credits. These key academic indicators help assess whether the predicted graduation timelines for still-studying students align with real-world graduation patterns. The following sections present these case study comparisons in detail.

6.4.1. Case Study 1: Fast Track Group

This case study examines students who have currently studied for 2 years and are predicted to graduate within the next 3 years, comparing them to students who graduated with 2.25-5 years of total education length. According to Figures 31 and 32:

Study Pace: Average Credits Per Term The study pace, measured by the average credits per term, provides information on the workload taken by the students in each group. Still-studying students have an average credit load of 31.04 credits per term, with a range from 24 to 62.75 credits. Graduated students completed their degrees with a slightly higher 34.23 credits per term, ranging from 20 to 70.30 credits. While the still-studying students are progressing with a slightly lower study pace, their minimum credit intake per term is higher than that of graduates, suggesting a relatively steady academic engagement.

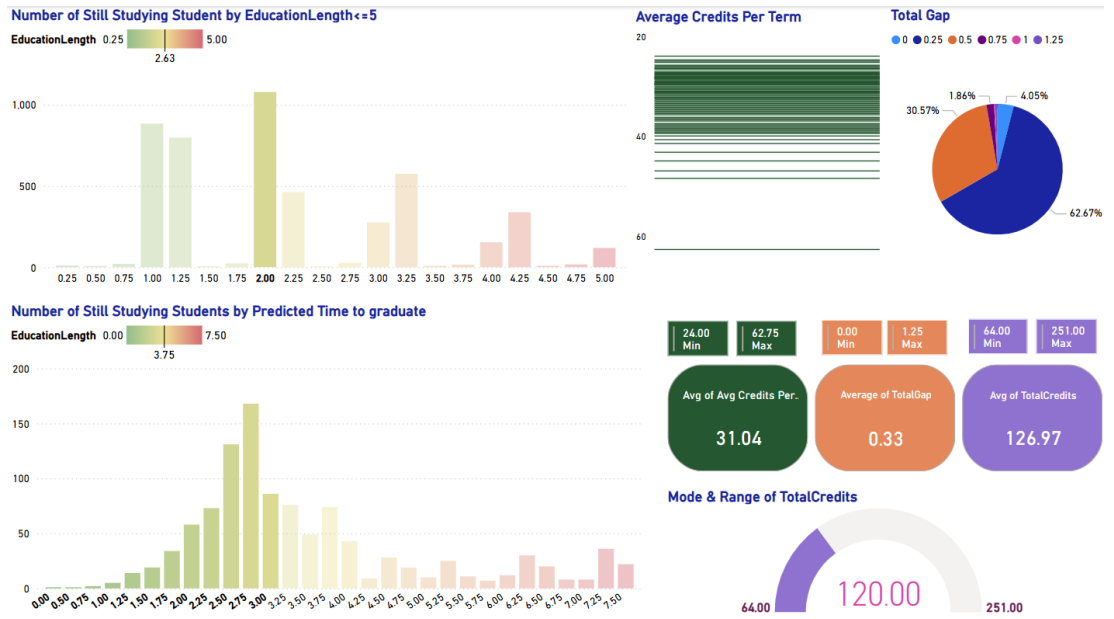


Figure 31. Fast Track Group: Still-Studying Students' Behavioural Dashboard

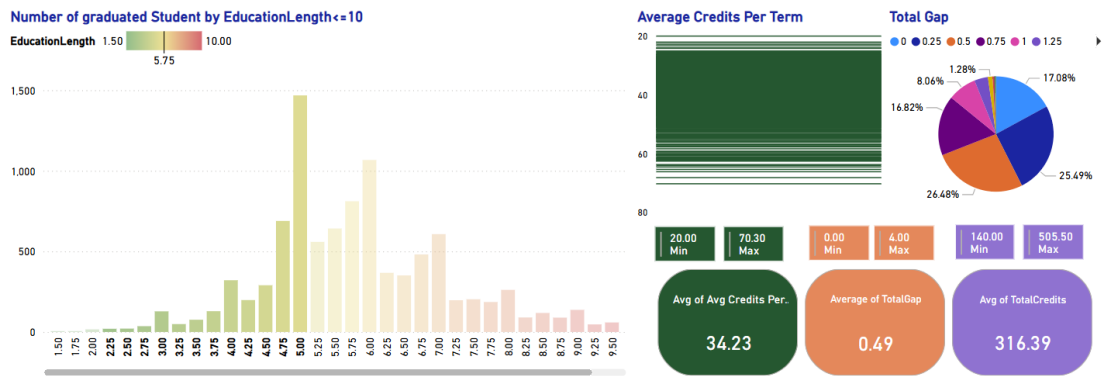


Figure 32. Fast Track Group: Graduated Students' Behavioural Dashboard

Study Interruptions: Total Gap in Education Interruptions in study progress can delay graduation. A comparison of total gap lengths between both groups reveals that still-studying students have a lower average total gap of 0.33 years, with a maximum of 1.25 years. Graduated students had a higher average total gap of 0.49 years, with some experiencing gaps as long as 4 years. Despite experiencing longer gaps, graduated students still completed their degrees, while still-studying students have maintained more continuous enrollment.

Total Credits Earned The total number of credits accumulated indicates academic progress and preparation for graduation. Still-studying students have earned an average of 126.97 total credits, with a range between 64 and 251, and a mod of 120 Total Credits. Graduated students completed their degrees with an average of 316.39 Total Credits, suggesting a similar academic trajectory. This indicates that many still-studying students have already reached or exceeded the credit levels of past graduates, supporting the prediction that they may graduate within the next 3 years.

The still-studying students' academic behaviour closely resembles that of past graduates, particularly in terms of total credits earned and study pace. Still-studying

students have had fewer interruptions, which may indicate a stronger commitment to timely graduation. Since some of these students have already accumulated credits comparable to past graduates, academic advising could help optimize their course selection to ensure they meet graduation requirements within the expected time frame.

6.4.2. Case Study 2: Moderate Progress Group

This case study examines students who have currently studied between 2.25 and 3 years and are predicted to graduate within the next 3 to 4 years, comparing them to students who graduated with 5.25-7 years of total education length. According to Figures 33 and 34:

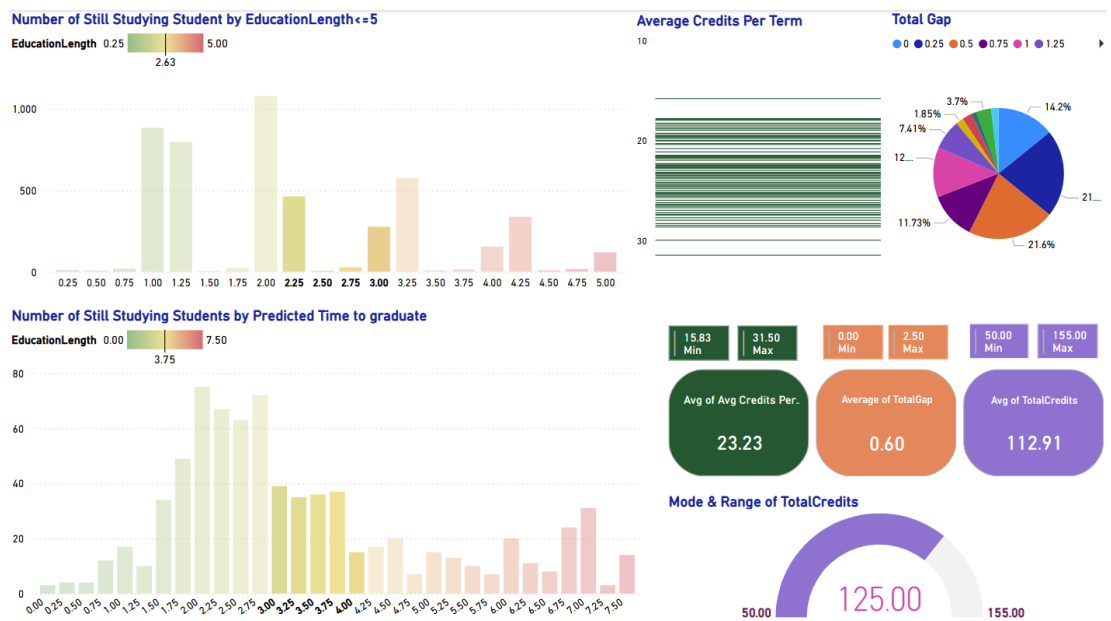


Figure 33. Moderate Progress Group: Still-Studying Students' Behavioural Dashboard

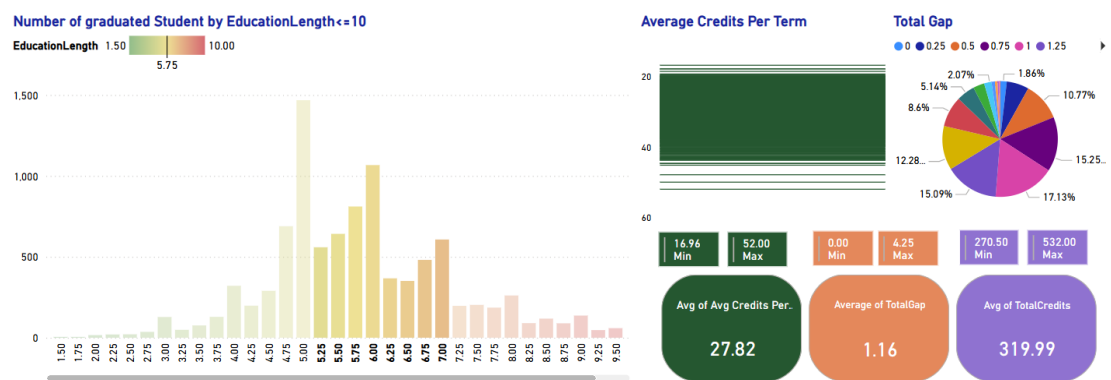


Figure 34. Moderate Progress Group: Graduated Students' Behavioural Dashboard

Study Pace: Average Credits Per Term Still-studying student's average credit per term is 23.23, with a range from 15.83 to 31.50. The average credit per term for the graduated student is 27.82, ranging from 16.96 to 52.00. Graduated students had

a slightly higher study pace on average, with a wider range. This suggests that, in general, students who completed their studies in 5.25 - 7 years were taking more credits per term than those still studying.

Study Interruptions: Total Gap in Education Study interruptions provide insight into the consistency of academic engagement. Still-studying students exhibit an average total gap of 0.60, with values ranging between 0.00 and 2.50. Graduated students have a higher average total gap of 1.16, ranging from 0.00 to 4.25. These results indicate that graduates experienced more interruptions in their studies, but were still able to complete their degrees within a reasonable time frame. In contrast, the lower gaps among still-studying students suggest more continuous study patterns; however, their ability to sustain this consistency while progressing toward graduation remains uncertain.

Total Credits Earned Still-studying students have earned an average of 112.91 credits, with values ranging between 50 and 155. While graduates have accumulated a substantially higher average of 319.99 credits, ranging from 270.50 to 532. This significant difference highlights the substantial academic workload still required for still-studying students to reach graduation. While they have demonstrated consistent study patterns, they must increase their credit accumulation rate to meet degree completion requirements.

Graduated students had a higher study pace, taking more credits per term on average, which contributed to their timely degree completion. Still-studying students exhibit fewer interruptions, suggesting a more consistent study pattern; however, they might need to increase their study pace to ensure timely graduation. The large gap in total credits earned emphasizes that still-studying students must significantly accelerate their credit accumulation to reach the graduation threshold. This comparison suggests that while maintaining study consistency is beneficial, ensuring a sufficient credit accumulation rate is equally crucial for degree completion. Future academic support efforts could focus on helping still-studying students optimize their course loads to align with successful graduation trajectories.

6.4.3. Case Study 3: Extended Study Group

This case study examines students who have currently studied between 3.25 and 4.25 years and are predicted to graduate within the next 4.25 to 5.75 years, comparing them to students who graduated with 7.5 to 10 years of total education length. According to Figures 35 and 36:

Study Pace: Average Credits Per Term The analysis reveals that still-studying students have an average credit load per term of 13.79, ranging from 6.67 to 25.00. In contrast, graduated students took an average of 22.98 credits per term, ranging from 14.98 to 50.57. This significant difference suggests that students who completed their degrees maintained a more intensive study pace, enabling them to reach graduation sooner. While, still-studying students exhibit a slower accumulation of credits per term, which may extend their time to graduation unless their study pace increases.

Study Interruptions: Total Gap in Education Still-studying students have an average total gap of 1.75, ranging between 0.25 and 3.75. On the other hand, graduates experienced longer interruptions, with an average total gap of 2.72, extending up

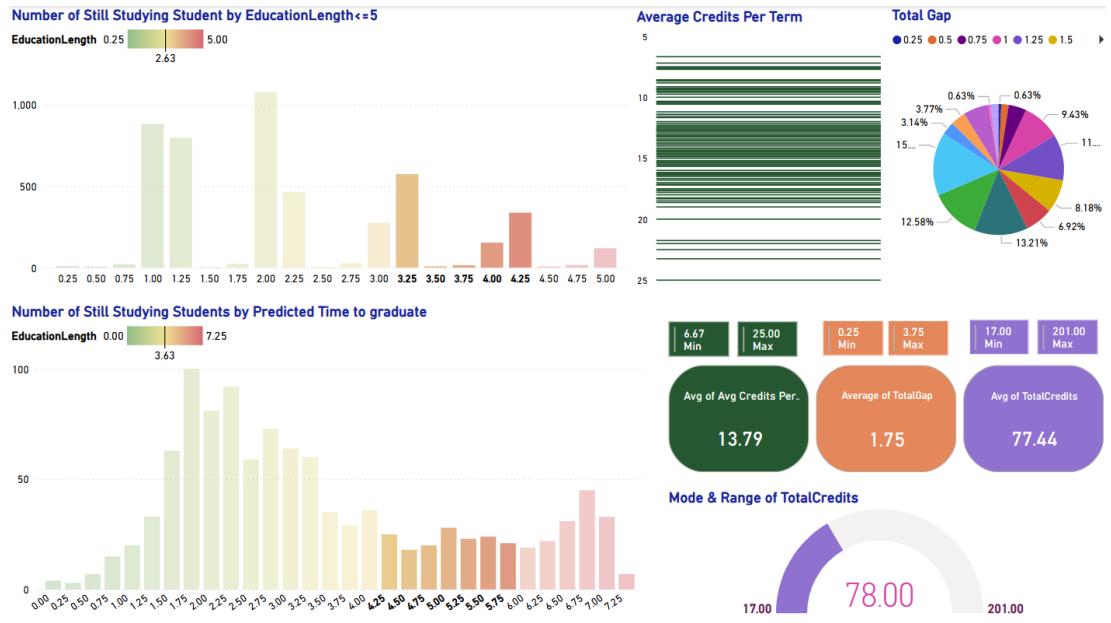


Figure 35. Extended Study Group: Still-Studying Students' Behavioural Dashboard

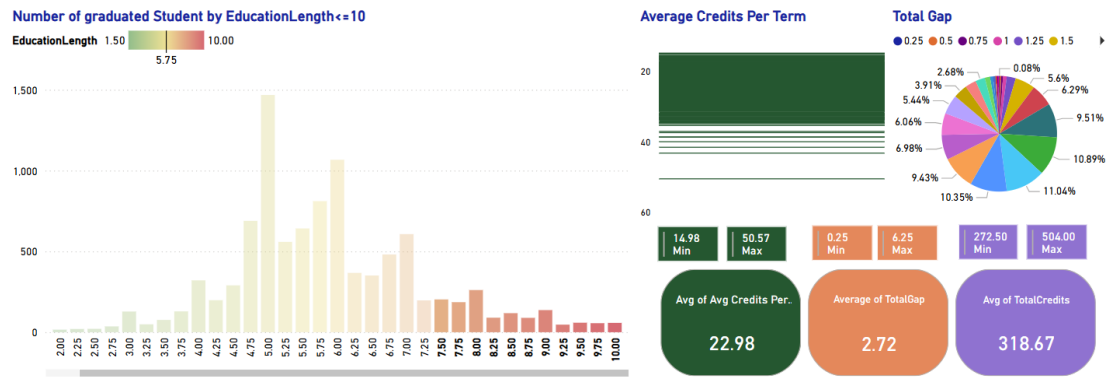


Figure 36. Extended Study Group: Graduated Students' Behavioural Dashboard

to 6.25. This finding suggests that occasional breaks in studies did not necessarily hinder graduation, as students who took longer to complete their degrees were still able to graduate. However, the shorter gaps observed among still-studying students may indicate a more consistent study pattern, which could be an advantage if paired with a sufficient credit accumulation rate.

Total Credits Earned The total number of credits earned further highlights the gap between these two groups. Still-studying students have an average total credit accumulation of 77.44, ranging from 17.00 to 201.00, whereas those who graduated accumulated 318.67 credits on average, with a range from 272.50 to 504.00. This stark difference underscores the fact that students who are still studying will need to make substantial progress in accumulating credits to reach the required threshold for graduation.

Overall, this comparison suggests that while still-studying students tend to have fewer interruptions in their studies, they also accumulate credits at a significantly slower rate. The findings highlight the importance of increasing both the pace of credit completion and the total number of earned credits to align with the patterns

observed among graduated students. If these students continue with their current study behaviour, they may face extended study durations beyond their predicted graduation time frame.

6.5. Summary of Findings

The results provided valuable insights into the predictive performance of the XGBoost model, the importance of key features, and the behavioral patterns of different student groups. The prediction results revealed a wide distribution of expected graduation timelines, with a significant portion of students projected to complete their studies within 1.75 to 4 years. Residual analysis confirmed that the model exhibited a balanced error distribution, reinforcing its reliability.

SHAP analysis further highlighted the nonlinear relationships between TotalGap, Average Credits per Term, and Total Credits Earned, identifying them as the most influential factors in predicting students' time to graduate. The findings confirmed that higher study gaps and lower credit accumulation rates tend to extend graduation timelines.

The case study comparisons provided deeper insights into different student trajectories. Across all cases, total credit accumulation emerged as the key differentiator between students on track for timely graduation and those facing delays. These findings underscore the importance of study pace, credit accumulation, and study interruptions in shaping students' academic progress.

The insights derived from this study can help educational institutions identify students at risk of delayed graduation, optimize resource allocation, and implement targeted interventions to improve student retention and completion rates.

7. DISCUSSION

This study explored the use of machine learning models to predict the remaining time to graduation for students in a combined degree program. The findings demonstrate the effectiveness of data-driven approaches in identifying academic patterns and generating actionable insights for higher education institutions.

7.1. RQ1: What Academic Behavioural Features Most Significantly Influence a Student's Time to Graduate?

The model interpretation revealed that *Total Gap*, *Average Credits per Term*, and *Total Credits* were the most significant academic behavioural features influencing a student's time to graduate. Students with prolonged or frequent gaps between study periods tended to have longer overall study durations, particularly when such gaps were not offset by higher performance in other periods. Conversely, students with minimal gaps and steady credit accumulation were more likely to complete their degrees on time.

These findings are supported by prior studies emphasizing the importance of study consistency and credit intensity in academic success [8, 7, 16]. For example, term-over-term performance has been shown to be a more reliable graduation predictor than demographic variables alone [16]. Similarly, course load regularity has been highlighted as a primary indicator of retention and timely completion [8].

Furthermore, the results of this study contribute a nuanced perspective: while study gaps generally correlate with delays, they are not always detrimental. Students who compensated for gaps with higher intensity during active periods often stayed on track, suggesting that *academic momentum* and *adaptive credit pacing* are equally critical factors.

7.2. RQ2: Which Machine Learning Model Provides the Most Accurate Prediction of Graduation Timelines for Still-Studying Students?

Among the models tested, **XGBoost** demonstrated superior predictive performance across all evaluation metrics. It consistently outperformed Random Forest, achieving lower mean absolute error (MAE), lower mean squared error (MSE), and higher R^2 scores on both training and test sets. This performance was further validated through k-fold cross-validation, confirming the model's robustness and generalization capability.

While Random Forest showed signs of overfitting, indicated by a notable gap between training and testing R^2 , XGBoost maintained stable accuracy, making it the most reliable model for predicting time to graduate. These results align with previous findings that have demonstrated XGBoost's strong performance on structured data tasks [34]. Other educational data mining studies have also reported high effectiveness of gradient boosting methods in predicting student outcomes [15, 9].

The inclusion of SHAP analysis further enhanced XGBoost's utility by providing transparent explanations for individual predictions. This explainability is especially valuable in educational settings, where data-driven interventions must be interpretable by academic staff and advisors.

7.3. RQ3: How Can the Model's Predictions Be Used to Identify Patterns and Support Students at Risk of Delayed Graduation?

The model outputs were used to compare still-studying students with historically graduated cohorts, allowing the identification of three behavioural clusters: the *Fast Track Group*, the *Moderate Progress Group*, and the *Extended Study Group*. These groupings provided actionable insights into the behavioural patterns of students predicted to graduate on different timelines.

In the *Fast Track Group*, students with two years of current education and a predicted graduation in the next three years showed strong similarities to graduates who completed their degrees within 2.25 to 5 years. These students exhibited high average credits per term, minimal study interruptions, and a comparable range of total accumulated credits, suggesting a solid trajectory toward timely graduation.

In the *Moderate Progress Group*, students with 2.25 to 3 years of study and expected to graduate within the next 3 to 4 years displayed more variation. Although they maintained lower study gaps, their average credit accumulation per term and total earned credits lagged behind those of graduates who completed their degrees in 5.25 to 7 years. This indicates that consistency alone is not sufficient, without increased credit loads, delays may occur.

The *Extended Study Group*, comprising students who had studied for 3.25 to 4.25 years and were predicted to graduate in 4.25 to 5.75 years, revealed the greatest divergence. These students had significantly lower study pace and credit accumulation compared to graduates who completed their degrees in 7.5 to 10 years. Although they experienced fewer study gaps, their slow progression suggests a risk of falling behind their predicted timelines unless intervention strategies are implemented.

These insights highlight how **model-based predictions can inform personalized advising**. For instance, still-studying students with consistent attendance but low performance may require intensive support, while those with gaps but strong output may need greater flexibility or counseling. Prior studies have emphasized the value of early-warning systems and personalized dashboards in improving retention [13, 12], and this study reinforces those findings with detailed behavioural segmentation.

From an institutional perspective, the integration of these predictions into academic platforms offers considerable benefits. For example, integrating visual dashboards into platforms such as *Peppi*⁶ could show students their predicted timelines compared to historical averages, encouraging reflection and goal setting. Coordinators could also receive alerts when students deviate from expected progress, enabling timely and targeted support. The Power BI dashboards developed in this study demonstrate how such predictive tools can be operationalized for real-time academic advising.

7.4. Limitations and Future Work

Despite promising results, the study has several limitations. The dataset was limited to students from the University of Oulu, which may restrict the generalizability of

⁶Peppi is the academic management system used by Finnish universities for study planning, course registration, and transcript tracking.

findings to other contexts. Additionally, the dataset lacked demographic and socio-economic variables such as employment status, family responsibilities, or mental health indicators, all of which may affect academic progression.

Furthermore, educational structures vary internationally. In Finland, the flexibility of degree programs and the absence of high tuition costs create conditions that may not apply elsewhere. In contrast, financial and regulatory pressures in countries such as the United States or Singapore may lead to different behavioural patterns. These contextual factors should be considered before applying this model to other institutions or educational systems.

Future research could explore the transferability of the model across different academic environments, incorporate broader contextual data, and refine predictions for students with non-traditional or highly irregular trajectories.

7.5. Conclusion

By addressing the three research questions, this study demonstrates the value of interpretable machine learning for educational planning. The model not only accurately predicts graduation timelines but also offers behavioural insights that support data-informed academic advising. The integration of predictive models, explainability tools, and interactive dashboards lays the foundation for institutional practices that proactively support student success and reduce delayed graduations.

8. SUMMARY

This study aimed to predict the remaining time to graduation for still-studying students enrolled in combined degree programs by applying machine learning techniques. By leveraging historical data from previously graduated students, the research developed a predictive framework that integrates feature engineering, exploratory analysis, and model evaluation.

Among the models tested, XGBoost demonstrated the highest predictive accuracy, achieving a Mean Absolute Error (MAE) of 0.2419, Mean Squared Error (MSE) of 0.1115, and an R-squared (R^2) value of 0.9596. These metrics confirm its robustness in forecasting graduation timelines. Feature importance analysis using SHAP values revealed that Total Gap, Average Credits per Term, and Total Credits were the most influential factors affecting graduation time. These findings were further supported by case study comparisons, which illustrated clear behavioural differences between still-studying and graduated students, particularly in credit accumulation and interruption patterns.

The study's results offer practical value for higher education institutions by providing a predictive framework that can support academic advising and decision-making. Universities can utilize the model to identify students at risk of delayed graduation and implement proactive interventions, such as personalized counselling, workload planning, or academic support services. The interactive dashboards developed as part of this research enhance these capabilities by allowing decision-makers to explore trends and student trajectories in real time.

Despite the promising results, the study faced certain limitations. The dataset included only students registered from 2004 onward, potentially limiting longitudinal insights. Furthermore, the model did not incorporate personal or socioeconomic variables, such as financial status, employment, or life events, that may also impact study progression. Future research could benefit from integrating these additional factors and extending the model to other degree types or institutions to assess generalizability.

In conclusion, this thesis demonstrates the value of machine learning in educational planning by providing interpretable, accurate predictions and identifying the academic behaviours most associated with extended study durations. The framework developed contributes not only to academic research but also offers actionable tools for improving student retention and graduation outcomes.

9. REFERENCES

- [1] M. F. Musso, C. F. R. Hernández, and E. C. Cascallar, “Predicting key educational outcomes in academic trajectories: a machine-learning approach,” *Higher Education*, vol. 80, no. 5, pp. 875–894, 2020.
- [2] S. Garmpis, M. Maragoudakis, and A. Garmpis, “Assisting educational analytics with automl functionalities,” *Computers*, vol. 11, no. 97, 2022.
- [3] S. Viswanathan and S. V. Kumar, “Study of students’ performance prediction models using machine learning,” *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 2, pp. 3085–3091, 2021.
- [4] O. Edeh, K. Almuzaini, F. Onu, D. Verma, G. Ugboaja, M. Puttaramaiah, and R. Kwasi, “Prospects and challenges of using machine learning for academic forecasting,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, 2022. Article ID 5624475.
- [5] N. Sghir, A. Adadi, and M. Lahmer, “Recent advances in predictive learning analytics: A decade systematic review (2012-2022),” *Education and Information Technologies*, vol. 28, pp. 8299–8333, 2023.
- [6] N. Suhaimi, S. Abdul-Rahman, S. Mutalib, N. Hamimah, A. Hamid, A. Md, and A. Ab Malik, “Review on predicting students’ graduation time using machine learning algorithms,” *I.J. Modern Education and Computer Science*, vol. 7, pp. 1–13, 2019.
- [7] R. Geetha, T. Padmavathy, and R. Anitha, “Prediction of the academic performance of slow learners using efficient machine learning algorithm,” *Advances in Computational Intelligence*, vol. 1, no. 4, 2021.
- [8] D. Kabakchieva, “Business intelligence systems for analyzing university students data,” *CYBERNETICS AND INFORMATION TECHNOLOGIES*, vol. 15, no. 1, 2015.
- [9] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, “Predicting student dropout and academic success,” *Data*, vol. 7, no. 11, p. 146, 2022.
- [10] Y. Ariani, M. Masrizal, and R. Muti’ah, “Prediction of student graduation rates using the artificial neural network backpropagation method,” *SinkrOn*, vol. 8, no. 2, pp. 1169–1177, 2024.
- [11] H. Yuliansyah, R. A. P. Imaniati, A. Wirasto, and M. Wibowo, “Predicting students graduate on time using c4.5 algorithm,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 7, no. 1, pp. 67–73, 2021.
- [12] L. Perkhofer, C. Walchshofer, and P. Hofer, “Does design matter when visualizing big data? an empirical study to investigate the effect of visualization type and interaction use,” *Journal of Management Control*, vol. 31, pp. 55–95, 2020.

- [13] G. B. Mandava, K. T. Phani, S. Kiran, D. Rajeswara Rao, and G. Mandava, "Analysis and design of visualization of educational institution database using power bi tool," *Global Journal of Computer Science and Technology*, vol. 18, no. 4, pp. 1–10, 2018.
- [14] W. Villegas-Ch, X. Palacios-Pacheco, and S. Luján-Mora, "A business intelligence framework for analyzing educational data," *Sustainability*, vol. 12, no. 5745, 2020.
- [15] S. Boopathy and P. Kumar, "Predictive analytics with data visualization." Preprint, Research Square, 2022. Preprint available at Research Square. DOI: <https://doi.org/10.21203/rs.3.rs-803205/v1>. Accessed: May 29, 2025.
- [16] D. Arora, "Comprehensive analysis of factors influencing the real-world application of machine learning for student success rate calculation and their impacts on student achievement educational institutions," *World Journal of Advanced Research and Reviews*, vol. 19, no. 03, pp. 942–953, 2023.
- [17] L. Al-Alawi, J. Al, A. Tarhini, and A. Al-Busaidi, "Using machine learning to predict factors affecting academic performance: The case of college students on academic probation," *Journal of Educational Data Mining*, vol. 15, no. 1, pp. 1–20, 2023.
- [18] A. C. Lagman, L. P. Alfonso, M. L. Goh, J. P. Lalata, J. P. H. Magcuyao, and H. N. Vicente, "Classification algorithm accuracy improvement for student graduation prediction using ensemble model," *International Journal of Information and Education Technology*, vol. 10, no. 10, pp. 723–726, 2020.
- [19] H. Brdesee, "Predictive model using a machine learning approach for enhancing the retention rate of students at-risk," *International Journal on Semantic Web and Information Systems*, vol. 18, no. 1, 2023.
- [20] V. Riyanto, A. Hamid, and R. Ridwansyah, "Prediction of student graduation time using the best algorithm," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 2, no. 1, 2019.
- [21] T. Ojha, *Prediction of Graduation Delay Based on Student Characteristics and Performance*. Phd thesis, University of New Mexico, 2017.
- [22] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, "Predicting time to graduation at a large enrollment american university," *PLoS ONE*, vol. 15, no. 11, p. e0242334, 2020.
- [23] N. Gilbert, "Predicting success: an application of data mining techniques to student outcomes," *International Journal of Data Mining & Knowledge Management Process*, vol. 7, no. 2, pp. 01–20, 2017.
- [24] K. Hynynen, "Predicting the on-time graduation of university students based on their study performance," master's thesis, Lappeenranta-Lahti University of Technology LUT, LUT Business School, Business Administration, 2023.

- [25] R. Yusof, N. Hashim, N. A. Rahman, S. Y. M. Yunus, and N. A. A. Fadzillah, "Academic performance prediction model using classification algorithms: Exploring the potential factors," *International Journal of Academic Research in Progressive Education and Development*, vol. 11, no. 3, pp. 706–724, 2022.
- [26] A. Jokhan, A. A. Chand, V. Singh, and K. A. Mamun, "Increased digital resource consumption in higher educational institutions and the artificial intelligence role in informing decisions related to student performance," *Sustainability*, vol. 14, no. 4, p. 2377, 2022.
- [27] M. Awaji, *Evaluation of machine learning techniques for early identification of at-risk students*. Doctoral dissertation, Nova Southeastern University College of Engineering and Computing, 2018.
- [28] F. Salas and J. Caldas, "Predicting undergraduate academic performance in a leading peruvian university: A machine learning approach," *Educación*, vol. 33, no. 64, pp. 55–85, 2024.
- [29] C. Özkurt, "Assessing student success: The impact of machine learning and xai-bbo approach," *Journal of Smart Systems Research*, vol. 5, no. 1, pp. 40–54, 2024.
- [30] A. K. Hamoud, M. B. M. Kamel, A. S. Gaafar, A. S. Alasady, A. M. Humadi, W. A. Awadh, and J. M. Dahr, "A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 2, pp. 1105–1116, 2022.
- [31] F. F. Ananna, R. Nowreen, S. S. R. Al Jahwari, E. A. Costa, L. Angeline, and S. R. Sindiramutty, "Analysing influential factors in student academic achievement: Prediction modelling and insight," *International Journal of Emerging Multidisciplinaries: Computer Science & Artificial Intelligence*, vol. 2, no. 1, pp. 1–20, 2023.
- [32] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, 3rd ed., 2020.
- [33] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2nd ed., 2019.
- [34] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
- [35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

- [36] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. B. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [37] C. Molnar, *Interpretable Machine Learning*. 2022. Available at: <https://christophm.github.io/interpretable-ml-book/>. Accessed: May 29, 2025.
- [38] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer, 2001.
- [39] P. Probst, M. N. Wright, and A.-L. Boulesteix, “Hyperparameters and tuning strategies for random forest,” *arXiv preprint arXiv:1804.03515*, 2019.
- [40] P. I. Frazier, “A tutorial on bayesian optimization,” *arXiv preprint arXiv:1807.02811*, 2018.
- [41] T. O. Hodson, “Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not,” *Geoscientific Model Development*, vol. 15, pp. 5481–5487, 2022.
- [42] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation,” *PeerJ Computer Science*, vol. 7, p. e623, 2021.
- [43] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Statistics Surveys*, vol. 4, pp. 40–79, 2010.
- [44] Z. Zhang, “Residuals and regression diagnostics: focusing on logistic regression,” *Annals of Translational Medicine*, vol. 4, no. 10, p. 195, 2016.