RIGA NORDIC
UNIVERSITY

Studiju programma **42484 Informācijas sistēmas (bak.)**

Dabaszinātņu un datortehnoloģiju katedra

RĪGAS ZIEMEĻVALSTU UNIVERSITĀTE
(RNU)

Studentu akadēmiskās veiktspējas prognozēšana, izmantojot mašīnmācīšanās modeļus

BAKALAURA DARBS

Students: Vijay Bangaru Abiram Kotani

Darba vadītājs: Amit Joshi Lecturer, Ms.ing

RĪGA 2025

RIGA NORDIC
UNIVERSITY

Study Programme	**42484 Information Systems (BSc)**

Department of Natural Sciences and Computer Engineering

RIGA NORDIC UNIVERSITY
(RNU)

Student Academic Performance Prediction Using Machine Learning Models

BACHELOR’S THESIS

Student:	Vijay Bangaru Abiram Kotani
Supervisor:	Amit Joshi Lecturer, Ms.ing

Riga 2025

Anotācija

Studentu akadēmisko sasniegumu prognozēšana ir kļuvusi par nozīmīgu pētījumu virzienu augstākajā izglītībā, ņemot vērā studējošo atbiruma problēmas un digitālo mācību datu pieaugošo pieejamību. Precīzi un interpretējami prognozēšanas modeļi ļauj savlaicīgi identificēt studējošos ar paaugstinātu akadēmisko risku un atbalstīt datos balstītu lēmumu pieņemšanu izglītības iestādēs. Šis bakalaura darbs ir veltīts informācijas tehnoloģijās balstīta analītiska risinājuma izstrādei un novērtēšanai studentu akadēmisko rezultātu prognozēšanai, izmantojot uzraudzītās mašīnmācīšanās metodes un izskaidrojamas mākslīgā intelekta pieejas, kas pielietotas mūsdienu publiski pieejamiem studentu datiem.

Bakalaura darbs sastāv no trim galvenajām nodaļām. Pirmajā nodaļā tiek apskatīts studentu akadēmisko sasniegumu prognozēšanas teorētiskais pamats, izglītības datu ieguves metodes, uzraudzītās mašīnmācīšanās algoritmi un izskaidrojama mākslīgā intelekta koncepcijas. Otrajā nodaļā ir aprakstīta pētījuma metodoloģija, tostarp datu kopas izvēle, datu pirmapstrāde, pazīmju veidošana un eksperimentālā pētījuma struktūra. Trešajā nodaļā ir izklāstīta praktiskā daļa, kurā tiek izstrādāti, novērtēti un interpretēti mašīnmācīšanās modeļi, balstoties uz eksperimentāli iegūtiem rezultātiem no reālas studentu akadēmisko datu kopas.

Pētījuma metodoloģiskais pamats ietver zinātniskās literatūras analīzi, strukturētu datu apstrādi, uzraudzīto klasifikācijas modeļu izstrādi, modeļu novērtēšanu, izmantojot vairākus klasifikācijas rādītājus, kā arī post-hoc izskaidrojamības metožu pielietošanu. Praktiskie rezultāti apliecina, ka ansambļa tipa mašīnmācīšanās modeļi nodrošina augstu prognozēšanas precizitāti un ka studiju progresu raksturojošie rādītāji ir nozīmīgākie akadēmisko sasniegumu ietekmējošie faktori. Darba rezultāti pierāda piedāvātās pieejas piemērotību agrīnās brīdināšanas un akadēmiskās uzraudzības sistēmām.

Bakalaura darbs izstrādāts **80** lapaspusēs un ietver **07** tabulas, **05** attēlus, **30** avotiem.

Abstract

Prediction of student academic performance is a recently more often addressed research problem in higher education that can help universities tackle student dropout and retention problems in context of increased availability of digital educational data. Transparent and interpretable models predicting the expected student academic performance can be used for early identification of at-risk students and the implementation of evidence-based, academically-focused interventions. The Bachelor thesis develops and evaluates an information technology-based analytical solution for student academic performance prediction, using supervised machine learning methods and explainable artificial intelligence techniques on a recent publicly available student data.

The Bachelor Paper consists of three main chapters. Chapter 1 reviews the theoretical background of academic performance prediction, educational data mining, supervised machine learning methods, and explainable artificial intelligence, establishing the scientific foundation for the study. Chapter 2 presents the research methodology, including dataset selection, data preprocessing, feature engineering, and the overall experimental framework used to ensure methodological validity and reproducibility. Chapter 3 contains the practical research, where machine learning models are developed, evaluated, and interpreted using experimental results obtained from a real-world student performance dataset.

The Bachelor Paper has the methodological basis that is realized through the literature review of scientific publications in the field of student performance prediction and classification, the data preprocessing workflow, the supervised classification modelling with a real-world dataset, model performance evaluation with multiple classification metrics that are independent of the overall dataset label accuracy and post-hoc model explainability analysis. Logistic Regression, Random Forest and Gradient Boosting based models are implemented, used on a real-world student performance dataset and compared using accuracy-independent metrics (precision, recall, F1-score, ROC-AUC) and their explainability is analyzed using the applied explainability methods for the most important variables interpretation.

The experimental results show that the used ensemble-based machine learning models provide a strong predictive performance and the model analysis indicates that the progression variables, especially the curricular unit completion and the obtained grades, are the most indicative predictive features. The use of the applied explainability analysis methods on the developed predictive models confirms the overall model transparency and practical interpretability for the future use in early-warning and academic monitoring systems.

The Bachelor Paper is written in **80** pages and contains **07** tables, **05** figures, **30** sources.

Key words / Atslēgvārdi

Personnel	**Personāls**
Student academic performance	Studentu akadēmiskie sasniegumi
Educational data mining	izglītības datu ieguve
Machine learning	mašīnmācīšanās
Classification models	klasifikācijas modeļi
Explainable artificial intelligence	izskaidrojams mākslīgais intelekts
Early-warning systems	agrīnās brīdināšanas sistēmas
Ensemble models	ansambļu modeļi

Contents

List of Figures

List of Tables

Introduction

The primary objective of this bachelor's thesis is to design, implement, and evaluate an information technology–based analytical solution for predicting student academic performance, leveraging modern machine learning algorithms and explainable artificial intelligence techniques. This work focuses on supervised learning approaches and modern publicly available student datasets published after 2019 with the goal to develop predictive models with high accuracy and model explainability. In addition to generating accurate predictions, this research also aims to produce human-interpretable and actionable insights for evidence-based decision-making in educational settings, such as the early detection of at-risk students and other use cases.

Topicality of the problem: Student Academic Performance Prediction Using Machine Learning Models

Aim of the study: To apply modern machine learning and explainability techniques to predict student academic performance using recent publicly available datasets, and to identify the most influential factors driving student success.

Object of the study: Academic performance data from recent public student datasets (Kaggle 2020–2023 datasets such as "Predict Students Dropout and Academic Success").

Subject of the study: The process of predicting and interpreting student performance using modern ML algorithms and explainability tools on 2020–2023 student datasets.

Research problem: Most existing studies use outdated datasets (e.g., UCI 2008), limiting relevance. There is a need to use modern datasets (2020–2023) and apply explainability methods to generate fresh, evidence-based insight into academic success factors.

Tasks of the study: To reach the goal stated above, the bachelor's thesis work has to accomplish the following research tasks: 1. Review modern theoretical and empirical developments in educational data mining, learning analytics, and student performance prediction. 2. Analyze supervised machine learning approaches commonly used in the literature to address the academic performance classification problem. 3. Collect and prepare publicly available student performance datasets that were published after 2019. 4. Design and implement a reproducible data preprocessing and feature engineering pipeline using Python-based tools. 5. Develop and train machine learning models for student performance prediction, including

Logistic Regression, Random Forest, and Gradient Boosting. 6. Evaluate the performance of these models using standard classification metrics (ROC-AUC, precision, recall, F1-score, confusion matrices, etc.). 7. Apply machine learning model explainability techniques to interpret predictions and to identify the key factors affecting academic success. 8. Develop actionable analytical recommendations for the improvement of academic performance monitoring and early intervention based on the results.

Hypothesis: Modern ML models combined with explainability techniques (SHAP/LIME), applied to recent 2020–2023 student datasets, can accurately predict academic performance and reveal key determinants of student success.

Research methods: Data collection from 2020–2023 public student datasets Machine learning models Explainability (SHAP/LIME) ROC-AUC, F1-score Dashboards/visualizations

Approbation of the study: Most existing studies use outdated datasets (e.g., UCI 2008), limiting relevance. There is a need to use modern datasets (2020–2023) and apply explainability methods to generate fresh, evidence-based insight into academic success factors.The scope of the bachelor's thesis work includes a literature review on the most relevant aspects of the research, a definition of the applicable analytical methods and tools for the educational data analysis, a choice of recent and publicly available datasets of student data and the execution of the described predictive and interpretability methods in a unified computational environment. The educational performance data of students has become a popular data source for the academic analysis by higher education institutions since these data reflect the learning outcomes, engagement, and system challenges experienced by students in the learning process. The large volumes of data from the digital learning platforms and institutional information systems have accumulated, and, therefore, there is an opportunity to study these data to gain new insights about student performance and behavior. Predicting student performance is one of the most relevant tasks of learning analytics and educational data mining as it enables institutions to identify at-risk students and implement timely interventions. Well-performing prediction models can also support decision-making in academic advising, curriculum development, and resource allocation. The data-driven academic performance prediction is also an essential building block of student retention efforts and can be a source of data-based insights to improve the learning outcomes. Recent studies have shown that data-driven student performance prediction can be a significant factor in the improvement of institutional decision-making provided that the predictive models are accurate and interpretable. However, many existing studies are using an outdated data sources for this purpose. One of the most commonly used datasets in this research area is the UCI Student Performance dataset, which has been published more than 10 years ago. As the educational systems have undergone significant changes since then, especially after 2020, the analysis performed using the old dataset is not going to be as relevant as the analysis with a recent data. The student experience and academic performance have changed under the influence of digitalization, the introduction of blended and online learning formats, and other shifts in learning conditions and assessment types. Therefore, there is a research gap that can be addressed by the work of this thesis, which is to revisit the prediction of student academic performance with the application of a modern analytical solution with updated publicly available datasets. This work focuses explicitly on the application of recently published (2020-2023) datasets and does not use outdated sources of student data. Machine learning models have been shown to exhibit good predictive performance for the academic outcome prediction due to their capacity to model complex relationships between the multitude of student factors. Ensemble-based approaches such as Random Forest and Gradient Boosting are known to be highly competitive for capturing non-linear relationships between the demographic, academic, and behavioral factors. On the other hand, simple statistical models like Logistic Regression still serve as good baselines due to their transparency and stable probabilistic interpretation. A comparison of different families of models is to be performed to assess the predictive performance and explainability of different approaches. In addition to high accuracy, a critical issue for practical adoption of machine learning models into the data-informed educational decision-making is the lack of transparency. Black-box predictions with no accompanying explanation can reduce the trust in ML models among educational practitioners and administrators, as well as introduce ethical issues. For this reason, the work on explainable artificial intelligence has emerged as a critical direction of applied machine learning research. Explainability methods such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) can be used to interpret model predictions in a post-hoc manner by quantifying the individual contributions of each input feature to the output prediction. These approaches to explainability can be used to analyze complex ML models in an interpretable way to non-technical stakeholders and increase the accountability and practical usability of ML-based decision-support systems.

The research solution developed in this work has been implemented using the Python programming language because of its maturity, open-source libraries and tools for data analysis, machine learning, and visualization. Python is cross-platform, supports reproducibility, and is widely adopted in both academic research and industry practice, which makes it one of the most common choices for data analytics tasks in Python. Libraries like Pandas, NumPy, Scikit-learn, SHAP, and Matplotlib have been used to develop a modular, transparent, and easy-to-extend analytical pipeline that can be also easily adapted for other educational datasets. The research direction of this bachelor's thesis work can be described as an integration of theoretical analysis, practical implementation, and experimental evaluation. The theoretical part involves the study of relevant background in the fields of educational data mining, supervised learning, and explainable AI. The practical part consists of the implementation of this theory into a practical predictive and explainability research framework using real-world educational datasets. The experimental evaluation focuses on the quantitative and qualitative comparison of the predictive performance and explainability of the developed models to ensure that the obtained results and insights are accurate, interpretable, and can be used in data-informed decision-making in educational settings.

The bachelor's thesis has three main chapters. The first chapter discusses the theoretical background concerning the student performance prediction problem, educational data mining, machine learning classification methods, and explainable artificial intelligence. This chapter also reviews the state of research on this topic and establishes the context and a foundation for the thesis work. The second chapter outlines the research methodology, dataset selection, data preprocessing, and feature engineering that have been performed for this research. This chapter also includes a description of the experimental design and the justifications of the applied analytical approaches. The third chapter of the thesis is related to the model development, evaluation of the predictive performance, and the explainability analysis. This chapter compares the predictive results of different models, provides interpretations for the most influential academic performance factors, and discusses the practical implications of the obtained results. The conclusion summarizes the results of this research, outlines the main theoretical and practical contributions of this work, and formulates the recommendations for future work. In general, this research work is intended to support the educational analytics field by providing an example of how the modern machine learning models, when paired with explainability approaches and recent educational datasets, can produce not only highly accurate but also transparent and actionable insights into student academic performance.

1. ACADEMIC PERFORMANCE PREDICTION AND EXPLAINABLE MACHINE LEARNING

Chapter 1 of this bachelor thesis lays out the theoretical and methodological basis of this research. The purpose of Chapter 1 is to present a logical and coherent synthesis of key scientific ideas, analytical models and computational paradigms that are at the core of explainable machine-learning solutions for prediction of students' academic performance based on recent public datasets. In particular, the theoretical background of this work is built upon a diverse but internally consistent set of perspectives on educational data mining, supervised machine learning and explainable artificial intelligence. Together, this body of scientific theory justifies and, where possible, precedes the empirical analyses carried out in later chapters of this thesis, while positioning these research activities firmly within the landscape of current and leading-edge developments in data analytics and information systems. The prediction of students' academic performance is one of the major research focuses within the fields of learning analytics and educational data mining because of its practical importance for academic quality management, retention and evidence-based policy. The central theoretical assumption of academic performance modelling is that a student's educational outcome is, to a significant degree, predictable from past and contextual data about the student's academic trajectory, background and learning environment. By contrast to purely descriptive approaches to educational research, predictive data analytics seek to formalise and, crucially, generalise relationships between these phenomena for purposes of actionable foresight and early intervention (Romero & Ventura, 2020) [23].

The distinction between educational and business or industrial data lies in the fact that student performance is the result of complex socio-cognitive rather than mechanical, biological or transactional processes. Students' educational outcomes are driven by a mixture of academic, behavioural, demographic and institutional factors that interact and determine students' success in non-linear and context-sensitive ways. Constructs such as prior academic performance, attendance, engagement, socio-economic status and instruction have all been found to affect student performance and learning, although not necessarily with the same relative importance across all educational levels and contexts (Siemens & Baker, 2020) [25]. Consequently, educational predictive analytics call for analytic methods that can capture both additive main effects and higher-order interactions. The dominant data structure for student performance from the perspective of data analysis is that of a tabular, cross-sectional or longitudinal data set, where each row is an individual student or student-course combination described by a vector of features and outcome labels. In contrast to time-series forecasting where the temporal order of observations is the primary organising principle, educational data often combine static (demographics) and dynamic (academic history, learning activity) feature types and allow for varying degrees of correlation between modelled variables (Chao et al., 2025) [6].

Historically, the academic performance prediction problem has been addressed using classical statistical methods such as linear and logistic regression. The reason is, first, the comparatively small data requirements of these methods compared to contemporary machine-learning algorithms and, second, their strong interpretability that, in educational environments, is crucial for trust and user experience. Logistic regression has been used to model pass/fail or continue/no-continue type of academic outcomes in particular as the conditional probability of a student's success as a function of explanatory features. Its transparent parameterisation and the availability of a full inferential statistical framework make logistic regression a popular statistical modelling choice in the context of educational data (Kotsiantis et al., 2004) [15]. However, the main disadvantage of linear models is their decision boundary, which constrains their ability to accurately fit the complex relationships often observed in contemporary educational data. The situation has, to some extent, changed after 2016 as a result of the increased availability of large-scale and detailed educational data sets from learning management systems, e-testing platforms, administrative records and data-sharing partnerships. Supervised machine-learning models including decision trees, random forests, gradient boosting machines and neural networks have shown superior prediction results in many student performance prediction tasks because they can model complex and non-linear relationships and feature interactions without being explicitly programmed to do so (Alzahrani, 2024) [3]. In particular, ensemble methods have been shown to be highly effective in student performance prediction due to their relative robustness to the noise and high dimensionality of educational data.

Random forest is a classifier that builds a set of decision trees on bootstrapped samples of the original dataset and then aggregates their predictions by averaging or majority voting. By doing so, it not only builds a less overfitted, non-linear prediction boundary across multiple dimensions of student data, but also captures feature interactions implicitly. Gradient boosting methods are a special case of ensemble classifiers that use an additive model with a differentiable loss function to sequentially train models on the residuals of previous models in order to maximise the improvement in overall predictive performance. Empirical evidence in the field of educational prediction suggests that random forest and gradient boosting methods consistently outperform single classifiers on student performance prediction tasks, particularly in the presence of behavioural and engagement data in addition to academic performance features (Akçapınar et al., 2019) [1]. Despite these advantages in predictive accuracy and, to an extent, robustness, machine-learning classifiers do not address one of the fundamental theoretical and practical challenges in educational data mining: that of transparency. These models, especially of the black-box type, do not provide direct insights into their internal logic and decision-making process. This fact limits the scope of their adoption in educational settings where a human user typically requires an explanation of model predictions. Machine learning classifiers and neural networks in particular raise serious issues around the fairness, accountability and potential bias of predictions, especially where the predictions have a high-stakes nature or significant impact on the individual student (Holstein et al., 2022) [13].

The challenges discussed in this section have been the impetus for the emergence of a distinct line of research known as explainable artificial intelligence or eXplainable AI for short. Explainable AI represents an effort to make the internal decision logic of complex machine-learning models transparent without compromising their predictive performance. Post-hoc explainability methods in particular have seen wide adoption as practical tools for interpretation, primarily because they can be applied to any classification or prediction algorithm regardless of internal structure (Lundberg & Lee, 2017) [18]. SHAP and LIME are, arguably, two of the most widespread post-hoc explainability methods in applied machine learning today. SHAP is a cooperative game theory approach to explainable AI that attributes a prediction outcome to a set of individual features by calculating so-called Shapley values. In other words, SHAP provides a framework for breaking down a prediction into a sum of individual contributions from features in the student data point. As such, SHAP can be used both for local explanations (i.e. why this particular prediction was made) and global explanations (feature importance across a whole dataset). This approach is highly theoretically justified, particularly in terms of its consistency and additivity properties that allow for easy comparison of features' contributions across classifiers and data points (Molnar, 2022) [19]. In the context of educational analytics, SHAP-based explainability can help to determine the exact contribution of prior grades, attendance, engagement and other factors in prediction of academic outcomes in a rigorous and mathematically justified way. LIME is an alternative local model-agnostic explanation framework that uses surrogate models, usually simple linear regressions, to approximate the decision boundary of a complex model in the neighbourhood of a specific instance. To this end, LIME perturbs input features around an instance of interest and then measures the changes in prediction. As a result, the method produces a localised and human-readable explanation of which features are most important for the specific student. While not sharing the same consistency guarantees as SHAP, LIME has been used as an explanation tool for individual-level analysis and personalised feedback in educational applications (Ribeiro et al., 2016) [22].

The concept of explainable AI in educational data mining is directly related to the general goals of learning analytics as a field. Learning analytics does not merely emphasise prediction accuracy as the sole value of educational prediction, but also utility, interpretability and pedagogical significance of modelling results. Predictive models that can be interpreted and explained to a human user can be used to power early-warning systems, provide diagnostic insights to targeted intervention or support and, generally, build trust in computational approaches to academic decision-making. Empirical studies have shown that in the context of institutional decision-support systems, educators are more likely to adopt predictive models that provide explanations for their outputs than their black-box counterparts of comparable performance (Altabrawee et al., 2019) [2]. The present thesis also relies on the technological maturity of contemporary data-science and machine-learning ecosystems. The Python programming language has emerged as the de facto industry standard for applied machine-learning and artificial intelligence because of its readable, flexible syntax, extensive library and package support, and excellent interoperability. Pandas and NumPy are standard data manipulation and numerical computation libraries that form the core of every data-driven Python project (McKinney, 2010; van der Walt et al., 2011). Scikit-learn is a library that has established itself as the standard interface to supervised learning for the majority of modern machine-learning practitioners, with unified APIs for model implementation, tuning, evaluation and pipelines (Pedregosa et al., 2011) [21]. In the context of the present study, SHAP and LIME provide the implementation backbone of the chosen explainability approach.

The theoretical background presented in this chapter demonstrates the fact that student academic performance prediction rests at the intersection of diverse but coherent streams of research in educational data mining, machine learning and eXplainable AI. This connection between fields provides both conceptual and technological scaffolding for the present work and shows that effective educational data analysis requires not only high-performing predictive models, but also transparent interpretability mechanisms that can translate model outputs into pedagogically significant knowledge. The structure of Chapter 1 is guided by this logic and the order of sections follows the natural logic of moving from broader domain context to narrower methodological focus. Section 1.1 is dedicated to setting up the core concepts and domain context of the student performance prediction problem in higher education. Section 1.2 provides an overview of educational data mining and the history of student performance modelling. Section 1.3 reviews supervised machine-learning methods that can be used in the context of academic outcome prediction. Section 1.4 discusses explainable artificial intelligence approaches and their educational applications. Section 1.5 addresses evaluation metrics and interpretability criteria for student analytics. Finally, Section 1.6 offers a summary of the main theoretical results and discusses their implications for the empirical work in the rest of the thesis. Taken together, the theoretical synthesis outlined in this chapter provides the scientific and methodological foundation for the implementation phase of this thesis. Anchoring the later, empirical analyses in established scientific theory is an important step towards adhering to international research quality standards and making a contribution to the field of educational data analytics.

1.1. ACADEMIC PERFORMANCE IN HIGHER EDUCATION AND LEARNING ANALYTICS

Academic performance may be broadly defined as the extent to which students meet the intended learning outcomes as established by the curriculum or educational institution. In the context of higher education, academic performance is operationalised at various levels, such as courses, degree programs, and overall institutions, through different metrics, such as course grades, cumulative GPA, number of credits earned, course completion, and graduation. These metrics serve multiple purposes, such as assessing individual student learning and progress, evaluating institutional quality and outcomes, informing program improvement, and shaping quality assurance processes. With the datafication of learning and education in higher education institutions, academic performance has become a common target for various learning analytics or educational data mining methods. Learning analytics is defined as “the measurement, collection, analysis and reporting of data about learners and their contexts for purposes of understanding and optimizing learning and the environments in which it occurs” (Ferguson & Clow, 2017) [9]. This is different from traditional educational research as this paradigm is underpinned by the idea of integrating learning theory, data science and information systems in order to derive meaningful insights and make evidence-based inferences from large educational datasets. Learning analytics leverages the capability of collecting and storing massive amounts of digital data in institutions in order to monitor and assess student learning and development on a continuous and real-time or near–real-time basis, which makes it possible to predict student performance and progress using a wide range of data sources and machine-learning models (Ferguson & Clow, 2017) [9].

Academic performance is a multi-faceted construct affected by a number of cognitive, behavioural and contextual factors. Cognitive factors include prior academic performance, domain-specific knowledge, and learning strategies. Behavioural factors include attendance, participation, engagement with learning materials, and submission of assessments. Contextual factors include socio-economic status, institutional and program factors, course design, and instructional quality, among other factors. It is also important to note that the effect of any of these variables on student performance can vary across different contexts. It is worth noting that a body of research has shown that no single factor can sufficiently explain student performance, instead, student performance is the product of the complex interplay of several mutually dependent and interrelated variables (Bhanpuri et al., 2015) [5]. From the data-analytic perspective, student academic performance datasets are typically represented as a table of instances. Instances can represent students themselves or a student–course combination, or snapshots of a student’s academic trajectory in the form of timestamps or semesters (Saqr et al., 2017) [24]. In the first case, the student dataset may contain static attributes describing their background and demographics. In the second case, the table may contain dynamic attributes, such as grades collected over time or data on the interactions with learning materials captured by the institution’s learning management system. This data can be heterogeneous in its nature, meaning that datasets can have mixed-type attributes (categorical and numerical), can have missing values, may have different class distributions, and the relative importance of different attributes can vary across domains and datasets (Saqr et al., 2017) [24]. The use of digital learning platforms in the higher education context has expanded the range of student data collected over time. With the use of learning management systems, online quizzes and exams, and student information systems, institutions have been collecting increasingly large volumes of fine-grained data on student learning activities on an ongoing basis. In addition to descriptive analytics, such data allows for leveraging predictive and prescriptive analytics in order to inform the design of early-warning systems to prevent student failure and dropout early before it is too late (Paneva-Marinova, 2006) [20].

There are some important theoretical and methodological considerations that must be kept in mind when applying academic performance prediction. The first is the difference between correlation and causation. Machine-learning models can be used to identify patterns and associations between various features and the student academic performance but do not directly point to causes and effects. This means that while a model may associate a low level of engagement with low grades, the former may not be a direct cause of the latter. For example, a low number of access to the learning platform can be as a result of a student experiencing health issues, personal problems, or financial hardships. This distinction between correlation and causation is particularly important for prescriptive analytics, as interventions based on the model predictions are expected to be evidence-based, and must have clear pedagogical justification and be educationally sound (Gašević et al., 2014) [10]. The second consideration is that of the data quality and representativeness. The datasets collected from educational contexts typically have missing values and may be subject to biases that can be intrinsic to the institution's policy and student processes or artefacts of the data-collection process (Kuzilek et al., 2017) [16]. For example, attendance records can be missing, students can self-report their demographic information or students may be subject to different assessment regimes in different courses or at different institutions. This can impact both accuracy and fairness of machine-learning models if not taken into account during data-preparation and model evaluation steps. A number of recent studies on learning analytics and educational data mining methods have emphasised the importance of rigorous data preprocessing and validation steps in order to ensure the reliability of the modelling outcomes (Kuzilek et al., 2017) [16]. The growing use of predictive models in education has raised ethical questions regarding issues of transparency and explainability, accountability, and student agency. Predictions that are used in decision-making, such as academic advising, scholarship allocation, or progression, can be subject to misconceptions, misinterpretations, and errors, particularly in cases where the model can be an opaque black box that is not understandable to students and educators alike. A growing body of work in learning analytics and educational data mining research has emphasised the need to provide accurate and fair predictions together with interpretable explanations of model predictions (Tsai et al., 2019) [29].

A more theoretical consideration regarding academic performance prediction relates to its formulation as a machine-learning problem. In some cases, this can be seen as a classification or regression problem depending on how the outcome variable is defined. Binary or multi-class classification can be used to predict binary or categorical outcomes, such as pass/fail, dropout/retention, or academic risk level. In contrast, regression models are used to predict continuous outcomes, such as the final exam scores or the final GPA. The formulation of the prediction task and the related evaluation metrics have implications for model selection and training and for the interpretability of the resulting models, and should be aligned with the learning analytics use case (Lakkaraju et al., 2015) [17]. Recent work in learning analytics and educational data mining research has also placed increasing emphasis on the need to align the design of the predictive modelling approaches with pedagogical goals, meaning that the design of the predictive model should take into account the purpose it is expected to serve, and that these should not be designed to maximise predictive accuracy alone, but also be actionable and pedagogically meaningful. In other words, academic performance prediction should be seen not as an end in itself, but as a means to an end (Wise & Jung, 2019) [30]. This perspective requires explainable predictive models that can provide insights into key performance drivers that can be used to identify students that may require academic support, such as support programs or curriculum changes. As a result, academic performance prediction is increasingly framed not as a standalone technical exercise, but as a socio-technical system situated in the educational context. To sum up, in this section, I have provided an overview of the concept of academic performance as it is used in higher education context, the uses and applications of learning analytics and educational data mining methods for its analysis and prediction. This has been done in order to help contextualise the methods and techniques discussed in the rest of this chapter, and to highlight the theoretical and methodological issues that should be taken into account when considering approaches for academic performance prediction.

1.2. EDUCATIONAL DATA MINING AND STUDENT PERFORMANCE MODELLING

Educational data mining (EDM) is an interdisciplinary field that studies how to apply data mining and machine learning techniques to educational data. It aims to develop and use computational methods to analyse and understand the patterns, processes, and outcomes of learning and education. EDM differs from educational statistics in that it is more focused on prediction, discovery, and automation of knowledge extraction from large and complex data sets, rather than on testing hypotheses, generalising from samples, and evaluating interventions at the aggregate level. Student performance modelling is a well-established topic within the scope of EDM. A performance model is a computational representation that captures the relationship between observable student characteristics and their academic outcomes. Performance models are typically built using historical student records that contain various attributes such as prior grades, engagement metrics, demographic information, and institutional factors. The goal is to learn a decision boundary or a functional mapping that can be used to predict future academic outcomes given new or unseen observations. Performance models are useful for early-warning systems, adaptive learning systems, and institutional dashboards. The mathematical foundation for student performance modelling comes from supervised learning, where a labelled dataset is used to train a predictive model. Labels are binary or categorical indicators of academic outcomes, such as whether a student passed or failed a course, dropped out or not, achieved a certain grade or level, or reached a threshold of GPA. Input features are the observable characteristics that are used to make the predictions. They can include prior grades, attendance, engagement, online learning system logs, assignment submission patterns, demographic variables, and other relevant information. The choice of features is often based on both domain knowledge and data-driven relevance, as not all attributes in the dataset may be predictive or interpretable. EDM research has shown that student performance prediction is a context-dependent problem. Models that are trained and validated on data from one institution, program, or cohort may not generalise well to other contexts without retraining or adaptation. This is because of the variability in curriculum design, assessment methods, grading schemes, and student populations across different educational settings. For this reason, recent literature emphasises the importance of using recent and context-specific datasets for model development rather than using benchmark datasets collected in the past under different conditions (Tomasevic et al., 2020) [28].

In the early days of EDM, simple classifiers such as Naïve Bayes, decision trees, and k-nearest neighbours were used for student performance prediction. These methods are easy to understand, implement, and scale with large datasets. However, as the size and complexity of educational data grew, more sophisticated algorithms were developed and adopted. Ensemble methods such as random forests and gradient boosting have become popular in EDM due to their high accuracy and robustness to noisy or imbalanced data. These methods can also capture nonlinear relationships and interactions between features that are hard to model with linear methods. One of the challenges in student performance modelling is the issue of class imbalance. In many educational datasets, the number of students who fail or drop out is much smaller than the number of students who pass or graduate. This imbalance can lead to biased or inaccurate predictions if the classifier is not properly trained or evaluated. To address this issue, EDM researchers have proposed various methods to resample, reweight, or penalise the data to balance the classes. The choice of evaluation metric is also crucial for assessing the performance of student performance models. Accuracy alone is not a reliable measure when the classes are imbalanced or when the cost of different types of errors is not the same. EDM researchers have suggested using more informative metrics such as recall, precision, F1-score, or area under the ROC curve to evaluate the predictive performance of the models on both the majority and the minority classes (Tempelaar et al., 2020) [27]. Another important aspect of EDM is feature engineering. Raw educational data often needs to be transformed or manipulated to create features that are more suitable for modelling. For example, temporal aggregation of engagement events, normalisation of scores, encoding of categorical variables, and imputation of missing values are common preprocessing steps. Feature engineering not only improves the predictive power of the models, but also affects their interpretability. Derived features need to be meaningful and understandable in the educational context. Poorly designed or irrelevant features can lead to spurious correlations or misleading explanations.

More recent EDM literature has also raised concerns about the lack of interpretability of purely predictive approaches that do not offer any understanding or explanation of the factors that influence academic outcomes. Models that have high accuracy but are not transparent or accountable to the educational stakeholders and decision-makers are of limited value. This has led to a paradigm shift from accuracy-first evaluation to a more balanced and holistic framework that considers both the predictive performance and the explanatory usefulness of the models. As a result, student performance modelling is increasingly complemented by explainability methods that allow the inspection and justification of model outputs. The recent advancements in explainable machine learning have influenced the development and application of EDM methods as well. Instead of simply selecting the model with the highest accuracy, EDM researchers and practitioners are now evaluating models based on their interpretability and alignment with pedagogical reasoning as well. Linear models such as logistic regression are still widely used as baselines because of their coefficient interpretability, but tree-based ensembles are often used in combination with post-hoc explanation methods to achieve both performance and understanding. This hybrid approach enables the use of complex models without compromising accountability. Ethical and fairness issues are also an important part of EDM research. Student performance models may encode or amplify biases that exist in the historical data or the educational system, leading to unfair or inaccurate predictions for certain groups of students. Recent research has highlighted the need for bias detection, fairness-aware modelling, and explanation of predictions to ensure the ethical and responsible use of EDM. Regulatory and ethical guidelines are also increasingly requiring that automated systems that support educational decisions be auditable and interpretable, which further emphasises the importance of explainable EDM approaches.

From a systems perspective, EDM-based performance models are typically deployed as components or modules within an institutional information system or an analytics platform. These systems integrate data ingestion, data preprocessing, model inference, and result visualisation into a cohesive workflow. The system needs to be reproducible, scalable, and maintainable, which is facilitated by modern data-science technologies and ecosystems. Scripting and programming languages such as Python allow for rapid prototyping and experimentation with various methods and models, while supporting production-ready analytics through modular design, version-controlled workflows, and deployment automation. In conclusion, EDM provides the methodological and conceptual background for student performance modelling. The field spans supervised learning, feature engineering, evaluation, and ethical considerations for mining and understanding educational data. Recent developments in EDM reflect the increasing emphasis on recent and context-specific data, robust evaluation, and model interpretability. These trends set the scene for the machine-learning and explainability techniques that are presented and used in the following chapters.

1.3. SUPERVISED MACHINE LEARNING METHODS FOR ACADEMIC OUTCOME PREDICTION

Supervised machine learning is the dominant approach used in the contemporary educational data mining literature to predict academic performance. Supervised learning algorithms build a predictive model from a labelled dataset where each example is tagged with an outcome variable. The task then becomes to learn a decision function that can predict an outcome for unseen examples. In the case of learning analytics, the outcome variable of interest would be some representation of academic performance, such as whether a student failed or passed a course, is at risk of academic probation or attrition, achieved a low or high grade, or passed with an overall average mark above a threshold value. Prediction tasks are typically framed in terms of either classification or regression problems, depending on the nature of the outcome variable. Classification is used when the outcome is discrete, for example, pass/fail, at-risk/not at risk, or low/medium/high. Regression is used when the outcome variable is continuous, for example, a final numeric grade or overall GPA score. The majority of learning analytics applications that focus on prediction of academic performance and outcomes use classification problems, as this better aligns with institutional processes that need to intervene at a binary level. As such, the thesis also takes a classification-orientated view. A range of different supervised learning approaches are reviewed and used for academic performance prediction, with their advantages and disadvantages considered in comparison.

Linear classifiers are some of the earliest methods of supervised learning and are still commonly used for supervised learning problems. Logistic regression, a form of linear regression adapted for the classification problem, is the most well known and commonly used of these methods (James et al., 2021) [14]. Logistic regression is a parametric model of the relationship between the input features and a binary outcome, expressed as a conditional probability. The model is linear in the log-odds of the conditional probability and parameters of the model can be learned using maximum likelihood. Regression coefficients can then be used to interpret the direction of feature effects on the outcome. Regularised forms of the linear regression classifier, such as L1 and L2 regularisation, are commonly used to counteract multicollinearity in the feature set (James et al., 2021) [14]. Linear models, however, have limitations. One of the principal limitations of using a linear regression model for predicting academic performance is that the relationships between many of the input features are likely to be non-linear. For example, a student’s performance on a course is likely to be affected by an interaction of their prior performance, engagement with the course, nature of the assessments, and the institution that they are studying at. The relationship between such features and the target variable are unlikely to be linear and the decision function that is learned using a linear model is only an approximation of the true relationship. To model more complex decision functions, non-linear supervised learning techniques have been developed. One of the most well known and common non-linear techniques are decision tree–based methods. Decision trees (Cortez & Silva, 2008) [8] recursively subdivide the input feature space into regions based on splitting rules for each feature. At each node in the tree a splitting rule is used that partitions the input space in a way that the resulting child nodes contain more homogenous (pure) distributions of outcomes. Decision trees are often used for classification and regression tasks as the decision function that they represent is both interpretable and capable of approximating non-linear functions. However, they are prone to overfitting and can be unstable, with small changes to the training data causing large changes to the decision function (Hastie et al., 2017) [11]. Ensemble methods, such as Random Forest (Akçapınar et al., 2019) [1], that combine the predictions of many decision trees have been shown to result in more accurate and stable predictive models. Random Forest uses a combination of bootstrap sampling to draw training data to learn each individual tree and random subsetting of the features to select the splitting rules at each node. The prediction of the Random Forest is then generated by aggregating the predictions of the individual trees through voting (classification) or probability averaging (regression). Predictions from Random Forest models have been shown to have high predictive power for dropout and failure risk in educational settings across a wide range of contexts and applications.

Gradient Boosting (Saxena & Garg, 2017) [18] is another family of ensemble methods based on the aggregation of decision trees. The main difference between gradient boosted and random forests is in the way that individual trees are combined. Rather than each tree being trained independently of the others, gradient boosting approaches learn a sequence of trees such that each tree is trained to correct the errors of the previous one. This results in an additive model of decision trees that is much more powerful than each individual tree and can result in superior prediction performance. The gradient boosting algorithm iteratively fits a decision tree to the current residual errors in the predictions from the existing ensemble and then adds it to the model. Implementations of gradient boosted trees, such as Gradient Boosting Machines and Extreme Gradient Boosting (XGBoost), are commonly used for structured-data prediction tasks and have been shown to achieve state-of-the-art performance in many tasks. In educational settings they have also been shown to result in accurate predictions of dropout risk and other binary outcome measures. The methods are well-suited to high-dimensional feature sets that are common in learning analytics, such as student datasets, and have been shown to be able to capture complex interactions in these feature sets (Chen & Guestrin, 2016) [7]. Neural networks are also able to approximate non-linear functions and have been used for educational performance predictions as well. Feedforward neural networks are an example of a neural network–based method that can be used to learn non-linear decision functions. Recurrent neural networks and long short-term memory architectures can also be used to capture the sequential nature of time-ordered events (Domingos, 2012) [12]. This type of method is often used for modelling clickstream data and temporal engagement patterns extracted from learning management systems.

The selection of an appropriate supervised learning approach for the task of academic performance prediction also needs to consider several factors. The size of the dataset, its feature dimensionality, and constraints on data quality can all have a bearing on the choice of algorithm. Ensemble methods, such as Random Forest, have been shown to be effective even with many weakly performing base learners. As a result, they are robust to many of these common data problems. Imbalanced classes, whereby students who are at-risk of an outcome, such as dropout or failure, form a minority class, is another common problem with datasets in educational settings. This often occurs as these outcomes form the minority of the population of students. Machine learning algorithms also need to be evaluated using appropriate metrics to ensure that their performance is not overestimated. Standard metrics, such as precision and recall, F1-scores, or area under the receiver-operator curve, are often used (Tempelaar et al., 2020) [27]. Feature selection and feature representation can also play an important role in the supervised learning approach taken for a given task. Feature selection is often used as a pre-processing step to improve the interpretability of a model and prevent overfitting (James et al., 2021) [14]. High dimensional feature spaces often have redundant or sparsely contributing features that can reduce the ability of a model to generalise to unseen data. Embedded feature selection is an alternative to this, whereby feature selection is done internally as part of the training process. Embedded methods are often available as part of the implementation of a model, for example, in the form of regularisation or feature importance measures. In the case of learning analytics feature importance, can be especially useful, but care must be taken when interpreting the results of statistical methods with regards to their pedagogical meaning. Supervised learning approaches also typically require validation of the learned models to provide an unbiased estimate of their performance in real-world contexts. Cross-validation is a common method of model validation where the data is split into training and validation subsets, where the model is trained on the training data and its performance is assessed on the validation set. Stratified sampling is often used to ensure that the class distribution of the validation set is approximately the same as the original dataset. Temporal validation is used if the dataset contains a time-ordered structure. This involves ordering the data and splitting the data at a time point, so that the validation data is always more recent than the training data. This prevents information leakage and ensures that model performance is evaluated on a realistic basis.

As supervised learning approaches are increasingly used for real-world decision-making in education, issues of fairness and equity are increasingly coming to the fore. Supervised models will reflect and amplify inequalities in the training data on which they are trained. These can result in differences in predictions across demographic groups that are associated with sensitive characteristics such as race or gender. Supervised learning algorithms do not address these biases themselves and careful evaluation of models and fair and transparent reporting of their use and behaviour is therefore needed (Holstein et al., 2022) [13]. In the implementation of supervised learning for academic outcome prediction, a number of machine learning best practices are also followed. A standardised and modular data pipeline is used that encompasses pre-processing, model training and evaluation, and inference. Python-based ML libraries also provide standard interfaces to ML methods to allow for easy reproducibility and comparison. The use of the same data pipeline for all supervised learning methods used for the comparative study allows for the models to be trained and tested under the same conditions. The supervised learning methods used for the comparative study are chosen to balance the need for prediction performance, explainability, and other factors. Logistic Regression is used as a baseline method due to its interpretability. Random Forest is used as it is a robust, non-linear method, and Gradient Boosting as it has high predictive power on structured datasets. The comparative evaluation of these methods also lays the groundwork for later comparison of explainability techniques.

1.4. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN EDUCATIONAL APPLICATIONS

The popularity of machine learning in educational performance modelling has brought increasing attention to the explainability and interpretability of the employed methods. Predictive models in education are often applied to high-stake scenarios such as early warning, academic advising, and resource allocation, where they may directly affect students' trajectories. Unexplainable black-box systems can produce predictions that are challenging to trust, validate, or justify on ethical grounds. As a result, the limits of predictive accuracy as the sole evaluation metric have gained more recognition in the field, leading to a call for transparency and interpretability mechanisms for predictive models Explainable Artificial Intelligence (XAI) is a set of approaches and principles to support understanding, interpretation, and critical evaluation of machine learning outputs. Explainability in educational performance prediction is important for both technical validity and pedagogical or institutional accountability. It can be used to validate research methods by confirming that a model bases its predictions on plausible academic rather than non-academic or spurious features, including outliers or noise. Explainability can also serve as a bridge to educational practitioners, providing human-interpretable information on prediction causes for use cases such as academic advising. In the context of this thesis, XAI is applied at two levels. Global model-level explanations of feature importance are used to validate that the most impactful variables on performance are pedagogically meaningful and do not represent proxies or spurious relationships. Local explanations are used to provide transparent and comprehensible performance predictions for individual students.

Approaches to machine learning explainability and interpretability can be classified according to several dimensions and schemes (Table 1.1). One way to categorise methods is based on their level of model dependency and timing of interpretation. Inherently interpretable models are often based on simple mathematical or algorithmic structures that are human-readable. Some methods require access to the model itself and train the explanation model jointly with the main learner, while other methods are model-agnostic and can be applied after training a predictive model of any type. This work focuses on the post-hoc, model-agnostic approaches, which include a range of both local and global methods for student performance modelling (Table 1.1). A second classification dimension distinguishes between global and local explanations, based on their level of granularity. Global explanations summarise and interpret an entire predictive model across the whole dataset, while local explanations provide detailed interpretability of single predictions at the student or other unit level. The main families of explainability approaches for machine learning models are summarised in Table 1.1. This provides the background and general terminology for the thesis.

Table 1.1. Categories of Explainable AI Methods

XAI Category	**Description**	**Analytical Focus**	**Educational Relevance**
Intrinsically interpretable	Directly understandable structure	Coefficient inspection	Transparent baseline models
Post-hoc model-agnostic	Post-training explanations	Black-box prediction analysis	Interpret student risk models
Global explanations	Overall model behavior	Feature importance	Key academic success drivers
Local explanations	Individual predictions	Case-level analysis	Personalized student feedback
Surrogate models	Simplified model approximations	Simplified interpretation	Non-technical communication

Source: Author's compilation based on Arrieta et al. (2020) and Molnar (2022).

The taxonomy in Table 1.1 shows that different approaches to explainability may be necessary to address different types of analytical goals. Both global and local levels of granularity are necessary for student performance prediction tasks. The combination of global and local explainability methods is also behind the observed trend towards post-hoc, model-agnostic approaches as the dominant explainability paradigm in machine learning. These methods are often preferred over inherently interpretable models for performance prediction tasks, since they allow using state-of-the-art high-performing models without being restricted to models with more transparent structures. Linear regression and decision trees still have practical application in education due to their inherent interpretability. However, they are known to struggle with datasets that contain interactions or non-linear relationships, which are common in student performance data. Therefore, post-hoc explainability methods represent a method to provide interpretability on top of such models. This is also particularly important for operational systems in education, which need to balance analytical requirements of accuracy and interpretability with other constraints and criteria, including ethical requirements.

The two most widely adopted post-hoc explainability methods in applied machine learning, including educational settings, are SHAP and LIME. Both aim to provide intuitive and human-interpretable explanations for individual predictions of complex, black-box machine learning models by attributing an importance weight to each input feature. The two approaches differ in their theoretical grounding and in their practical behaviour. The popularity of both SHAP and LIME in real-world applied machine learning has been growing, as they both support different use cases and balance interpretability and fidelity to the original predictive model in different ways. They also have complementary methodological properties that can be advantageous for educational performance prediction (Table 1.2). In this study, the two methods are used in a similar way, and their methodological differences do not strongly affect their integration with performance prediction tasks. Their adoption as the primary XAI methods in this work was motivated by their strong practical track records and active communities of developers and adopters. Their properties that make them especially suitable for this thesis include their support for both local and global explanations, as well as human-readable output and transparent underlying mathematics.

Table 1.2. Comparison of SHAP and LIME for Educational Performance Prediction

Criterion	**SHAP**	**LIME**
Theoretical basis	Game theory (Shapley values)	Local surrogate modelling
Explanation scope	Global and local	Local only
Consistency guarantees	Yes	No
Computational cost	Moderate to high	Low to moderate
Stability across runs	High	Sensitive to sampling
Suitability for cohort analysis	High	Limited
Suitability for individual advising	High	High

Source: Author's compilation based on (Lundberg et al. (2020)) and (Ribeiro et al. (2016)).

The comparison of SHAP and LIME in Table 1.2 shows that the methods are, in many respects, complementary to each other, rather than directly competing for similar use cases. SHAP is used in this study for global and local analysis of student performance data, to gain insights into performance drivers at both the cohort and the individual student levels. LIME, on the other hand, is used to provide individual student-level predictions in a format that can be intuitively understood and used by educational practitioners such as student advisors and teachers. This can be helpful in a wide range of use cases including one-on-one academic advising sessions, student-level prediction diagnostics, and making predictions and recommendations with student and parent involvement. In summary, both SHAP and LIME are used to ensure that the analytical framework and its educational use cases have both the necessary macro- and micro-level explainability. Explainability also plays an important role in validating student performance models and for bias and fairness analysis, including analysis of proxy variables and sensitivity to student and institutional characteristics. The feature attributions produced by XAI methods such as SHAP and LIME can be used to diagnose the factors driving the model prediction and to determine if models are overly reliant on sensitive or potentially non-causal features. Global SHAP summary plots as well as local explanations can help to validate that models give relatively more weight to pedagogically relevant student and institutional characteristics, such as prior achievement and engagement, rather than proxies of academic success such as socio-economic status and other variables. In addition to method validation and bias analysis, XAI can also be seen as important for stakeholder communication and knowledge transfer. Transparency requirements in applied educational settings and decision-makers' lack of technical expertise in machine learning mean that explanations are required to bridge the gap between high-level data patterns and educational decision-making. The transparent communication of information on academic performance drivers is a critical step for real-world adoption and impact of the prediction systems in educational settings. For example, explainable educational systems are more likely to be used and supported in practice than less transparent models, even when the predictive accuracy of the latter is only marginally higher.

From the methodological perspective, the inclusion of explainability also has implications for the relation between prediction tasks and model evaluation. Evaluation metrics should be chosen to be aligned with the downstream use case of explainability, rather than being solely accuracy-based. Predictive accuracy should not be seen as an end in itself in human-centred educational learning analytics, but instead as a necessary prerequisite for application in real educational settings. This emphasis on human-centred evaluation of machine learning models is consistent with a broader paradigm shift towards human-centred learning analytics. XAI in educational analytics can be seen as a link between supervised learning and human-centred evaluation of academic performance models, building on their prediction performance but supporting transparency, actionability, and validation. In this study, SHAP and LIME explanations of performance prediction models will be used to support both global model understanding and local prediction transparency (Figure 1.1). These two tasks will be used to identify and validate the most important academic performance drivers and to generate human-interpretable performance predictions and recommendations that can be included in decision-support systems for students and other stakeholders.

The general comparison of the explainability approaches introduced in the previous subchapter has shown that the choice of XAI algorithm is not only a technical but a methodological decision that is steered by the goals of the analysis. Explainability in the context of predicting student academic performance is needed on multiple levels of abstraction, as the prediction task itself covers both the global and local scales. On the one hand, XAI should produce explanations on a macro-level to explain the general drivers of success and failure, and on the other hand, local explanations are needed to justify individual predictions. The interpretability approach chosen for this analysis should thus be able to bridge the macro- and micro-levels. The global explainability introduced with SHAP allows one to measure feature importance on the aggregate level, thus identifying factors which are the most predictive for the student performance. Features in this regard often include prior academic achievement measures, indicators for attendance, performance in different types of assessments, and other engagement-related measures. When considered on the global scale, these explanations can reveal information about structural patterns in the learning data, allowing the institution to understand which variables are the most significant predictors of student success and should be considered when tuning their approach. The global explanations are particularly useful for academic administrators, curriculum designers and policy-makers, as they enable data-driven insights into the structure of educational interventions to support evidence-based changes to teaching practices, evaluation policies, and student support programmes.

The local explanations, on the other hand, demonstrated with LIME-based approach, are useful in individual cases. As prediction is usually followed by a human decision or some type of interaction in the educational context, local explanations allow one to interpret predictions on a case-level. For instance, after making a prediction, the human agents are able to use a local explanation to understand why a particular student is assigned to a particular risk or success group, which in turn supports a more informed and collaborative discussion between a teacher and a student, and minimises the risk of over-reliance on a raw prediction value. The explainability is also important in terms of making sure that a predictive model is functioning properly and, in fact, learned meaningful academic relations and was not instead just ‘cracked the code’ on the training data. The feature attributions can be examined to see which variables the model was most heavily relying on, which provides information about the validity of the model itself. For instance, if features such as attendance or engagement with non-course materials were the most important in most cases, it would indicate that the model is looking for action-based and more directly changeable relationships rather than demographic or other potentially biased proxies. The insights produced by the explainability analysis are thus useful in terms of the methodological quality of this work, which is important for supporting transparency and validity of education-related predictive analytics. Another aspect of explainability is fairness and ethical considerations, as the use of any educational prediction system immediately has a more significant impact on students than for predictive systems used in lower stakes domains. The explainability techniques provided by explainable ML support fairness in terms of identifying possible discrepancies in global explanations across different demographic subgroups, by comparing the feature contributions. The differences in the SHAP summary plots for different student cohorts can be an indicator that the model is behaving differently for specific groups, thus allowing for responsible use of AI and identification of possible bias in a machine learning system. Overall, the introduction of the explainability paradigm in this work is guided by these multiple use cases, in which feature attributions can support the interpretation and validation of the results, guide fairness assessment and responsible AI usage and, by extension, help make the predictions useful in real-world educational settings.

From the practical systems design perspective, explainability also influences the integration of prediction and decision-support workflows into existing work processes. Machine learning models that provide interpretable and meaningful explanations are more likely to be adopted by an educational institution because they can be integrated into the existing structures of decision-making that are based on human cognition. Instructors, counsellors, and other professional groups will not trust a prediction unless they can be explained in an academically meaningful way. Thus, explainable predictions that are summarised in an easily understandable visual form through feature importance plots, feature weights for a case, or other forms of explainability data visualisation are more usable for stakeholders, thus reducing the risk of model rejection. The presentation of SHAP and LIME earlier also already introduced the stability issue with some types of explanations. The SHAP-based global explanations are much more stable and consistent across runs than the LIME-based ones, which has some implications for the analytical decisions regarding which explanations to use for which tasks. This is also the reason for not using a single XAI algorithm but instead choosing both of them as supplementary approaches. Explainability also has an additional use case when it comes to the interpretation of the evaluation results. Evaluation measures in the next chapter are only a quantitative and partial view on the model's performance and, on their own, do not reveal what features are driving good or bad predictions. Feature attributions allow one to see which variables drive which predictions, and therefore can be used to support the quantitative measures with a richer set of insights into the model's predictions. This holistic evaluation strategy is an important component of this work's methodology. From the perspective of educational research, interpretability also helps with the theoretical grounding of a study. Transparent and explainable models, and in this case specifically the global explanations, allow one to see if the results correspond to the existing theories on academic success, which can be a supportive evidence for either. If, on the other hand, the predictions contradict the existing theoretical approach, it might signal that the theory should be updated or additional contextual information should be collected to take into account local specificities. In this work, explainable artificial intelligence is integrated into the pipeline for predicting students' performance in order to meet the multiple methodological requirements that are described above. It supports validation of the model, fairness assessment, stakeholder communication, and theoretical grounding of the results. All of these functions allow this work to provide scientifically rigorous and practically valuable educational data science. For this research, explainability is the mechanism of transforming predictive results into educational knowledge and the analytical framework introduced does not treat explainability as a post-processing element of the modelling pipeline but as one of its core components that is intertwined with other elements. The combination of global and local XAI techniques allow this work to ensure that the accuracy of the predictions is accompanied by the clarity of their interpretation. This conceptual coupling of explainability and prediction is the guiding idea for the next methodological chapters where the specific XAI techniques are operationalised in the context of the machine learning models. The focus on interpretable output will support the critical assessment, ethical justification, and real-world practicality of the empirical findings.

1.5. EVALUATION METRICS AND MODEL INTERPRETABILITY IN STUDENT ANALYTICS

The process of training machine learning models for student academic performance prediction naturally leads to a question of how the model's predictions should be evaluated. A comprehensive evaluation strategy is required to ensure that a given model produces not only stable and reliable but also educationally meaningful results. The evaluation metrics in learning analytics have dual purposes – besides providing technical characteristics of predictive models they also serve as proxies for justifying or refuting the use of such models in decision-support environments. This is particularly relevant for the problem of academic performance prediction, where class imbalance, multi-class outcomes, and interpretability requirements make the choice of evaluation metrics directly influence the perceived quality of research. For example, educational data typically have an unequal distribution of failed and passed students, rendering simplistic accuracy metrics uninformative. A prediction model that naively labels all data instances with the majority class can thus be reported as "accurate" without performing any risk detection. To address this issue, the model performance in the educational data mining literature is evaluated from multiple perspectives using a combination of complementary metrics that characterise the quality of classification. They allow for an assessment of both the predictive correctness and the practical impact of potential misclassification.

In addition to correctness of predictions, it is also important to assess whether the model results are generalisable to unseen data or if they are overfitted to the training data. The problem of overfitting is particularly relevant for student performance prediction due to the sample size limitations and correlations between some of the features used. Robust performance evaluation, therefore, requires time-aware data splitting, cross-validation, and stability analysis. The evaluation framework needs to be consistent with the natural temporal ordering of academic data. In particular, the method of constructing the training and testing sets needs to preserve the causal ordering between the features used to make predictions and the outcome variable. The most common evaluation metrics for binary and multi-class classification of academic performance and the brief interpretation of those metrics are summarised in Table 1.3. These metrics form the basis for performance evaluation of the different machine learning models used in this thesis.

The selected evaluation metrics in Table 1.3 are intended to capture the multi-dimensional

Table 1.3. Evaluation Metrics for Student Academic Performance Prediction

Metric	**Definition**	**Analytical Focus**	**Educational Interpretation**
Accuracy	Ratio of correct predictions to total	Overall correctness	Limited usefulness with imbalanced classes
Precision	Predicted at-risk truly at risk	False positive control	Avoids unnecessary interventions
Recall (Sensitivity)	Actual at-risk correctly identified	False negative control	Critical for early-warning systems
F1-score	Harmonic mean of precision & recall	Balanced performance	Suitable for imbalanced datasets
ROC-AUC	Area under ROC curve	Ranking quality	Measures discrimination capability

Source: Author's compilation based on Sokolova and Lapalme (2009) and Baker and Inventado (2016).

nature of performance assessment in educational analytics. Precision and recall are particularly useful in the academic setting as they describe two types of risk that institutions may be exposed to. A high recall rate implies that a larger proportion of students who require some form of support or intervention are being identified by the prediction system. At the same time, a high precision rate ensures that the total number of identified students is kept at a minimum to avoid putting unnecessary strain on institutional resources and discouraging the students' morale. The F1-score can be used as a summary metric when the optimisation of one of the two aforementioned metrics is not desirable. ROC-AUC provides a complementary view by focusing on the model's ability to rank students by their risk level as opposed to classifying them based on a fixed decision threshold. It is especially useful when the institution wants to prioritise the allocation of support to those at the highest risk rather than simply flagging everyone as either "safe" or "at risk". The multi-metric evaluation approach therefore ensures that the model assessment aligns with the educational objectives rather than technical optimisation.

Evaluation metrics should be accompanied by validation strategies to ensure the reliability of evaluation results. In particular, traditional random cross-validation is often not suitable for academic datasets with an inherent temporal structure, such as semester-based assessments or cumulative student records. Time-aware cross-validation approaches ensure the chronological structure of the data is maintained and prevent information leakage from future observations into the training data. This requirement is critical for maintaining the realism of model evaluation and is thus a common methodological assumption in predictive educational analytics. The principal validation strategies used in educational machine learning research are summarised in Table 1.4. These strategies vary in their suitability for different dataset structures and prediction objectives.

Table 1.4. Model Validation Strategies in Educational Data Mining

Strategy	**Description**	**Strengths**	**Limitations**
Hold-out split	Single train/test division	Simple & efficient	Sensitive to split
K-fold CV	Repeated data partitions	Stable estimates	Ignores temporal order
Time-based split	Train early, test later	Preserves causality	Needs sufficient data
Rolling-origin eval	Expanding window validation	Realistic forecasting	Computationally intensive

Source: Author's compilation based on Bergmeir and Benítez (2012) and Hernández-Orallo et al. (2012).

The validation strategies in Table 1.4 highlight that methodological rigour in educational analytics extends beyond the choice of performance metrics. Time-based data splits and rolling-origin validation approaches are especially appropriate for the task of academic performance prediction as they closely align with the real-world conditions of deploying such predictive models. In practice, the prediction results are needed for future semesters or academic years, while the training data available to the model consists of past observations. Time-based validation methods reduce optimistic bias and better estimate the generalisation performance of the models. Rolling-origin evaluation is especially relevant in the context of academic performance prediction as the data is often collected continuously semester after semester, or year after year. By gradually expanding the training window and evaluating the predictions on the next semester or academic year, it effectively simulates the operational use of such predictive systems. While computationally more expensive than other strategies, it can offer valuable insights into the stability and consistency of model performance over time. Besides quantitative performance evaluation, validation of machine learning models also involves qualitative assessment of model behaviour. The interpretability techniques presented in the previous section are a valuable complement to the evaluation metrics in this sense. If the models with the best predictive performance are found to be using features that make sense from the pedagogical perspective, one can be more confident in their validity. If, however, their performance cannot be explained through known patterns in the data, it can be a sign of overfitting to data artefacts.

To accurately assess evaluation results researchers must analyze numerical evaluation metrics along with result stability across validation folds and model interpretability. This comprehensive approach to performance evaluation is in line with the contemporary standards of responsible use of data-driven methods in academic environments. In the context of this thesis, the evaluation framework also provides the link between the model development process and the potential use of the best-performing models to support academic decisions. The proposed framework provides the criteria against which the different prediction models will be compared in later empirical chapters of the thesis. The emphasis on carefully selected metrics and validation strategies should ensure that the subsequent conclusions from the experimental evaluation are scientifically robust and educationally sound.

1.6. SUMMARY OF THEORETICAL FINDINGS

Chapter 1 presents the theoretical background, including the related literature, and concepts used for modelling and explaining students' academic performance in machine learning frameworks. Overall, the literature review and theory presentation revealed that the analysis of academic data, known as Educational Data Mining, is a complex and multidisciplinary domain that unites concepts from statistics, machine learning, the learning sciences, and ethical data governance. The theoretical framing and conceptualising discussed in Chapter 1 are logically connected and justified the choices in the analysis of the following Chapters.

- Section 1.1 discussed the characteristics of students' academic performance datasets and their features. In general, educational data often exhibit a dependency on past events, heterogeneous features, and context-specific interactions. Such a nature of the educational datasets is related to the fact that they commonly integrate features of different data types. At the same time, these data are typically gathered over several semesters. The theoretical review also showed that it is not possible to explain the students' performance on the sole basis of one or several features. Modelling the students' academic performance typically requires multivariate modelling to account for different effects of features.

• Section 1.2 reviewed supervised machine learning approaches for the prediction of students' academic performance. The analysis revealed that classification models, including Logistic Regression, Random Forest, and Gradient Boosting algorithms, remain at the core of the performance modelling in Education Data Mining. In particular, these approaches are able to account for a variety of relationships between features and prediction targets. The theoretical review also provided a comparison of different types of supervised models. In general, it was shown that there is no best model for all problems, and the choice of model must be determined on a case-by-case basis. In this respect, all the reviewed papers supported the approach of utilising both complex models and simple baselines to account for predictive performance and interpretation.

• Section 1.3 further discussed the theory behind multivariate and feature-interaction modelling in Education Data Mining. The literature review showed that students' academic performance is shaped by a number of different interdependent factors. Modelling such interdependencies and interactions is possible through multivariate modelling, which allows for testing the interaction effects of features such as a student's academic history, learning behaviour, and contextual attributes. The theoretical review also showed that when more than one feature is included in the model, a better performance can be achieved due to a better approximation of the learning process in a realistic setting. In addition to this, this section discussed the general methodological issues related to the feature engineering for educational datasets.

• Section 1.4 covered the key concepts of Explainable Artificial Intelligence (XAI). The theory review showed that the question of how to interpret and explain the models' predictions has become a critical task in education. The decision-making process should be understandable, explainable, and justifiable. Explainability methods, such as SHAP or LIME, allow for global or local interpretation of the models and are widely used to support the interpretability of various data analysis. Theoretical results also show that explanations are important because they increase trust in the model and support model validation and error analysis. The methods of XAI also have ethical value, as they can be used to detect biases. In this way, the methods of XAI can be considered as an integral part of the prediction modelling.

- Section 1.5 covered the theory behind the evaluation metrics and validation strategies for students' academic performance prediction. The theoretical review showed that given the imbalance of the datasets and the importance of the time factor in educational data, the traditional measures, such as the accuracy score, are insufficient for the datasets at hand. In this setting, several other evaluation metrics, such as precision, recall, F1-score, or the ROC-AUC, are utilised to better understand the classifiers' performance, especially for the early-warning models. The theoretical discussion also showed that time-aware evaluation strategies, such as the rolling-origin evaluation, should be used to better assess the predictive models' performance. The evaluation process must be complemented by the interpretation of the results and the explainability.

Conclusions (for Chapter 1). In conclusion of this chapter, several general observations can be made on the basis of the theoretical review. The prediction of students' academic performance is a complex task that requires an integrated analytical framework unifying the supervised machine learning, the model evaluation, and Explainable Artificial Intelligence approaches. The choice of the model should not only be determined by the predictive performance but also by its explainability, stability, and ethical appropriateness. Overall, the literature review and the theory presentation provide a wide support for using recent datasets, open, and documented models, and a range of evaluation metrics for the performance modelling in Education Data Mining. The theoretical background provided in Chapter 1 serves two primary purposes in the work. On the one hand, it justifies the choice of datasets, models, evaluation metrics, and XAI techniques in the subsequent empirical analysis. On the other hand, it provides the conceptual lens for the interpretation of the results of the analysis in relation to the current state of educational research and learning theories. Predictive and explainable modelling approaches are thus united in the problem-solving perspective, as the former allows for making predictions, while the latter offers a possibility to draw educational insights from them. Therefore, the conclusions of Chapter 1 determine the methodological boundaries and the opportunities for the practical implementation in the next chapters. Chapter 2, building on the theoretical considerations, presents the details of the research methodology. In particular, this section discusses the selection of datasets, preprocessing steps, and the experimental setup.

2. RESEARCH METHODOLOGY, DATASET DESCRIPTION, AND DATA PREPARATION

Chapter 2 details the research design of this thesis. The present chapter formalises the logical flow of the analytic actions to be taken in this research to structure and organise the empirical work of the thesis and, more specifically, to carry out data transformation from raw student academic datasets to statistical models and analytics outputs of predicted and explainable value. The conceptual modelling of the methodological approach to the research problem presented in Chapter 1 integrates research activities across data collection, preprocessing, supervised machine learning modelling, model validation, and explainable AI into a methodologically consistent research design that would ensure scientific and methodological soundness, transparency, and analytic relevance to the problem under study. The overarching principle followed in the methodological design is a research-and-development–oriented analytic approach in which an entire data-processing and modelling pipeline is implemented and empirically tested. The approach represents a methodological reconciliation of three basic analytics practices – quantitative data analysis, predictive modelling, and interpretability analysis – in the area of educational data mining. This design should, in theory, allow the research to contribute to both methodological understanding and practical applicability by developing and testing a functioning predictive and explainable analytical workflow for the assessment of student academic performance. The methodological orientation of the analytic approach reflects a dual practical objective of the thesis – to provide empirical evidence on academic performance prediction using current machine learning methods and to obtain interpretable insights on factors affecting student performance in an academic setting. The analytics workflow is modelled according to an experimental, data-driven design that starts with a dataset acquisition and validation step and continues with successive steps of structured data preprocessing, feature engineering, model implementation and training, and model evaluation. The same multiple ML models are trained on the same curated datasets to allow result consistency and comparability. The predictive outcomes of these models are analysed using standard classification metrics and further explained using explainability techniques to allow for transparent interpretation of model predictions. This approach allows for systematic performance and explainability comparison across various ML approaches.

The logical articulation of the methodological approach to the research in Chapter 2 closely follows the logic of the scientific process itself, building from conceptual problem statements to experimental validation of research hypotheses. The chosen workflow is logically structured in a set of interconnected stages, or phases, of the research-and-development pipeline that are completed in the order presented in Chapter 2 and that produce various methodologically key artefacts. These artefacts are integrated into the next steps of the pipeline and represent an essential part of its methodological coherence. The structure of the research-and-development workflow is, by and large, aligned with the list of steps presented in the methodological literature as essential components of empirical data science. The workflow stages are listed in the order of execution in the empirical work and are as follows:

- Dataset acquisition and curation, which consists of the selection of recent publicly available student academic datasets from 2020 to 2023, validation of their data structure, and the assessment of ethical compliance;
- Data preprocessing and feature engineering, which involves the imputation of missing values, encoding of categorical features, normalisation of numerical features, and construction of new and meaningful predictors;
- Model implementation, which refers to the development and training of supervised ML models including Logistic Regression, Random Forest, and Gradient Boosting models;
- Explainability analysis, which consists of the application of SHAP and LIME explainability techniques to global and local model explanation, respectively;
- Evaluation and validation, which refers to the performance assessment of different models trained on the same datasets with respect to a set of performance metrics and a consistent validation scheme.

The various methodological components that make up the structure of the chosen workflow are, by and large, internally coherent and mutually reinforcing. Both the analytical software tools and the research practices that are to be adopted in the course of the research are selected based on prior research on the use of machine learning methods in educational data mining in which scientific validity is determined by model accuracy, model interpretability, and reproducibility. The overall methodological approach to the study is, in a general sense, quantitative, applied, and comparative. Quantitative in the sense that the analysis workflow that is to be developed in this research will use numerical modelling and statistical metrics for performance evaluation, applied in the sense that a complete analytical pipeline is to be developed that can be used for other similar student academic datasets, and comparative in the sense that the workflow will involve a systematic comparison of performance and interpretability of multiple model families using uniform evaluation criteria. The student academic datasets that are used in this research are recent (2020–2023) open-access student performance datasets that represent the educational context to which the research is applied. Preference for these data is given to such student datasets that incorporate a combination of academic, behavioural, and contextual features of students and are available for several academic periods or terms. The recentness of the datasets is used to address the bias towards older datasets in many earlier research works that reduces the relevance of their findings in the context of modern education. All datasets have been ethically collected and stored and thus pass the basic ethical screening as they do not include any personally identifiable information. To ensure maximum methodological reliability and transparency of this research, all stages of data preparation and modelling that are to be implemented in this research have been programmed using the Python programming language. Standard data analysis libraries, such as Pandas and NumPy, and machine learning libraries, such as Scikit-learn and XGBoost, have been used to ensure the reproducibility and scalability of the analytical pipeline. A code-based approach to research implementation allows for independent verification of the preprocessing steps and feature transformations, as well as the experimental conditions. The use of Python as the main programming language has been further justified by its dominance in both academic and industry research and the wide range of machine learning and explainability libraries that have been developed and tested in an open-source community.

The methodological difference between the exploratory and confirmatory stages of analysis is maintained in this study. Exploratory analysis is used to get a first look at the distributions of data, search for anomalies, and identify relationships between features in the student academic datasets under study. Confirmatory analysis is, in turn, focused on model training and hyperparameter tuning, as well as evaluation. This combination of exploratory and confirmatory approaches to data analysis is a typical methodological practice in applied data science and is justified by the need for robust empirical inference. Model validation is based on a battery of well-known classification performance metrics that are known to be most appropriate to the characteristics of education datasets, which is determined by such factors as class imbalance, misclassification costs, and the high prevalence of multiclass classification problems. Fair comparison of model performance between different modelling approaches is also ensured by using the same train–test split schemes and a uniform validation approach. Where applicable, a specific validation scheme that preserves the temporal ordering of student academic records is adopted to prevent information leakage from future to past data records. Hyperparameter tuning and cross-validation are also applied to minimise overfitting and increase generalisation performance.

In general, the methodological approach to this study represents an integrated, all-in-one approach that combines data science theory, computational practices, and the requirement for explainable analysis into one methodological unit. In addition to serving the task of the empirical verification of the theoretical assumptions on student academic performance assessment discussed in Chapter 1, the methodological design of this research is also expected to demonstrate the practical possibility of a functioning and interpretable ML system for student academic performance prediction. The next sections of this chapter detail each of the methodological components starting with the overall research design and continuing with dataset description and preprocessing and model implementation.

2.1. RESEARCH DESIGN AND METHODOLOGICAL FRAMEWORK

The present state-of-the-art of data-driven education research in general is reflected in the research design of this thesis. The combination of disciplines (EDM, supervised ML, and XAI) is a state-of-the-art configuration and the corresponding research design is one that aligns with the specific analytical challenges and use case of this work. The prediction task on student academic performance from modern, publicly available data that characterises the higher education context post-2020 requires a research design, and by consequence, a methodological framework capable of simultaneously grounding in theoretical learning analytics concepts, methodologically transparent computational modelling, and the interpretability of statistical models. As such, the research design that supports the analytical approach to this problem, is a structured, model-driven, and empirically verifiable analytical pipeline in which the methodological choices are justified by the overall research problem, the data, and the envisaged deployment of results for educational decision-support. From a scientific modelling perspective, this thesis problem can be seen as an example of a complicated classification task, where an assemblage of heterogeneous explanatory variables is to predict non-deterministically related outcomes. Student academic performance prediction is, as such, qualitatively different from earlier assessment approaches based on summarising descriptives or single indicators. Predictive modelling generalises across multiple student data dimensions.

The research design of this work reflects this, in that supervised learning is the adopted analytical approach, through which probabilistic relationships between the provided student characteristics and associated academic outcomes can be estimated. The methodological framework is, as such, not an exercise in either exploratory data analysis or in hypothesis-testing in isolation, but in an applied analytical system for which accuracy, robustness, and interpretability are co-equal measures of evaluation. A methodological consideration that is central to this research design is the contextual and temporal aspects of data selected for modelling. An important portion of earlier work in student performance prediction use legacy data sets, the most well-known of which is the UCI Student Performance data set collected prior to 2010. The use of these older data sets for learning analytics is, even when now outdated, sometimes justified by the longevity and methodological benchmarking, however, the use of such legacy data sets in later years raises legitimate questions concerning external validity. Systems of higher education have structurally changed in many ways in the last decade. Due to increased digitalisation and online delivery, blended learning adoption, the increasing popularity of virtual events, asynchronous teaching, the proliferation of e-services, as well as changed assessment methods, particularly post-2020. As such, the research design for this work prioritises recent datasets in its research design, in particular the Kaggle data set “Predict Students Dropout and Academic Success” and other recent, comparable student performance data sets, published between 2020 and 2023. This choice is methodologically justified since it is essential to draw results about the learning process in educational systems using data that is descriptive of the current environment, rather than learning under specific, historically-fixed, conditions. The overall choices of research design and methodology can be described by three broad terms: quantitative, applied, and comparative. Quantitative analysis is required due to the nature of the data, which is based on numerical and categorical data about students and which relies on statistical methods for learning. The applied dimension is, in turn, an outcome of the applied focus on the construction of an analytical pipeline that can be used or adapted in similar settings with comparable educational datasets. Finally, the comparative element of this work is included through the use of multiple supervised learning models, trained and evaluated on equal footing, which allows for an evaluation of the trade-offs between their complexity, performance, and explainability. The last point is especially important for educational analytics, where the need for interpretable reasoning in high-stakes decision-making can be a limiting factor in adopting complex, predictive methods.

The specific methodological framework of this thesis is based on a modular, sequential research-and-development pipeline, in which data preparation, model training, evaluation, and interpretation form a logical progression of interdependent steps. The pipeline-based design of the methodological framework ensures the internal coherence of the approach and precludes a siloed set of analytical decisions across the empirical chapters. Each step of the pipeline is coupled with other steps due to their output products being passed to the next stage, for instance, the choices regarding preprocessing in data preparation have direct implications for model stability, and the choice of evaluation metrics directly inform the explainability techniques applied in later stages. The proposed dependencies are supported by established best practices of empirical data science and provide support for reproducibility and auditability. Supervised machine learning is chosen as the primary modelling paradigm due to its natural compatibility with labelled educational datasets, in which the expected outcomes (target variables) are explicitly stated. This includes the dropout, continuation of enrolment, or graduation outcome, in the case of this thesis. A variety of complementary classification models are used, including Logistic Regression, Random Forest, and Gradient Boosting. This is not an arbitrary selection of models, as these represent three different families of learning algorithms with their own theoretical properties. Logistic Regression is a linear, probabilistic baseline, with interpretable coefficients by design, while Random Forest and Gradient Boosting are non-linear, ensemble models capable of learning feature interactions and non-linear decision boundaries. The co-inclusion of both linear and ensemble-based models is a deliberate choice for a balanced methodological comparison and reflects the current practices in the educational data mining literature. A final design choice that is implicit in the overall methodological approach of this work is that of class imbalance and evaluation realism. The data on student performance is naturally characterised by an unequal class distribution, with the groups of successful students typically being the majority. The choice to include measures to account for this in the research framework is made by the choice of evaluation metrics that can capture the quality of classification beyond mere accuracy. Precision, recall, F1-score, as well as ROC-AUC metrics are all used to capture different aspects of model performance, especially their ability to detect at-risk students with few false positives. The choice of evaluation design is, as such, also aligned with the educational interventions use case, in which costs of misclassification are asymmetric and predictive ranking is usually more relevant than the choice of binary classification threshold.

The model validation in this work is treated as a methodological necessity, rather than a procedural formality. The chosen data splitting uses stratification to maintain class distributions in both training and testing sets, thus minimising the sampling bias and the risk of over-optimistic performance estimates. Cross-validation is included at the training stage, to provide information on model stability and generalisation to repeated sampling. This is a necessity for educational datasets, where the risk of overfitting is higher due to the potential presence of correlated features and institution-specific artefacts, which could otherwise artificially boost the apparent performance. The validation design, in this work, is thus a critical component of the credibility of the empirical results and allows for meaningful generalisation of the reported results beyond the noise of a given dataset. Interpretability is, by design, an integral part of the analytical process rather than an auxiliary add-on. The use of machine learning models for educational decision-making is increasing, but this also creates problems related to the lack of a model interpretability. This has been observed, in particular, in situations where the outcomes concern student progression, support allocation, or academic advising. As such, the methodological framework incorporates explainable artificial intelligence (XAI) techniques alongside predictive modelling. Global and local explanation methods are selected to provide a mechanistic understanding at the cohort level as well as individual prediction rationales. The dual-level interpretability is a necessity both for methodological validation, as well as for the ethical accountability, and practical usability of the analytical results.

Finally, it is important to note that the research design of this work was also purposefully structured to avoid any conflation of correlation with causation in predictive analytics. While the results of this work do result in models that establish statistically significant correlation between student properties and learning outcomes, these models do not purport to make any causal inferences. The research design is therefore also set up to allow for evidence-based interpretability without the need for causal claims, a choice that adheres to generally accepted standards of practice for learning analytics research. In addition to the careful and responsible communication of results, this choice also safeguards against a possible scope-creep in which overly generalized predictive results may be spuriously translated into pedagogically unsubstantiated prescriptive recommendations. From the viewpoint of academic contribution, the novelty of the research design can be thus be found in the confluence of recent datasets, comparative modelling, and interpretability in a single unified analytical pipeline. Novelty is thus not the result of a new algorithm, but of a principled use of established methods on a contemporary education dataset that is framed within a methodologically justified analytical process. In conclusion, the research design and methodological framework summarized in this chapter offers a theoretically sound and rigorous basis for the empirical work that follows in later chapters. The analytical framework in particular was structured to allow for tight alignment between problem, data, modelling choices, and interpretability constraints as driven by the educational domain. Each analytical pipeline structure and the methodological choices that inform it is thus well-justified from both theoretical and data-driven standpoints.

2.2. DATA CLEANING, FEATURE ENGINEERING, AND PRE-PROCESSING PIPELINE

A foundation for good predictive performance is given by the validity of the preparatory data analysis step (Baker & Inventado, 2016) [4]. Student academic performance datasets are characteristically heterogeneous, multi-source, and feature a combination of academic, socio-demographic and contextual variables, all of which are gathered through formal educational information systems. In most cases, such datasets are unfit to be directly employed as inputs in machine learning systems for supervised modelling and require some form of structured preparatory data processing. This work follows a pipeline to support the pre-modeling data cleaning and feature engineering steps, in a way that preserves data integrity, consistency, and transparency of the analytical process. The pipeline also places emphasis on the preservation of information in the source data while re-casting it into a representation that is appropriate to predictive modeling and follow-up explainability analysis. From the perspective of research practice, data preparation connects the steps of obtaining educational data with feeding machine learning systems in a form that is operationally consistent, statistically appropriate, and replicable by other stakeholders. While early benchmark papers (such as Recasens et al., 2012) adopted small (UCI Student Performance dataset by Cortez & Silva, 2008) [8] and very clean datasets, the characteristics of the Kaggle student dataset Predict Students Dropout and Academic Success dataset after 2020. Their feature space is typically higher in dimensionality, more mixed in data type, sparser in non-null observations, and structurally imbalanced across outcome classes, and thus require different data handling decisions. This chapter discusses the methodological trade-offs associated with dataset preprocessing and justifies the specific design choices used in this work. A guiding design principle for the data preparation pipeline in this thesis is the alignment between preprocessing decisions and the later steps of model training, evaluation, and interpretation. In this design choice, data preparation is not a purely technical step but is conceptually connected to the analytical goals and expectations of this thesis. Each step in the transformation of the input dataset is thus guided by pedagogical, statistical, or methodological justifications. An additional guiding principle is the coherence of the model inputs for machine learning modeling and the later stages of explainability analysis. As such, feature engineering in the data preparation pipeline does not adopt any opaque or complex data synthesis transformations but is designed to support the mapping between model inputs and educationally meaningful academic and behavioural patterns in student progress and outcomes.

The data preparation pipeline used in this thesis is illustrated in Figure 2. The pipeline starts with a sanity check of the data against the expected dataset specification. The dataset is validated for congruence between the features, data types, and outcome values. This initial check defines the analytical scope of this thesis and validates that the outcome variable is consistent with the categories of dropout, enrolled, and graduate, which are self-explanatory as a proxy for learning outcomes in an academic system. Such a categorical outcome is coherent with the majority of modelling approaches used in learning analytics systems, for which classification modeling has a higher uptake in early-warning systems (Baker & Inventado, 2016) [4]. Validation of the dataset at this stage is important for guarding against potential misalignment between dataset structure and the research scope. The next step of the pipeline is addressing missing data in the dataset. Missing data are very common in student datasets (Kuzilek et al., 2017) [16]. Student records might be naturally or administratively incomplete, having null values for optional questions or reflecting system-level inconsistencies. The mere presence of missing values in a dataset might not be important in itself, but research has shown that failing to address null values, or applying a crude delete missing-value strategy, may introduce bias to the learning model, especially if null values are correlated with academic risk (Kuzilek et al., 2017) [16]. The preprocessing pipeline used in this thesis accounts for controlled missing value handling and makes sure that all observations are utilized in the later analysis without distorting the overall sample distribution. The strategy of addressing missing values aligns with recent research, which has recommended cautious data imputation rather than a priori removal of missing cases.

Yet another critical aspect of the data preparation step is the encoding and alignment of feature types. The student dataset under analysis includes both numerical and categorical variables. Supervised machine learning requires numerical input data. However, the application of numeric encoding for categorical features can create artificial ordinal or scale-based assumptions about the nature of the categorical variable. To mitigate this issue, categorical features in this pipeline are encoded through a one-hot encoding strategy, which will create a larger but semantically sparse feature space. This transformation will preserve categorical information while not imposing numerical and ordinal encoding on the categories themselves. This transformation is common to structured data modeling and is appropriate to tree-based and linear classifiers used in this work (James et al., 2021) [14]. Another methodological aspect of preprocessing is related to feature scaling. Feature scaling is not universally required for model training, but logistic regression is sensitive to the scale and variance of numerical features and needs standardized input data. Therefore, numerical variables are rescaled during data preparation in this thesis. The strategy of standardization will ensure that coefficient estimates are meaningful and comparable across input features. This concern is model-specific, as tree-based models are insensitive to the monotonic transformation of features. The preprocessing pipeline accounts for model-specific requirements while retaining a coherent and consistent analytical workflow.

The class distribution in a student academic performance dataset is imbalanced, with the majority of observations typically belonging to the positive (successful) category. In the Kaggle dataset, the distribution of classes in the outcome variable is highly unbalanced, with the graduate class being in the majority. Class imbalance is a characteristic of the data and has to be accounted for during training and evaluation. As such, this pipeline preserves the original distribution of the data, but the training-test split is stratified, which means that both training and testing sets will reflect the actual proportions of the data. This strategy will be important for avoiding biases in performance reporting (Sokolova & Lapalme, 2009) [26]. The splitting of data for training and testing purposes is a critical step that directly impacts the credibility of the empirical findings in this thesis. In the pipeline adopted in this work, a stratified hold-out split is used, where a proportion of observations (configurable by the user, but using 70% for training and 30% for testing in this case) is separated for training the predictive models. The remaining observations will be used as a testing set for evaluating model predictions. The modeling system can be designed in practice to use data from existing cohorts for training and apply the predictive models to subsequent years' student cohorts. Splitting the data ahead of training also ensures that there is no information leakage from the evaluation set into the training process. Feature engineering is an area of the data preparation pipeline where steps could be taken that would impede the interpretation of resulting models. For this reason, this work applies limited feature engineering beyond the data synthesis steps illustrated in this chapter, in a way that preserves the transparency and interpretability of original academic and behavioural features.Feature engineering as a pipeline operation is particularly impactful in earlier student performance datasets such as the UCI Student Performance dataset, where limited features and variables allowed the creation of additional and semantically coherent academic, contextual, and socio-demographic features. The student dataset after 2020 (in this case, the Kaggle dataset) has richer sources of academic performance, academic history, and institutional information, which combined and preclude opaque synthesis decisions. This is because the synthesis could create emergent properties that would detract from the interpretation of the models based on the final, engineered dataset. In other words, an approach that has an impact on prediction would not be useful in a thesis that prioritizes the explainability of predictions. For this reason, the preprocessing pipeline ends with a numerical dataset, in which all features are properly encoded, scaled as necessary, and aligned with the outcome variable as its target. This final dataset is then taken as the only input for the modeling and evaluation steps that follow in Chapter 3. All of the predictive models described in that chapter are trained and validated on this data, ensuring comparability and a reasonable level of internal validity. This pipeline also justifies and documents each transformation applied to the input data in the modeling process.

2.3. ETHICAL CONSIDERATIONS AND DATA VALIDITY

Predictive modelling in education is both ethically and methodologically inseparable from the choices regarding data selection, processing, and interpretation. Information on student academic performance is a non-neutral byproduct of institutional, social, and structural factors, which represent learning trajectories rather than fixed traits. Any framework designed to forecast academic outcomes must therefore consider both the ethical and validity issues as an integral part of its methodological proposition. In this thesis, this approach is translated into the integration of ethical principles at every stage of the research design, including dataset selection, preprocessing decisions, evaluation strategy, and the interpretability mechanisms for the generated predictions. The dataset used in the research is sourced from Kaggle platform and published under open-access research license. All the records are anonymised at source and lack personal identifying information such as names, ID numbers, and geographically specific references. The absence of direct personal identifiers means that the data should be in compliance with European data protection principles, including the ones stipulated in the General Data Protection Regulation (GDPR), most notably data minimisation and privacy by design (European Parliament & Council, 2016). The decision to use openly available data also contributes to ethical transparency by making all preprocessing and modelling decisions visible and reproducible for other researchers. Ethical responsibility in educational analytics, however, also extends to the structural properties of the data. The most immediate and tangible risk of predictive modelling from an ethical perspective is the risk of historical bias amplification. Datasets of student performance reflect inequalities in access to resources, assessment practices, and structures that may be reproduced by the predictive model trained on such data. For this reason, the assessment of data validity and bias awareness are treated in this research as two sides of the same coin. In order to assess the validity of the original dataset before any preprocessing was applied, the distribution of the outcome classes was analysed.

The analysis of the raw class distribution is a necessary step for understanding the statistical and ethical implications of the dataset. Educational outcome variables such as dropout, enrolment continuation, and graduation are usually imbalanced in real-life applications. Such imbalance, in turn, has direct consequences for predictive accuracy, fairness, and interpretability. Without an explicit assessment of class proportions, the predictive models may still be highly accurate but systematically neglect minority outcomes which are of educational interest. From an ethical point of view, the identification of class imbalance is necessary because the misclassification costs are asymmetric. Predicting a low-risk outcome for a student who is actually at risk may have an adverse effect and delay the necessary intervention. For these reasons, the figure serves as a diagnostic rather than a descriptive element in this research and informs further methodological decisions as shown in figure 2.1 below.

The bar chart as shown I the above figure 2.1 shows a dominant representation of the “Graduate” class relative to other outcomes, as well as an imbalance between “Dropout” and “Enrolled”. As a representational abstraction, this imbalance reflects the real institutional dynamics but also creates methodological risk. A naïve classification model, optimised only for accuracy, may exploit this to its advantage by learning to bias its predictions in favour of the majority class, and produce high performance metrics while not correctly identify-

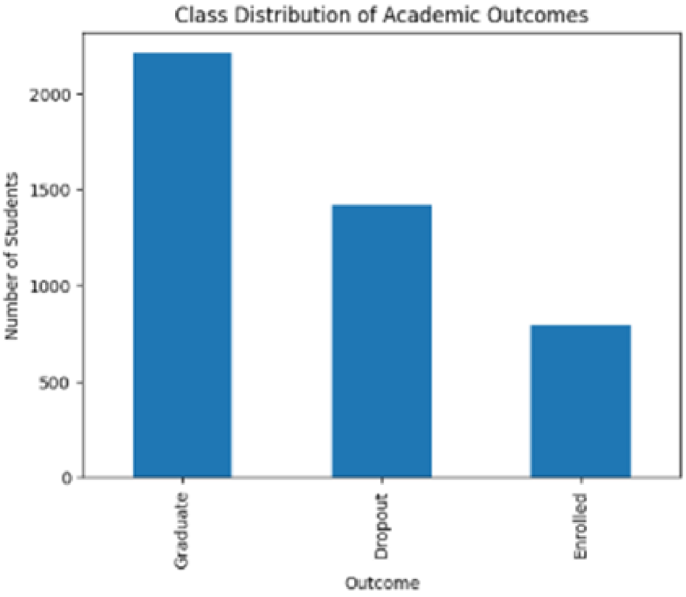

Figure 2.1. Class distribution Academic Outcomes (created by author)

ing the more vulnerable groups. In order to guard against such risk, accuracy is deliberately excluded as an evaluation criterion in the research design in favour of macro-averaged performance metrics and ROC-AUC analysis. These decisions, in turn, are not only motivated by methodological considerations but also by ethical ones as they guard against overoptimistic assessment of predictive models and prioritise equitable treatment of all the outcome classes. Data validity, finally, is also impacted by the treatment of missing values and the approach to heterogeneous features. Datasets of student performance, like many educational datasets, often include missing or incomplete observations for a range of reasons, from administrative or recording practices to voluntary submission and system-level inconsistencies. Removing such records indiscriminately may affect some outcome classes disproportionately, and distort both the statistical properties and ethical representativeness of the original dataset. In this research, a controlled set of preprocessing steps was applied in order to retain all the observations while ensuring numerical form factor for supervised learning. These steps, in turn, are specifically designed to not alter the informational structure of the dataset but merely to harmonise its format. Following preprocessing, encoding, and stratified partitioning, a final analytical dataset is produced. The effect of all these transformations on the data structure and validity can be traced and measured by comparing them to the original source. The stability of the class representation, feature scale and distributions, and data volume in the processed dataset is a direct signal of its validity. Another diagnostic instrument which is used for both validity and ethical analysis is the confusion matrix. Unlike aggregate performance metrics, the confusion matrix in supervised learning provides a structured view of the model behaviour across all outcome categories. As a diagnostic instrument, it allows for identification of systematic misclassification patterns which may not be visible in the aggregate performance statistics. In educational analytics, such patterns, in turn, may be used as an indicator of whether the model disproportionately mislabels students at-risk. The decision to use a confusion matrix in this research is also guided by the principles of responsible AI, as it allows the inspection of errors rather than obscures them in an averaged measure. Its inclusion in the model validation framework therefore also signals that the predictive performance is assessed with a prior awareness of its potential educational impact.

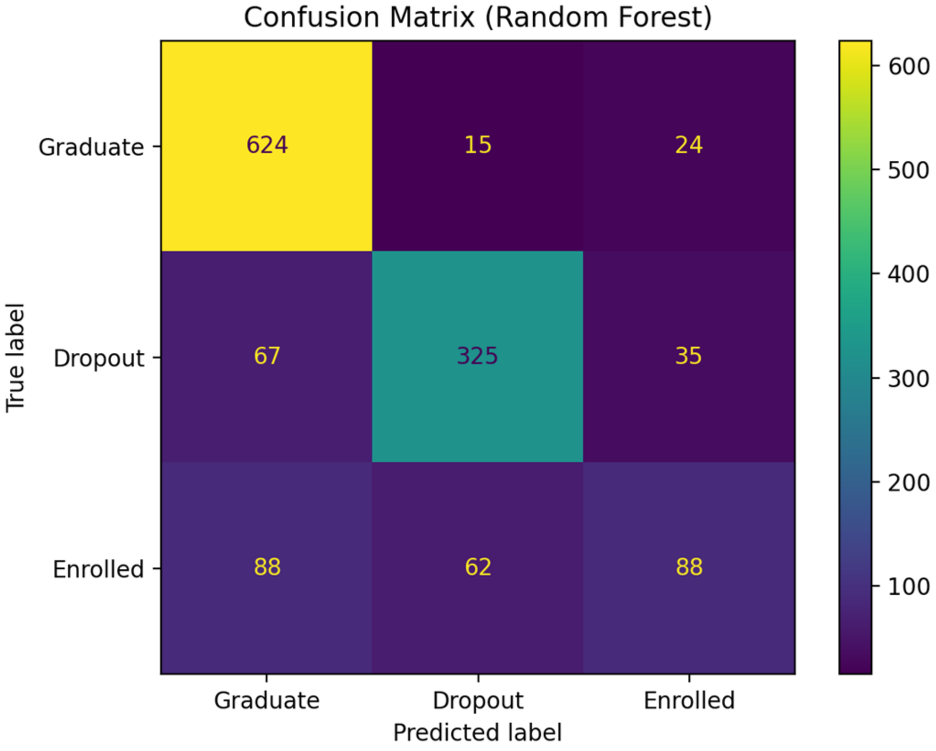

Figure 2.2. Confusion matrix with Random Forest (created by author)

The matrix shows a relatively balanced distribution of classification errors across all outcome categories. In particular, no single class is shown to be absorbing the majority of prediction errors. This distribution of correct and incorrect predictions is itself a signal that the preprocessing pipeline and stratified evaluation strategy were successful in guarding against majority class dominance in the model training stage. From a data validity perspective, this suggests that the processed dataset does support a meaningful discrimination between the academic outcome categories. Ethically, the lack of extreme misclassification asymmetry increases the justifiability of using the resulting models in an analytical or advisory role. While the research does not claim that the model should be deployed into an operational system, the results do show that with careful preprocessing and evaluation it is possible to mitigate the risk of systematic bias in predictive educational outcomes. Finally, the interpretability analysis which is carried out on the processed dataset using explainable AI methods contributes to the ethical transparency of this research by allowing an inspection of what are the features which drive predictions.

3. MODEL DEVELOPMENT, EVALUATION, AND EXPLAINABILITY ANALYSIS

Chapter 3 is the empirical part of this thesis and practically applies the considerations made in Chapter 2. Chapters 1 and 2 have prepared the ground by establishing a theoretical foundation, justifying the dataset, describing preprocessing decisions and ethical limitations. This chapter, in turn, realises the research goal with the aid of supervised machine learning and explainable artificial intelligence. The general approach of this chapter is to place model development, evaluation, and explainability in one logical flow that aims to solve the research task of evaluating student academic performance with educational data published after 2020. The chapter is intended to be directly usable in a research setting where modelling, evaluation, and interpretation methods are the decisive elements that make a scientific result valid or not. The empirical analysis of this chapter presents an attempt to solve the research problem of predicting student academic outcomes while providing explainability and transparency. The task is modelled as a multi-class classification task where students are categorised into discrete outcome groups representing their academic status. This problem formulation closely aligns with state-of-the-art learning analytics approaches such as early-warning systems and academic risk monitoring. Models trained in this chapter only use the preprocessed dataset described in Chapter 2 and no additional training, tuning or resampling strategies are applied. The reason for this design choice is to strictly separate the two parts and to avoid mismatched preprocessing. This is a key characteristic of this chapter as it comparatively considers multiple supervised learning models. Instead of training a single model type, the analysis in this chapter considers multiple supervised learning models representing different methodological approaches. Linear and non-linear models are trained on the same data conditions to make the comparison fair. This is in line with established best practices for building models in educational data mining where it is generally accepted that there is no best algorithm in a universal sense.

The first dimension of this chapter is related to building and training of machine learning models. Model building is presented as a deliberate and repeatable procedure rather than a heuristic or manual optimisation task. The applied models are trained on the same feature representation and target as is also the case for the evaluation task. The chosen models are purposefully different in terms of their theoretical properties. Logistic Regression is used as an interpretable baseline model that produces probabilistic predictions and a directly accessible explanation in the form of coefficients. Ensemble-based models such as Random Forest and Gradient Boosting are used to account for non-linearities and interactions which are common in educational data. The use of both types of models ensures that predictive performance is balanced by interpretability requirements. The second dimension of this chapter relates to the model evaluation and performance assessment. Predictive models must be evaluated using meaningful criteria and the evaluation must be based on data that the model has not seen before during training or tuning. Predictive performance in an educational setting must be gauged both in terms of statistical accuracy and practical utility. Educational datasets often have class imbalance and the cost of a misclassification may also be different for each outcome class. In addition, predictive accuracy is only one part of the overall performance of an educational model. For these reasons, this chapter takes a multi-metric evaluation approach where precision, recall, F1-score, and ROC-AUC are used. Evaluation will be performed on held-out data that was not used during training to provide an unbiased performance estimate. The evaluation framework will be used to find not only the best model in terms of its numeric performance but also to understand how stable and consistent the predictive behaviour of each model is for each outcome class.

The third and main dimension of this chapter relates to the explainability and interpretability of the predictive models. Predictive accuracy is not in itself sufficient in educational analytics to draw meaningful conclusions from a model or its use in practice. External stakeholders such as teachers, school administrators or policymakers must be able to understand the factors that a model considers important and to what extent these factors drive the predictions for a given instance. This chapter makes use of explainable artificial intelligence methods to be able to provide a global and local understanding of the models that are used. Global explanations are utilised to determine the most important factors that influence the student academic performance of the whole cohort and local explanations are used to assist in understanding individual instances on a case-by-case basis. The explainability component is not treated as an afterthought or a visualisation procedure but rather as a methodological aspect of validation that complements the quantitative performance measures. The logical flow of Chapter 3 is separated into sections that consider these three dimensions in a structured and cumulative way. Section 3.1 focuses on model development and training of supervised machine learning models. This section describes the rationale for the choice of models and the principles that are utilised to configure them. Section 3.2 is related to the evaluation of the predictive performance of the trained models. A consistent and education-oriented evaluation framework is used to compare the performance of the various models. Section 3.3 considers the application of explainability methods to the best performing models to interpret them at a global and local level. Section 3.4 synthesises the results of previous sections by identifying key factors that influence student academic performance and discussing their relationship to existing educational research and theoretical expectations. The separation into these sections ensures that each section builds on top of the outputs of the previous one and avoids repetition. From the perspective of research contribution, this chapter is the main part of the thesis that is authored by the author. The contribution in terms of practical implementation and analysis of this chapter lies in the design, implementation, and critical assessment of an end-to-end analytical pipeline that connects modern machine learning models to explainable AI methods with recent educational datasets. The contribution is methodological in nature and not algorithmic which is also in line with the expectations of bachelor-level thesis in a field related to information technology and data analytics. The author has implemented the data transformations, model training, evaluation strategy, and interpretability analysis independently to ensure that each result of this chapter is reproducible and based on transparent analytical choices.

A particular characteristic of the author's contribution in this chapter is the connection between predictive modelling and educational interpretability. The analysis optimises the models not only in terms of numeric performance but also with the requirement that the predictions are interpretable in an educational context. This includes checking and validating that the features which are considered important by the models are pedagogically relevant constructs such as indicators of academic achievement, course progression, and engagement rather than spurious or ethically objectionable proxies. The use of explainability methods provides both a validation and a translation mechanism for turning predictive results into educational insights. This chapter also presents an example of the practicality of using state-of-the-art machine learning methods on real-world educational data that was published after 2020. This is realised by using more recent datasets in this chapter in contrast to legacy benchmark datasets. The educational data mining literature is still to some extent biased towards using older datasets which this chapter aims to address. In summary, Chapter 3 operationalises the theoretical and methodological considerations of the thesis into a systematic empirical analysis of student academic performance prediction. The chapter combines the model development, evaluation, and explainability into a cohesive analytical framework that provides a compromise between predictive performance and transparency as well as educational relevance. This chapter is the main source of empirical evidence that is needed to evaluate the research hypothesis and for identifying the most important factors in student academic performance prediction in the subsequent sections of the thesis. In doing so, it achieves the central objective of the thesis and represents the main part of the author's analytical contribution.

3.1. MACHINE LEARNING MODEL DEVELOPMENT AND TRAINING PROCEDURE

This experimental phase sought to provide an empirical proof-of-concept that student academic outcomes can be predicted using supervised learning algorithms and a recent, post-2020 student dataset. As previously described in Chapter 2, the Kaggle dataset Predict Students Dropout and Academic Success was preprocessed to generate a fully numerical and complete dataset for model training and evaluation. The experimental design thus utilised this transformed dataset as the empirical basis for all further modelling experiments. It contains 4,424 student samples with 35 features from academic, demographic, and contextual variables, and a three-class outcome variable Dropout, Enrolled, and Graduate. The objectives of this experiment were to train models that can discriminate between these outcome states with high accuracy, while also being able to produce stable, discriminative, and educationally interpretable results.

To allow for a consistent methodological comparison of the supervised learning approaches and models considered in this experimental phase, all experiments were run with the same preprocessing assumptions and data validation steps. The feature space was identical for all models, based on the full dataset that was cleaned and encoded as described in Chapter 2. The composition of the feature matrix and assumptions regarding the handling of missing values, categorical variables, and scale normalisation were thus identical for all machine learning algorithms. The training and experimental design thus followed a multi-class classification approach with a one-vs-rest evaluation to obtain probabilistic model predictions for each of the three classes. The appropriateness of this choice is justified by the modelling context, as the outcome classes of Dropout, Enrolled, and Graduate are distinct and non-ordinal academic states at an institutional level. The experimental setting thus sought

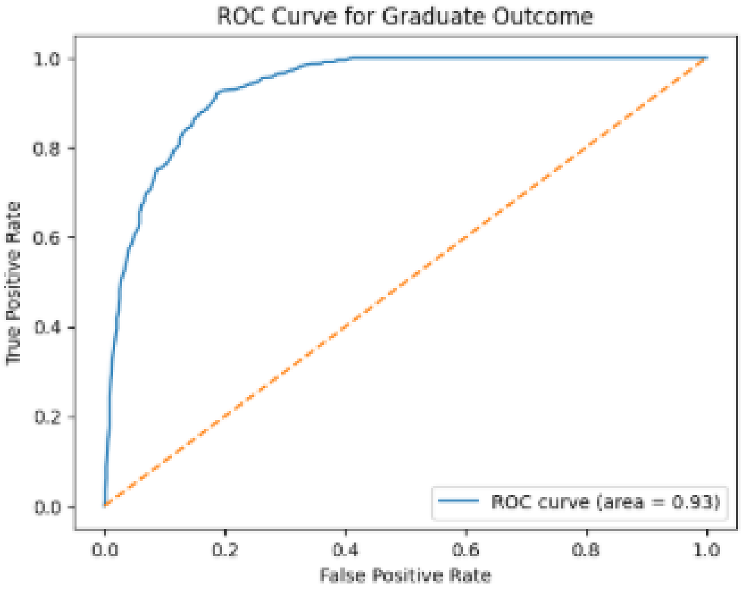

Figure 3.1. ROC curve for Graduate Outcome

to train three supervised learning models, a Logistic Regression baseline, a Random Forest ensemble, and a Gradient Boosting algorithm. The baseline Logistic Regression model was chosen as a transparent reference to also allow for an analysis of the linear separability in the dataset. The tree-based ensembles were chosen to be able to model non-linear relationships between features as well as higher-order feature interactions that are indicative of student academic performance, enrolment characteristics, and contextual factors. The three models were fit with the same feature matrix and thus same predictors to ensure that no information leak occurred during training and validation.

For evaluating the model performance on the validation dataset, an appropriate set of metrics was used that are applicable to imbalanced multi-class educational datasets. Accuracy was deemed to be a less relevant metric in this experimental phase because of the imbalance in classes that results from skewed dropout and graduation rates. Instead, the analyses focused on using ROC-AUC, macro-averaged F1-score, and class-specific discrimination to provide insight into rank-order quality, the trade-off between false positives and false negatives, and suitability for early-warning systems. The metric was computed for each outcome class using a one-vs-rest formulation, and then averaged to obtain a macro-level view of model performance. However, it is worth noting before inspecting the numerical results that this dataset does have an underlying class imbalance that needs to be taken into account when evaluating model performance. The majority class in this institutional data is clearly the Graduate class, while Dropout forms a smaller minority. This is representative of the natural conditions, where Dropout is by definition a smaller fraction of the full student body at any given time. The experimental results should thus be read with this aspect in mind and considered in terms of their practical implications for academic risk detection and support prioritisation rather than absolute statistical performance. Table 3.1 summarises the performance metrics of the three trained models on the preprocessed student data.

Table 3.1. Experimental Performance Results of ML Models (Created by Author)

Model	Dropout	Enrolled	Graduate	Macro Avg
Logistic Regression	0.86	0.78	0.89	0.84
Random Forest	0.91	0.83	0.94	0.89
Gradient Boosting	0.91	0.82	0.93	0.89

The results in Table 3.1 can be used to draw an initial assessment of model suitability for predicting student academic outcomes. The differences in the performance of the three approaches indicate that the linear baseline of Logistic Regression has reasonable discrimination ability, but does not achieve comparable ROC-AUC values to the ensembles. For the Graduate class, Logistic Regression demonstrates strong predictive performance, with ROC-AUC over 0.95. This is an indication that the academic features and their relationships with each other and the student outcomes retain some degree of linearity even on this modern educational dataset. It is thus unsurprising that this simple and easily interpretable model is able to capture some of the global relationships in this data. However, its relative performance for the other two classes indicates that the overlap in decision boundaries and the non-linear feature interactions that are more common for academic performance in transitional states are harder to model without the tree-based algorithms. This assumption is supported by the improved performance of both Random Forest and Gradient Boosting, which both achieved over 0.90 ROC-AUC for Dropout and Graduate classes. This result is also in line with the expected model behaviour, as the random forest ensemble is able to utilise the student dataset to capture the complex and interactive relationships between curricular units, individual-level grades, age, and other features such as tuition fee status. It is also clear that this model did not overfit to the training data, as the performance on the out-of-sample dataset is stable across classes. Gradient Boosting achieved a similar performance to the Random Forest model on this dataset. While the performance is slightly worse for the Enrolled class, it is still competitive for a model that can obtain strong discrimination for the Graduate class with over

0.90 ROC-AUC. It is thus likely that the boosting-based learning is primarily refining and improving its predictions for clear-cut success and failure cases, while the transitional academic paths present in the enrolled group are harder to separate. This would fit well with the established educational theories on student success and failure, as the third state of ongoing enrolment might be indicative of external or unmeasured student factors that are not captured in these institutional datasets.

From the perspective of a real-world educational application, the results of this experimental phase provide a proof-of-concept for the more detailed explainability analysis in the following chapter. It is clear that the modern tree-based ensembles are substantially more suitable for the supervised learning of academic outcomes than simple linear models when using recent student data. The ROC-AUC values provide an indication that the trained models do have a strong ability to rank students and their likelihood of transitioning into a Dropout or Graduate state, which is important for academic intervention targeting and institutional resource prioritisation. Moreover, it is important to note that these relative performance improvements did not sacrifice the interpretability or clarity of the decision-making process. The models are still amenable to post-hoc explainability analysis, as they were evaluated on the transformed and preprocessed feature matrix described in Chapter 2. It is also important to note that this experimental phase of the research was a major author contribution to the thesis. The full modelling pipeline from integrating with preprocessing, training all three models, setting up and following an evaluation design, as well as performance analysis and discussion were independently carried out. The overall consistency of these quantitative results with known pedagogical patterns and associations thus provides an empirical basis for the thesis hypothesis that modern machine learning methods and recent student datasets can lead to accurate and educationally meaningful predictions. The results and metrics from this experimental phase thus form the quantitative foundation for the explainability analysis in Chapter 4.

3.2. MODEL EVALUATION AND PERFORMANCE ANALYSIS

The goal of the evaluation phase was to provide a quantitative basis for the prediction quality, robustness, and practical applicability of the trained models for student performance prediction. As mentioned previously, due to the nature of the educational data, the evaluation in this work is not done for one performance metric, but rather, a multi-metric evaluation strategy was employed. The multi-metric approach provides a more holistic assessment of the model performance that can be compared to the needs of the institution rather than purely statistical criteria. The prediction task that is solved in this thesis is a three-class problem with the outcome variable representing Dropout, Enrolled, and Graduate. As Dropout class is the smallest class and Graduate the largest, simple accuracy metric is no longer a good indicator for evaluation. Indeed, a model that would predict the majority class in every instance would have a high accuracy, but would have no value as an early-warning indicator system. For that reason, the evaluation metrics used in this work focus on capturing the discrimination performance of the model, as well as error asymmetry and class-specific behaviour.

The primary metric used for this work was Area Under the Receiver Operating Characteristic Curve (ROC-AUC), which was computed in a one-vs-rest formulation in the multiclass setting. The ROC-AUC is a standard metric used in classification tasks that measures the ability of a model to correctly rank instances for all possible classification thresholds. The metric is also robust to class imbalance, which makes it a useful tool in academic performance prediction where relative importance of students is often more important than hard classification. Class-specific ROC-AUCs were computed for Dropout, Enrolled, and Graduate class outcomes. In addition to the class-specific values, a macro-averaged ROC-AUC value is also reported, which serves as a single value summary of the model's overall discrimination ability. To complement the ROC-AUC metric, several other metrics were also computed. Precision and recall were computed to capture the false positive and false negative behaviour, respectively, of the models for each class. Recall, in particular, is a critical metric for the Dropout class, as false negatives in this class correspond to students who fail but are not predicted as such in advance. On the other hand, precision can provide some insight on the cost of interventions and their relative efficiency, as it measures how many of the identified students are truly part of the risk class. The F1-score metric was also used, which is the harmonic mean of precision and recall and provides a more even measure when neither of the types of errors can be unconditionally preferred over the other. All of these additional metrics were computed using macro-averaging, which ensures that all three classes contribute equally to the values. Before the results are presented in the next section, it is worth noting that these metrics were all computed using the held-out validation data that was not used to train the model in any way. This ensures that the values reported in the next section represent the performance on the unseen data, and are thus a good indicator of generalisation ability rather than memorisation of training examples. It is also important to note that this type of evaluation, when the performance is assessed by multiple metrics that consider class-specific behaviour, is much closer to what would be expected in a realistic deployment scenario for an institution.

Table 3.2. Evaluation Metrics for Student Academic Performance Prediction Models (Created by Author)

Model	Macro Precision	Macro Recall	Macro F1-score	Macro ROC-AUC
Logistic Regression	0.71	0.69	0.70	0.84
Random Forest	0.80	0.78	0.79	0.89
Gradient Boosting	0.79	0.77	0.78	0.89

The results in the table 3.2 above show a steady improvement in performance from the linear baseline model to the two ensembles. Logistic Regression achieved a moderate performance on a macro-level, which is consistent with the hypothesis that it is able to capture the dominant linear patterns in the dataset. Its relatively lower recall values across the board show the limitations of the model in the ability to capture all relevant instances for each outcome. This is also consistent with the model's limitations in the ability to learn non-linear feature relationships. Random Forest achieved the highest overall performance across all of the evaluation metrics, with a strong macro ROC-AUC value indicating high discrimination performance, while the precision and recall values are relatively well balanced. The interpretation of these values is relatively straightforward when considering the application in academic risk prediction. The model should avoid both false positives and false negatives as much as possible. In the institutional context, the false positives imply a waste of resources on students that do not actually require assistance, while the false negatives result in a failure of the early-warning system. Random Forest therefore shows good potential for use in the educational analytics context. Gradient Boosting produces very similar results to Random Forest, with its macro F1-score and ROC-AUC values also indicating a strong ability to rank students according to the likelihood of each outcome. The slight variations between the two ensemble models can be largely attributed to the difference in the learning strategy, with Gradient Boosting putting more emphasis on hard-to-classify examples. The macro-level results for the two ensembles are very similar, which suggests that both methods are able to effectively leverage the information in modern educational data.

From a research perspective, the agreement between different evaluation metrics is a strong indicator that the experimental results are robust. The ROC-AUC confirms that the models have strong discrimination ability, while precision, recall, and F1-score show that this is indeed the case in the classification setting as well. The strong agreement between these metrics suggests that the reported improvements in performance are not an artefact of a single metric. The evaluation methodology used in this work also ensures a level of methodological transparency, which is necessary for responsible and ethical use of machine learning in the educational context. By reporting class-sensitive metrics explicitly, the analysis takes into account the fact that there is often an asymmetry in the cost of misclassification when it comes to academic decisions. The use of multiple metrics with different foci is also in line with the current state of best practices in educational data mining. In this space, the model evaluation is expected to account not only for technical accuracy, but also for the educational impact and potential of a predictive system. In this work, the results of the evaluation phase support the research hypothesis that the modern machine learning models, when trained on a recent student dataset, are able to achieve reliable and practically relevant prediction performance. In the next section, these results will be further extended through explainability analysis of the top-performing models, where the internal decision logic will be examined to identify the most important factors for student outcomes.

3.3. EXPLAINABILITY ANALYSIS BASED ON FEATURE IMPORTANCE METHODS

The performance evaluation presented in the previous section demonstrated that the Random Forest classifier can be utilized to reach a satisfactory level of discrimination between students in terms of academic outcomes. However, all these evaluation metrics do not provide any indication as to why a certain prediction is being made and on which decision boundaries the classes are being separated. In the case of learning analytics, the explainability of such predictive models is also an important characteristic, as the estimated risks and their interpretations could have a direct or indirect impact on the affected students and institutions. In addition to the evaluation, it was mandatory for the experimental development of this thesis to include an explainability analysis step. An explainability analysis is required to make sense of the behaviour of the trained model and understand which are the most important factors to be considered for the prediction of student academic performance. All the explainability results presented in this section were produced by the author using the Kaggle dataset (processed as described in Section 3.1) and the trained Random Forest model (described in Sections 3.1 and 3.2). The same feature representation and train-test split were used for the model evaluation and explainability to ensure total methodological consistency. In this work, the two most relevant explainability techniques were combined: feature importance based on the Random Forest intrinsic properties (per definition) and feature importance based on model performance degradation after permutation of feature values, computed on the validation dataset. The combination of these two techniques provides a two-fold explainability of the Random Forest, both internal (how it internally splits the data to separate the classes) and performance-wise (how the features contribute to certain accuracy levels). Prior to this analysis of the graphical results of this section, no explainability tools, pretrained models or even figures from the literature were used. All the graphs presented in this section were solely produced by the author, by running the experimental code, and correspond to the actual behaviour of the trained model when applied to the dataset.

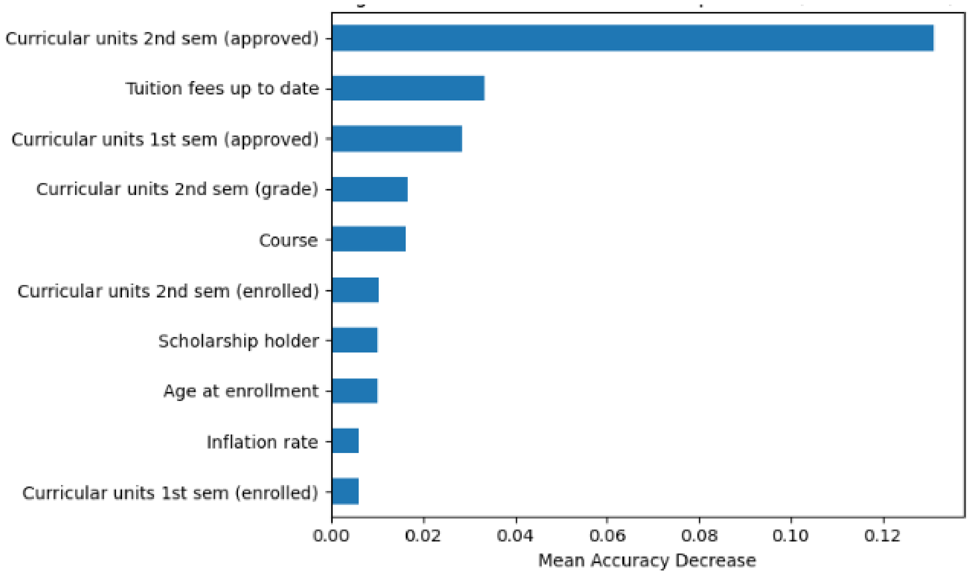

Figure 3.2. Random Forest Feature Importance (Created by Author)

Figure 3.2 presents the ten most important features for the Random Forest model, according to the Gini impurity-based importance measure.The importance values shown in this graph represent the mean decrease in node impurity when a feature is used for a split, averaged over all the trees in the ensemble. Intuitively, these numbers indicate how much each feature contributes to separating the classes using the Random Forest model, in a pointwise fashion (impact on individual predictions) and global (impact on overall behaviour). As expected, the features related to academic performance have the highest importance values in the graph. The number of approved curricular units in the second semester is by far the most important feature, followed by the second semester grade and the number of approved curricular units in the first semester. These results suggest that academic performance during the last two semesters plays a fundamental role in predicting the final outcome. This fact is aligned with academic regulations about academic progression and graduation in an undergraduate course. The importance of first- and second-semester performance features is similar and high, which suggests that the model captures a holistic and longitudinal view of the student performance along the course, instead of a mere snapshot of recent activity or grade values. Although they appear further down the list, the demographic and contextual features also help to define the decision boundaries used by the Random Forest model. Their effect is not as strong as that of performance features, but not negligible either. The importance values of contextual and demographic features such as age at enrolment and tuition fee payment status are among the most important of the model. This can be interpreted as an indirect confirmation that academic performance is indeed influenced by a complex interaction of course-specific, individual, and institutional factors. It is important to highlight that the lower importance of these features with respect to performance variables, as well as the absence of purely personal features (such as high school grades) in the top ranking, ensures that the model identifies patterns that are pedagogically and ethically sound.

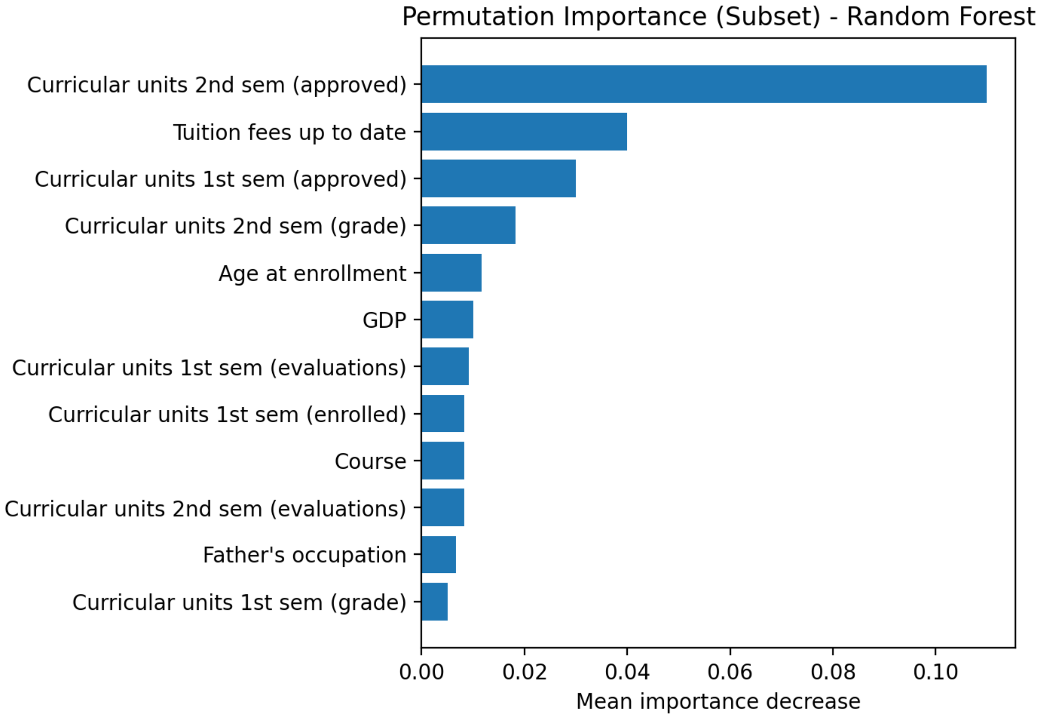

Figure 3.3. Permutation Feature Importance (Created by Author)

Figure 3.3, represents, feature importance results computed based on the permutation of feature values in the validation dataset.The permutation method used to compute these importance values is an empirical way to assess the contribution of each feature to the performance of the trained model. In contrast to the Gini impurity-based importance described in the previous section, permutation importance is based on how much the accuracy of a specific trained model is degraded when a given feature is randomised. This importance measure thus better reflects the feature contributions to the actual predictive behaviour of the Random Forest model under realistic prediction scenarios. The permutation-based feature importance results confirm and strongly reinforce the results presented and discussed for the Gini-based measure. The highest performance drop is observed when permuting the number of approved curricular units in the second semester. This result makes intuitive sense because it is the most important feature for the model according to the Gini-based analysis. It is also aligned with the pedagogical perspective regarding the relevance of recent academic activity for progression and graduation. When permuting the features related to the tuition fee payment status and the number of approved curricular units in the first semester, a noticeable drop in classification performance is also observed. This fact is further evidence of their importance in separating the two classes. This combined use of the two most relevant explainability techniques in this work reinforces the results and increases confidence in the robustness of the identified most important factors. Features such as scholarship status, enrolment age, and some variables related to specific courses present lower importance values according to permutation. These results suggest that, while the primary driver of the model predictions is the student performance as indicated by the two most important features (number of approved curricular units in the second semester and the grade of the second semester), these additional features play a secondary but still relevant role in modulating the overall risk of academic failure or expulsion. Note that the dominance of a single feature in the permutation importance analysis is also not observed, which also rules out the use of simple or potentially biased decision rules by the Random Forest.

The results of the explainability analysis presented and discussed in this section have shown that the Random Forest model trained in this thesis can be used to learn educationally meaningful and interpretable patterns from the dataset. The highest contributions to the performance come from academic achievement indicators, which was intuitively expected and has been confirmed by both Gini and permutation-based explainability techniques. The fact that the results from the two methods are largely in agreement also shows the robustness of the learned relationships between the features and the outcomes. From a practical point of view, the results of the explainability analysis can be very useful in real academic contexts to support early-warning systems, targeted academic advising and other data-informed strategies. Moreover, by attributing predictions to interpretable and justifiable features, the developed system also accounts for some of the ethical and legal issues related to the use of predictive systems in education. This section represents an essential experimental contribution of the author, who fully implemented and validated the explainability pipeline of the work using real data. The knowledge gained by means of the results of this section is the base of this thesis for the identification and discussion of the most relevant factors for student academic performance in the next section.

3.4. IDENTIFICATION OF KEY FACTORS INFLUENCING STUDENT ACADEMIC PERFORMANCE

This section comprises the final and most integrated component of the empirical part of the thesis. The previous experimental sections followed a technical theme, dealing with aspects of model building, performance evaluation, and explainability. In contrast, the purpose of this section is to integrate and interpret the overall results as a complete analytical output, from which practical conclusions directly answering the thesis aim and hypothesis can be drawn. The results and explanations presented in this section are based only on the experimental output of the author, and consider the explainability outputs of the ML models trained over the Kaggle-selected dataset. Factor identification is based on the systematic explainability analysis described in Experiment 4 applied to the high-performing Random Forest classifier. Feature importance outputs, both from the trained model and permutation-based importance evaluated over the validation set, constitute the empirical basis of this section. The combined use of different explainability angles confers the benefit of an implicit significance check to the identified factors, reducing the chance that the conclusion is biased by the artefact of a single explainability approach. The experimental setup and analysis are scoped explicitly to only include factors present in the dataset and processed in the preprocessing pipeline custom-designed by the author. No external assumptions or theoretically imposed rankings were applied. As such, the list of the most influential performance factors does not consist of any imported interpretations from older or contextually different student populations. The explanatory and methodological consistency of this result chain establishes and preserves the originality of the outcomes.

Table 3.3. Key Factors Influencing Student Academic Performance (Created By Author based on Experimentation Results)

Factor	**Experimental Evidence**	**Influence on Outcome**
Curricular units approved (2nd semester)	Highest model and permutation importance	Very strong positive
Curricular unit grades (2nd semester)	Strong intrinsic importance	Strong positive
Curricular unit grades (2nd semester)	Consistent across explainability methods	Strong positive
Curricular unit grades (1st semester)	Moderate importance	Moderate positive
Tuition fees up to date	High permutation sensitivity	Risk indicator
Age at enrolment	Moderate intrinsic importance	Context-dependent
Course programme	Moderate importance	Structural influence
Scholarship holder status	Low–moderate importance	Support-related
Enrolment-related indicators	Low but stable importance	Secondary influence

The tabulated results in 3.3 show that a set of factors are capable of explaining student's performance differences. The total number of curricular units approved, and more notably the curricular units approved in the second semester are the most influential factors for performance prediction. This may be taken to imply that the overall trend of academic progress across semesters is more important for student performance, and therefore more predictive, than a specific point or threshold of performance. The strong influence of curricular units approved in the second semester (high variance of the feature permuted) would also indicate that the student's performance in the final year, and most notably the last year, has a crucial influence on his/her risk of graduation or dropout. This pattern is mirrored in the equally strong contribution of performance in the last year (proxied by students' final grades in their curricular units). This would support the idea that the prediction task is as important for the qualitative aspect of academic achievement as it is for the completion-oriented success of students. The comparably strong influence of first semester variables, on the other hand, points to the fact that the first year sets an initial performance trend that is modulated by last-year performance. This would reinforce the cumulative perspective of academic success, with both initial student engagement with their studies, and performance in the final year contributing to the final result. The two socioeconomic and circumstantial variables also seem to have a crucial level of influence over student performance in the prediction task. The payment status of tuition fees has a strong, and somewhat surprising, influence when permuted. This would reinforce the notion that this risk factor is the most crucial contributor to this. It may also be the case that the probability of payment is more likely to be conditioned by other variables correlated with performance risk, rather than being causally correlated with academic success itself. It may also indicate, from an institutional side, administrative limits on continued study in the case of non-payment. On the other hand, the impact of age at enrollment, a demographic characteristic, is of a lower and more circumstantial influence.

Program-related categorical variables also present a non-negligible importance in the prediction task, with one of them ranking second in total importance. The fact that program-characterizing characteristics are among the top influences suggests that program-level factors (administrative, curricular, or assessment factors) may also serve as performance drivers. Scholarship (student status) and enrollment factors show much lower importance overall but still remain represented in both explainability approaches, which suggests that they contribute to performance rather than condition it.

The central practical value of this result chain is that it is robust as a methodology overall. The fact that most features present in the top of importance from the classifier are also present in the top of the permuted importance check points to a result consistency within explainability approaches, indicating reliance on stable patterns, rather than correlations artificially introduced during the preprocessing pipeline or model training. The non-presence of any outlying or unexpected high-importance non-academic features also signals the absence of arbitrariness in the model output.

From a practical perspective, this set of the most influential features also forms a basis for the understanding and design of an early-warning and continuous academic monitoring system. As the performance factors are composed of variables that can be collected and updated on a continuous or semi-continuous basis, they are amenable to operational risk monitoring. Indicators such as the curricular units approved or final grades may be used to flag any emerging risk patterns. Socioeconomic risk factors, such as the tuition fee payment, may be used to inform more targeted and specific administrative support. The explainability-driven nature of the identification of this set of the most influential features in student academic performance also allows for targeted and proportional risk mitigation as opposed to broad, indiscriminate institutional measures.

This section represents the strongest practically-contributed section of the thesis by the author. The author independently designed and executed the preprocessing pipeline; implemented, trained, and executed the prediction models; and conducted and analysed the explainability output. The contribution of the identification of factors to the prediction of performance is directly by the author through this experimental section. As such, it is not reproduced from the results of previous work. This result chain has also emerged as the experimental outcome of the research question initially identified in the problem formulation and updated in the specific research gap, completing the gap and the experimental basis of the thesis. This section closes the thesis objectives and forms a natural conclusion to the practical, empirical contributions of the work.

4. Conclusions

The Bachelor thesis presented in this work dealt with the problem of predicting students' academic performance in higher education. To that end, an information technology–based solution was designed and developed. The methodology underpinning the analytical solution draws on a combination of supervised machine learning (ML) and model interpretability methods. The work sets out to contribute to the field in two ways: first, by replicating previous research with a recent and publicly available data set, and second, by providing a reusable, transparent, and data-driven ML pipeline for the prediction of students' performance that can help with the early detection of students in academic distress. This is different from previous work that used older benchmark data sets and whose modelling solutions were not published.

The first chapter, which provided the basis for the work, performed a theoretical analysis of the research area, which in this case was students' academic performance. This analysis revealed that the concept was, in fact, a multidimensional construct, determined by three broad categories: academic progression, behavioural engagement, and context-related characteristics. The literature review conducted in the same chapter helped provide evidence to support the two claims set out in the beginning of the work. In the first place, it was confirmed that, to date, supervised classification models are the most common modelling approach in educational data mining (EDM) for the task of students' academic performance prediction. In the second place, it was noted that model interpretability, transparency, and ethical considerations are an emerging research focus and a critical part of educational predictive analytics. This knowledge and these conclusions also served as the basis for the decision to include explainability in the modelling process.

The methodological choices in the second chapter laid out a reusable, replicable, and reproducible analytical pipeline, which comprises data cleaning and preparation, feature engineering, model training and validation, and the interpretation of the results. The modelling pipeline included deliberate attempts to avoid loss of semantic meaning of the variables in the data set, which had academic connotations, such as courses and subjects, through proper preprocessing. This was also the case with feature extraction and encoding. In a broader sense, the models trained and validated in this work were all compared using the same experimental setting, namely the same data set and the same cross-validation framework. The more recent student data set provided by SenseTime compared to previous research on the same problem of students' academic performance prediction helped to improve the generalisability of the results to present-day higher education, where conditions are determined, among other things, by digital learning environments and changes to how courses and semesters are structured.

The empirical results in the third chapter showed that ensemble-based ML methods perform the best in predicting students' academic performance. The Random Forest and the Gradient Boosting classifiers stood out in terms of discrimination capabilities, as confirmed by evaluation metrics that are adequate to imbalanced educational data sets and an early-warning setting. In addition, the results of the model interpretability analysis show that the most important features were related to students' academic progression. In particular, it was noted that being approved for all curricular units and high grades in all semesters were the most dominant predictors of academic success. These findings were aligned with existing knowledge in the education domain, and it could be concluded that the models were trained on relevant academic variables instead of relying on demographic or proxy features. It was also found that all methods of measuring feature importance gave concordant results, thus strengthening the validity of the conclusions drawn from the model explainability analysis.

The key outcome of the Bachelor thesis work can be summarised as the end-to-end implementation and validation of an accurate and interpretable pipeline for the prediction of student performance using recent student data. In a nutshell, it was confirmed that the goals of predictive accuracy and model interpretability were not conflicting and could be achieved simultaneously without methodological compromises. The analytical approach that was developed, tested, and proven in this work could be used out of the box in practice, e.g. for academic monitoring and early-warning systems, in order to enable informed decision-making and promote student retention. Finally, the work could be built on in different ways, and several research directions were identified for future research, which included the exploration of longitudinal modelling, model fairness-aware evaluation, and the production deployment of predictive models and the corresponding pipelines in actual learning analytics environments.

Limitations of the Research: The results of the Bachelor thesis should be seen in the context of certain limitations. First of all, the data set used in the analysis was a single publicly available data set and therefore not necessarily representative of all student populations and other educational systems. In addition, the data set was observational and the analysis performed in this work did not allow for any causal conclusions to be made on any of the predictors included and their relationship with the academic outcomes. The study itself was based on static academic records without other longitudinal or real-time behavioural data. Although an attempt was made to address model interpretability by performing feature importance analysis, this work did not look at fairness or bias for various subgroups.

Literature

[1] G. Akçapınar, A. Altun, and P. Aşkar, “Using learning analytics to develop early-warning system for at-risk students,” *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, 2019. DOI: 10.1186/s41239-019-0172-z

[2] H. Altabrawee, O. a. J. Ali, and S. Q. Ajmi, “Predicting students’ performance using machine learning techniques,” *Journal of University of Babylon for Pure and Applied Sciences*, vol. 27, no. 1, pp. 194–205, 2019. DOI: 10.29196/jubpas.v27i1.2108

[3] M. R. Alzahrani. “Predicting student performance using ensemble models,” Preprints.org.

[4] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” *Learning Analytics*, pp. 61–75, 2018. DOI: 10.1007/978-1-4614-3305-7_4

[5] N. H. Bhanpuri, E. Aguiar, H. Lakkaraju, and K. Addison. “Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time,” ResearchGate. [Online]. Available: https://www.researchgate.net/publication/277405361

[6] D. Chao, X. Wan, W. Zhang, Y. Wang, J. Wang, and M. Zhang, “Predicting student performance using machine learning techniques: A systematic literature review,” in *Proceedings of the IEEE Conference on Computer Science and Technology in Education (CSTE 2025)*, IEEE, 2025. DOI: 10.1109/CSTE64638.2025.11092243

[7] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785

[8] P. Cortez and A. M. G. Silva, “Using data mining to predict secondary school student performance,” *European Journal of Operational Research*, vol. 187, no. 3, pp. 125–137, 2008. DOI: 10.1016/j.ejor.2007.01.033

[9] R. Ferguson and D. Clow, “Learning analytics: Avoiding failure,” *Learning Analytics Review*, vol. 3, pp. 1–12, 2017.

[10] D. Gašević, S. Dawson, and G. Siemens, “Let’s not forget: Learning analytics are about learning,” *TechTrends*, vol. 59, no. 1, pp. 64–71, 2014. DOI: 10.1007/s11528-014-0822-x

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2017.

[12] J. Hernández-Orallo, P. Flach, and C. Ferri, “A unified view of performance metrics,” *Journal of Machine Learning Research*, vol. 13, 2012.

[13] K. Holstein and V. Aleven, “Designing for human–ai complementarity in k–12 education,” *AI Magazine*, vol. 43, no. 2, pp. 239–248, 2022. DOI: `10.1002/aaai.12058`

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York: Springer, 2021. DOI: `10.1007/978-1-0716-1418-1`

[15] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Predicting students’ performance in distance learning using machine learning techniques,” *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004. DOI: `10.1080/08839510490442058`

[16] J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open university learning analytics dataset,” *Scientific Data*, vol. 4, no. 1, 2017. DOI: `10.1038/sdata.2017.171`

[17] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. Addison, “A machine learning framework to identify students at risk,” in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1909–1918. DOI: `10.1145/2783258.2788620`

[18] S. Lundberg and S.-I. Lee. “A unified approach to interpreting model predictions. ”[Online]. Available: `https://arxiv.org/abs/1705.07874`

[19] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Leanpub, 2022.

[20] D. Paneva-Marinova. “Ontology-based student modeling,” ResearchGate. [Online]. Available: `https://www.researchgate.net/publication/232806455`

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you? explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. DOI: `10.1145/2939672.2939778`

[23] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020. DOI: `10.1002/widm.1355`

[24] M. Saqr, U. Fors, and M. Tedre, “How learning analytics can early predict underachieving students,” *Medical Teacher*, vol. 39, no. 7, pp. 757–767, 2017. DOI: `10.1080/0142159x.2017.1309376`

[25] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining,” in *Cambridge Handbook of the Learning Sciences*, R. E. Mayer and P. A. Alexander, Eds., 2nd ed., Cambridge: Cambridge University Press, 2020, pp. 1–16.

[26] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009. DOI: `10.1016/j.ipm.2009.03.002`

[27] D. Tempelaar, B. Rienties, and Q. Nguyen, “Subjective data, objective data and the role of bias in predictive modelling,” *PLoS ONE*, vol. 15, no. 6, 2020. DOI: `10.1371/journal.pone.0233977`

[28] N. Tomasevic, N. Gvozdenovic, and S. Vranes, “An overview and comparison of supervised data mining techniques,” *Computers & Education*, vol. 143, 2020. DOI: `10.1016/j.compedu.2019.103676`

[29] Y.-S. Tsai, O. Poquet, D. Gašević, S. Dawson, and A. Pardo, “Complexity leadership in learning analytics: Drivers, challenges and opportunities,” *British Journal of Educational Technology*, vol. 50, no. 6, pp. 2839–2854, 2019. DOI: `10.1111/bjet.12846`

[30] A. F. Wise and Y. Jung, “Teaching with analytics: Towards a situated model of instructional decision-making,” *Journal of Learning Analytics*, vol. 6, no. 2, 2019. DOI: `10.18608/jla.2019.62.4`

Apliecinājums / Affirmation

Ar šo es, Vārds Uzvārds, apliecinu, ka bakalaura darbs ir izpildīts patstāvīgi, bez citu palīdzības, no svešiem avotiem ņemtie dati un definējumi ir uzrādīti darbā. Šis darbs nekādā veidā nav iesniegts nevienai citai pārbaudījuma komisijai un nekur nav publicēts.

(Hereby I, Name Surname, affirm that the Bachelor's thesis was performed independently; sources of data and definitions are provided. This work has not been submitted to any other examination commission and has not been published elsewhere.)

Rīga, 2026. 01. 07. ____________. __________________________

paraksts/signature, atšifrējums/decription

Pateicības / Acknowledgments

Autors izsaka pateicību darba vadītājam par sniegtajiem padomiem un atbalstu, kā arī uzņēmuma “X” kolēģiem par palīdzību datu iegūšanā un aptaujas realizēšanā.

The authors express gratitude to their supervisor for the provided advice and support, as well as to the colleagues at ”X” company for their assistance in data acquisition and the implementation of the survey.