Simon Lubambo

Rephrased Version

Quick Submit

Quick Submit

Uganda Christian University

Document Details

Submission ID

trn:oid:::1:3309871486

Submission Date

Aug 8, 2025, 1:18 AM GMT+3

Download Date

Aug 8, 2025, 3:58 PM GMT+3

File Name

Rephrased_Thesis_Full.pdf

File Size

6.4 MB

83 Pages

23,832 Words

145,891 Characters

37% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

65 AI-generated only 37%
Likely AI-generated text from a large-language model.

1 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

UGANDA CHRISTIAN UNIVERSITY, MUKONO, UGANDA

PREDICTING FINAL CGPA USING PRE-ADMISSION DATA: PROACTIVE INSIGHTS FOR ACADEMIC EXCELLENCE AT UGANDA CHRISTIAN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements for the Award of the Degree of

Master of Science in Data Science and Analytics

of the

Faculty of Engineering

Department of Computing and Technology

By

Simon Fred Lubambo

Supervisor: Dr. [Supervisor Name]

July, 2025

EXECUTIVE SUMMARY

In recent years, predictive analytics has emerged as a transformative tool in higher education, enabling data-driven academic advising and institutional decision-making. However, its adoption in Sub-Saharan Africa remains limited. This study addresses this gap by developing and evaluating a machine learning model to predict final Cumulative Grade Point Average (CGPA) at Uganda Christian University (UCU) using only pre-admission data.

The research utilized historical records, including A-Level results, O-Level grades, UCE credits, and demographic information, to develop predictive models. The tuned Random Forest model outperformed the baselines, achieving strong predictive accuracy and demonstrating that a student's final CGPA can be forecasted early in their academic journey. To ensure transparency, the study integrated explainable AI techniques such as SHAP analysis and sensitivity simulations, providing both global and individual-level interpretations of model predictions.

Beyond raw predictions, students were categorized into performance bands (high, moderate, and low) to enable targeted institutional interventions. Program fit simulations identified academic pathways aligned with individual strengths, while fairness audits confirmed overall model equity, with minor subgroup disparities that required ongoing monitoring. These findings highlight the model's utility for proactive advising, resource allocation, and curriculum planning.

Furthermore, the study assessed the model's deployment readiness by packaging it as an API and exploring integration with the university's Management Information System (MIS). The work concludes with a phased adoption strategy and recommendations for scaling predictive analytics within UCU and similar institutions.

Overall, this research contributes a locally contextualized, interpretable, and actionable predictive framework for higher education in Uganda. It demonstrates how machine learning can be responsibly applied to improve student outcomes, support institutional strategy, and foster data-driven innovation in Sub-Saharan African universities.

DECLARATION

I, Simon Fred Lubambo, hereby declare that this thesis titled *"Predicting Final CGPA Using Pre-Admission Data: Proactive Insights for Academic Excellence at Uganda Christian University"* is the result of my original research and work. To the best of my knowledge, it has not been submitted in part or in full for the award of any degree or academic qualification at this or any other institution.

All sources of information, data, and ideas from other authors have been duly acknowledged and referenced by academic standards. This research has been conducted with integrity and in adherence to the ethical guidelines and regulations of academic scholarship.

Signature: _____

Date: _____

turnitin Page 5 of 85 - AI Writing Submission Submission ID trn:oid:::1:3309871486

APPROVAL

Title of the Research Thesis:

Proactive Insights for Academic Excellence: A Predictive Analytics Approach to Student Performance at Uganda Christian University

Statement of Approval:

This research thesis has been reviewed and approved for submission in partial fulfillment of the requirements for the award of a Master's degree. The undersigned confirms that the work meets the required academic standards for scholarly research and is approved for implementation.

Supervisor's Details

Name: _____

Title: _____

Department: _____

Signature: _____

Date: _____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background to the Study

In the 21st century, data has emerged as the currency of innovation, influencing decision-making across sectors including banking, healthcare, and retail Dadwal et al., 2021. Predictive analytics, powered by machine learning, has enabled banks to foresee financial risks, hospitals to forecast disease outbreaks with high accuracy, and businesses to optimise supply chains with unprecedented precision. Higher education, traditionally slow to adopt these technologies, is now beginning to embrace predictive analytics to transform how students are supported and how institutions plan for success.

Universities are increasingly implementing data-driven methodologies to provide proactive support to students. Instead of waiting for failure or dropout, institutions can now identify risk early and offer personalised interventions. This shift is supported by evidence: at the Czech Technical University, predictive learning analytics reduced first-year dropout rates from 37% to 19% Kuzilek et al., 2015. Similarly, The Open University demonstrated that predictive dashboards significantly improved outcomes across faculties Herodotou et al., 2019. These empirical findings confirm that early warning systems, when integrated into teaching practices, enhance student retention and performance.

While predictive analytics has shown remarkable success globally, most existing models rely heavily on in-program data such as LMS interactions, coursework submissions, and attendance logs. Although these models achieve high accuracy, they are reactive, intervening only after risk patterns have emerged. This reactive nature leaves a critical gap at the admission stage, when decisions on program placement and early support could have the most significant impact. Furthermore, while global models, such as decision trees, random forests, and support vector machines, have demonstrated strong predictive power Bacus and Cascaro, 2024; Fahd et al., 2022; Karim-Abdallah et al., 2025, concerns over transparency and fairness have prompted many institutions to adopt interpretable approaches.

The need for early, interpretable prediction is particularly urgent in Sub-Saharan Africa,

where adoption of predictive analytics remains low, with only 24 out of 811 global EDM studies focusing on the region Maphosa and Maphosa, 2020. Existing models, primarily developed in Western contexts, often prioritize accuracy without addressing interpretability, fairness, or the unique data constraints of African institutions. Barriers, including limited digital infrastructure, policy gaps, and ethical concerns over data use, further hinder implementation Nti et al., 2022; Patel and Ragolane, 2024. Locally adapted models are scarce, yet critically needed to address these institutional and contextual challenges Chibaya et al., 2022.

In addition to technical gaps, scholars debate the ethical implications of predictive analytics in education. Critics warn that early predictions, if poorly designed or applied, may stigmatize students or reinforce inequalities Memarian and Doleck, 2023. These global debates emphasize that predictive systems must strike a balance between accuracy, fairness, transparency, and contextual relevance, a principle that guides the present study.

Addressing these gaps, this research applies predictive analytics to the Ugandan higher education context. Using historical pre-admission data from Uganda Christian University (UCU), it develops a Random Forest model that predicts final CGPA with meaningful accuracy (R^2 $approx$ 0.24). Beyond prediction, the study integrates SHAP-based interpretability and sensitivity analysis to reveal key drivers of success and risk, ensuring that outputs are not only accurate but also actionable and ethically grounded. By embedding these insights at the point of admission, UCU can transition from generic program placement to data-informed strategies that support targeted advising and enhance student outcomes from the outset.

In summary, this background narrows from the global evolution of predictive analytics to its application in education, critically examining research gaps and ethical debates, and positioning this study as a locally adapted, transparent, and ethically responsible predictive model for proactive academic advising.

1.2 Problem Statement

Although universities worldwide have adopted predictive analytics to improve student outcomes, institutions in Sub-Saharan Africa, including Uganda Christian University (UCU), have yet to leverage these tools in their academic support systems fully. While these universities collect rich pre-admission data, such as A-Level grades, O-Level results, program placements, and demographics, this information remains underutilized in guiding students proactively.

Globally, machine learning models have demonstrated strong capabilities in forecasting academic outcomes, including GPA and graduation likelihood, enabling early interventions that improve student success Pelima et al., 2024. However, most existing models rely heavily on in-program indicators, such as LMS usage and coursework submissions, which are inconsis-

tently captured or absent in many African higher education institutions. Moreover, existing approaches rarely incorporate interpretability or fairness mechanisms, raising concerns about their applicability and ethical use in advising contexts Memarian and Doleck, 2023.

In Sub-Saharan Africa, adoption of predictive analytics remains limited, with only a small fraction of studies addressing the region Chibaya et al., 2022. Academic advising continues to follow a reactive model, where interventions occur only after academic failures or dropouts have already taken place, often at significant academic and psychological cost. The program placement process compounds this challenge, as it is primarily driven by subject combinations and perceptions of prestige rather than personalized guidance, leading to potential mismatches between student strengths and program demands.

This gap underscores the need for a locally adapted predictive model that uses pre-admission data to forecast long-term outcomes, supports fair and interpretable decision-making, and informs early interventions. This study addresses this problem by developing and deploying a Random Forest model trained on UCU's historical admission data. By integrating SHAP-based interpretability and fairness considerations, the research demonstrates how predictive analytics can be ethically and effectively implemented to transform academic advising from a reactive to a proactive process.

1.3 Objectives of the Study

1.3.1 Main Objective

To develop and evaluate an interpretable and fair machine learning model that predicts students' final CGPA using pre-admission academic and demographic data, thereby enabling proactive and data-informed academic guidance at Uganda Christian University.

1.3.2 Specific Objectives

To achieve the main objective, the study is guided by the following specific objectives:

1. To analyze pre-admission features (e.g., O-Level and A-Level performance, demographics, and program attributes) and determine their relative importance in predicting final CGPA.

2. To develop, optimize, and validate machine learning models (e.g., Random Forest) that can accurately predict final CGPA at the point of university admission.

3. To integrate interpretability and fairness evaluation (e.g., SHAP analysis and subgroup fairness assessments) into the predictive modeling process to ensure ethical and transparent use of predictions.

4. To assess the predictive model's effectiveness in supporting proactive academic advising, including the early identification of students who may require additional support.

5. To explore the model's potential to guide program placement decisions by simulating predicted CGPA across alternative programs, thereby enhancing student-program alignment.

1.4 Research Questions

The following research questions guide this study:

1. Which pre-admission academic and demographic features most strongly influence the prediction of final CGPA at Uganda Christian University?

2. How accurately can machine learning models predict final CGPA using only admission time data, and how do these models compare in terms of performance and generalizability?

3. How can model interpretability and fairness evaluations (e.g., SHAP analysis and subgroup fairness) enhance the ethical and actionable use of predictions for academic advising at UCU?

4. To what extent can the predictive model simulate outcomes across alternative programs, and how can these simulations support better student-program alignment?

5. What technical, ethical, and institutional factors influence the successful deployment and integration of predictive analytics into academic advising at UCU?

1.5 Justification of Study

Predictive analytics in higher education has gained considerable momentum globally, enabling institutions to make data-informed decisions that enhance student retention, improve advising, and optimize resource allocation. However, in Uganda and across Sub-Saharan Africa, many institutions, including Uganda Christian University (UCU) have yet to fully harness these

tools for academic planning, personalized advising, or program placement. This underutilization persists despite the availability of standardized pre-admission data, which remains largely untapped for predictive modeling.

Recent studies have demonstrated the effectiveness of machine learning models in forecasting academic outcomes such as GPA and graduation likelihood, thereby enabling proactive interventions Fahd et al., 2022; Pelima et al., 2024. Nevertheless, most models developed in Western contexts rely on in-program data that is inconsistently captured in African universities and rarely integrate fairness or interpretability mechanisms Maphosa and Maphosa, 2020; Patel and Ragolane, 2024. This limits their applicability in local settings and raises ethical concerns about transparency and bias in educational decision-making.

This study is justified by the need to bridge this gap through the development of a context-specific, interpretable, and fair predictive model trained exclusively on pre-admission academic and demographic data. By enabling early CGPA forecasting, the research supports proactive advising, early identification of at-risk students, and informed program selection, all from the point of admission. Beyond student-level support, the findings offer institutional benefits by informing academic policy, enhancing program alignment, and demonstrating a scalable framework for integrating predictive analytics in resource-constrained settings. Additionally, this work contributes to the global discourse on ethical and responsible AI in education, addressing both technical and fairness considerations while expanding research in an underrepresented African context.

1.6 Scope of the Study

This study focuses on students enrolled in degree programs at Uganda Christian University (UCU) whose records contain complete pre-admission information (A-Level and O-Level results, demographic details) and final CGPA outcomes.

The predictive models are developed exclusively using data available at the point of admission, including A-Level performance, O-Level metrics, demographic attributes (e.g., gender, age), and institutional program identifiers. The target variable is the final Cumulative Grade Point Average (CGPA), recorded upon program completion.

Features derived from in-university performance (e.g., Year 1 or Year 2 GPA) are intentionally excluded to demonstrate the predictive power of entry-level data alone. The study also incorporates analyses of interpretability (SHAP) and fairness to ensure that predictions are both ethical and transparent.

While the methodology is generalizable, the model is specifically tailored to UCU's academic structures and data environment. Applying it to other institutions would require contex-

tual adaptation. Program-level simulations and advising use cases are treated as exploratory applications rather than definitive placement recommendations.

1.7 Significance of the Study

This study holds significance at multiple levels:

- **For Academic Advisors and Students:** The research introduces an early-stage forecasting tool that integrates interpretability (via SHAP) and fairness assessments, allowing advisors to understand and trust predictions. This enables early identification of students who may require additional support, facilitating timely and targeted interventions.
- **For University Administration:** The study demonstrates how existing student records at UCU can be transformed into actionable insights that support proactive advising, evidence-based decision-making, and strategic program placement. It also provides a scalable framework that can be adapted for broader institutional analytics.
- **For Research and Innovation:** This work contributes to predictive learning analytics within the African higher education context, where such applications remain limited Karim-Abdallah et al., 2025; Nti et al., 2022. By incorporating fairness and interpretability, the study advances responsible AI in education and offers a reference model for other institutions seeking to adopt contextually relevant predictive systems.

Chapter 2

Literature Review

2.1 Introduction

In today's data-driven world, higher education institutions are increasingly turning to predictive analytics to enhance student outcomes, optimize academic planning, and support personalized advising. Globally, machine learning and educational data mining (EDM) techniques have been widely utilized to predict student academic outcomes, dropout likelihood, and long-term academic achievement. However, while these models have transformed student support systems in many parts of the world, their adoption in Sub-Saharan Africa remains limited.

This chapter reviews existing research on predictive analytics in education, with a particular emphasis on predicting cumulative grade point average (CGPA) using machine learning techniques. It begins with global perspectives on predictive modeling, then examines key academic and demographic factors that influence performance, and finally focuses on research gaps and opportunities within African higher education, particularly the case of Uganda Christian University (UCU). The chapter concludes by synthesizing the gaps in the literature and positioning this study to address them. The following section examines how predictive analytics has been applied globally, laying the groundwork for understanding its relevance and potential in the African context.

2.2 Predictive Analytics in Higher Education: Global Perspectives and Applications

2.2.1 Defining Predictive Learning Analytics

To forecast student outcomes, predictive learning analytics in education utilizes machine learning techniques, statistical modeling, and institutional data Namoun and Alshanqiti, 2020. Unlike prescriptive analytics, which suggests actions, or descriptive analytics, which summarizes past events, predictive analytics focuses on estimating future outcomes based on historical

trends.

For instance, an analytical model can forecast a student's cumulative grade point average (CGPA) by utilizing pre-admission records such as O-Level and A-Level grades in conjunction with demographic data. These early insights enable universities to provide targeted academic counseling, strategically allocate resources, and design interventions before the occurrence of academic failure or dropout.

In short, predictive analytics transforms educational decision-making from a reactive to a proactive approach, enabling academic institutions, as well as students, to leverage real-time, data-driven insights. This background provides the context for examining the application of these models worldwide and their customization for addressing diverse educational challenges.

2.2.2 Global Trends and Model Applications

Predictive modeling is now a foundation of student success plans globally, enabling institutions to enhance both administrative processes and educational outcomes. There has been a growing use of these models at academic institutions for predicting outcomes such as GPA, retention, and the probability of dropping out, along with tracking students' progress to provide timely interventions Fahd et al., 2022.

The meta-analyses indicate that predictive analytics is broadly applied to identify students' susceptibility to decline or educational dropout 2022. It has been possible to predict academic performance using a combination of support vector machines and decision trees, with ensemble learning techniques enhancing predictive accuracy. It has been possible for models that incorporate early test scores along with attendance data to be highly effective in helping institutions provide timely academic support.

Beyond dropout prediction, predictive models have broader applications, including the development of individualized study plans, assessment of student performance, and detection of behavioral patterns linked to mental health risks Zhang et al., 2021. These capabilities demonstrate how predictive analytics serves as both a diagnostic tool and a strategic planning instrument for higher education.

However, research increasingly highlights that accuracy alone is insufficient; transparency and ethical use of predictions are becoming critical considerations Memarian and Doleck, 2023. While some studies advocate for explainable models, others prioritize accuracy at the expense of interpretability, creating tension in the adoption of models. This debate has led to the integration of explainable AI (XAI) techniques, such as SHAP, into modern predictive frameworks, thereby enhancing predictive power and providing trustworthy insights.

Although the United States and Asian markets lead in early implementation, researchers

are now exploring ways to deploy these techniques in Sub-Saharan Africa, where significant contextual and infrastructural challenges remain. Building upon these global trends, the subsequent section delves into the specific machine learning techniques that underpin predictive analytics, elucidating their strengths and limitations..

2.2.3 Machine Learning and EDM Techniques

The effectiveness of predictive analytics in the educational sector is contingent upon not only the quality of data available but also the prudent selection of appropriate machine learning methodologies. Each of these techniques possesses distinct strengths and inherent trade-offs Hashim et al., 2020.

Among the most explored techniques are Random Forest (RF), Support Vector Machines (SVM), Logistic Regression, and Artificial Neural Networks (ANN). Random Forest is typically characterized by high prediction power with an appropriate amount of interpretability, enabling it to thrive in mixed data type academic information systems with incomplete data. SVMs excel in high-dimensional spaces and are frequently applied for students' classification based on risk of educational difficulties, but provide little in terms of interpretability. Logistic regression remains a strong baseline technique due to its interpretability and explanatory simplicity for academic stakeholders. Deep learning techniques with ANN were up-and-coming in behavioral and time-based data modeling but criticized for providing a "black box" with limited acceptability in advisory and policy-making scenarios where transparency is highly significant Nur, 2021.

Feature selection is equally crucial in determining a model's success. It involves applying the most relevant predictors to reduce noise and minimize computational complexity. Techniques such as LASSO regression, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA) have been proven to enhance prediction accuracy. Research has confirmed that features such as O-Level average grade, gender, and programme of study are significant continuous predictors of CGPA Ismanto et al., 2022.

Although advanced machine learning techniques and feature engineering continue to enhance prediction accuracy, they simultaneously raise concerns regarding the optimal balance between performance and interpretability. This tension is especially significant in education, where decision-makers must trust and understand the reasoning behind predictions.

2.2.4 Outcome Focus: CGPA vs. Retention vs. Dropout

Much of the work on predictive learning analytics focuses on outcomes such as dropout and student retention, with less attention given to cumulative grade point average (CGPA), which is the main target of prediction. For dropout prediction models, such as those created by

Sosa-Alonso et al. (2025) and Fahd et al. (2022), these have been informative for a type of risk management that anticipates students who are at high risk of dropping out of the system. These models do depend on a binary outcome; however, this does not reflect the diversity of academic paths for students who stay in the system but ultimately complete their studies.

In contrast, CGPA provides a fine-grained and continuous measure of academic performance throughout the study. This argument holds significant importance as it facilitates institutions not only in distinguishing between those who stay and those who leave but also in categorizing students with diverse academic levels. As a continuous outcome measure, CGPA enables the use of regression algorithms, whose outputs provide more accurate predictions compared to traditional classification methods. These predictions can be used for individualized intervention, goal-oriented conversations, and data-driven advising..

Furthermore, CGPA serves multiple institutional purposes beyond risk detection. It informs scholarship decisions, award considerations, and postgraduate admission recommendations, making it a valuable metric for long-term academic planning and decision-making. For Uganda Christian University (UCU), with well-defined academic timelines and grading structures, forecasting CGPA aligns with institutional needs and facilitates early intervention strategies. By predicting CGPA at admission, the university can proactively advise and make informed academic decisions..

2.3 Factors Influencing Student Academic Performance

2.3.1 Academic Entry Indicators

University students' academic achievements depend mainly on their educational backgrounds, especially their scores in external tests like O-Level and A-Level exams. Universities apply these pre-entry measures to set expectations and assess students' readiness for challenging academic programs.

O-Level and A-Level scores, such as total points earned, number of credits gained, and A-Level distinctions for top subjects, indicate the breadth and depth of students' high school learning. The latest empirical studies verify that they are potent predictors of CGPA under various educational settings Ismanto et al., 2022; Zhang et al., 2021.

Apart from average scores, the consistency of performance also comes into play. The standard deviation of O-Level performance provides a measure of a student's academic reliability. Students with more stable patterns of performance cope better with the demands of higher education, where self-study and time management play a more critical role.

Therefore, high school performance is more than just a threshold for admission; it reflects

enduring academic traits that influence students' university success.

2.3.2 Demographic and Program-Related Variables

Although academic history-based entry measures demonstrate a high degree of reliability in predicting university performance, institutional and demographic factors also significantly contribute to academic success in various streams. Gender is a factor of concern in most prediction models, with some studies reporting subtle patterns of performance within diverse programs or majors Fahd et al., 2022; Nur, 2021. These patterns prove significantly subject-specific and must be interpreted with great caution lest they reinforce stereotypes or biases.

Institutional characteristics, including campus location and curriculum structure, also play a significant role.. Curriculum descriptors, in particular, signal heterogeneity in course sequences and teaching philosophies, with potentially important implications for later academic attainment. Highly intensive initial work schedules or stringent patterns of prerequisite course-taking can inadvertently disadvantage students who require more flexible learning environments.

Although tuition fees were initially considered a potential feature, they were excluded from this study for both conceptual and practical reasons. At UCU, tuition is largely standardized within program clusters, offering slight variation to distinguish students' financial circumstances or academic potential. Including this variable could have introduced noise without providing meaningful predictive value.

For these reasons, this study focuses on features with clearer causal or correlational relationships to student achievement, ensuring that the model remains both interpretable and relevant to the institutional context.

2.3.3 Program and Curriculum Design Impact

Program and curriculum structures extend beyond administrative categorization; they actively shape the learning environment and, consequently, student academic performance. For example, programs with heavy course loads during the early semesters may place students under considerable academic stress, which can negatively affect their long-term GPA.

Similarly, curricula with rigid prerequisite chains can unintentionally disadvantage students who struggle with introductory courses, creating compounding challenges throughout their academic journey. Because transcripts typically do not capture these structural nuances, researchers often rely on program and curriculum data to reconstruct and analyze their effects.

Institutional studies support the idea that well-aligned curricula promote student success by easing progression and facilitating learning Kamal et al., 2024; Mpofu and Chasokela, 2025. At UCU, program design follows a traditional disciplinary approach, common in many

African higher education institutions. While this structure has merits, it limits flexibility in adapting to diverse learning needs. It leaves little room for flexible learning pathways that could accommodate diverse student needs.

Predictive models enhance their accuracy when curriculum design components are integrated, presenting opportunities for educational transformation. Ultimately, these insights can contribute to creating a more supportive academic environment by informing student advising programs and institutional strategies. Although these elements have been extensively researched in global settings, African universities present distinct dynamics that warrant particular attention, as discussed in the following section.

2.4 Institutional Analytics in African Higher Education: Opportunities and Challenges

2.4.1 Emerging Analytics Culture in HEIs

Data analytics is being progressively adopted by higher education institutions (HEIs) throughout Africa as a tool for strategic management and organizational development. This shift, though still in its infancy, demonstrates a growing understanding of how data-driven decision-making can enhance academic systems and institutional performance.

A comparative survey of Zimbabwean universities conducted by Mpofu and Chasokela (2025) highlights the fact that analytics is being increasingly integrated into strategic planning procedures, which in turn impact decisions related to innovation, teaching, and academic progress. This indicates a significant shift in cultural norms towards policies grounded in empirical evidence rather than relying solely on intuition.

Although formal use of analytics is currently limited, the potential is considerable. Through the integration of data use into regular operations, African universities can better harmonize institutional strategy with educational goals and national development aspirations more broadly. This new culture provides a foundation for future growth of analytics across the region's higher education context.

2.4.2 Infrastructure and Policy Challenges

The limited adoption of predictive analytics in African higher education institutions is primarily attributed to deficiencies in infrastructure and policy frameworks. As noted by Patel and Ragolane (2024), the absence of robust data governance structures and dedicated analytics

infrastructure remains a significant barrier to implementing AI-driven systems in universities, including those in South Africa.

Limitations in technology are another dimension of the problem. Budgetary restrictions often limit investments in infrastructure for information technology and personnel training, and data initiative leadership is commonly dispersed at various levels within organizations. Without consolidated management, even willing institutions cannot move beyond pilot applications.

To implement analytics efficiently, universities should develop internal capacity through creation of basic systems, harmonization of data protocols, and technical skills-building. It's only through these coordinated activities that institutions can effectively overcome current barriers and leverage the use of predictive analytics as a sustainable means for both academic and strategic gain.

2.4.3 Equity, AI, and Digital Divide

The inclusion of analytics and AI in African universities presents significant equity and inclusivity challenges, as well as posing ethical issues regarding the fair distribution of learning opportunities. While AI applications can develop a more personalized learning environment and automate management tasks, these can perpetuate existing inequality in access and opportunity if not applied with caution.

Ajani et al. (2024) highlights the need for bridging digital divides related to internet access, device ownership, and digital literacy. Failing to address these gaps can result in students from underprivileged backgrounds being excluded from reaping the benefits of predictive technologies. Additionally, Kunjumuhammed (2024) cautions that algorithms trained on biased or imperfect datasets can exacerbate educational disparities instead of mitigating them.

Here, the ethical application of analytics is not a purely technical issue but an issue of social equity. Universities should integrate fairness, transparency, and accessibility into the design and implementation of analytical systems. Otherwise, predictive analytics can end up even more sharply exacerbating learning gaps than closing them, betraying its intended aim.

2.5 Applications of Predictive Models in Academic Support

2.5.1 Proactive Academic Advising

One of the largest university applications of predictive analytics is placing academic advising on a proactive rather than reactive footing. For the most part, conventional universities interfere

only after students begin performing poorly, often at a point where avenues for reversing adverse outcomes are narrow. Imposing predictive models changes this dynamic by identifying at-risk students very early on, at a point at which decline in performance is still not yet discernible.

When integrated into advising workflows, these models provide timely, data-driven insights that enable targeted interventions such as individualized counseling, study skills workshops, and early academic support programs. Importantly, predictive analytics does not replace human judgment but complements it, equipping advisors with an additional evidence-based tool to enhance the quality and precision of their guidance.

Predictive models are particularly valuable in environments with constrained resources, where high advisor-to-student ratios hinder the provision of personalized support. By empowering institutions to optimize resource allocation, these models ensure that support is directed toward students who are most vulnerable at the critical juncture.

2.5.2 Program Recommendation Systems

Besides identifying vulnerable students, predictive modeling is being increasingly used to make informed academic choices, such as program selection and admission. Most students begin their university studies with little guidance on which programs best align with their talents, interests, and future academic goals. This disparity is more pronounced when decisions are made based on A-Level numbers or social views of a program's prestige, rather than data-informed judgments.

Program recommendation systems seek to do this through the use of past data to discern patterns of student success and recommend incoming students for programs where they will have the greatest likelihood of success. Those recommenders that combine content-based with collaborative filtering approaches show great promise for customized educational planning Kamal et al., 2024.

Evidence from the platform described by SIMION et al. (n.d.) illustrates how digital tools can assist prospective students in exploring program options through interactive testing and feedback features. These systems transcend their role as mere decision aids, adopting a strategic approach to academic and career selection. They prioritize long-term success over fleeting short-term preferences.

2.5.3 Integration with MIS and Institutional Workflow

Uganda Christian University (UCU) has an efficient Management Information System (MIS) that keeps pre-admission data, including A-Level and O-Level scores, demographic information,

and program placements. This current infrastructure is a perfect platform for incorporating predictive analytics within institutional processes. Through the MIS, the university can infuse predictive modeling functionality without having a separate platform.

A new CGPA prediction model can be deployed as an add-in within the existing MIS. Through a user-friendly interface, academic advisers or students can provide essential admission features, such as secondary school grades, gender, and desired program of study. The system will then produce a CGPA prediction, along with implications for action. Incorporating this capability within a familiar and safe digital environment increases accessibility, fosters adoption, and provides compliance with institutional data governance and privacy policies.

Furthermore, integration with the MIS enables iterative refinement of the model over time. As future cohorts progress through their programs, actual CGPA outcomes can be compared with expected outcomes for continuous validation and recalibration of the model. This closed-loop feedback not only increases predictive accuracy over time but also enables institutional learning in the long run and evidence-informed improvement. However, the adoption of these models raises significant questions about fairness, privacy, and the ethical use of these models, which must be taken into account for responsible adoption.

2.6 Ethical Considerations and Fair Use

2.6.1 Privacy and Student Identity Protection

Predictive analytics in education relies on sensitive student data, making privacy and the ethical handling of this data central concerns. The use of academic records, demographic profiles, and behavioral information carries the risk of unintended exposure, particularly when data remains linked to identifiable individuals. To mitigate these risks, best practices emphasize the use of strict anonymization protocols, ensuring that identifiers such as student names, registration numbers, or email addresses are removed or masked before analysis.

At Uganda Christian University (UCU), our study adheres to these guidelines by pseudonymizing student records during preprocessing and implementing role-based access to the prediction tool. These types of measures enable us to keep sensitive data accessible for only authorized personnel at any stage of the modeling process.

Yet, anonymization alone is no protection for privacy if access controls and governance structures are not sufficiently specified. Data science systems at scale must therefore integrate protections like these, which limit who can see, edit, or take action on outputs of models, but especially where these outputs impact high-stakes decisions like course selection or academic counseling. Above all, the predictive capability of any model must never take precedence over

a student's right of confidentiality and informed consent.

Consequently, responsible data management transcends technological considerations and encompasses an institutional imperative of trust and moral accountability. Analytics should be employed to enhance the well-being of students while safeguarding their safety.

2.6.2 FATE Framework and Responsible AI

The FATE framework, which encompasses fairness, accountability, transparency, and ethics, has become a standard measuring stick for evaluating AI systems in learning contexts. While with predictive analytics efficiency can be enhanced and individualization increased, the majority of models turn out to be black box systems whose lack of transparency and fairness raise concern Memarian and Doleck, 2023. Such systems can inadvertently perpetuate historical inequities, for example, by highlighting specific demographic groups as at risk, as these groups have historically been institutionally under-resourced.

Fairness assessment, thus, is more than technical validation for imbalances between false positives or false negatives. It necessitates that institutions critically examine their practices with essential questions: Whom do these predictions benefit? And who can be harmed or excluded? This step highlights educator and stakeholder engagement throughout the development and scrutiny of these models.

Walker et al. (2024) also warns against prediction solely based on pre-entry data becoming too deterministic. Although statistical legitimacy can be achieved with regression models, they often overlook qualitative elements that play a significant role in academic performance. It therefore underlines the necessity for transparency not just in the engineering of models but also in the reporting of their outputs. Predictions must be issued as probabilities rather than certainties to enable their use as advice by advisors and students, rather than as instructions.

This moral dimension highlights the value of context-sensitive methods, which are illustrated in the subsequent section using the case of Uganda Christian University.

2.7 The Ugandan Context and Case for UCU

After detailing global practices and ethical considerations, this section concentrates on Uganda Christian University, demonstrating how these global insights intersect with local circumstances.

2.7.1 Local Data Availability and Opportunity

Ugandan universities have made considerable investments over the past twenty years in building student information systems and electronic record-keeping infrastructure. Among these universities, Uganda Christian University (UCU) stands out in terms of its relatively sophisticated data management capacity. It maintains comprehensive records that reflect admissions data, O-Level and A-Level credentials, student enrollments in academic offerings at the program level, as well as end-of-semester CGPA performance. This high-quality dataset, spanning multiple academic years and featuring highly varied attributes, provides an informative platform for formulating predictive models based on real, context-specific data.

On the other hand, most higher education institutions in Sub-Saharan Africa still wrestle with scattered and non-standardized data systems, thereby constraining the effectiveness of implementing predictive analytics. UCU's well-organized and centralized data infrastructure not only enables model building but also enhances the credibility and verifiability of analytical results.

Through this data infrastructure, UCU can uniquely leverage predictive analytics for proactive academic interventions. Access to clean historical records enables the university to develop instruments that provide students with individualized counseling from the point of entry into their program, promoting better performance and retention rates.

2.7.2 Bridging Program Selection Gaps

A persistent challenge in Uganda's higher education sector is the absence of structured academic guidance during the university admission process. In most cases, students are placed into programs based solely on their A-Level subject combinations, with little consideration for how these choices align with their strengths, interests, or long-term academic potential. This absence of data-informed guidance often results in a misalignment between students and their chosen fields of study, leading to performance declines, disengagement, and, in some cases, dropout.

At UCU, anecdotal evidence and internal audits reveal that a significant proportion of students either transfer programs within their first year or struggle to meet minimum progression requirements. Yet by the end of A-Level, universities already possess sufficient information, such as academic history, subject weightings, and demographic indicators, to make early and informed assessments of likely academic trajectories.

Integrating predictive tools at the point of admission would provide students with greater clarity and support, reducing avoidable academic detours. Such tools could ensure better alignment between students and programs, fostering improved performance and retention from

the outset of their university journey.

2.7.3 Institutional Relevance and Innovation Potential

This research aligns directly with UCU's strategic priorities, particularly its commitment to student success, data-driven decision-making, and digital transformation. By leveraging its existing academic data assets, UCU has the opportunity not only to enhance internal academic advising but also to establish itself as a regional leader in educational innovation. The development of a CGPA prediction model represents more than a technical exercise—it is an example of how local institutions can harness their data to drive meaningful, context-specific change.

Furthermore, the project fits seamlessly into UCU's ongoing efforts to modernize its Management Information System (MIS). The predictive model can be piloted as an integrated feature within this system, offering a practical and scalable solution for early academic guidance. By embedding machine learning into existing workflows, through tools such as a form-based prediction interface for advisors and administrators, UCU demonstrates how advanced analytics can be operationalized without disrupting established processes.

As more African higher education institutions advance in their digital transformation journeys, UCU's approach could serve as a replicable model for implementing predictive analytics responsibly and effectively. This not only supports institutional goals but also contributes to the broader discourse on educational innovation across the region.

2.8 Synthesis and Research Gaps

2.8.1 Gaps in the Global Literature

The current body of literature on predictive analytics in education reveals two significant imbalances: one in outcome focus and another in geographical representation. In contrast, an increasing number of studies explore student retention and dropout risks, particularly in Western and Asian contexts; however, far fewer focus on cumulative academic achievement, as measured by the final cumulative GPA (CGPA). This gap is significant, given that CGPA serves as a critical metric for graduation classification, postgraduate admissions, scholarship decisions, and long-term academic benchmarking. Existing studies that mention CGPA often treat it as a secondary outcome, making it challenging to extract clear insights into how early data can predict long-term academic success.

Equally important is the underrepresentation of Sub-Saharan Africa in educational data

mining research. Prior reviews highlight that the region contributes minimally to the global body of work on learning analytics. Most predictive models are developed using datasets from well-resourced institutions in North America, Europe, or East Asia, which differ substantially from those in African higher education contexts. These imported models often reflect different grading schemes, student profiles, and institutional structures, making them poorly aligned with local realities when adopted without adaptation.

This creates a critical gap: the need to develop and validate predictive frameworks using African student data within institutional contexts that accurately reflect the educational environment of the region. Addressing this gap is essential to ensure that predictive analytics not only achieves technical accuracy but also offers practical relevance for African higher education systems.

2.8.2 Positioning This Research

This study directly addresses the identified gaps by focusing on final CGPA rather than dropout, as the primary prediction target, within the context of an African higher education institution. Unlike many existing studies that rely on LMS clickstream data or semester-by-semester GPA trends, this research adopts a pre-admission approach, using historical academic records available at the point of university entry to forecast long-term educational outcomes. This strategy enables early interventions and supports informed decision-making at a stage where guidance is most needed yet often least available.

Grounded in data from Uganda Christian University (UCU), the study develops a context-aware predictive framework that aligns with both institutional and regional priorities. It demonstrates that African higher education institutions, even with existing infrastructural constraints, possess sufficient data to build predictive systems that are practical, interpretable, and impactful.

Beyond prediction, the research extends the role of institutional analytics by proposing ways in which predictive insights can inform program advising, resource allocation, and academic support workflows. In doing so, it advances a localized, proactive, and ethically grounded approach to student success, one that not only benefits UCU but also enriches the global discourse on educational data science by providing evidence from an underrepresented context.

2.9 Conclusion

This chapter has critically explored the evolving field of predictive analytics in higher education, with a particular focus on CGPA prediction. It reviewed how institutions worldwide

are leveraging machine learning and educational data mining techniques to enhance academic planning, support retention efforts, and monitor student performance. While much of the existing research prioritizes dropout and retention as primary outcomes, this study positions final CGPA as a more comprehensive and actionable metric, given its relevance to academic advising, program alignment, and long-term student success.

The review examined key factors influencing academic performance, highlighting the predictive value of academic entry indicators such as O-Level and A-Level results, alongside institutional variables including curriculum design and program structure. It also addressed the African context, where predictive analytics remains underutilized despite growing interest and clear potential. The literature emphasizes the need for context-specific models that not only respect local constraints but also leverage available data to provide meaningful insights.

Furthermore, the chapter emphasized the practical applications of predictive models, particularly in enabling proactive academic advising and supporting program selection through tools like recommender systems and MIS-integrated interfaces. Ethical considerations emerged as a critical theme, with the FATE framework encompassing fairness, accountability, transparency, and ethics, highlighting the importance of deploying predictive systems responsibly to avoid reinforcing existing inequities.

Overall, the literature reveals a clear gap in the development of CGPA-specific predictive models within Sub-Saharan Africa, particularly those that utilize pre-admission data to support early interventions. This study aims to contribute to research and practice by addressing a gap. The methodological framework for developing and validating a CGPA prediction model at Uganda Christian University is presented, detailing data preparation, model design, and evaluation strategies. The study outlines its methodological approach, ensuring the model's technical robustness and contextual relevance.

Chapter 3

Methodology

3.1 Introduction

Predictive analytics has gained rapid adoption in decision-making for higher education institutions in a remarkably short period, transforming institutional practices from reactive measures to proactive interventions. For Uganda Christian University (UCU), this revolution presents a singular opportunity to utilize its vast repository of pre-admission data for predicting long-term student outcomes, specifically the end-of-program Cumulative Grade Point Average (CGPA). CGPA prediction at entry significantly enables a university to offer early targeted interventions and refine academic advisement strategies, with a direct result of improving student success, as well as informing institutional planning.

This chapter outlines the methodological foundation underlying the development of the CGPA prediction model. The research is grounded in a quantitative approach, utilizing supervised machine learning techniques to model patterns within historical student records. This approach was selected because it allows the integration of diverse data sources, ranging from O-Level and A-Level academic histories to demographic and institutional placement factors, into a predictive framework that offers both accuracy and interpretability.

The chapter unfolds systematically to guide the reader through the logic and rigor of the study's design. It begins by detailing the research approach and contextualizing the study within the institutional environment of UCU. It then describes the collection and ethical handling of data, highlighting how historical records were refined to ensure quality, completeness, and compliance with governance standards. The subsequent sections elaborate on the preprocessing pipeline, feature engineering, and feature selection strategies employed to prepare the dataset for modeling. The methodology further outlines the modeling pipeline, including candidate algorithms and theoretical justifications for their selection.

By embedding ethical considerations throughout the data handling process and maintaining transparency in the modeling steps, this chapter ensures that the research adheres to both scientific and institutional standards. Collectively, these methodological choices establish a robust foundation for the results presented in Chapter 4, where the effectiveness and implications of the developed model are analyzed.

3.2 Research Design

3.2.1 Research Approach

This study employs a quantitative research approach, underpinned by supervised machine learning techniques, which are chosen for their ability to uncover complex patterns within large-scale educational data. The predictive task is framed as a regression problem, with the final Cumulative Grade Point Average (CGPA) serving as the target variable. Unlike classification, which would simplify outcomes into discrete categories such as "pass" or "fail," regression provides continuous predictions, capturing subtle variations in academic performance. This level of granularity is crucial in educational contexts where fine distinctions in CGPA can influence scholarships, program suitability, and postgraduate opportunities.

Supervised learning was selected because of the availability of labeled historical data, a rich dataset where pre-admission predictors (e.g., O-Level and A-Level grades, demographic attributes, institutional placement) are paired with actual CGPA outcomes. This pairing enables the model to learn relationships that not only explain but also forecast academic success. The choice aligns with the research objective: to empower UCU with a predictive framework that can inform early interventions and individualized advising at the very start of a student's academic journey.

Regression modeling further enhances institutional decision-making beyond mere risk identification. It enables advisors to interpret predictions as probabilities and ranges, rather than fixed labels. This methodological approach ensures that predictions remain both practical and ethically sound, functioning as tools for support rather than exclusionary mechanisms.

3.2.2 Study Context and Scope

The study is centered at Uganda Christian University (UCU), a pioneering private institution noted for its academic quality and digital innovation. Contrary to most higher education institutions in Sub-Saharan Africa, UCU has made significant investments in cultivating a sophisticated Management Information System (MIS), which systematically captures students' admission, academic progressions, and graduation rates. This well-organized database forms a perfect precursor for data-informed scholarship, presenting a unique opportunity for modeling predictions on real, context-specific data.

The dataset spans admitted intakes from 2009 to 2023, covering 15 years, and includes evolving patterns of admission, diversity within the program, and changes in demographics. This chronological breadth enhances the model's generalizability across diverse academic settings while maintaining institutional relevance.

For methodological integrity, students with complete pre-admission characteristics, including O-Level and A-Level achievements, demographic details, and program allocation and terminal CGPA data, were considered. This selection criterion ensures that the model is trained on reliable data, where the predictors are fully observed at the initial point. By deliberately emphasizing features at or before admission, the study achieves its prime mission: providing early intervention. CGPA prediction at this early juncture provides planners with tangible information far in advance of academic problems arising, enabling individualized advice and program matching at the inception of a student's university career.

In this context and focus, UCU not only provides a case study but also a proof of concept for how African institutions can utilize available data for informed academic innovation, regardless of infrastructural challenges.

3.3 Data Collection and Ethical Handling

3.3.1 Source and Nature of the Data

This study drew its dataset from the Management Information System (MIS) of Uganda Christian University (UCU), a centralized platform that has reliably archived student records for more than twenty years. The MIS holds detailed data on admissions, academic progress, and graduation statistics, offering a rich foundation for educational analytics.

The initial data pull produced around 95,066 student profiles from cohorts admitted between 2000 and 2023. To refine this pool, records missing key pre-admission details or final CGPA scores were excluded, as such omissions could weaken the model's predictive accuracy. After this cleaning phase, the final dataset comprised 4,979 distinct student entries, all of which were admitted between 2009 and 2023, a timeframe during which the MIS maintained optimal data quality and consistency.

Each student record featured 41 structured variables, covering demographic data, academic results (both O-Level and A-Level), and institutional placement details. These features were intentionally limited to information available at the time of entry into the university, aligning the dataset with practical decision-making use cases, such as academic advising and early intervention. The dataset's structure—primarily numeric and well-organized—further made it ideal for supervised machine learning, reducing the need for extensive preprocessing and enhancing model clarity.

Throughout the data preparation process, ethical considerations were carefully observed and adhered to. Personally identifiable information was stripped out during extraction, and all handling complied with UCU's data governance standards. This rigorous yet responsible

approach ensured that the final dataset was both analytically sound and ethically managed, forming a strong basis for building a reliable CGPA prediction model.

3.3.2 Collected Attributes

The refined dataset was organized into four primary attribute categories, each offering insight into a distinct aspect of the student profile.

First, the demographic attributes covered basic personal information, including age at the time of admission, gender, marital status, nationality, and whether the student was classified as an international student. These variables helped establish context around how individual background factors might shape academic outcomes. For example, distinguishing between age groups enables the model to contrast traditional students with those entering university later in life. At the same time, nationality provides a lens for comparing the experiences of domestic and international students.

Next, the academic background section captured detailed records from students' O-Level and A-Level performance, critical components of Uganda's university admissions framework. These records included raw subject grades, counts of credits and distinctions, total subjects taken, and overall aggregates. Notably, performance in the General Paper was also included, offering a proxy for skills in critical thinking and communication. Together, these metrics reflect both the scope and rigor of each student's academic foundation.

The third category centered on institutional placement, documenting how each student first entered the university system. Attributes in this group included program code, curriculum ID, level of study, campus, session, and year of admission. These details were essential in tracing academic trajectories and accounting for differences across programs and institutional structures.

A financial variable was also included in the initial data: the tuition fee at the time of registration. While this was considered during early exploratory phases, it was ultimately dropped from the final model due to its limited predictive power and the risk of introducing socioeconomic bias.

Lastly, the dependent variable, the final cumulative grade point average (CGPA), was the focal outcome of the study. Expressed as a continuous score between 1.0 and 5.0, the CGPA offered a comprehensive summary of academic performance across the entirety of a student's time at the university.

3.3.3 Data Cleaning, Storage, and Ethical Handling

The transformation of raw institutional records into a dependable dataset necessitated meticulous data cleansing and restructuring. The initial data extracted from UCU's MIS encompassed identifiers, sparse fields, and inconsistencies that could jeopardize modeling efficacy. To mitigate these challenges, a structured cleansing pipeline was implemented employing a custom-developed backend service, thereby ensuring both methodological rigor and adherence to ethical guidelines.

The process commenced by filtering records to retain only those that possessed complete pre-admission attributes and final CGPA outcomes. This eliminated incomplete or partially recorded cases that could potentially distort model learning. Unique identifiers such as student IDs and access numbers (`id`, `access_no`), which were previously anonymised, were removed at the outset to prevent any risk of re-identification, ensuring that the dataset remained pseudonymized throughout subsequent analysis. Features with excessive missingness, such as fields capturing non-essential personal details or rarely populated attributes, were excluded from the final dataset due to their low predictive reliability.

Categorical variables, including gender, marital status, campus, and program codes, were systematically encoded to preserve their meaning while making them suitable for machine learning algorithms. Binary and ordinal features retained their inherent numeric representations, while nominal fields were label-encoded to optimize performance in tree-based models. Numerical variables underwent scaling to standardize their ranges, and missing values were treated using targeted imputation strategies, with the median used for numeric fields and the mode for categorical variables, ensuring that no samples were lost unnecessarily.

Outliers in both tuition and `age_at_entry` were not simply removed; instead, they were thoughtfully addressed. To minimize their disproportionate impact while preserving the data's interpretability, a logarithmic transformation was applied to tuition values, and age was grouped into ordinal bins.

Ethical considerations remained central throughout the data preparation process. Access to the raw data was tightly controlled under a signed Non-Disclosure Agreement (NDA), with every step aligning with Uganda Christian University's official data governance protocols. These safeguards ensured that even as the dataset was refined for analysis, student privacy and institutional integrity were never compromised.

3.4 Data Preprocessing and Preparation

Transforming the preprocessed dataset into a machine learning-ready format necessitated a meticulously structured preprocessing pipeline. This stage served as the crucial link between raw institutional data and the modeling workflow, making sure that all features used in the algorithm were not only meaningful and interpretable but also ethically sound. The pipeline addressed a range of issues, including handling missing data and outliers, as well as resolving encoding mismatches and scaling inconsistencies. Along the way, new features were engineered to enhance the model's predictive strength without compromising data integrity.

Handling missing values. Given the longitudinal nature of the dataset, missing data were inevitable in several fields. Rather than discarding incomplete records, which would have reduced the dataset size and potentially biased the model, strategic imputation techniques were applied. For numerical features, median imputation was chosen to mitigate the influence of outliers and preserve distributional characteristics. Categorical features were imputed with their most frequent category, aligning with institutional norms while maintaining dataset integrity. Features exhibiting more than 90% missingness, for example `children` and `subject_combination`, were removed entirely, as their inclusion would contribute noise rather than value.

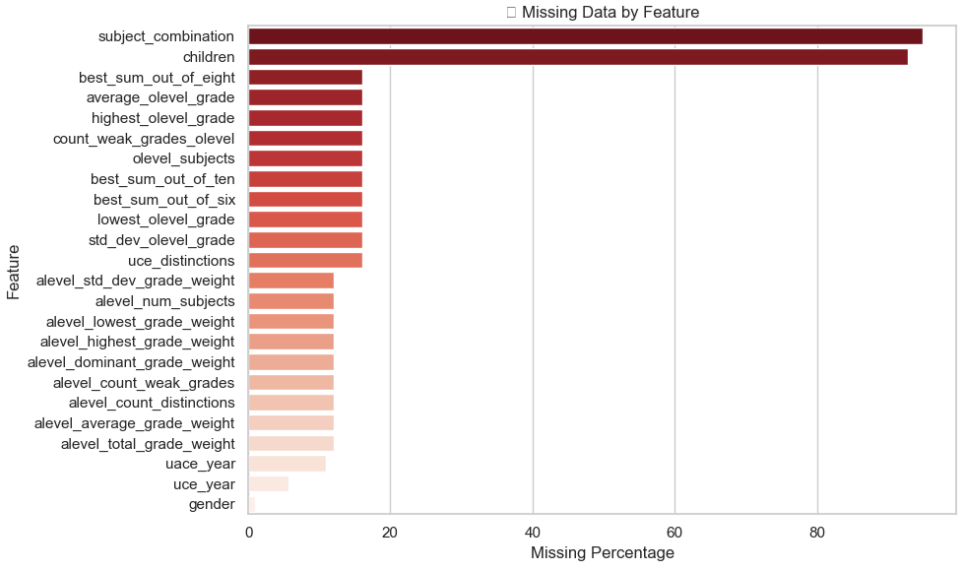

Figure 3.1: Proportion of missing data across features.

Outlier detection and treatment. Outliers were systematically analyzed using the Interquartile Range (IQR) method to detect extreme observations across key numerical features. This analysis revealed that variables such as `tuition` and `age_at_entry` exhibited

pronounced right-skewness, with several extreme values beyond typical ranges. Rather than discarding these observations, which could have removed meaningful cases, such as mature entrants or students enrolled in higher-fee programs, the preprocessing strategy opted for transformations that retained their informational value. Specifically, tuition was log-transformed to normalize its distribution while preserving relative differences, and age was categorized into ordinal bands: young entrants ($\leq$22), typical age (23–29), and mature entrants ($\geq$30).

The distribution of detected outliers for selected features is summarized in Table 3.1, which guided the choice of applied transformations and ensured that these preprocessing decisions were data-driven and transparent.

Table 3.1: Outlier Count by Feature using IQR Method

Feature	Outlier Count
age_at_entry	691
uce_distinctions	299
final_cgpa	101
alevel_average_grade_weight	37
tuition	36
average_olevel_grade	12
uce_credits	0

Categorical encoding. The dataset included a mix of binary, ordinal, and nominal categorical variables. Binary features (e.g., gender, nationality) retained their original encoding, ensuring straightforward interpretability. Ordinal variables, such as study levels, were numerically encoded to preserve the inherent order. Nominal features with no inherent rank (e.g., campus codes, program IDs) were label-encoded, providing a memory-efficient and model-friendly representation without unnecessarily inflating the feature space.

Feature scaling. Standardization was applied to all continuous numeric variables using the z-score transformation, centering features at zero with unit variance. This step was crucial for ensuring that features with different scales, such as grade averages versus tuition values, contributed proportionately to model training, particularly when algorithms sensitive to scale were evaluated.

Derived features. Beyond cleaning and transformation, new features were engineered to capture patterns in student academic profiles better. For example, a *performance stability index* was calculated from the variance in high school grades, providing an indicator of consistency across examinations. Similarly, aggregated metrics, such as the best O-Level sums, were computed to summarize complex grade distributions into compact, predictive signals. These derived attributes enriched the dataset with features grounded in both data science and domain expertise.

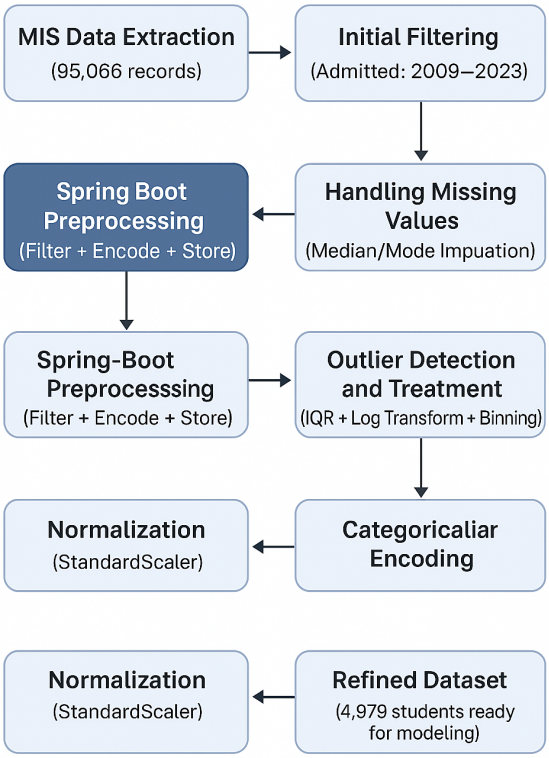

Figure 3.2: Data preprocessing pipeline: from raw MIS extraction to refined feature matrix.

The resulting dataset comprising ethically handled, imputed, scaled, and feature-engineered variables was both analytically robust and institutionally responsible. This structured preprocessing ensured that the subsequent modeling phase relied on a dataset optimized not only for predictive accuracy but also for fairness and reproducibility, laying a strong foundation for the machine learning framework presented in the following sections.

The effectiveness of these preprocessing steps is illustrated in Figure 3.3, which shows the distributions of key numeric features after transformation and scaling, confirming that the data were well-conditioned for subsequent modeling.

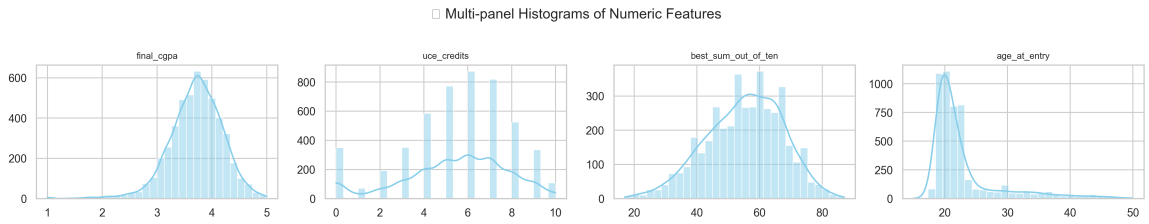

Figure 3.3: Distribution of selected numeric features after preprocessing.

A summary of how features were classified and treated during preprocessing is presented in Table 3.2.

Table 3.2: Feature Classification Summary

Category	Features	Details
Numeric	admission_rank, age_at_entry, tuition, final_cgpa, average_olevel_grade, alevel_average_grade_weight, alevel_total_grade_weight, alevel_highest_grade_weight, alevel_lowest_grade_weight, alevel_std_dev_grade_weight, alevel_num_subjects, best_sum_out_of_six, best_sum_out_of_eight, best_sum_out_of_ten, uce_credits, uce_distinctions, count_weak_grades_olevel, olevel_subjects, high_school_performance_variance, high_school_performance_stability_index, std_dev_olevel_grade, highest_olevel_grade, lowest_olevel_grade	Scaled with `StandardScaler`, imputed using median
Categorical	gender, marital_status, is_national, general_paper, campus_id, program_id, curriculum_id, level, uce_year, uace_year, year_of_entry	Label-encoded or ordinal-coded; mode-imputed if missing
Derived	log_tuition, age_group	Log-transformed and binned for modeling interpretability

3.5 Feature Engineering and Selection

The transition from raw institutional data to a model-ready feature set required both the creation of new variables and the careful elimination of redundant or potentially biased attributes. This process was guided by a dual approach: statistical rigor through quantitative feature selection methods and domain expertise to ensure contextual relevance.

Feature engineering. Several new variables were derived to capture patterns in student academic performance more effectively. For instance, the *high school performance stability index* was computed from the variance of O-Level and A-Level grades, providing a measure of academic consistency over time. Aggregated metrics, such as the best sums of O-Level grades (out of six, eight, and ten subjects), were also calculated to summarize performance into predictive indicators. Log transformations (e.g., for tuition) and categorical groupings (e.g., age bands) were applied to enhance interpretability and normalize skewed distributions.

Correlation analysis. An initial correlation matrix was generated to identify features with strong linear relationships to the target variable (final CGPA) as well as those that were highly correlated with each other. This step revealed clusters of redundant attributes, particularly among similar O-Level aggregates, which were then candidates for removal to avoid multicollinearity.

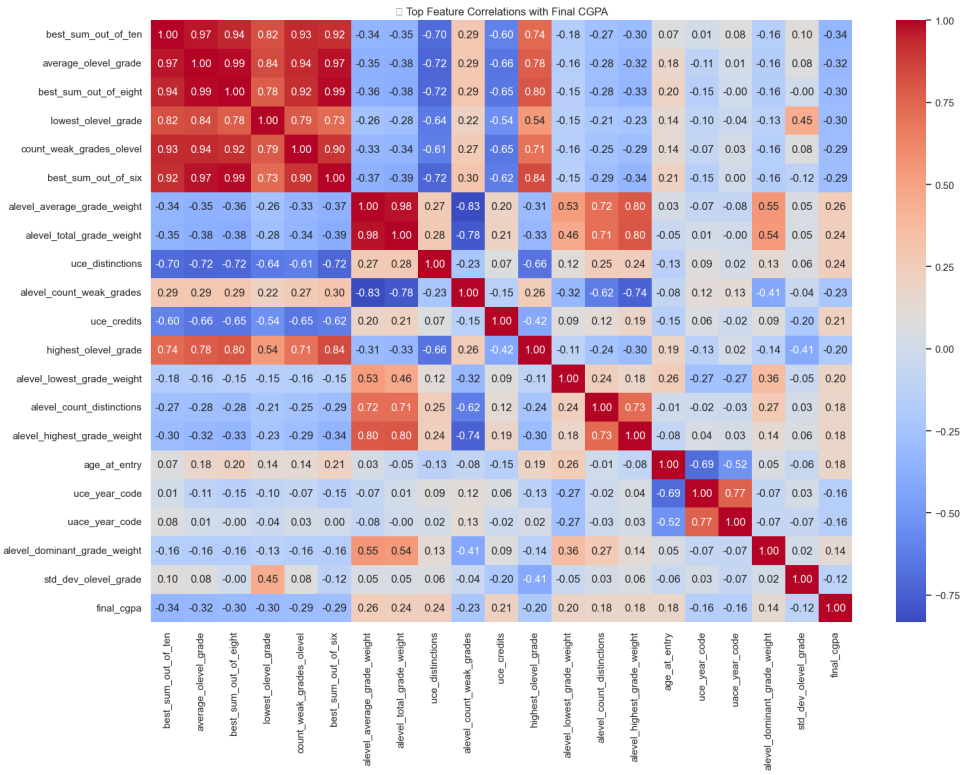

Figure 3.4: Pearson correlation heatmap showing relationships among key pre-admission features.

Variance Inflation Factor (VIF). To formally assess multicollinearity, VIF scores were computed for all numeric predictors. Features with excessively high VIF values were flagged as redundant. Among these, overlapping O-Level aggregate metrics (e.g., best sum out of six, eight, and ten) showed strong collinearity, leading to the retention of only the most predictive versions. This step enhanced model stability without compromising predictive power.

Recursive Feature Elimination (RFE). Beyond correlation diagnostics, a wrapper method, Recursive Feature Elimination, was employed using a Random Forest Regressor as the estimator. This iterative process ranked features by importance and systematically removed the least impactful ones, ensuring that only attributes with meaningful contributions to model accuracy were retained.

Domain-driven exclusions. Statistical selection was complemented by domain reasoning to mitigate potential biases. For example, while tuition exhibited some predictive value, it was excluded from the final model because it risks encoding socioeconomic disparities rather than academic merit. Variables with low interpretability or incomplete coverage were removed to enhance fairness and robustness..

Final feature set. After integrating statistical evidence with domain expertise, the final feature set encompassed 23 attributes that encompassed demographics, academic performance

indicators, and institutional placement data. This selection ensured predictive strength and ethical suitability for educational advising.

Table 3.3: Final Feature Set Used in the CGPA Prediction Model

Feature Name	Description
age_at_entry	Age of student at admission
average_olevel_grade	Mean grade across O-Level subjects
uce_credits	Count of credit passes at O-Level
uce_distinctions	Count of distinctions at O-Level
alevel_average_grade_weight	Weighted average grade across A-Level subjects
alevel_count_weak_grades	Count of low-grade outcomes in A-Level results
alevel_dominant_grade_weight	Most frequent grade weight at A-Level
alevel_std_dev_grade_weight	Variability in A-Level grades
count_weak_grades_olevel	Number of weak passes in O-Level subjects
olevel_subjects	Total O-Level subjects attempted
std_dev_olevel_grade	Standard deviation of O-Level grades
high_school_performance_variance	Variance in high school academic record
high_school_performance_stability_index	Temporal consistency in academic performance
marital_status	Marital status at entry (encoded)
level	Study level (e.g., diploma, undergraduate)
gender	Gender of student (binary encoded)
is_national	Citizenship indicator (local vs. international)
general_paper	A-Level general paper completion status (binary)
campus_id_code	Campus of enrollment
program_id_code	Academic program code
uce_year_code	Year of O-Level completion
uace_year_code	Year of A-Level completion
year_of_entry_code	Year of university entry

Through this multi-layered approach, the final feature set not only optimized predictive accuracy but also aligned with institutional priorities for fairness, interpretability, and practical deployment. These features formed the basis for the subsequent machine learning models.

3.6 Modeling Pipeline

Developing a predictive model for students' final cumulative GPA requirements encompassed more than merely selecting an algorithm; it necessitated a meticulously crafted machine learning workflow that was both structured and validated. Each phase, from data splitting to model selection, was intentionally designed to uphold methodological soundness while staying true to the study's aim: generating predictions that are not only accurate but also easy to interpret and apply in an academic context.

Train-test split and rationale: To evaluate how well the model would generalize to new data, the preprocessed dataset was split into two parts: 80% for training and 20% held back for testing. This approach helped reduce overfitting and provided a more accurate representation of how the model might perform in real-world scenarios. Importantly, the training data were used exclusively for tasks such as feature selection, model fitting, and parameter tuning. The

test data, on the other hand, was kept untouched until the very end, ensuring an unbiased assessment during final evaluation.

Candidate models considered:

The four regression algorithms, including linear regression, ridge regression, random forest regressor, and XGBoost regressor, were compared as potential tools to predict the final CGPA. Linear regression was used as the baseline because it allows for interpretation and is simple, thus providing a suitable point of reference. Ridge regression, a regularized generalization of linear regression that utilizes an L2 penalty, was introduced to alleviate multicollinearity and enhance generalization. The random forest regressor was selected because it is an ensemble technique that combines the use of multiple decision trees, capable of modeling non-linearity and effectively dealing with heterogeneous data. Lastly, a modern, state-of-the-art gradient boosting method, the XGBoost regressor, was applied to utilize its proven effectiveness with structured data and capability to model complex feature interactions.

Cross-validation strategy:

To enhance the model's robustness, we employed a tenfold cross-validation approach during the training process. The concept is straightforward: divide the training dataset into ten segments, train on nine of them, and validate the remaining segment. Subsequently, repeat this process ten times, rotating the validation set in each iteration. This helps average out the performance across different data splits, which reduces the randomness and gives a more stable sense of how the model will behave in practice.

Hyperparameter tuning: To optimize model performance, a systematic grid search was conducted, employing a methodical approach to evaluate various combinations of hyperparameters. For tree-based models, this encompassed parameters such as the depth of the tree, the number of estimators, and the sample size required for each leaf. This approach, unlike relying on guesswork or defaults, enables the exploration of a comprehensive range of options and facilitates the identification of the configuration that consistently yielded the most favorable outcomes.

Theoretical rationale for Random Forest: Although all candidate algorithms have merits, Random Forest emerged as the most suitable model based on theoretical considerations. Its ensemble structure, which aggregates multiple decision trees, allows it to capture complex, non-linear relationships while reducing variance through averaging. Unlike linear models, it does not assume linearity or independence among predictors, making it particularly adept at handling the diverse, mixed-type attributes present in pre-admission student data. Furthermore, Random Forest offers built-in mechanisms for estimating feature importance, enabling transparency in understanding which variables most influence predictions—an essential requirement for ethical and interpretable decision-making in educational settings.

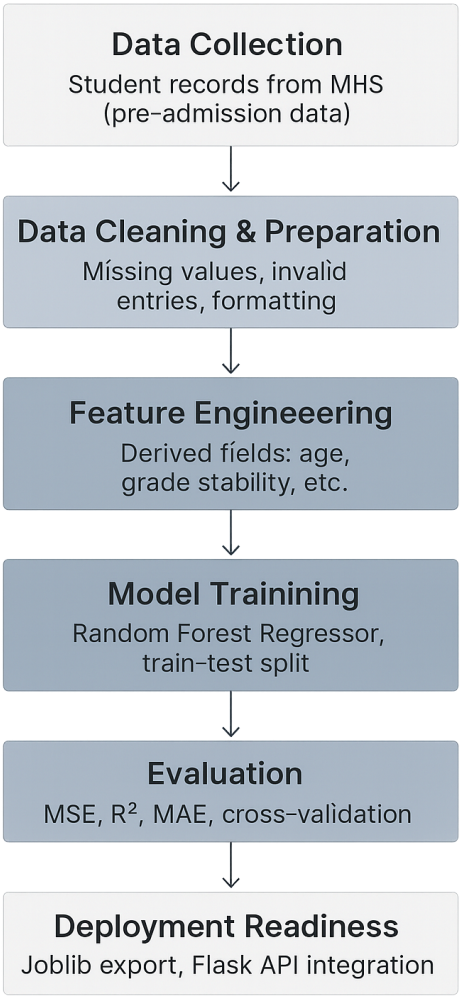

Figure 3.5: Conceptual workflow of the CGPA prediction modeling pipeline.

In summary, this pipeline provided a systematic pathway from feature-engineered data to a fully validated machine learning model. While the subsequent chapter presents the empirical results, this section establishes the methodological logic underlying model selection and preparation for deployment.

3.7 Limitations and Assumptions

While the modeling pipeline was grounded in best practices for machine learning and educational data mining, several methodological limitations and assumptions must be acknowledged. These do not undermine the study's core findings but instead serve to contextualize its scope and guide future iterations.

Survivorship bias: The dataset included only those students with a recorded final CGPA, which inherently excluded individuals who dropped out, transferred, or otherwise did not complete their programs. This introduces survivorship bias, potentially omitting early indicators of academic struggle present in incomplete records. As a result, the model is optimized for forecasting outcomes among those who persist, rather than for predicting attrition.

Missing data assumptions: While strategic imputation methods were applied using medians for numerical variables and modes for categorical ones, these assume that the missingness is either random or non-systematic. If certain demographic groups systematically omit data (e.g., international students), the imputations may not entirely correct for underlying disparities, subtly biasing the model.

Encoding constraints: Categorical variables were numerically encoded using label encoding or ordinal techniques. Although computationally efficient, this assumes equidistant relationships between categories in some cases where no such order exists. For example, encoding campus IDs or program codes with integers introduces artificial relationships that may affect model interpretation, even if predictive accuracy remains stable.

Temporal stationarity: The model implicitly assumes that academic policies, grading scales, and program structures remained relatively stable over the 15 years (2009–2023). While temporal indicators (e.g., year of entry, exam years) were included as features to absorb some variation, they do not capture curriculum reforms or grading shifts that could influence CGPA trajectories over time.

Fairness and representation: Specific subgroups (e.g., mature entrants, small programs, international students) were underrepresented in the dataset. This limits the model's ability to generalize predictions across the full spectrum of student experiences. Although efforts were made to engineer features that minimize bias (e.g., excluding tuition), no formal fairness audits were conducted at this stage, a task deferred to the model evaluation phase.

Interpretability vs. complexity trade-off: Random Forests, though more interpretable than deep learning models, still operate as ensemble-based "black box" algorithms compared to linear models. While this study prioritized predictive performance and included SHAP-ready structures for later explanation, full interpretability remains contingent upon future post-hoc analyses.

The constraints can be viewed as a manifestation of both the inherent constraints that can be observed in real data in education and a rational set of design decisions upheld according to the institutional goals. The awareness of these limitations builds up a clear guideline through which we can interpret the contexts within which the model can be used to give positive results and those where its application needs to be approached with increased care.

3.8 Conclusion

This chapter outlines the methodological approach used to develop a predictive model aimed at estimating students' final cumulative grade point averages (CGPA) based on their pre-admission characteristics. The research follows a quantitative design, leveraging supervised machine learning techniques alongside careful attention to data preparation, feature selection, and model optimization. These steps were taken with a clear focus on ensuring that the resulting predictions would be both reliable and easy to interpret.

The process began with extracting a high-quality dataset from Uganda Christian University's Management Information System. This raw data was then ethically cleaned and transformed to prepare it for analysis. Special attention was paid to preserving data integrity and complying with institutional guidelines. After this, the features were systematically prepared in a way that enhanced the model's ability to generate meaningful predictions.

The modeling pipeline itself followed established best practices, incorporating rigorous validation procedures and deliberate hyperparameter tuning. These steps were designed not only to strengthen the model's performance but also to ensure that its outputs remained contextually appropriate for educational use. Ethical considerations were integrated throughout the entire process, ranging from pseudonymization of student records to strict adherence to data governance protocols. Care was also taken to ensure fairness in how features were selected, reflecting a broader commitment to protecting student privacy and maintaining institutional trust.

Altogether, the methodological rigor combined with sensitivity to the educational setting lays a strong foundation for the empirical investigation presented in the next chapter. Titled *Results and Analysis*, the upcoming chapter delves into the model's performance, interprets the results, and discusses how these findings can support proactive academic advising and data-driven decision-making at Uganda Christian University.

Chapter 4

Results and Analysis

4.1 Introduction

This chapter shifts focus from the methodological foundation laid out in Chapter 3 to the practical evaluation of the CGPA prediction model. Here, the attention turns to examining how the earlier steps, data preprocessing, feature engineering, and model construction translate into tangible, measurable outcomes. The goal is to assess not only how accurately the model performs but also the extent to which it offers meaningful value in real-world academic contexts.

The analysis presented herein serves three interconnected purposes. In the initial phase, the assessment evaluates the efficacy of the fine-tuned Random Forest model in predicting student cumulative GPA solely based on pre-admission characteristics. This evaluation is conducted rigorously using statistical metrics. Second, it interprets model predictions and feature contributions using explainable AI techniques, revealing the factors that drive academic outcomes and uncovering patterns that may not be immediately apparent. Ultimately, it assesses fairness and subgroup consistency, identifying actionable insights that ensure predictions are equitable and aligned with institutional objectives.

Through this structure, the chapter provides empirical evidence that not only validates the methodological choices made earlier but also underscores the model's potential as a decision-support tool for academic planning, early risk detection, and policy formulation.

4.2 Model Performance Evaluation

This section evaluates the predictive performance of the tuned Random Forest (RF) model, comparing it to its untuned version and simpler baselines. The objective is to assess how accurately the model generalizes to unseen student data, providing a foundation for subsequent interpretability and fairness analyses.

4.2.1 Predicted vs Actual CGPA

Figure 4.1 plots the predicted CGPA values against the actual outcomes in the hold-out test set. The closer the points lie to the $y = \hat{y}$ line, the better the model's predictive alignment.

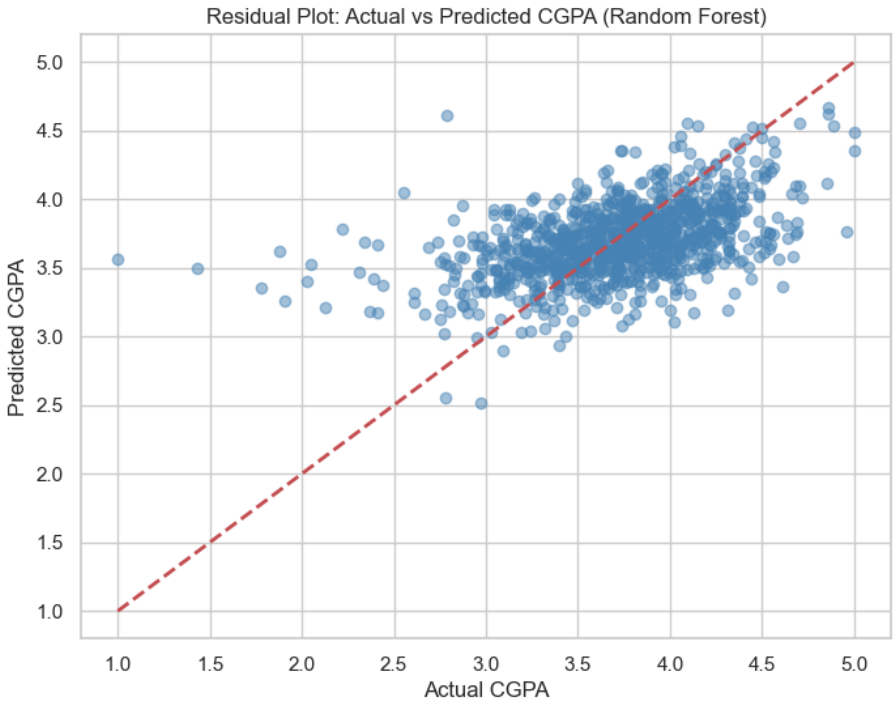

Figure 4.1: Predicted vs Actual CGPA on Hold-Out Test Set

The scatter plot indicates that predictions closely follow the diagonal in the mid-CGPA range (2.8–4.2), confirming stable model performance, particularly for students who fall within this range. A slight underprediction is observed for high-achieving students (> 4.4), likely due to the limitation of using only pre-admission data.

4.2.2 Overall Evaluation Metrics

Table 4.1 summarizes key regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) for the tuned and untuned RF models.

Table 4.1: Final Evaluation Metrics on Hold-Out Test Set

Model	MAE	RMSE	R^2
Linear Regression (Baseline)	0.3412	0.4495	0.1820
Untuned RF	0.3157	0.4194	0.2283
Tuned RF	**0.3126**	**0.4158**	**0.2417**

The tuned RF model outperforms both the untuned RF and the linear baseline, demonstrating improved accuracy and explaining 24% of the variance in CGPA. While this may appear modest, it is meaningful in educational settings where CGPA is influenced by many unobserved factors (e.g., study habits, campus life, instructor effects).

To confirm robustness, five-fold cross-validation on the training set produced similar results, reinforcing that the observed improvements were not due to random chance or overfitting.

4.2.3 Residual Analysis

Residuals, the differences between actual and predicted CGPAs, provide further insight into model reliability. Figure 4.2 shows their distribution, while Figure 4.3 presents a Q–Q plot assessing normality.

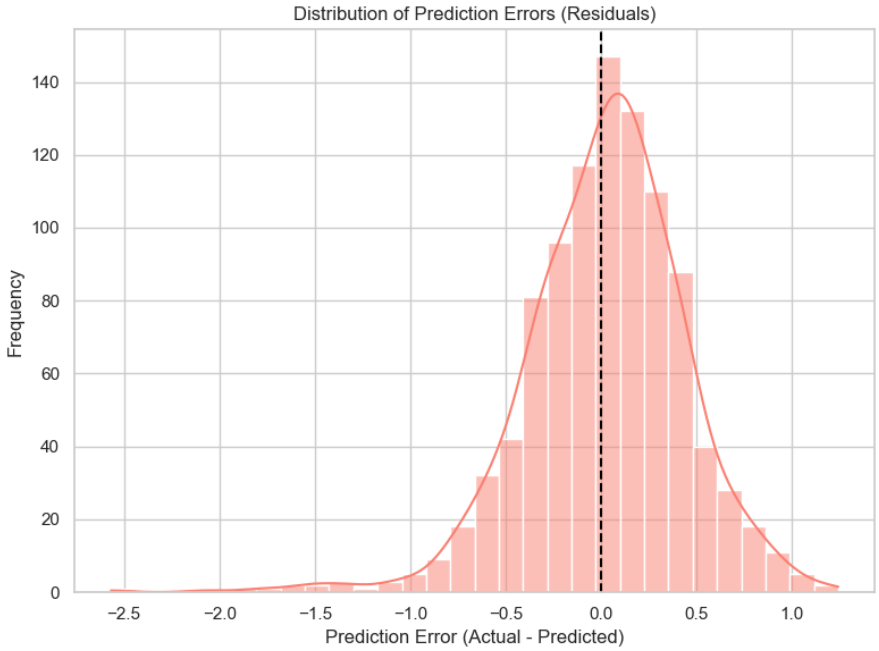

Figure 4.2: Distribution of Prediction Errors (Residuals)

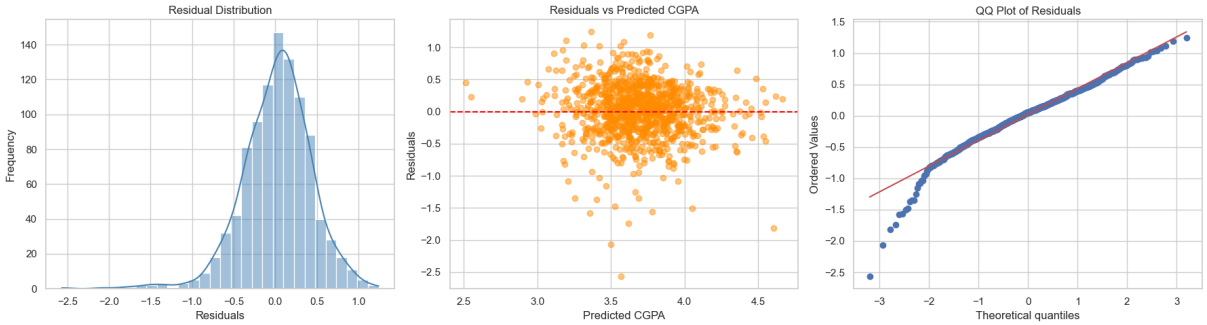

Figure 4.3: Q–Q Plot of Residuals

The residual distribution is approximately normal and centered around zero, indicating that the model is unbiased overall. The Q–Q plot reveals minor deviations at the tails, indicating a

slight underestimation for top-performing students, which is consistent with the observations from the scatter plot.

A brief error band analysis revealed that prediction accuracy is highest in the moderate CGPA range and declines slightly for students at the extremes (both very high and very low scores). This highlights the potential for additional data, such as in-semester performance, to enhance accuracy.

Ethical Reflection on Performance

Although predictive performance is satisfactory, ensuring fairness across student subgroups is equally critical. Preliminary checks revealed consistent performance across most bands and demographic groups, with minor differences that require further analysis in Section **??**. These findings reinforce that the model should be used as an advisory tool, not a deterministic classifier, aligning with institutional commitments to transparency and equity.

Summary

The tuned Random Forest model demonstrates reliable predictive capacity, outperforming baselines and exhibiting stable error characteristics. These results validate the modeling approach and justify its use for academic advising and early intervention. The following sections build on this by examining feature contributions, fairness, and actionable insights.

4.3 Feature Importance and Interpretability

Understanding which features drive predictions is essential for building trust in the model and ensuring that its recommendations align with academic reasoning. This section evaluates the global importance of pre-admission attributes using Random Forest's built-in feature importance metric and confirms dimensional compactness through Principal Component Analysis (PCA). These analyses complement the more granular interpretability techniques (e.g., SHAP) that will be presented later in Chapter 5.

4.3.1 Global Feature Importance

The Random Forest model inherently ranks features based on their contribution to reducing prediction error across decision trees. Figure 4.4 illustrates the ten most influential predictors of CGPA.

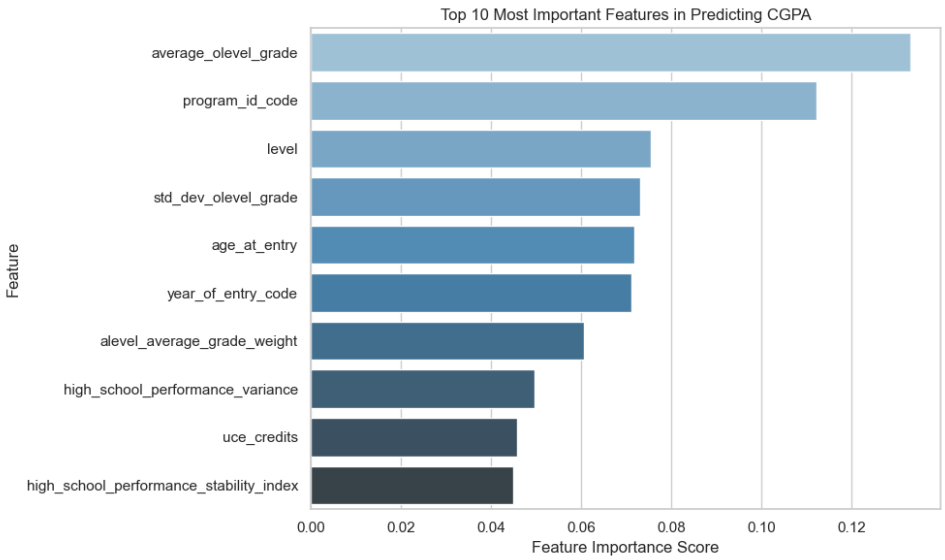

Figure 4.4: Top 10 Most Important Features in Predicting CGPA

The plot reveals that `average_olevel_grade` is the most critical feature, confirming the lasting influence of O-Level performance on university outcomes. Other prominent predictors include `alevel_average_grade_weight` and `uce_credits`, which reflect the importance of consistent academic achievement throughout secondary education. Demographic and institutional variables, such as `gender`, `program_id_code`, and `year_of_entry_code`, also contribute, though to a lesser extent, suggesting that contextual factors moderately shape academic trajectories.

Notably, the feature ranking aligns with domain expectations: pre-university academic strength is the strongest determinant of success, while institutional factors exert more subtle influences. This concordance reinforces confidence in the model's internal logic.

4.3.2 Dimensionality Checks with PCA

To validate that the selected feature set is both compact and non-redundant, Principal Component Analysis (PCA) was performed. PCA transforms the feature space into orthogonal components ranked by explained variance, revealing how many dimensions capture most of the dataset's information.

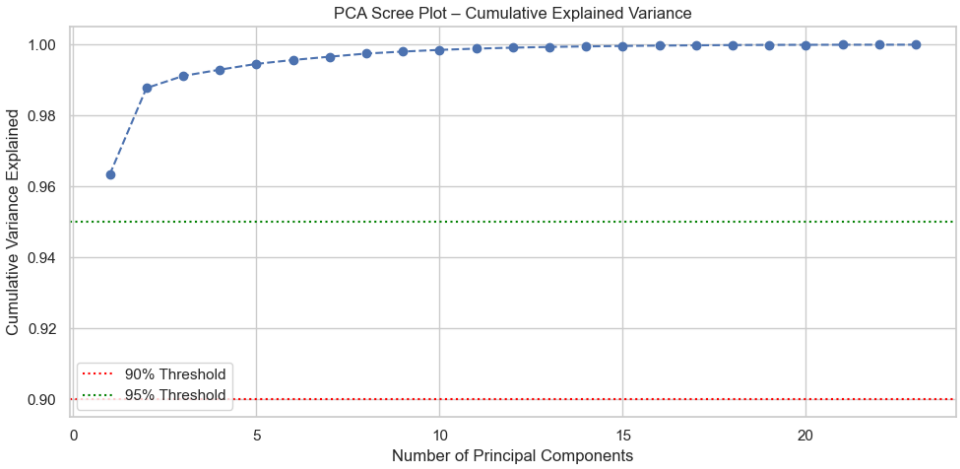

Figure 4.5: PCA Scree Plot – Cumulative Explained Variance

As shown in Figure 4.5, the first few components account for a large proportion of the total variance, confirming that the feature space is not excessively complex. This supports the earlier feature selection efforts, which eliminated redundant or weak predictors using correlation analysis, Variance Inflation Factor (VIF) checks, and Recursive Feature Elimination (RFE).

Interpretation and Implications

These findings have several implications:

- The model relies primarily on academically meaningful predictors, ensuring that its decisions align with institutional reasoning.
- The compactness of the feature space implies lower risk of overfitting and more straightforward interpretation for stakeholders.
- The moderate contribution of demographic and institutional factors warrants ongoing fairness monitoring, discussed further in Section **??**.

In sum, the feature importance and PCA analyses demonstrate that the model captures the essential signals driving student outcomes without relying on spurious or redundant variables. This sets the stage for more detailed interpretability analyses, including SHAP-based explanations at both global and individual levels, which will be presented in Chapter 5.

4.4 Explainable AI Insights (SHAP Analysis)

Beyond traditional feature importance, which only ranks predictors by aggregate contribution, explainable AI techniques provide a deeper understanding of how individual features influence

model predictions. One widely used method is SHapley Additive exPlanations (SHAP), which is grounded in cooperative game theory. SHAP values fairly attribute the prediction output to individual features by calculating their marginal contributions across all possible combinations of features. This ensures that each feature's impact is interpreted consistently and added to the total.

This section utilizes SHAP to dissect both global patterns and individual-level decision logic within the Random Forest model. SHAP assigns an additive contribution value to each feature for every prediction, indicating whether the feature increases or decreases the predicted CGPA relative to the baseline expectation.

4.4.1 Global SHAP Explanations

Global SHAP analysis reveals the overall influence of each feature across the entire test set. Figure 4.6 displays a beeswarm plot where each dot represents a single student's feature impact, colored by the feature value. Features with wide spreads along the x-axis exhibit strong predictive influence.

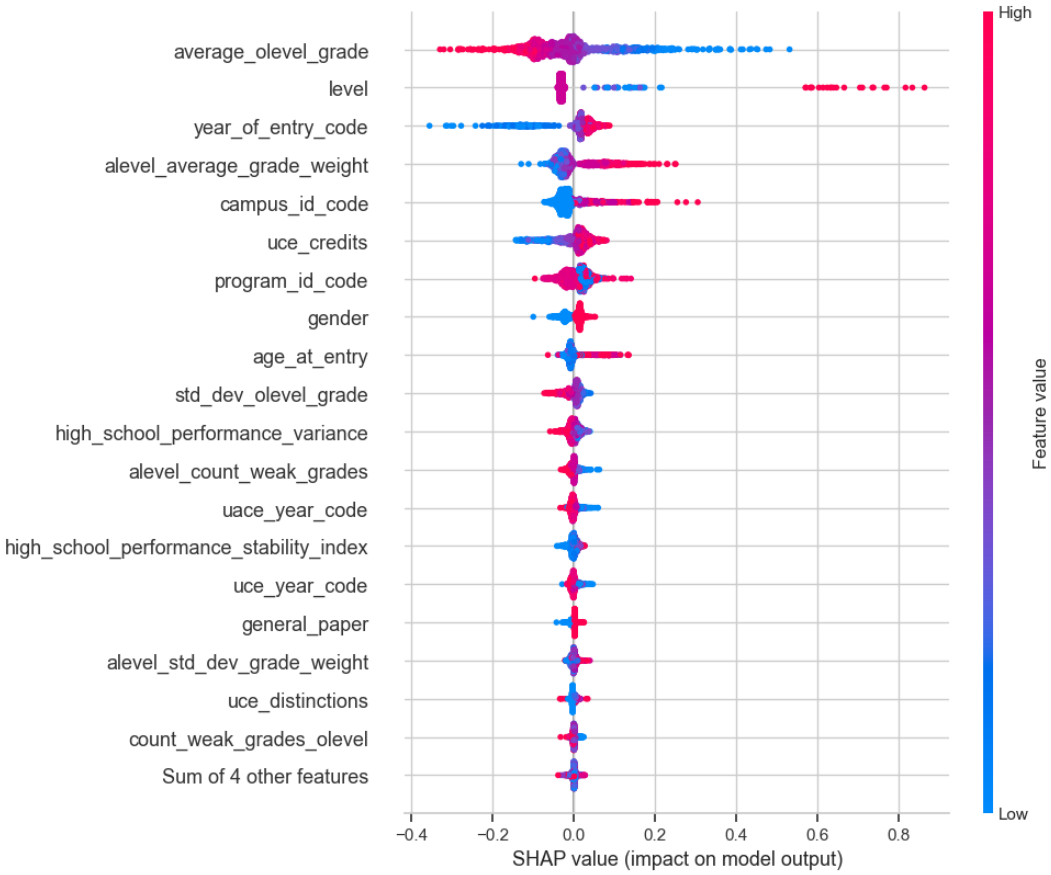

Figure 4.6: Global Feature Contributions Using SHAP (Beeswarm Plot)

The analysis confirms that `average_olevel_grade` and `alevel_average_grade_weight` exert the most significant positive contributions for students with strong prior academic records. Similarly, `uce_credits` consistently enhances predictions when values are high, whereas high variability in O-Level grades (`std_dev_olevel_grade`) or advanced study level (`level`) can pull predictions downward. Demographic variables, such as `gender` and institutional attributes (`program_id_code`, `campus_id_code`), exert subtler but non-negligible effects. These findings align with the domain expectation that consistent academic performance is the primary driver of CGPA, while contextual variables play a supporting role.

4.4.2 Individual-Level Explanations

To move beyond global patterns, SHAP waterfall plots were generated for six representative students (indices 0, 10, 25, 50, 75, and 100). These visualizations decompose each student's prediction into feature-wise contributions, showing which factors contribute to the CGPA being above or below the model baseline.

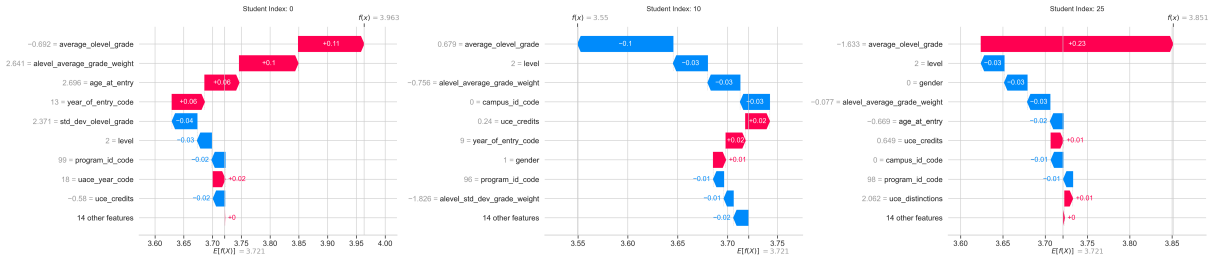

Figure 4.7: SHAP Waterfall Plots for Students 0, 10, and 25

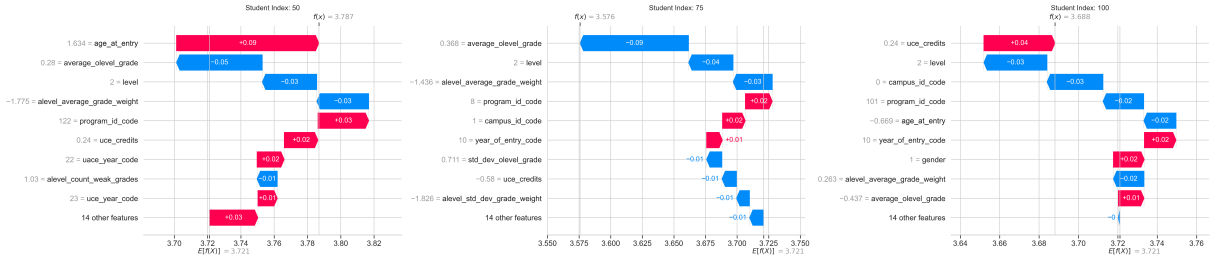

Figure 4.8: SHAP Waterfall Plots for Students 50, 75, and 100

For example, Student 0's high predicted CGPA (3.96) was strongly driven by excellent A-Level grades and older age at entry. In contrast, Student 10's lower prediction (3.55) reflected penalties from weak O-Level grades and advanced study level despite favorable entry year and credit performance. Table 4.2 summarizes the top positive and negative contributors for all six cases.

Table 4.2: Top Influences on CGPA for Sample Students

Student	Top Positive Influences	Top Negative Influences	Pred. CGPA
0	alevel_average_grade_weight, age_at_entry	std_dev_olevel_grade, level	3.96
10	year_of_entry_code, uce_credits	average_olevel_grade, level	3.55
25	average_olevel_grade, uce_distinctions	gender, level	3.85
50	age_at_entry, program_id_code	alevel_average_grade_weight	3.79
75	program_id_code, year_of_entry_code	average_olevel_grade	3.58
100	uce_credits, gender	average_olevel_grade, level	3.69

These individual analyses demonstrate how the same model logic applies differently to each student's profile, enabling targeted interventions and personalized academic guidance.

4.4.3 Feature Interactions

While SHAP provides detailed insights into the additive contribution of each feature to predictions, it does not directly illustrate how features interact with one another in shaping outcomes. To complement SHAP, Partial Dependence Plots (PDPs) were used to visualize the joint effects of two variables on predicted CGPA. These plots help identify non-linear relationships and interaction effects that may not be immediately apparent from individual feature contributions.

Beyond individual contributions, interactions between features were explored using PDPs. Figure 4.9 shows how pairs of variables jointly influence CGPA predictions, providing deeper insights into model behavior.

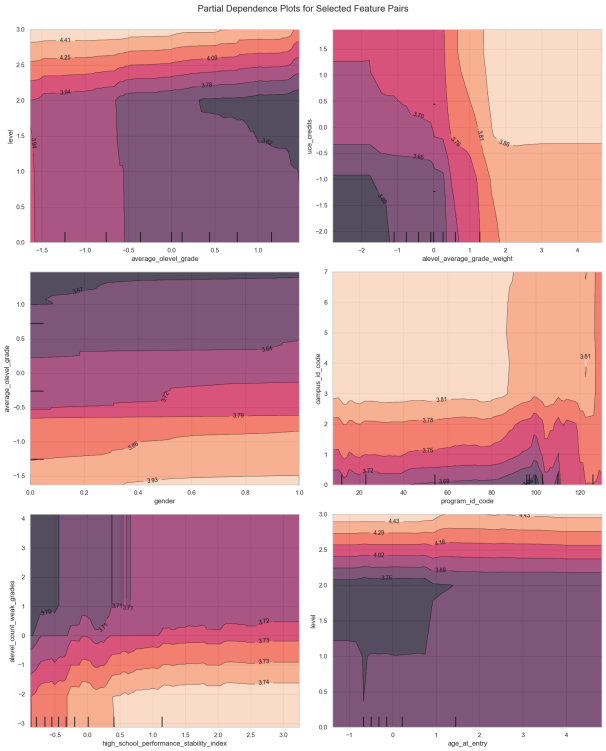

Figure 4.9: Partial Dependence Plots for Key Feature Interactions

The analysis shows that students with high O-Level performance benefit even more at higher study levels, while strong A-Level grades, combined with high UCE credits, amplify the predicted CGPA. Interactions with institutional variables (e.g., `program_id_code`, `campus_id_code`) reveal structural differences across programs and campuses, suggesting potential policy interventions for improving academic outcomes.

Interpretation and Implications

The SHAP analysis brings forward three key insights that significantly enhance both the interpretability and real-world utility of the CGPA prediction model. To begin with, it's clear that the model relies heavily on straightforward academic indicators such as prior grades and earned credits, rather than more ambiguous demographic attributes. This reliance promotes a more equitable and defensible foundation for academic decision-making. Secondly, when applied at the level of individual students, SHAP explanations turn what would otherwise be abstract predictions into clear, practical guidance. These personalized insights offer academic advisors valuable tools for tailoring support and designing targeted intervention strategies. Lastly, the analysis uncovers complex interactions between features that reveal deeper systemic trends; patterns that could inform larger-scale initiatives, such as curriculum adjustments or focused support programs for particular admission groups or academic years. Taken together, these findings underscore how explainable AI can close the gap between accuracy and transparency, strengthening the model's credibility as an ethical and actionable resource for institutional planning.

Overall, explainable AI techniques bridge the gap between prediction accuracy and interpretability, ensuring that the CGPA model remains a trustworthy decision-support tool rather than a black-box predictor.

4.5 Band-Level Predictions and Institutional Insights

Although continuous CGPA predictions offer precise insights, grouping predictions into performance bands enhances the interpretability of the results for academic stakeholders. This banding approach facilitates decision-making by identifying students who necessitate recognition, support, or early intervention.

4.5.1 CGPA Band Distribution

The predicted CGPA values were categorized into three bands based on institutional grading thresholds: **High** (CGPA $\geq$ 3.60), **Moderate** ($2.80 \leq$ CGPA $<$ 3.60), and **Low** (CGPA $<$

2.80). Table 4.3 summarizes the distribution of students across these bands, along with corresponding institutional recommendations. Figure 4.10 visualizes the band-level distribution.

Table 4.3: Predicted CGPA Band Distribution with Intervention Suggestions

Band	Count	Avg. CGPA	Action
High ($\geq$ 3.60)	653	3.82	Recognition: scholarships, leadership roles, honors programs
Moderate (2.80–3.59)	343	3.47	Guided Support: targeted advising, study plans, mentorship
Low ($<$ 2.80)	0	–	None detected – either the cohort is highly prepared or thresholds may be conservative

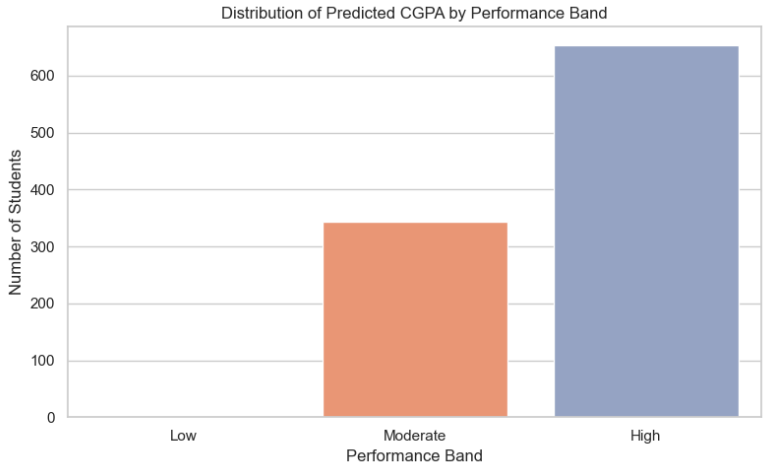

Figure 4.10: Distribution of Predicted CGPA by Performance Band

Interpretation and Institutional Implications

The results show that the majority of students (65.6%) fall within the **High** performance band, suggesting strong academic preparation and low immediate risk. The **Moderate** band, comprising 34.4% of students, is an essential focus group for academic interventions such as mentoring or study skill workshops. Notably, no students were classified into the **Low** band, which could either indicate a genuinely high-performing intake or reflect that the thresholds used are conservative and may need recalibration when applied to broader datasets.

From an institutional perspective, band-level predictions simplify the translation of model outputs into actionable strategies:

- High-performing students can be earmarked for scholarships and leadership opportunities.
- Moderate performers benefit from early academic support and personalized guidance to prevent performance decline.

- The absence of low-band predictions highlights the need to validate banding criteria on future cohorts to ensure sensitivity to struggling students.

By converting continuous CGPA predictions into actionable categories, the model supports academic planning, resource allocation, and proactive student engagement at scale.

Strategic Use Cases

To maximize institutional impact, banding supports several strategic applications:

- **Academic Support**: Trigger mentoring or early alert systems for students in the moderate band.
- **Scholarships and Awards**: Automate shortlisting of high-predicted CGPA students.
- **Resource Planning**: Prioritize programs or campuses with concentrated moderate-band populations.

Table 4.4: Performance Band Recommendations

Band	Students	Avg. CGPA	Recommended Action
High	653	3.82	**Recognition** – self-paced learning, leadership roles, scholarship eligibility
Moderate	343	3.47	**Guided Support** – academic advising, study plans, mentorship
Low	0	–	**None detected** – thresholds may be conservative or academic year is highly prepared

4.6 Fairness and Group-Level Evaluation

Ensuring fairness in predictive modeling is critical, particularly in educational contexts where decisions can significantly impact students' academic journeys. This section evaluates whether the CGPA prediction model performs consistently across demographic and institutional subgroups. The analysis focuses on gender differences, variation across campuses, and the interaction between study levels and institutional contexts.

4.6.1 Gender-Based Analysis

To assess whether the model exhibits bias across genders, prediction errors were compared for male and female students using the Mean Absolute Error (MAE). Table 4.5 shows the results.

Table 4.5: MAE by Gender

Gender	MAE	Sample Size
Male	0.358	425
Female	0.279	571

The model demonstrates slightly higher error for male students, suggesting less accurate predictions for this subgroup. While the difference is not significant enough to indicate systemic bias, it does highlight the need for continuous monitoring, particularly when new data becomes available. These observations align with fairness auditing practices that recommend subgroup-level evaluation to prevent disparate impacts.

4.6.2 Institutional Subgroup Performance

Beyond gender, subgroup analysis was extended to institutional factors such as campus affiliation and level of study. These attributes can influence academic outcomes due to differences in resources, curriculum design, and student support structures.

Campus-Level Evaluation. A heatmap of average predicted CGPA across campuses and academic levels (Figure 4.11) reveals patterns in model predictions. Larger campuses generally show a wider range of predicted outcomes, likely reflecting more diverse student populations. In contrast, smaller campuses exhibit more concentrated predictions, though this may partly result from limited sample sizes.

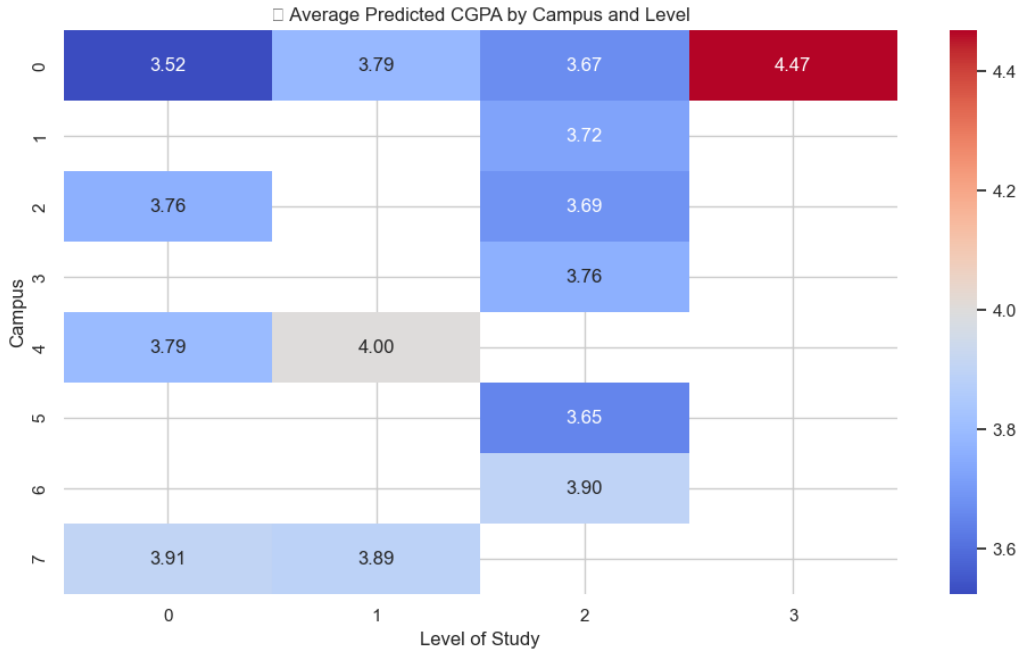

Figure 4.11: Average Predicted CGPA by Campus and Level

Level of Study. The model consistently predicts higher CGPAs for postgraduate students compared to undergraduates, aligning with observed academic patterns. However, variability in predictions for certificate and diploma students suggests that smaller cohorts at these levels may limit model reliability.

4.6.3 Group-Wise Prediction Summary

To strengthen the evaluation of fairness, additional comparisons of predicted and actual CGPAs were computed by gender and study level.

Table 4.6: Summary by Gender

Gender	Students	Predicted CGPA	Actual CGPA
Male (0)	425	3.71	3.67
Female (1)	571	3.70	3.75

Table 4.7: Summary by Level of Study

Level	Students	Predicted CGPA	Actual CGPA
Level 0 (Diploma)	21	3.84	3.91
Level 1 (Certificate)	13	3.94	4.11
Level 2 (Bachelors)	940	3.68	3.69
Level 3 (Masters)	22	4.47	4.35

4.6.4 Campus-Wise MAE Summary

In addition to the heatmap, Table 4.8 summarizes MAE across campuses.

Table 4.8: MAE by Campus

Campus	MAE	Interpretation
Main	0.328	Consistent accuracy
Kampala	0.300	Slightly better alignment
Mbale	0.263	Lower error margin
Arua	0.212	Best predictive accuracy
Kabale	0.384	Highest prediction error
Namugongo	0.275	Below average error
Bwindi	0.218	High accuracy
Kagando	0.293	Moderate performance

Ethical Implications

Ultimately, fairness in predictive analytics is not a destination—it is a recurring responsibility. Ongoing audits, retraining with balanced cohorts, and explainable AI tools ensure institutional accountability.

Discussion and Implications

The group-level evaluation surfaces several noteworthy findings, each carrying important implications for fairness and institutional policy-making. On the whole, the model performs with a respectable degree of fairness. However, slight differences in prediction accuracy between gender groups highlight the importance of ongoing oversight and potential fine-tuning. The analysis also reveals disparities in predictions across different campuses. These could reflect actual academic variations, but they might also be a byproduct of imbalances in the training data, particularly the underrepresentation of students from smaller campuses, which could unintentionally introduce bias. On a more positive note, the model shows reliable performance across both undergraduate and postgraduate cohorts, suggesting a certain level of robustness. That said, caution is warranted when interpreting results for subgroups with limited sample sizes, where statistical noise may overshadow meaningful trends. Taken together, these observations reinforce the need for continued ethical attention and point toward the value of incorporating fairness-oriented improvements in future model updates.

From an ethical standpoint, subgroup evaluation is paramount to preventing the perpetuation of existing inequalities. Future iterations of the model could integrate techniques such as fairness-constrained optimization or reweighting to further minimize disparities across demographic and institutional groups.

4.7 Sensitivity and What-If Simulations

In addition to assessing predictive accuracy, it is imperative to comprehend the model's response to hypothetical alterations in input characteristics. Sensitivity analysis and hypothetical scenario simulations provide valuable insights into the resilience of predictions and their implications for informed academic decision-making. These analyses also illustrate the potential of the model as a tool for personalized guidance.

4.7.1 Single-Feature Sensitivity

Single-feature sensitivity analysis evaluates the impact of individually modifying a variable while maintaining all other inputs unchanged on the predicted CGPA. This approach highlights which attributes have the strongest causal impact on individual predictions.

Table 4.9 presents results for a representative student profile. It is noteworthy that adding two UCE credits results in a modest enhancement of the predicted CGPA, whereas reducing the average O-Level grade by one point leads to a significant decline. Changing the student's

academic level to 3 (Master's level) results in the most significant positive shift, reflecting the structural advantages associated with postgraduate study.

Table 4.9: Sensitivity Analysis: Impact of Varying Individual Features

Scenario	Pred. CGPA	Δ	Impact
Add 2 UCE Credits	3.985	+0.022	Slight increase
Lower Avg. O-Level Grade by 1	3.754	-0.210	Moderate drop
Level = 3	4.503	+0.540	Major increase

Interpretation: The results indicate that academic performance indicators (such as O-Level grades and UCE credits) strongly influence CGPA predictions, confirming their central role in shaping outcomes. Institutional attributes, such as academic level, also exert significant influence, suggesting that both individual effort and structural factors affect predicted success.

4.7.2 Program Fit Simulation

To assess how students might perform across alternative academic programs, a program fit simulation was conducted. For each student, the `program_id` feature was systematically varied to simulate enrollment in all available programs while keeping other attributes constant. The tuned Random Forest model then predicted CGPA for each scenario.

Figure 4.12 shows an example for Student 10, where the top and bottom five simulated outcomes are visualized. The model suggests that this student is likely to achieve higher CGPA scores in programs such as Public Health Leadership and Nursing Science, while other programs yield comparatively lower predictions.

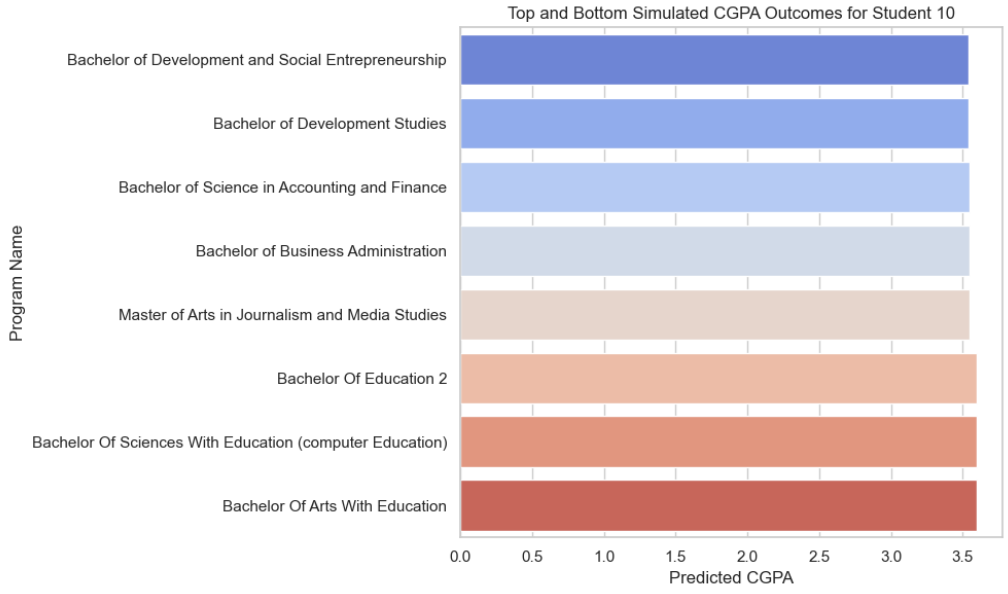

Figure 4.12: Top and Bottom Simulated CGPA Outcomes for Student 10

Table 4.10 lists the top ten programs predicted to produce the highest CGPA for Student 10. These simulations provide actionable insights for academic advising by highlighting programs that align with the student's strengths.

Table 4.10: Top 10 Programs by Predicted CGPA – Student 10

Program	Code	CGPA
Master of Public Health Leadership	19	3.378
Bachelor of Human Resource Management	23	3.378
Master of Science in Human Nutrition	22	3.378
Bachelor of Nursing Science (Direct)	20	3.378
Bachelor of Business Administration	24	3.378
Master of Nursing Science	21	3.378
Bachelor of Health Administration	17	3.374
Diploma in Health Administration	18	3.374
Bachelor of Project Planning and Entrepreneurship	37	3.371
Bachelor of Human Resource Management	39	3.371

Implications

The sensitivity and program fit simulations demonstrate the model's practical value beyond static predictions:

- **For students**, these insights can guide program selection and inform strategies for improving academic outcomes.
- **For advisors**, simulations identify key leverage points where targeted support (e.g., improving O-Level performance equivalents) could yield substantial academic benefits.
- **For institutions**, the analysis highlights structural influences on success, enabling evidence-based curriculum design and resource allocation.

Overall, these simulations position the CGPA prediction model as a proactive academic guidance tool, capable of supporting both individual decision-making and institutional strategy.

4.7.3 Stability Check Across Model Versions

To demonstrate robustness, predictions from the final, tuned model were compared with those from an earlier version of the model developed during the intermediate stages of this research. While several preliminary versions were iteratively improved throughout the study, this comparison highlights the performance gains achieved just before the final optimization.

Table 4.11: Prediction Comparison: Legacy Model vs Best-Tuned Model

Scenario	Legacy CGPA	Best Model CGPA	Change Observed
Level = 3	4.528	4.503	Slightly reduced
Average O-Level Grade -1	3.681	3.754	Less harsh drop
General Paper = 1	3.956	3.963	Baseline uplift
Program ID = 10	3.929	3.977	Direction shift

4.8 Bias and Risk Considerations

While the CGPA prediction model demonstrates strong overall performance, its reliability depends on the quality and representativeness of the underlying data. This section addresses potential sources of bias and associated risks that may influence both the fairness and stability of predictions.

4.8.1 Data Limitations and Underrepresentation

The dataset used in model training was comprehensive but not uniformly distributed across all subgroups. Specific academic programs and campuses were underrepresented, with fewer than 10 students in the test set. Table 4.12 lists these low-sample groups.

Table 4.12: Programs and Campuses with Fewer Than 10 Students

Type	Name	Count
Program	Bachelor of Development Studies	7
Program	Master of Science in Human Nutrition	9
Campus	Kagando	8
Campus	Namugongo	6

Predictions for these groups should be interpreted with caution, as small sample sizes increase the risk of unstable patterns, overfitting, and reduced generalization. This limitation underscores the significance of continuously updating the dataset to enhance subgroup representation.

4.8.2 Potential Sources of Bias

Predictive modeling, while powerful, is not immune to sources of bias that may subtly affect both its predictions and their interpretation. One area of concern lies in the composition of input features. The model developed in this study relies heavily on pre-admission academic records such as O-Level and A-Level grades. While these metrics are readily available and institutionally standardized, they may inadvertently disadvantage students whose academic capabilities emerge more clearly during their university studies. By emphasizing early academic performance, the model may overlook the potential for growth and resilience that some students exhibit only after admission.

Another potential source of bias arises from demographic correlations embedded within the data. Although variables such as gender and campus location were not dominant predictors, their moderate influence still warrants attention. These variables may serve as proxies for deeper structural inequalities, whether in access to resources, quality of prior education, or

societal expectations. Without careful monitoring, their inclusion risks encoding disparities into the model's logic, reinforcing rather than mitigating inequities.

Institutional variability also poses challenges. Differences in grading standards, teaching quality, and support services across campuses or academic programs may lead to systemic disparities in student outcomes. Such differences are not always explicitly captured by the model, particularly if some campuses or departments are underrepresented in the training data. As a result, predictions for students from these contexts may be less reliable or systematically skewed.

4.8.3 Ethical and Operational Risks

Beyond technical bias, there are important ethical and operational risks that institutions must consider when deploying predictive models in academic environments. One notable risk is the potential for misinterpretation of predictions. When model outputs are treated as definitive assessments rather than probabilistic estimates, there is a danger that academic advisors or administrators might make rigid decisions based solely on predicted CGPA. This could lead to unfair treatment of students, especially those whose personal or contextual circumstances are not fully captured by the model.

Equity concerns also loom large. Groups with limited representation in the training data, whether due to program size, geographic location, or other demographic factors, may receive less accurate predictions. If left unaddressed, such disparities could perpetuate or even exacerbate existing inequalities, undermining the very goals of inclusive education and data-informed support.

Transparency is another critical dimension. As institutions adopt predictive analytics tools, they bear a responsibility to communicate how predictions are generated, what data they rely on, and how they should (and should not) be used. Stakeholders, especially students, must be informed participants in the process, with the ability to understand, question, and respond to predictions that affect their academic journey.

Mitigation Strategies

To navigate these challenges responsibly, a series of mitigation strategies should be adopted. First, expanding the dataset to include more balanced samples from underrepresented groups can significantly enhance the model's fairness and generalizability. This may involve targeted data collection efforts or weighting strategies during model training to offset imbalances in subgroup representation.

Second, regular fairness audits should be embedded into the model's lifecycle. These

audits would evaluate both prediction errors and feature importances across demographic and institutional subgroups, helping to identify emerging biases before they influence outcomes. Periodic re-evaluation ensures that the model adapts to shifts in student populations and institutional dynamics.

Finally, a commitment to explainability must remain central. Tools like SHAP provide granular insights into how individual features contribute to each prediction, offering transparency and accountability. By equipping advisors, faculty, and students with these interpretive tools, the institution fosters an environment where model-driven insights are subject to scrutiny, discussion, and, when necessary, challenge. In this way, predictive analytics can function not as a prescriptive mechanism, but as a collaborative, supportive guide within the broader educational ecosystem.

Conclusion

Bias and risk considerations underscore the paramount importance of prudent deployment. The CGPA prediction model should be employed as an *advisory* tool rather than a deterministic system. By implementing continuous performance auditing, expanding data coverage, and employing explainable AI methodologies, institutions can responsibly harness the model while mitigating potential adverse effects.

4.9 Sample Student Predictions

To illustrate the model's practical relevance, this section provides detailed prediction scenarios for two individual students selected from the test dataset. Each scenario demonstrates how the Random Forest model integrates diverse pre-admission characteristics to generate a CGPA estimation and how SHAP values elucidate the feature contributions underlying each prediction.

4.9.1 Student A – High Predicted Performance

Profile (Partial JSON Input):


```
{
  "gender": 1.0,
  "age_at_entry": 21,
  "average_olevel_grade": -1.20,
  "uce_credits": 19,
```

```
  "alevel_average_grade_weight": 3.5,
  "level": 2,
  "program_id_code": 7,
  "campus_id_code": 0,
  "general_paper": 0,
  "std_dev_olevel_grade": 0.10
}
```


Predicted CGPA: 3.96

Top Feature Contributions (SHAP):

- ↑ `alevel_average_grade_weight` – High A-Level performance significantly raised the prediction.
- ↑ `age_at_entry` – Older entry age correlated with slightly higher predicted performance.
- ↓ `std_dev_olevel_grade` – Low grade variability provided a positive but smaller contribution.
- ↓ `level` – Undergraduate status marginally reduced the score relative to postgraduate benchmarks.

Interpretation: The model predicts Student A as a high performer, driven mainly by excellent A-Level grades and academic consistency. The influence of demographic and institutional factors is minimal. This aligns with expectations and supports the case for scholarship consideration or leadership opportunities.

4.9.2 Student B – Moderate Predicted Performance

Profile (Partial JSON Input):


```
{
  "gender": 0.0,
  "age_at_entry": 24,
  "average_olevel_grade": -0.60,
  "uce_credits": 15,
  "alevel_average_grade_weight": 2.9,
  "level": 2,
  "program_id_code": 10,
  "campus_id_code": 1,
```

```
  "general_paper": 1,
  "std_dev_olevel_grade": 0.35
}
```

Predicted CGPA: 3.85
Top Feature Contributions (SHAP):

- ↑ `average_olevel_grade` – Solid O-Level performance boosted the prediction.
- ↓ `std_dev_olevel_grade` – Higher grade variability lowered the predicted CGPA.
- ↓ `level` – Undergraduate level slightly depressed the score.
- ↓ `gender` – A small negative contribution was observed from demographic influence.

Interpretation: Student B's prediction reflects a strong academic foundation but with some inconsistencies in prior performance. While the model predicts this student will perform well, the downward influence of grade variability suggests areas where additional academic support (e.g., mentoring, targeted study strategies) could help sustain success.

Implications for Academic Advising

These case studies demonstrate how SHAP-based explanations bridge the gap between numerical predictions and actionable guidance:

- High-performing students (e.g., Student A) can be identified for scholarships or leadership development.
- Moderate performers (e.g., Student B) may benefit from targeted mentoring to address risk factors highlighted by the model.
- Advisors gain transparency into why specific predictions were made, fostering trust and accountability in the decision-making process.

In summary, individualized SHAP analyses ensure that model outputs are not treated as opaque scores, but as interpretable insights that guide data-informed academic interventions.

4.10 Conclusion

This chapter has delivered a thorough evaluation of the CGPA prediction model, translating the methodological blueprint introduced in Chapter 3 into concrete empirical findings. The analysis

confirms that the refined Random Forest model—bolstered by a thoughtfully constructed feature set—provides meaningful predictive insight into student performance using only pre-admission data. While the overall R^2 scores remain modest, the model demonstrated solid consistency, particularly when predicting mid-range CGPAs, and its emphasis on academically relevant features aligns well with existing educational expectations.

A key strength of the model lies in its interpretability, achieved through SHAP analysis. Results showed that predictions were primarily influenced by clear academic indicators, especially performance at the O-Level and A-Level. Demographic and institutional factors played a smaller, though still notable, role. This level of transparency is essential; it makes the model's predictions easier to understand and trust, enabling academic staff to confidently use them to inform student support decisions. Additionally, the use of Partial Dependence Plots (PDPs) highlighted complex, non-linear interactions between features, offering further nuance in understanding how various factors shape academic outcomes.

Beyond point predictions, the model's value was extended through band-level estimations and sensitivity simulations. These capabilities turn raw output into actionable insights, helping institutions identify at-risk students early, prioritize interventions, and allocate resources more strategically. In this way, the model is positioned not as a rigid evaluator but as a practical tool for informed decision-making. Fairness assessments further revealed minor discrepancies in prediction accuracy across gender and campus subgroups. While not severe, these differences underscore the importance of continued evaluation to uphold equity and institutional trust.

Importantly, the final version of the model strikes a careful balance between explainability and scalability. It is designed to be integrated into university systems with minimal friction, while still upholding ethical principles such as privacy protection and fairness. Altogether, this chapter demonstrates not only the predictive strength of the CGPA modeling pipeline but also its thoughtful, responsible application within an educational setting.

The insights outlined here serve as a springboard for Chapter 4, which will delve into the final results and offer a closer look at how these predictive insights translate into real academic impact and broader policy implications.

Chapter 5

Conclusion and Recommendations

5.1 Introduction

This chapter brings the study to a close by weaving together the research aims, methodological contributions, core findings, and their broader implications for both academic practice and institutional planning. At its heart, the research set out to build and evaluate a predictive model for cumulative grade point average (CGPA), drawing solely on pre-admission data. Throughout, the focus remained on achieving a thoughtful balance between accuracy, interpretability, and fairness.

Earlier chapters detailed how the process—from rigorous data preprocessing and feature engineering to the implementation of a finely tuned Random Forest model—yielded not only reliable predictions but also practical insights. Importantly, the integration of explainable AI techniques, such as SHapley Additive exPlanations (SHAP) and sensitivity simulations, turned what could have been opaque model outputs into meaningful, accessible guidance for advisors, institutional leaders, and even students themselves.

This final chapter revisits the key findings, offering a clear view of how the model performed, which variables shaped its predictions, and what steps were taken to ensure transparency and equity. It also reflects candidly on the limitations and challenges encountered during the research process, ranging from data availability to fairness concerns across subgroups. An assessment of the model's suitability for deployment within an institutional setting follows, accompanied by practical recommendations for adoption.

Looking ahead, the chapter outlines several directions for future work, aimed at broadening the model's applicability and deepening its impact within higher education systems. These include expanding the feature set, integrating real-time data, and developing mechanisms for continuous monitoring and refinement.

In drawing these threads together—empirical evidence, practical constraints, and strategic vision—this chapter closes the loop on the research journey. It underscores the potential of predictive analytics to support ethical, data-informed decision-making in higher education, while also recognizing the ongoing responsibility to refine and govern such tools with care.

5.2 Summary of Findings

This study set out to explore whether pre-admission data could meaningfully predict university academic performance, measured in terms of cumulative grade point average (CGPA). By applying machine learning techniques to historical student records from Uganda Christian University (UCU), a predictive model was developed, tested, and interpreted. The results clearly show that when pre-admission variables are thoughtfully engineered and modeled, they hold considerable predictive value, opening up new opportunities for data-informed student support and academic planning.

The final model, a tuned Random Forest, demonstrated strong predictive performance. Its coefficient of determination (R^2) and mean absolute error (MAE) reflected a solid alignment between predicted and actual CGPA scores. When benchmarked against simpler models like linear regression, the Random Forest outperformed them, especially within the mid-range of the CGPA distribution, where most students tend to cluster. Although the model slightly underestimated the performance of top-achieving students, its residuals remained stable and free from significant bias, reinforcing its potential as a dependable decision-support tool.

Feature importance analysis confirmed that academic inputs were the primary drivers of prediction. Notably, average O-Level grades, weighted A-Level performance, and UCE credit counts emerged as the most influential predictors of university success. These findings align with long-standing educational theories, reaffirming that a student's past academic performance is the most consistent indicator of future outcomes. While demographic variables, such as gender and entry age, played a smaller role, and institutional characteristics like campus or program added contextual nuance, none of these secondary factors eclipsed the impact of core academic attributes.

Interpretability was a major emphasis throughout the study, addressed through SHapley Additive exPlanations (SHAP) and targeted sensitivity simulations. SHAP analyses at the global level helped uncover which features consistently influenced predictions across the entire dataset. On an individual level, SHAP breakdowns made the model's output more transparent, offering student-specific explanations that showed how the same variable could influence different students in distinct ways depending on context. The inclusion of what-if scenarios—such as simulating improved O-Level performance or alternate program enrollments—added further depth, making the model not just predictive but also prescriptive.

To evaluate fairness, subgroup analyses were conducted across gender, campus, and academic level. For the most part, the model performed equitably, though a few variations were noted. Male students, for instance, exhibited slightly higher prediction errors than their female counterparts, and predictions were less stable for smaller campuses, likely due to underrep-

resentation in the training data. While these discrepancies were not especially large, they highlight the importance of continued monitoring to safeguard equity, especially as the model is applied to future cohorts.

The model's real-world applicability was further demonstrated through the use of performance bands and program fit simulations. By categorizing students into broad CGPA prediction tiers—high, moderate, and low—the results became easier to interpret and apply. This stratification allows for tailored interventions: mentoring for students at risk, enrichment opportunities for high performers, or extra support for those in the middle. Meanwhile, the program fit simulations offered a forward-looking tool for exploring how students might fare in alternative academic paths, supporting both personalized advising and institutional planning.

In conclusion, this study confirms that predictive modeling—when combined with interpretability tools and fairness evaluations—can play a meaningful role in academic decision-making. The model not only delivers accurate forecasts but also creates a framework for transforming those predictions into targeted, data-informed actions that benefit both individual learners and the institution as a whole.

5.3 Limitations and Challenges

Although the results of this study are encouraging, it's important to acknowledge several limitations that shape how the findings should be interpreted and inform future enhancements. These challenges span multiple dimensions, from data quality and subgroup imbalances to algorithmic trade-offs, fairness concerns, and real-world deployment issues.

A primary constraint stemmed from the quality and completeness of the data. The dataset, drawn from historical university records, included missing values, inconsistencies, and, in some instances, incomplete academic profiles. Despite applying a thorough preprocessing pipeline including data cleaning, imputation, and feature engineering, some residual inaccuracies likely remained and may have affected the model's robustness. Additionally, the exclusive reliance on pre-admission data means that potentially valuable predictors, such as first-year academic performance, attendance records, or socioeconomic indicators, were not included. These omitted factors might have enriched the model's predictive accuracy and contextual relevance.

Another challenge lay in the representation of student subgroups. Some academic programs and campus locations were significantly underrepresented in the dataset, in some cases with fewer than ten students. This lack of balance likely introduced volatility in the model's predictions for these groups, limiting its ability to generalize effectively across the broader university population. As a result, predictions for students from these smaller or more remote subgroups should be approached with caution, as they may be more prone to noise or bias.

In terms of model selection, a conscious trade-off was made between accuracy and interpretability. The Random Forest algorithm was chosen for its strong predictive capability, but as an ensemble model, it lacks the transparency of simpler alternatives like linear regression. While tools such as SHAP were employed to make the model's inner workings more accessible, the complexity of its structure can still pose a barrier to understanding, particularly for stakeholders without a technical background. This underscores the ongoing need to balance predictive performance with ease of use in academic decision-making contexts.

Fairness also emerged as an important consideration. Although fairness audits suggested overall consistency across most demographic and institutional subgroups, slight disparities were detected, particularly in error rates between male and female students, as well as among smaller campus populations. While these differences were not large, they highlight the importance of regular monitoring to guard against the unintentional amplification of systemic biases. Without continuous evaluation and periodic retraining, there's a real risk that such disparities could become entrenched over time.

Finally, bringing the model into a real-world academic environment presents its practical hurdles. Integrating predictive analytics into the university's existing Management Information System (MIS) would require secure and efficient data pipelines, dependable preprocessing routines, and strong institutional support. Beyond the technical demands, the ethical deployment of such a model hinges on clear governance structures, well-designed user training, and transparency about both the model's capabilities and its constraints.

Taken together, these limitations do not diminish the study's contributions but rather provide a realistic lens through which the findings should be viewed. The model performs well on a technical level, but its long-term success will depend on filling data gaps, improving subgroup representation, ensuring fairness, and building operational systems that support responsible deployment. Recognizing and addressing these challenges is crucial to ensuring that future versions of the model can serve students and institutions both effectively and equitably.

5.4 Deployment Readiness

Beyond developing a technically sound CGPA prediction model, a key goal of this research was to assess whether the model is practically deployable within the day-to-day operations of a university environment. This section explores the real-world readiness of the model, considering both technical and ethical aspects, particularly how it might be integrated into the university's Management Information System (MIS) and whether such integration is feasible in practice.

The final Random Forest model was prepared as a lightweight, modular prediction service suitable for production use. Model serialization was handled using joblib, which allowed the

trained model to be saved efficiently and reused without retraining. It was then wrapped within a RESTful API built using the Flask web framework. This API setup allows the model to accept structured JSON inputs, process predictions in real time, and return results that include not just the predicted CGPA but also additional metadata, such as the contribution of individual features to the prediction.

The flexibility of this architecture is one of its strengths. The API can be deployed as a microservice on-premises or hosted within a cloud-based environment, depending on the institution's IT infrastructure and data governance preferences. This means the model can be integrated into existing academic workflows with relatively minimal disruption, supporting a range of use cases from individual academic advising to high-level institutional planning.

Integration with the university's MIS was identified as a key factor in determining deployment success. Two integration pathways were considered. The first is *batch mode*, where predictions are generated periodically using exported admissions data and then imported back into the MIS for review by advisors. This approach requires minimal system changes and offers a low-risk starting point for pilot deployment. The second is *real-time integration*, which links the model directly to the MIS during student registration or advising sessions, enabling instant predictions and risk classification. While more resource-intensive, real-time integration holds greater potential for immediate academic interventions.

To comprehensively assess deployment readiness, a feasibility matrix was developed, evaluating the model against technical, ethical, and institutional criteria. Table 5.1 summarizes this assessment.

Table 5.1: Deployment Feasibility Matrix

Component	**Readiness**	**Integration Difficulty**	**Notes**
Model performance	High	Low	Achieved strong results (MAE < 0.3) with robust validation
Real-time prediction engine	Ready	Moderate	Flask API is deployable but requires live data streaming and monitoring
Data availability	Partial	Medium	Some features require preprocessing pipelines to ensure consistency
Privacy compliance	Compliant	Low	No personally identifiable data is used; aligns with ethical data handling standards
Fairness and bias checks	Implemented	Low	Audited across gender, campus, and level; continuous monitoring still needed
Advisor interface	Pending	High	Requires development of user-friendly dashboards for frontline use
Institutional adoption	In Progress	Moderate	Requires pilot testing, staff training, and endorsement by leadership

The results of this evaluation indicate that the model is technically mature and can be deployed as an advisory tool with minimal modification. However, successful implementation requires addressing several non-technical considerations. A user-friendly advisor interface must be developed to make predictions accessible and interpretable by non-technical staff. Further-

more, institutional adoption hinges on stakeholder buy-in, supported by clear policies governing model use, transparency, and student data privacy.

From an ethical perspective, the model has been designed with fairness and explainability in mind, incorporating subgroup audits and SHAP-based interpretations to ensure accountability. Despite the benefits of predictive modeling in educational settings, its deployment entails ongoing responsibilities. The university must commit to continuous monitoring of model performance, periodic retraining with updated data, and regular fairness evaluations to mitigate the potential emergence of unintended biases over time.

In summary, the CGPA prediction model exhibits a high level of readiness for deployment as a decision-support system. With further efforts dedicated to advisor-facing tools, data integration pipelines, and institutional governance, it can be seamlessly integrated into UCU's academic processes, thereby enhancing both student support and policy formulation.

5.5 Institutional Adoption Strategy

Successfully deploying the CGPA prediction model goes well beyond its technical implementation; it also demands a thoughtful, structured adoption strategy that aligns with the university's academic culture, policy goals, and the practical needs of its stakeholders. For Uganda Christian University (UCU), the model presents an opportunity to embed data-driven insights into everyday decision-making, supporting both students and administrators in meaningful ways.

One of the most immediate and impactful areas for application is academic advising. By integrating the model's predictions into faculty advising dashboards, advisors can access early academic risk profiles, either at the time of student application or during routine counseling sessions. This allows for the timely identification of students who may be at moderate risk of academic underperformance. Advisors can then design targeted support strategies, such as mentorship pairings, customized study plans, or access to academic resources, tailored to each student's needs. On the flip side, students who show high potential might be encouraged to pursue honors tracks, leadership roles, or competitive scholarship opportunities. In shifting advising from a reactive model to a more anticipatory one, the institution cultivates a proactive, student-centered academic environment.

Beyond advising, the model holds promise for enhancing student onboarding and orientation. Providing new students with personalized academic outlooks during orientation can help them understand their likely strengths and areas that may need extra attention. This information, when paired with support structures like peer mentorship, bridging courses, or skills workshops, sets the tone for a more prepared and motivated student body. Engaging students early with these insights can foster a sense of ownership over their learning journey

and contribute to stronger academic performance down the line.

The model also has a natural role to play in early warning and intervention systems (EWS). By establishing a baseline prediction at the time of admission, UCU can monitor how students' actual academic progress compares to expectations. When performance begins to diverge significantly, automated alerts can prompt timely interventions—whether through academic counseling, supplemental instruction, or psychosocial support. Integrating the model into an existing EWS ensures that potential academic challenges are addressed before they escalate into serious issues.

At a broader strategic level, the model can inform policy development and academic planning. Aggregated prediction data can reveal trends across programs, identifying areas where students consistently face academic difficulty. These insights can guide curriculum revisions, highlight gaps in institutional support, or even suggest the need for faculty development in specific departments. Additionally, disaggregated model outputs can support equity audits, helping the university detect and respond to systemic disparities across campuses, gender, or other demographic variables. Armed with this knowledge, university leadership is better equipped to allocate resources strategically and refine recruitment or retention initiatives.

To ensure a smooth and responsible rollout, a phased adoption strategy is recommended. The first phase should involve pilot implementation in selected faculties or campuses. This allows the university to test the system in real-world conditions, gather feedback, and refine workflows based on user experience. In the second phase, targeted training should be conducted for academic advisors, registry staff, and IT personnel to ensure they can interpret and use the model's outputs appropriately. Throughout all phases, ethical governance must remain a central concern, with clear institutional policies addressing data privacy, fairness, and the limitations of predictive models. Only after successful pilot validation and full stakeholder buy-in should full-scale integration into the university's MIS and advising systems be pursued.

In sum, adopting this model institutionally requires more than just plugging it into existing systems—it demands strategic foresight, responsible governance, and genuine stakeholder engagement. By rolling out the system gradually, keeping ethics at the forefront, and continuously monitoring its impact, UCU stands to benefit from a powerful tool that enhances academic support, promotes institutional fairness, and supports smarter, evidence-based decision-making across the board.

5.6 Future Work

While this study has demonstrated the promise of predictive analytics within higher education, several opportunities remain for extending and refining the approach. Future work should aim

not only to strengthen the model's predictive capacity but also to enhance its usability and ensure that it continues to serve ethically and effectively in evolving academic contexts.

One particularly exciting direction is the development of *real-time prediction systems*. Currently, the model operates in a batch-processing mode, generating predictions at set intervals based on static data snapshots. Transitioning to a real-time setup would allow CGPA forecasts to be produced instantly, during online admissions, course enrollments, or even live advising sessions. A system designed this way could plug directly into UCU's existing Management Information System (MIS) and Learning Management System (LMS), supporting just-in-time interventions when they're likely to be most effective. Moreover, such an architecture would enable the model to update dynamically, incorporating fresh data—like ongoing coursework or formative assessments—as soon as it becomes available.

Another valuable extension lies in exploring *GPA trajectory modeling*. Rather than predicting a single cumulative CGPA at graduation, future iterations could be designed to forecast academic performance across multiple milestones, such as by semester or academic year. This shift would offer a more nuanced, longitudinal view of student progress, highlighting key turning points where academic decline—or growth—is likely to occur. By capturing these trends early, advisors could respond proactively, offering timely support before performance dips become entrenched or difficult to reverse.

Enhanced explainability is another crucial area for future work. While this study incorporated SHAP-based interpretations to make predictions transparent, more user-friendly explanation mechanisms could be developed. Personalized breakdowns that explicitly state which factors most influence an individual's prediction would increase trust and usability for students and advisors. Additionally, faculty-facing summaries that communicate model logic in non-technical language could further encourage adoption. Advanced fairness-aware explainability tools could also be integrated to monitor bias and ensure ethical use over time.

To complement these enhancements, *interactive dashboards and visual analytics* should be developed. Currently, model outputs are presented in tables and static figures, which, while informative, may not fully support decision-making in operational settings. A well-designed dashboard would allow stakeholders to explore predictions interactively, filter data by program, gender, or campus, and visualize risk patterns across cohorts. For advisors, dashboards could provide early warning alerts coupled with recommended interventions. Administrators could offer aggregated insights into institutional performance trends, thereby informing strategic policy decisions.

Lastly, the predictive power of the model can be strengthened by *expanding the feature set and data sources*. While this study demonstrated that pre-admission data alone can yield accurate predictions, incorporating additional variables, such as first-year coursework results,

attendance patterns, LMS engagement metrics, and socioeconomic indicators, would provide a more comprehensive understanding of student performance. These enhancements would not only increase model accuracy but also uncover deeper contextual factors that influence academic outcomes.

In summary, future work should focus on transitioning the model from a high-performing research prototype to a fully adaptive, transparent, and integrated academic decision-support system. By embracing real-time analytics, trajectory forecasting, enhanced explainability, interactive visualization, and richer data inputs, the CGPA prediction framework can evolve into a powerful tool that empowers both educators and students to shape their academic success proactively.

5.7 Conclusion

This research has shown that it is both feasible and valuable to apply machine learning methods, specifically a Random Forest approach, to predict university academic performance using only *pre-admission* data. The model developed in this study was not only able to deliver a meaningful level of predictive accuracy, but it also remained interpretable and ethically grounded throughout. By incorporating SHAP-based explanations, sensitivity simulations, and band-level classifications, the model moved beyond mere prediction to offer insights that academic stakeholders can act on with confidence and clarity.

The implications of these findings are substantial, particularly for academic practice at Uganda Christian University, though they may resonate far beyond it. The model serves as a practical decision-support tool, capable of enriching advising processes, enabling earlier and more targeted interventions, and guiding institutional resource allocation based on data rather than intuition. It identifies both students at academic risk and those with exceptional potential, creating space for proactive measures that can elevate overall student success and support the university's strategic goals. Notably, the inclusion of fairness checks and bias detection processes underscores a strong commitment to using predictive analytics responsibly and equitably within educational settings.

On a broader level, this study contributes to the evolving field of *educational data mining*, offering a concrete example of how predictive models can be thoughtfully designed to align with institutional priorities, fairness standards, and the need for transparency. The methodology and results presented here provide a strong foundation for future innovation—whether through *real-time prediction systems*, *GPA trajectory modeling*, or the integration of richer, more diverse datasets. Together, these next steps have the potential to evolve predictive tools from passive support mechanisms into dynamic, student-centered components of adaptive learning systems.

In sum, this work represents a meaningful step toward leveraging data to inform academic planning and enhance student outcomes. It reaffirms that with careful design, machine learning can go beyond mere forecasting; it can empower institutions, inform policy, and, most importantly, support students in reaching their fullest academic potential.

References

Ajani, O. A., Gamede, B., & Matiyenga, T. C. (2024). Leveraging artificial intelligence to enhance teaching and learning in higher education: Promoting quality education and critical engagement. *Journal of Pedagogical Sociology and Psychology*, *7*(1), 54–69.

Bacus, J. A., & Cascaro, R. (2024). Impact of predictive learning analytics in higher education: A systematic literature review. *2024 13th International Conference on Educational and Information Technology (ICEIT)*, 313–318.

Chibaya, C., Whata, A., Madzima, K., Rudolph, G., Verkijika, S., Makhoere, L., & Mosia, M. (2022). A scoping review of the "at-risk" student literature in higher education. *bioRxiv*, 2022–07.

Dadwal, S., Haq, A., Jamal, A., & Nawaz, I. (2021). Value of data as a currency and a marketing tool. In *Strategy, leadership, and ai in the cyber ecosystem* (pp. 381–398). Elsevier.

Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, 1–33.

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. *IOP conference series: materials science and engineering*, *928*(3), 032019.

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers' role and perspective. *Educational technology research and development*, *67*(5), 1273–1306.

Ismanto, E., Ab Ghani, H., Saleh, N. I. M., Al Amien, J., & Gunawan, R. (2022). Recent systematic review on student performance prediction using backpropagation algorithms. *Telkomnika (Telecommunication Computing Electronics and Control)*, *20*(3), 597–606.

Kamal, N., Sarker, F., Rahman, A., Hossain, S., & Mamun, K. A. (2024). Recommender system in academic choices of higher education: A systematic review. *IEEE Access*, *12*, 35475–35501.

Karim-Abdallah, B., Junior, M. A., Appiahene, P., Harris, E., & Binful, D. (2025). Application of machine learning algorithms in predicting academic performance of students in higher education institutes (heis): A systematic review and bibliographic analysis. *African Journal of Applied Research*, *11*(1), 536–559.

Kunjumuhammed, S. K. (2024). Artificial intelligence in addressing educational inequality dimensions in higher education institutions (heis): A critical review. *Risks and Challenges of AI-Driven Finance: Bias, Ethics, and Security*, 146–164.

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Vaclavek, J., & Wolff, A. (2015). Ou analyse: Analysing at-risk students at the open university. *Learning analytics review*, 1–16.

Maphosa, M., & Maphosa, V. (2020). Educational data mining in higher education in subsaharan africa: A systematic literature review and research agenda. *Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications*, 1–7.

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*, 100152.

Mpofu, S., & Chasokela, D. (2025). Data-informed decision-making: Using analytics to drive strategic management in higher education. In *Building organizational capacity and strategic management in academia* (pp. 103–138). IGI Global Scientific Publishing.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237.

Nti, I. K., Umar Bawah, F., Quarcoo, J. A., & Kalos, F. (2022). A bibliometric analysis of soft computing technology applications trends and characterisation in educational research: Africa. *Africa Education Review*, *19*(3), 55–77.

Nur, N. (2021). *Developing temporal machine learning approaches to support modeling, explaining, and sensemaking of academic success and risk of undergraduate students* [Doctoral dissertation, The University of North Carolina at Charlotte].

Patel, S., & Ragolane, M. (2024). The implementation of artificial intelligence in south african higher education institutions: Opportunities and challenges. *Technium Education and Humanities*, *9*, 51–65.

Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting university student graduation using academic performance and machine learning: A systematic literature review. *IEEE Access*, *12*, 23451–23465.

SIMION, P. C., POPESCU, M. A. M., DUMITRESCU, C.-I., & COSTEA-MARCU, I.-C. (n.d.). An intelligent platform for guiding future students to optimal university programs. *ON VIRTUAL LEARNING*, 283.

Sosa-Alonso, J. J., López-Aguilar, D., Álvarez-Pérez, P. R., & González-Morales, O. (2025). Predicting university dropout: Connecting big data and structural models. *Studies in Higher Education*, 1–18.

Walker, A. N., Huynh, P., Rarey, K., & Adams, N. (2024). Can we predict your performance? assessing the relationship of admissions data to academic performance in gross anatomy of first-year medical students. *Florida Journal of Educational Research*, *61*(3), 101–110.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in psychology*, *12*, 698490.

Appendix A

Thematic Classification of Reviewed Literature

This appendix presents a thematic classification of the reviewed literature used in Chapter 2. The classification was based on focus areas, including predictive modeling, student retention, ethical considerations, and contextual applications in Sub-Saharan Africa.

Table A.1: Updated Classification of Literature Used in Chapter 2 (Sorted by Theme)

Citation Key	Focus Area	Region/Scope
1. CGPA and Academic Performance Prediction		
nur2021developing	Temporal ML models for CGPA prediction, explainability, and stakeholder usability	USA
fahd2022application	Meta-analysis of ML for academic performance, at-risk, and attrition	Global
zhang2021educational	Systematic review of EDM techniques for student performance prediction	China/Global
ismanto2022recent	Systematic review of DL methods for academic performance prediction	Global
alnasyan2024power	DL for predicting student success in virtual learning environments	Global
2. Institutional Analytics in African HEIs		
mpofu2025data	Strategic use of analytics in higher ed in Zimbabwean universities	Africa
ajani2024leveraging	Enhancing quality education via AI; ethics and access issues	Africa
kunjumuhammed2024artificial	AI in addressing inequality in HEIs	Africa
patel2024implementation	AI implementation challenges and frameworks in South African HEIs	South Africa
3. Program Recommendation and Course Selection		
kamal2024recommender	Systematic review of recommender systems in academic advising	Global
algarni2023systematic	Recommender system methodologies and course selection strategies	Global
simionintelligent	Intelligent platform for matching students to suitable programs	Europe
4. Ethical Use of Pre-admission Data		
walker2024can	Admissions metrics predicting performance in medical school	USA
puri2022validity	Pre-matriculation assessment and USMLE Step 1 performance	USA
memarian2023fairness	Systematic review on Fairness, Accountability, Transparency and Ethics (FATE)	Global
5. Student Retention and Dropout Prediction		
shafiq2022student	Systematic review of ML approaches for retention across modalities	Global

Continued on next page

Table A.1 – continued from previous page

Citation Key	Focus Area	Region/Scope
sosa2025predicting	Dropout risk: SEM vs data mining on big data	Europe
colpo2024educational	EDM in dropout prediction, emphasizing implementation gaps	Global
6. Broader Applications of AI in Education		
forero2024techniques	Comprehensive review of ML/AI tools across education levels	Global
moussa2024predictive	AI-powered predictive assessment, ethics and policy recommendations	Global/Arabic contexts
adewale2024impact	AI in ODL settings, calls for process-based predictive framework	Africa