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

SIMON FRED LUBAMBO

J23M19/223

A DISSERTATION SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN, AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF UGANDA CHRISTIAN UNIVERSITY

April, 2026

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

EXECUTIVE SUMMARY

The higher education landscape is increasingly shaped by predictive analytics, which opens new opportunities for student support and institutional planning. While such approaches are well established in Western contexts, their adoption across Sub-Saharan Africa remains limited. This study addresses this gap by developing and evaluating a machine learning model to forecast the final Cumulative Grade Point Average (CGPA) of students at Uganda Christian University (UCU) using only pre-admission academic and demographic data. The model provides an interpretable and fair framework with direct application to proactive academic guidance.

The study drew on historical admission records comprising UCE and UACE results, subject-level scores, and demographic variables. Several machine learning models were trained and evaluated, with hyperparameter optimisation applied to improve performance. The Random Forest Regressor achieved the strongest results, confirming that CGPA prediction is feasible at admission. To enhance transparency, SHAP value analysis and sensitivity simulations were employed, offering both global insights into feature importance and individual-level explanations of predictions.

To facilitate practical application, predictions were categorised into three CGPA bands: high, moderate, and low, enabling targeted interventions. Programme-fit simulations were also tested to examine potential outcomes across alternative academic pathways. Fairness audits conducted across gender, campus, and programme groups showed no significant disparities; however, continuous monitoring is recommended to guard against bias.

For institutional adoption, the model was implemented as an application programming interface (API) and prepared for integration into UCU's Management Information System (MIS). A phased implementation plan was outlined to ensure that technical, ethical, and organisational considerations are addressed.

Overall, this research demonstrates that interpretable and fair machine learning can strengthen academic advising, programme alignment, and planning within Ugandan higher education. By leveraging pre-admission data already available in most universities, predictive analytics can provide actionable insights in resource-constrained environments. The study thus positions predictive modelling as a practical and ethical tool for advancing data-driven decision-making across African higher education.

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: Lubambo.

Date: 29 April 2026

APPROVAL

Title of the Research Thesis:

Predicting Final CGPA Using Pre-Admission Data: Proactive Insights for Academic Excellence 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:	Dr. Innocent Ndibatya
Title:	Head of Department
Department:	Departments of Computing and Technology
Faculty:	Faculty of Engineering, Design, and Technology
Signature:	______________________
Date:	29 April 2026

DEDICATION

I dedicate this thesis to my late father, James Mayanja, whose sacrifice and commitment gave me the opportunity to pursue education.

I also dedicate it to my mother, Margaret Nakiyingi, for her unwavering belief in me and her continued support throughout my academic journey.

To my young daughter, Miley Kirabo Lubambo, may this achievement stand as an assurance of my love for you and as a source of inspiration for your own journey in life.

To my wife, Miriam Namutosi, thank you for your support, encouragement, and belief in me throughout this process.

Above all, I give thanks to God for His guidance, strength, and faithfulness in leading me to this achievement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background to the Study

In the 21st century, data has become a central driver of innovation, shaping decision-making across sectors such as banking, healthcare, and retail Dadwal et al., 2021. Predictive analytics, powered by machine learning, has enabled banks to anticipate financial risks, hospitals to forecast disease outbreaks with greater accuracy, and businesses to optimise supply chains more efficiently. While these advances are well established in other domains, higher education has traditionally been slower to adopt predictive approaches.

Universities are now beginning to implement data-driven methods to provide proactive support to students. Rather than intervening only after failure or withdrawal, institutions can identify risk early and deliver targeted 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 findings suggest that early-warning systems, when embedded in institutional practices, enhance both retention and performance.

However, most predictive models rely heavily on in-programme data such as LMS interactions, coursework submissions, or attendance logs. While effective, these models are inherently reactive because they only act once risk patterns have already emerged. This leaves a critical gap at the admission stage, when programme placement decisions and early support can have the greatest impact. Beyond that, global studies highlight the predictive power of decision trees, random forests, and support vector machines Bacus and Cascaro, 2024; Fahd et al., 2022; Karim-Abdallah et al., 2025. At the same time, concerns about transparency and fairness have prompted calls for interpretable and contextually relevant approaches.

The need for early and 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. Models developed in Western contexts often prioritise accuracy without addressing interpretability, fairness, or the unique data limitations of African institutions. Barriers such as limited digital infrastructure, policy gaps, and ethical concerns around student data use further constrain uptake Nti et al., 2022; Patel and Ragolane, 2024. Locally adapted models remain scarce but are critically needed Chibaya et al., 2022.

Alongside technical challenges, ethical debates highlight the risks of poorly designed predictive systems, which may stigmatise students or reinforce inequalities Memarian and Doleck, 2023. These debates emphasise that predictive systems must balance accuracy, fairness, transparency, and contextual fit. These principles guide this 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 and evaluates a Random Forest model to forecast students' final CGPA. Beyond prediction, the study integrates SHAP-based interpretability and sensitivity analysis to reveal key drivers of academic success, 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 programme placement to data-informed strategies that support proactive advising and improved student outcomes.

In summary, this background traces the global evolution of predictive analytics, examines gaps and ethical debates in educational applications, and positions this study as a locally adapted, transparent, and ethically responsible predictive model for proactive academic advising in Uganda.

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 fully utilise these tools. While UCU collects rich pre-admission data, such as O-Level and A-Level results, programme placements, and demographic information, this data is underutilised for proactive student guidance.

Globally, machine learning models have shown strong capabilities in forecasting academic outcomes such as GPA and graduation likelihood, enabling early interventions that improve student success Pelima et al., 2024. Yet most existing models rely on in-programme indicators, such as LMS activity or coursework submissions, which are inconsistently captured or absent in many African institutions. Furthermore, few approaches integrate interpretability or fairness mechanisms, raising concerns about their ethical and practical application in advising contexts Memarian and Doleck, 2023.

In Sub-Saharan Africa, adoption of predictive analytics remains limited, with only a small fraction of EDM studies addressing the region Chibaya et al., 2022. Academic advising continues to operate reactively, with interventions occurring only after academic challenges or student withdrawals, often at significant academic and psychological cost. Programme placement further compounds this challenge, being shaped largely by subject combinations, institutional traditions, and perceptions of programme prestige, rather than personalised guidance informed by student strengths.

This gap highlights 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 that need by developing and deploying a Random Forest model trained on UCU's historical admission data. By embedding SHAP-based interpretability and fairness assessments, the research demonstrates how predictive analytics can transform academic advising from a reactive to a proactive process in a Ugandan context.

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

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

2. To develop, optimise, and validate machine learning models that can accurately predict final CGPA at the point of admission.

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

4. To assess how the model can support proactive academic advising, early identification of students who may require additional academic support, and evidence-informed programme guidance.

1.4 Research Questions

The following research questions guide this study:

1. Which pre-admission academic, demographic, and programme-related 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?

3. How can model interpretability and fairness evaluation, including SHAP analysis and subgroup fairness assessments, support the ethical and transparent use of CGPA predictions?

4. How can the predictive model support proactive academic advising, early identification of students requiring academic support, and evidence-informed programme guidance?

1.5 Justification of the Study

Predictive analytics in higher education has gained considerable momentum globally, enabling institutions to make data-informed decisions that improve retention, strengthen advising, and optimise resource allocation. Yet in Uganda, and across Sub-Saharan Africa, these tools remain underutilised despite the availability of standardised pre-admission data that could inform proactive advising and programme placement.

Studies confirm the effectiveness of machine learning models in forecasting academic outcomes such as GPA and graduation likelihood Fahd et al., 2022; Pelima et al., 2024. However, most models developed in Western contexts depend on in-programme data that is inconsistently captured in African institutions, and few integrate fairness or interpretability mechanisms Maphosa and Maphosa, 2020; Patel and Ragolane, 2024. This limits their applicability locally and raises ethical concerns around bias and transparency.

This study is justified by the need to bridge these gaps through the development of a context-specific, interpretable, and fair predictive model based solely on pre-admission data. By enabling early CGPA forecasting, it supports proactive advising, early identification of students requiring support, and informed programme placement. Beyond individual benefits, the findings offer institutional value by informing academic policy, supporting programme alignment, and providing a framework for predictive analytics in resource-constrained settings. The study also contributes to global debates on responsible AI in education, addressing both technical and fairness considerations while extending research in an underrepresented African context.

1.6 Scope of the Study

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

Although the study used available pre-admission academic, demographic, and programme-related variables, it did not include school categorisation, such as urban versus rural school background or traditional versus international curriculum pathway. These variables were not available in the extracted MIS dataset. Their absence is acknowledged as a limitation because school context and curriculum background may influence academic preparation and CGPA outcomes. They are therefore recommended for future data enrichment and model refinement.

The predictive models are developed exclusively using admission-time data, including secondary school performance, demographic details (e.g., gender, age), and programme identifiers. The target variable is the final CGPA recorded upon programme completion.

Features derived from in-university performance (e.g., Year 1 or Year 2 GPA) have been deliberately excluded to ensure that predictions rely solely on admission data. The study also incorporates SHAP-based interpretability and subgroup fairness analysis to ensure that outputs are transparent and equitable.

Although the methodology may be adapted elsewhere, the model is tailored to UCU's academic structures and data environment. Application in other institutions would require contextual adjustment. Programme simulations and advising use cases are presented as exploratory demonstrations rather than prescriptive placement recommendations.

1.7 Significance of the Study

This study holds significance at several levels:

- **For Academic Advisors and Students:** The research introduces an early-stage forecasting tool that integrates interpretability (via SHAP) and fairness assessments, enabling advisors to interpret and act on predictions with greater clarity. This allows for the early identification of students who may require additional support and the design of timely interventions.

- **For University Administration:** The study demonstrates how existing records at UCU can be transformed into actionable insights that support proactive advising, evidence-based decision-making, and programme placement. It also provides a framework that can be adapted and scaled within similar institutional contexts.

- **For Research and Innovation:** The study 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, it advances responsible AI in education and establishes a practical benchmark for other institutions seeking to adopt predictive systems suited to their context.

Chapter 2

Literature Review

2.1 Introduction

In today's data-driven world, higher education institutions are increasingly adopting predictive analytics to improve student outcomes, optimise academic planning, and enhance personalised advising. Globally, machine learning and educational data mining (EDM) techniques are widely applied to predict academic outcomes, dropout risk, and long-term achievement. Predictive tools are now common in sectors such as banking and healthcare, and higher education is gradually following suit, using these models to provide early warnings, improve retention, and optimise resources. However, while predictive models have transformed support systems in North America, Europe, and parts of Asia, adoption in Sub-Saharan Africa remains limited due to infrastructural, contextual, and policy challenges.

This chapter reviews existing research on predictive analytics in education, with emphasis on cumulative grade point average (CGPA) prediction using machine learning. It proceeds from global applications to methodological considerations, then examines outcome choice (CGPA), key predictors available at admission, and African institutional contexts. The chapter also considers applications of predictive models for advising and programme alignment, alongside the ethical debates on fairness, transparency, and privacy. It closes by identifying research gaps that this study addresses at Uganda Christian University (UCU).

2.2 Predictive Analytics in Higher Education: Global Perspectives

2.2.1 Defining Predictive Learning Analytics

Predictive learning analytics applies statistical and machine-learning methods to institutional data to forecast student outcomes Namoun and Alshanqiti, 2020. Its focus is on probabilistic estimation from historical patterns, such as predicting CGPA at graduation, rather than simply describing past trends or prescribing fixed interventions. By moving beyond descriptive reporting, predictive models allow institutions to anticipate risk and act before negative outcomes occur. For example, dashboards developed at The Open University in the UK have enabled faculty to anticipate disengagement, thereby reducing attrition rates and improving progression Herodotou et al., 2019. Similarly, studies at the Czech Technical University reported reductions in dropout rates when predictive analytics were embedded into advising workflows Kuzilek et al., 2015. These cases highlight how predictive analytics has moved from experimental to mainstream practice in global higher education.

2.2.2 Trends and Model Applications

Predictive models have been widely applied to estimate outcomes such as GPA, retention, and dropout risk Fahd et al., 2022. Early-warning systems allow advisors to act before failure becomes visible, while other models are used to design personalised study plans, assess long-term academic achievement, and even detect behavioural risk factors associated with mental health Zhang et al., 2021. However, debates increasingly stress that predictive accuracy alone is not sufficient. Models must also be transparent, interpretable, and ethically deployed Memarian and Doleck, 2023. The rise of explainable AI (XAI) techniques, such as SHAP values, has been critical in balancing performance with interpretability. These approaches help institutions move beyond "black-box" predictions, ensuring that advisors and students can trust the rationale behind outputs.

Yet most high-performing models in global contexts depend on in-programme data such as learning management system (LMS) activity, assessment submissions, or attendance logs. While effective in institutions with sufficient resources, these approaches are poorly suited to many African universities where such data is inconsistently captured. This gap highlights the need for admission-only prediction frameworks that can deliver actionable insights from data reliably collected at entry. Accordingly, this study shifts attention from dropout-focused models to CGPA prediction, and from in-programme indicators to admission-time predictors.

2.2.3 Machine Learning and EDM Techniques

The choice of modelling technique is central to the effectiveness and acceptance of predictive analytics in higher education. Random Forests have consistently achieved a balance between predictive accuracy and interpretability, making them particularly suitable where both technical performance and stakeholder trust are required. Support Vector Machines (SVMs) perform strongly in high-dimensional settings but provide little transparency, limiting their value in advising contexts where feature-level explanations are important. Logistic regression remains a common baseline, valued for its simplicity and ease of interpretation, though it often underperforms when relationships between variables are highly non-linear Hashim et al., 2020. Deep neural networks have shown promise in modelling behavioural or time-series data, yet their "black-box" nature has led to reluctance in academic advising and policy contexts, especially in resource-constrained environments Nur, 2021.

Feature selection further shapes model success. Techniques such as LASSO regression and Recursive Feature Elimination (RFE) are often applied to reduce noise and highlight key predictors. In admission-only contexts, however, the most salient variables remain aggregates of O-Level and A-Level scores, subject-level performance, demographic factors, and programme identifiers Ismanto et al., 2022. These predictors combine availability, interpretability, and predictive value. Careful management of multicollinearity among examination aggregates and programme indicators is also necessary to maintain model stability and ensure outputs remain usable in advising rather than serving purely as opaque technical predictions.

2.2.4 Outcome Focus: CGPA vs. Retention vs. Dropout

Most predictive studies in education have focused on dropout or retention, often framed as binary outcomes Sosa-Alonso et al., 2025. These models have proven valuable in reducing attrition, but do not capture the full spectrum of academic trajectories. CGPA, by contrast, offers a continuous and policy-relevant outcome measure. It enables granular profiling of risk across a spectrum, supporting differentiated interventions rather than binary classifications. Furthermore, CGPA underpins critical institutional decisions such as scholarship awards, eligibility for academic honours, and postgraduate admissions. For universities like UCU, where progression and completion rates are tied to both institutional reputation and student futures, CGPA represents a more holistic and actionable measure of long-term academic success. This motivates the present study's emphasis on CGPA as the principal outcome and situates admission-time predictors as key inputs for early guidance.

2.3 Admission-Time Predictors of Academic Performance

2.3.1 Academic Entry Indicators

Academic entry indicators remain among the strongest predictors of university performance. In many systems, O-Level and A-Level results serve as gatekeeping criteria for admission, but they also provide meaningful signals of academic readiness. Total points, subject distinctions, and credit aggregates are widely used by institutions to benchmark preparedness, and empirical studies confirm their correlation with eventual GPA outcomes Ismanto et al., 2022; Zhang et al., 2021. Beyond raw aggregates, stability in performance across subjects may also indicate readiness for self-regulated learning at university. Students with consistent grades across diverse subjects often demonstrate resilience and adaptability, traits that support long-term academic performance. However, evidence remains mixed, and caution is needed when over-interpreting dispersion measures. What is clear is that pre-admission academic indicators capture both the breadth and depth of prior learning, making them essential features in prediction models that aim to forecast CGPA.

2.3.2 Demographic and Programme Variables

In addition to academic history, demographic variables such as gender and age, alongside institutional variables such as programme identifiers or campus locations, contribute to prediction accuracy. Several studies have noted gendered patterns of achievement across disciplines, though these patterns must be interpreted carefully to avoid reinforcing stereotypes. Institutional structures, such as curriculum design or campus-level resourcing, also play roles in shaping outcomes. In this study, tuition fees were deliberately excluded as predictors, given their limited variation within UCU's structure and their low interpretive value. By focusing on variables with clear and plausible links to academic achievement, the modelling process maintains both interpretability and institutional relevance.

2.3.3 Curriculum Structures

While not included as predictors in this study, curriculum structures and programme design shape long-term student trajectories. Heavy course loads in early semesters, rigid prerequisite chains, or limited flexibility can compound academic risk, whereas well-aligned curricula promote smoother progression and improved outcomes Kamal et al., 2024; Mpofu and Chasokela, 2025. For UCU, recognising these downstream influences is essential for interpreting prediction outputs and situating them within broader institutional reforms and support systems.

2.3.4 Admission-Only Prediction Models

Most predictive learning analytics frameworks rely on in-programme variables such as coursework submissions, LMS activity, or semester GPA trends. While powerful, these indicators become available only after students are enrolled, thereby limiting opportunities for proactive interventions at the admission stage. An alternative strand of research has focused on *admission-only models*, which utilise pre-entry academic and demographic records to forecast long-term performance. Such approaches are particularly valuable in contexts where digital infrastructure for real-time data capture is weak, but where secondary school records are systematically collected.

From the post-graduate perspective, Muratov et al. 2017 demonstrated that Random Forest models trained exclusively on pre-admission variables, including prior GPA, entrance exams, and degree background, achieved strong predictive accuracy for PharmD programme outcomes. Their work underscores the viability of admission-only features in building interpretable models that can inform early academic counselling. Similarly, in Sub-Saharan Africa, Myburgh 2019 applied statistical and machine learning methods to South African university admissions data, including school-leaving results and national benchmark tests. The study not only confirmed the predictive value of admission-time variables but also explicitly tested subgroup fairness across race and gender, providing important evidence for the ethical deployment of such models in African higher education.

Other studies in medical and professional education echo the significance of pre-admission indicators. For example, Puri et al. 2022 showed that pre-matriculation assessments could reliably predict later academic performance in pharmacy programmes, offering evidence that admission-stage scores remain valuable for forecasting success. At the same time, Walker et al. 2024 cautioned against an overreliance on such metrics, noting that admission models may oversimplify complex academic trajectories if used rigidly or without contextual interpretation. Their review highlights the need to balance predictive power with transparency and ethical safeguards.

Together, these contributions underscore both the promise and the limitations of admission-only approaches. Despite encouraging results, the literature remains sparse, especially in African contexts, and tends to focus narrowly on specific professional programmes. This study builds on these precedents by modelling final CGPA at Uganda Christian University (UCU) using exclusively admission-time records (UCE, UACE, and demographic data), thereby addressing both a methodological gap and a regional evidence gap in predictive learning analytics.

2.4 Institutional Analytics in African Higher Education

2.4.1 Emerging Analytics Culture

African higher education institutions are beginning to adopt analytics for strategic planning, resource allocation, and academic management. In Zimbabwe, for example, analytics has been integrated into decision-making processes at the university level, demonstrating how data-driven cultures can enhance both academic and administrative outcomes Mpofu and Chasokela,

2025. Yet, the culture of analytics remains nascent, with adoption often limited to pilot projects or donor-funded initiatives. The unevenness of practice underscores the importance of models that are simple to implement, require minimal infrastructure, and provide immediate value to institutional stakeholders.

2.4.2 Infrastructure and Policy Constraints

Despite growing interest, infrastructural and policy limitations continue to constrain adoption. Bandwidth restrictions, limited budgets, and inconsistent IT governance remain common barriers in African HEIs Patel and Ragolane, 2024. In some cases, data ownership and governance are fragmented across faculties, further complicating implementation. These realities accentuate the value of models that avoid dependence on complex LMS or in-semester data and instead leverage admission data, which is consistently collected and centrally stored. For institutions like UCU, this approach provides a feasible entry point for predictive analytics without overburdening infrastructure.

2.4.3 Equity, AI, and Digital Divide

Equity considerations are central to responsible predictive analytics. Variations in internet access, device ownership, and digital literacy create risks of uneven adoption and benefit, potentially reinforcing existing inequalities Ajani et al., 2024. Additionally, biased datasets can amplify disparities if certain groups are underrepresented or misrepresented in training data Kunjumuhammed, 2024. Addressing these risks requires not only technical safeguards but also deliberate fairness evaluations and governance frameworks. In this study, subgroup fairness audits across gender, campus, and programme categories are integrated to ensure predictions remain equitable and to highlight the importance of continuous monitoring. This consolidated ethics thread complements the interpretability practices described earlier, aligning predictive accuracy with fairness and transparency Memarian and Doleck, 2023.

2.5 Applications of Predictive Models in Academic Support

2.5.1 Proactive Academic Advising

Traditional advising often intervenes after problems become visible, when options for remediation are limited. Predictive analytics enables proactive advising by identifying at-risk students early and targeting interventions such as counselling, study skills workshops, or structured peer mentoring. Evidence from institutions in Europe and North America shows that such systems improve retention and student satisfaction. For UCU, proactive advising informed by pre-admission predictions offers a chance to strengthen early academic support without adding significant data collection burdens. In practice, this means translating continuous CGPA predictions into actionable risk bands and combining them with advisor judgement and student preferences.

2.5.2 Programme Recommendation Systems

Programme recommendation systems extend predictive analytics into programme choice and alignment. By analysing historical data, these systems suggest degree pathways where students with similar profiles have succeeded. Evidence from recommender systems in Europe shows promise in guiding students toward programmes aligned with their strengths and goals Kamal et al., 2024; SIMION et al., n.d. In this study, programme-fit simulations are introduced as exploratory applications. Rather than prescribing placements, they provide an evidence-based basis for dialogue between advisors and students, supporting more informed programme selection and reducing mismatches that can hinder long-term success. Sensitivity analyses (e.g., varying A-Level subject combinations within realistic bounds) can further contextualise these simulations for advising.

2.5.3 Integration with MIS

Prior research emphasises the practicality of embedding predictive models into existing information systems. Integration ensures accessibility, aligns with governance frameworks, and facilitates institutional adoption. For UCU, the Management Information System (MIS) provides a natural platform for operationalising predictive analytics. While technical details of integration are presented in later chapters, the literature consistently highlights that embedding models within existing workflows increases sustainability and institutional trust. In addition, MIS integration supports ongoing model monitoring, critical for fairness drift checks and recalibration as admissions cohorts and programme structures evolve.

2.6 Ethical Considerations and Fair Use

2.6.1 Privacy and Data Protection

Predictive analytics relies on sensitive student records, raising legitimate concerns over privacy and data protection. Academic records, demographic details, and personal identifiers must be anonymised to avoid exposing individuals. Literature stresses best practices such as pseudonymisation, access controls, and role-based permissions to safeguard data integrity. Beyond technical measures, institutions must cultivate trust by ensuring transparency in how predictions are generated and used. In this study, these principles guide implementation, with specific safeguards outlined in Chapters 3 and 5.

2.6.2 Fairness, Accountability, Transparency, and Ethics (FATE)

The FATE framework provides a widely recognised lens for assessing responsible AI use in education Memarian and Doleck, 2023. Models must ensure fairness across subgroups, maintain accountability in design and deployment, provide transparency in how outputs are generated, and adhere to ethical principles that prioritise student well-being. Predictive outputs should be communicated as probabilistic estimates rather than fixed certainties, supporting human judgement rather than replacing it Walker et al., 2024. This framing is especially critical in advising contexts, where rigid or opaque predictions risk stigmatising students rather than empowering them.

2.7 The Ugandan Context and Case for UCU

Over the last two decades, Ugandan universities have invested in electronic record-keeping and student information systems. Uganda Christian University (UCU) maintains structured records of admissions, O-Level and A-Level results, programme placements, and final CGPA outcomes, providing a robust dataset for predictive modelling. While fragmented and inconsistent data environments remain common in the region, UCU's system enables context-specific analysis and offers a realistic platform for piloting predictive tools.

A persistent challenge in Ugandan higher education is the absence of structured guidance during programme placement. Students are often allocated programmes based primarily on A-Level subject combinations or perceptions of prestige, with limited consideration of individual strengths or long-term potential. At UCU, internal audits have shown that many students transfer programmes within their first year or struggle with progression despite adequate admission records being available. Predictive analytics could therefore play a transformative role by offering data-driven advising at the point of admission, reducing misalignments, and promoting better academic trajectories.

This research aligns closely with UCU's strategic priorities of student success, evidence-based planning, and digital transformation. By piloting a CGPA prediction model within its MIS, UCU may contribute to the regional discourse on responsible adoption of predictive analytics. While caution is warranted, the institution's relatively strong data infrastructure provides a unique opportunity to lead in contextualised, ethical, and impactful application of predictive models in Sub-Saharan Africa.

2.8 Synthesis and Research Gaps

The reviewed literature highlights several gaps that directly motivate this study. First, few studies explicitly target the final Cumulative Grade Point Average (CGPA) as the primary outcome. Most existing frameworks focus on dropout or retention risks Shafiq et al., 2022; Sosa-Alonso et al., 2025. While valuable, these binary outcomes provide limited granularity for long-term planning compared to CGPA, which forms the basis for scholarships, academic honours, and postgraduate admission. Addressing this outcome gap informs the study's first research question, which investigates the feasibility of predicting final CGPA from pre-admission data, and also relates to the fourth research question on the use of predictions in supporting academic advising.

Second, although admission-only models have demonstrated potential in professional education and selected African contexts Muratov et al., 2017; Myburgh, 2019; Puri et al., 2022, their application remains narrow in scope and uneven in regional coverage. Most work is concentrated in medical and pharmacy education or within well-resourced systems, leaving little evidence for broader undergraduate contexts in Sub-Saharan Africa. Beyond that, as Walker et al. 2024 cautions, admission-stage metrics risk oversimplification if not accompanied by interpretability and fairness safeguards. This gap motivates the second research question, which evaluates the predictive strength of pre-admission features, and the third research question, which addresses the ethical and interpretability dimensions of such models.

Third, interpretability and fairness remain underexplored in African higher education. Although global debates highlight the importance of transparency in predictive analytics Memarian and Doleck, 2023, few studies in the region have integrated subgroup fairness checks or interpretability tools such as SHAP. This limits both the practical and ethical legitimacy of predictive models in these settings. Responding to this gap aligns directly with the third research question, which examines how fairness and interpretability can be embedded in predictive frameworks.

Fourth, there are limited demonstrations of advising-ready artefacts tailored to local contexts. Many studies stop at predictive accuracy, rarely extending to outputs that can be directly used in advising, such as CGPA banding, programme-fit simulations, or sensitivity analyses. This gap informs the fourth research question, which explores how predictive insights can be translated into actionable tools for academic support and programme guidance.

Finally, institutional adoption pathways remain scarcely addressed. Very little literature discusses how predictive models can be embedded within Management Information Systems (MIS) or aligned with governance frameworks in resource-constrained African universities. This gap informs the fourth research question, which considers how predictive outputs can support proactive advising, early identification, and evidence-informed programme guidance within institutional workflows.

In sum, this study responds to these gaps by (i) modelling final CGPA from pre-admission data at Uganda Christian University (UCU), (ii) prioritising interpretability through SHAP and subgroup fairness evaluation, (iii) demonstrating advising-ready applications such as CGPA banding and programme-fit simulations, and (iv) outlining a deployment pathway integrated with UCU's MIS and governance structures. Beyond its institutional application, the research contributes to the global discourse on predictive learning analytics by presenting evidence from an underrepresented region, advancing the case for admission-only prediction models, and illustrating how fairness and interpretability can be operationalised in real-world higher education contexts. In doing so, it situates UCU's case as both a local innovation and a transferable framework for universities across Sub-Saharan Africa and comparable resource-constrained environments worldwide.

Chapter 3

Methodology

3.1 Introduction

Predictive analytics has become an essential instrument in higher education, enabling institutions to shift from reactive responses to proactive decision-making. At Uganda Christian University (UCU), this paradigm offers a unique opportunity to leverage its extensive repository of pre-admission data to forecast long-term student performance, particularly final Cumulative Grade Point Average (CGPA). Early forecasting of CGPA equips universities with actionable insights for targeted academic advising, improved resource allocation, and enhanced institutional planning.

This chapter presents the methodological foundations of the study. The design follows a quantitative approach that employs supervised machine learning techniques to uncover patterns in historical student data. By integrating O-Level and A-Level academic histories with demographic and institutional placement factors, the methodology seeks to construct a predictive framework that balances accuracy, interpretability, and fairness. The chapter is organised into sections covering the research design, study context, data collection and ethical handling, preprocessing and feature engineering, model development and validation, evaluation metrics, and limitations. Together, these steps establish the foundation for the results presented in Chapter 4.

3.2 Research Design

3.2.1 Research Approach

This study adopts a quantitative research design, employing supervised machine learning for predictive modelling. The problem is formulated as a regression task, with the final CGPA as the continuous target variable. Regression is preferred over classification since it provides granular predictions of academic outcomes, reflecting variations that may influence scholarships, honours, and postgraduate opportunities. Continuous predictions, therefore, offer more practical value for institutional planning and personalised advising.

Supervised learning is appropriate given the availability of labelled historical data in which pre-admission predictors (O-Level and A-Level grades, demographics, and institutional placement) are paired with observed CGPA outcomes. This pairing enables the algorithms to learn underlying relationships and apply them to new admissions. The emphasis is not only on predictive performance but also on interpretability and ethical application, ensuring the outputs function as supportive tools for advising rather than exclusionary labels.

3.2.2 Study Context and Scope

The study is situated at Uganda Christian University (UCU), a private institution in Uganda with established investments in digital infrastructure and a comprehensive Management Information System (MIS). Unlike many institutions in Sub-Saharan Africa with fragmented or incomplete records, UCU maintains structured data covering admissions, programme placements, and final academic outcomes, providing a suitable context for predictive analytics.

The dataset covers students admitted between 2009 and 2023. Only students with complete pre-admission information and the final CGPA were included. This scope ensures methodological completeness by focusing on predictors available at the point of entry, thus aligning to inform interventions before academic challenges manifest. The inclusion of more than a decade of records strengthens generalisability while maintaining relevance to current institutional demands.

3.3 Data Collection and Ethical Handling

3.3.1 Source and Nature of Data

Data was obtained from UCU's MIS, which has maintained digital records for over two decades. The first extraction produced 95,064 student profiles from academic years 2000–2023. After filtering for completeness, 4,979 student records from the 2009–2023 academic years were retained. These records included 41 structured variables spanning demographic information, O-Level and A-Level academic outcomes, and institutional placement details. Only variables observable at admission were retained to simulate real advising conditions, ensuring that the modelling outcomes are institutionally applicable.

The dependent variable was the final CGPA, measured on a 1.0–5.0 scale and recorded at graduation. Ethical safeguards were applied throughout: all personally identifiable information (e.g., student IDs) was removed, and pseudonymisation was ensured before analysis in line with UCU's governance policies.

3.3.2 Attribute Categories

The refined dataset was organised into four categories, each capturing a different aspect of the student profile:

- **Demographic attributes:** age at entry, gender, marital status, nationality, and international status.
- **Academic background:** O-Level and A-Level subject grades, credits, distinctions, subject counts, aggregate scores, and General Paper performance.
- **Institutional placement:** programme codes, curriculum identifiers, level of study, campus, session, and year of admission.
- **Target variable:** final CGPA as the dependent outcome.

A tuition fee variable was initially considered but later excluded due to its limited predictive contribution and potential to encode socioeconomic bias.

The dataset did not include structured prior-school variables, such as urban or rural school location, school ownership type, or traditional versus international curriculum pathway. These variables were therefore excluded from the modelling pipeline and are recommended for future data enrichment because they may provide useful context for interpreting CGPA predictions.

3.3.3 Data Cleaning and Storage

The cleaning process involved removing incomplete records, eliminating identifiers, and resolving inconsistencies. Features with excessive missingness (greater than 90%) were excluded. Imputation was applied strategically: medians for numeric variables and modes for categorical ones. Outliers in tuition and age were managed through log transformations and ordinal binning, ensuring preservation of their interpretive value. All records were stored in a pseudonymised dataset compliant with UCU's ethical and data governance standards.

3.4 Data Preprocessing and Preparation

The preprocessing pipeline transformed the cleaned dataset into a model-ready feature matrix. Key steps included imputation of missing values, outlier handling, categorical encoding, feature scaling, and creation of derived variables. Figure 3.1 summarises this workflow, highlighting how raw MIS records were systematically converted into an analysable feature set.

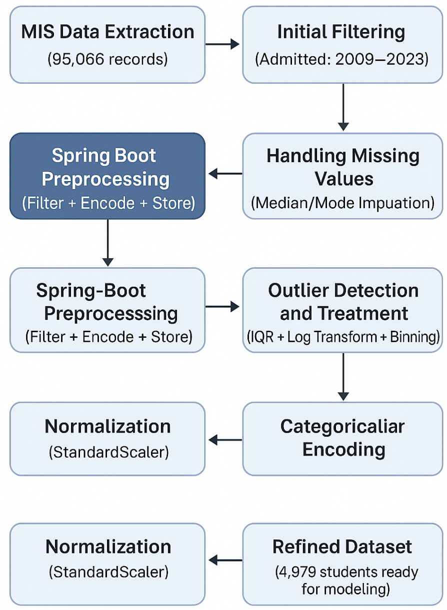

Figure 3.1: Data preprocessing pipeline

3.5 Feature Engineering and Selection

Feature engineering was used to capture underlying patterns in academic records, while selection ensured efficiency and fairness. Domain knowledge guided the derivation of variables such as aggregate best sums of O-Level subjects and the stability index of secondary school grades. Transformations, including log-scaling of tuition and binned age variables, improved the representation of skewed distributions, though tuition was excluded from the final model on fairness grounds.

Correlation analysis and variance inflation factors (VIF) were applied to reduce redundancy among overlapping variables. Recursive Feature Elimination (RFE) with a Random Forest estimator further ranked predictors by importance. The final feature set consisted of 23 attributes covering demographics, academic background, and institutional placement. Table 3.1 provides a compact overview of these retained predictors.

Table 3.1: Final Feature Set for CGPA Prediction

Feature	Description
Age at entry	Age of student on admission
Average O-Level grade	Mean grade across O-Level subjects
O-Level credits	Count of credit passes
O-Level distinctions	Count of distinctions
A-Level average grade weight	Weighted mean grade at A-Level
A-Level weak grades count	Number of low-grade outcomes
A-Level grade variability	Standard deviation of A-Level grades
High school performance stability index	Consistency across O-Level and A-Level
General Paper completion	Indicator of A-Level General Paper completion
Gender	Male/female (binary encoded)
Nationality	Local vs. international
Marital status	Encoded categorical attribute
Programme ID	Programme of study at entry
Curriculum ID	Curriculum identifier
Campus ID	Campus of enrolment
Level of study	Diploma/undergraduate
Year of entry	Year of admission
Years of O-Level/A-Level completion	Intake indicators

3.6 Modelling Pipeline

The predictive framework followed a structured machine learning pipeline. The dataset was divided into training (80%) and testing (20%) subsets. Within training, a 10-fold cross-validation was applied to reduce variance and mitigate overfitting. Candidate algorithms included linear regression, ridge regression, Random Forest, and XGBoost. Random Forest was ultimately selected due to its ability to capture non-linear relationships, handle mixed-type predictors, and provide interpretable feature importance measures.

Hyperparameter tuning was performed using grid search, optimising depth, number of trees, and minimum samples per leaf. Figure 3.2 illustrates the conceptual workflow from feature preparation to model validation.

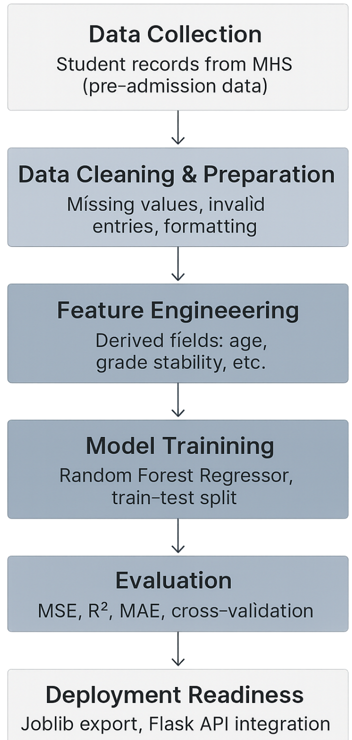

Figure 3.2: Conceptual workflow of the modelling pipeline.

3.7 Evaluation Metrics and Error Analysis

Model evaluation employed three complementary metrics:

- **Mean Absolute Error (MAE):** measures the average absolute deviation between predicted and actual CGPA, directly reflecting predictive accuracy for advising contexts.
- **Root Mean Squared Error (RMSE):** penalises larger errors more heavily, highlighting extreme deviations that could distort institutional planning.
- **Coefficient of determination (R^2):** captures overall explanatory power, indicating how much variance in final CGPA is explained by admission-time predictors.

Residual distributions were also examined to detect systematic bias across subgroups (e.g., gender, programme, campus). These analyses are reported in Chapter 4 as part of the fairness evaluation.

3.8 Limitations and Assumptions

The methodology is subject to several limitations. First, the dataset included only students with recorded final CGPA, excluding dropouts and transfers, which introduces survivorship bias. Second, imputation methods assume data was missing at random, which may not hold in all cases. Third, encoding nominal variables imposes artificial order, potentially distorting relationships. Fourth, the model assumes relative stability in grading practices and curricula across 2009–2023, though reforms may have introduced unobserved variation. Finally, while Random Forest supports feature importance measures, it still functions as an ensemble model requiring post-hoc explanation (e.g., SHAP) for full interpretability.

3.9 Conclusion

This chapter has outlined the methodology for developing and validating a predictive model of student CGPA at UCU. Through careful data preparation, feature engineering, model selection, and validation, the approach balanced predictive rigour with interpretability and fairness. Figures summarised the preprocessing and modelling pipelines, while a feature summary table ensured transparency in predictor selection. Evaluation metrics were explicitly tied to research questions, ensuring alignment with study objectives. The next chapter (Chapter 4) presents the results, including model performance, fairness evaluation, and advising-ready artefacts.

Chapter 4

Results and Analysis

4.1 Introduction

This chapter presents the empirical findings of the study, translating the methodological design in Chapter 3 into measurable outcomes. The analysis evaluates how well pre-admission information predicts final Cumulative Grade Point Average (CGPA), explains the contribution of individual predictors, examines subgroup fairness, and demonstrates uses of the model for academic counselling. Results are organised as follows: model performance and error characteristics; global and local interpretability (including principal component structure and SHAP analyses); band-level predictions to support institutional action; fairness and subgroup evaluation; sensitivity and *what-if* simulations; and illustrative student-level cases. The chapter concludes with a synthesis of implications for academic advising and institutional planning.

4.2 Model Performance Evaluation

4.2.1 Predicted vs. Actual CGPA

Figure 4.1 plots predicted against observed CGPA for the hold-out test set, with the $y = \hat{y}$ line indicating perfect agreement.

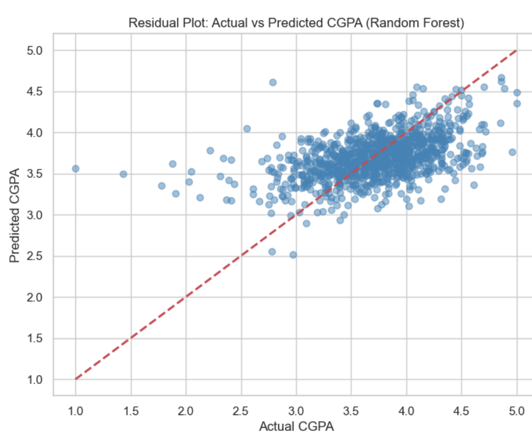

Figure 4.1: Predicted versus actual CGPA on the hold-out test set.

Predictions align closely with the diagonal in the mid-CGPA interval (approximately 2.8–4.2), indicating stable generalisation in the range where most students lie. Slight underestimation is observable for the highest achievers (> 4.4), consistent with a model constrained to admission-time predictors rather than in-programme data.

4.2.2 Overall Evaluation Metrics

Table 4.1 summarises the principal regression metrics on the hold-out set: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2).

Table 4.1: Evaluation metrics on the hold-out test set.

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

The tuned Random Forest improves on both the untuned variant and the linear baseline across all metrics, explaining approximately 24% of the variance in final CGPA. While the R^2 is moderate, reflecting unobserved in-programme factors (e.g., study behaviours, pedagogy, assessment regimes), the error magnitudes are consistent with prior work using pre-admission data alone and are actionable for early advising.

4.2.3 Residual Characteristics

Error distributions further characterise reliability. Figure 4.2 shows residuals centred near zero with mild tail deviations; Figure 4.3 provides the Q–Q plot.

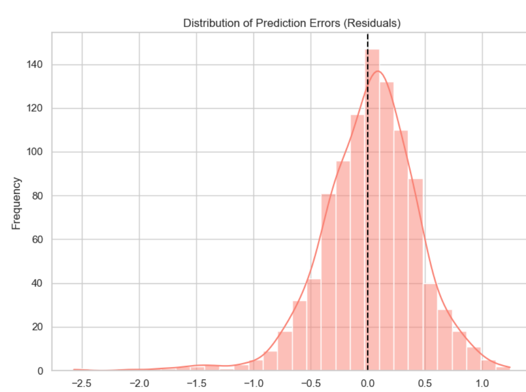

Figure 4.2: Distribution of prediction errors (residuals).

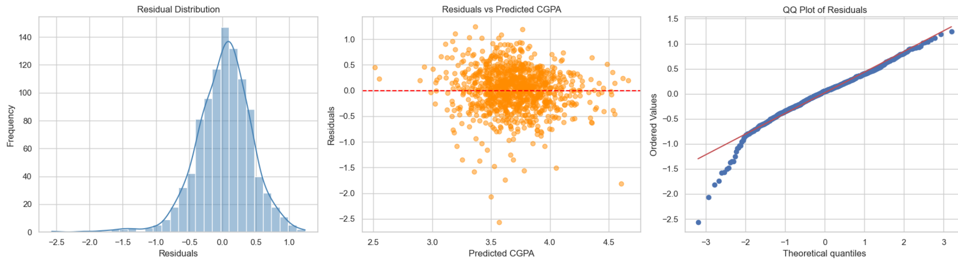

Figure 4.3: Q–Q plot of residuals.

Residuals are approximately normal and centred, indicating limited systematic bias. Slight tail divergence corresponds to underestimation among the highest performers, consistent with Figure 4.1.

4.3 Feature Contributions and Compactness

4.3.1 Global Importance Ranking

Random Forest feature importance highlights predictors that most reduce prediction error across the ensemble. Figure 4.4 displays the ten strongest contributors.

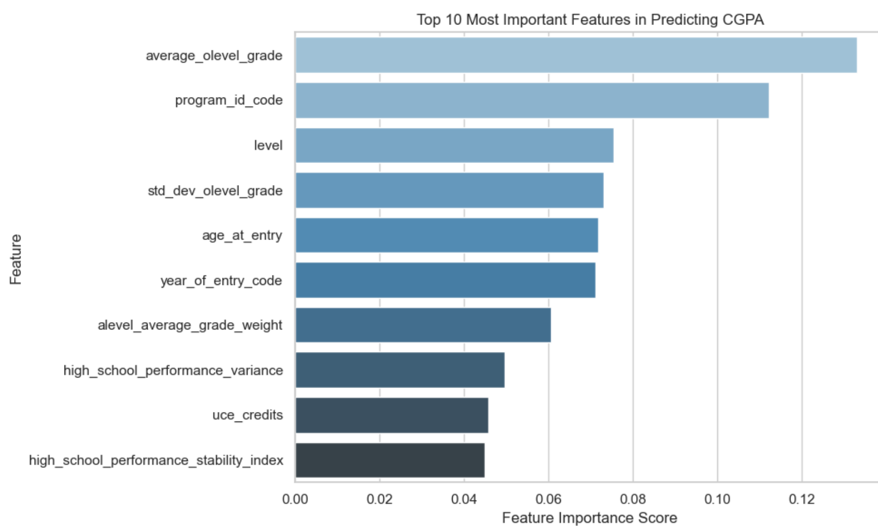

Figure 4.4: Top ten features contributing to CGPA prediction (test set).

The average O-Level grade emerges as the dominant predictor, followed by A-Level average grade weight and UCE credits. These results confirm that consistent secondary-school attainment provides the primary signal of subsequent academic performance. Contextual variables (e.g., `program_id_code`, `year_of_entry_code`, `gender`) contribute moderately, reflecting programme structure and historical intake differences.

4.3.2 Dimensionality Check (PCA)

Principal Component Analysis (PCA) was used to assess compactness of the final feature space. The scree plot in Figure 4.5 indicates that a small number of orthogonal components capture a large proportion of total variance, corroborating the feature selection choices.

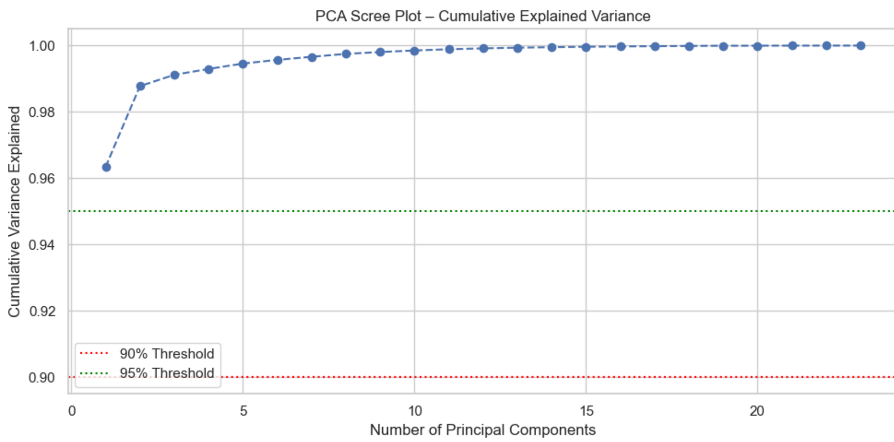

Figure 4.5: PCA scree plot: cumulative explained variance.

Reduced redundancy and a compact representation support model stability and facilitate interpretation.

4.4 Explainable AI Insights

4.4.1 Global SHAP Patterns

SHAP values provide additive, instance-level attributions consistent with cooperative game-theoretic principles. The beeswarm plot in Figure 4.6 summarises global effects: wider horizontal spread indicates stronger influence, while colour indicates feature magnitude.

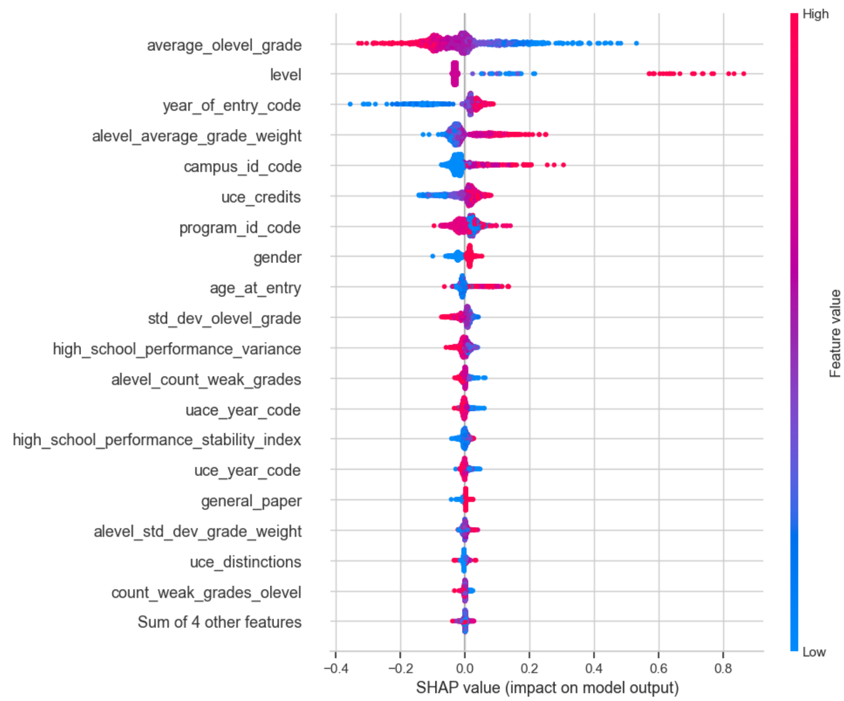

Figure 4.6: Global SHAP beeswarm plot of feature contributions (test set).

High average O-Level grade, strong A-Level average grade weight, and larger UCE credits systematically increase predicted CGPA. Greater O-Level grade variability (instability across subjects) tends to depress predictions. Contextual features exert smaller but consistent effects.

4.4.2 Predictive Factors for Higher and Lower CGPA Outcomes

The feature importance and SHAP analyses show that higher predicted CGPA was mainly associated with stronger prior academic attainment, particularly better average O-Level performance, stronger weighted A-Level performance, and a higher number of UCE credits. These patterns suggest that consistent secondary-school achievement provides an important signal of likely academic performance at university level.

Lower predicted CGPA, or increased need for academic support, was associated with weaker admission-time indicators, fewer UCE credits, lower A-Level grade weights, and greater variability in O-Level performance across subjects. These factors should not be treated as fixed measures of ability, but as early advising signals. For policy and practice, this means admission scores can support both recognition of strong students and timely interventions such as study-skills support, mentoring, bridging support, and closer firstyear monitoring for students who may need additional academic support.

4.4.3 Individual-Level Explanations

Waterfall plots (Figures 4.7–4.8) decompose predictions for six representative students, revealing how the same model logic produces personalised attributions.

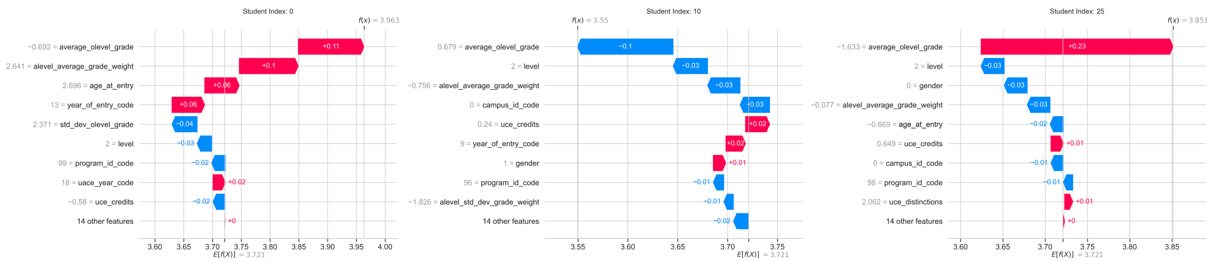

Figure 4.7: SHAP waterfall plots for Students 0, 10, and 25.

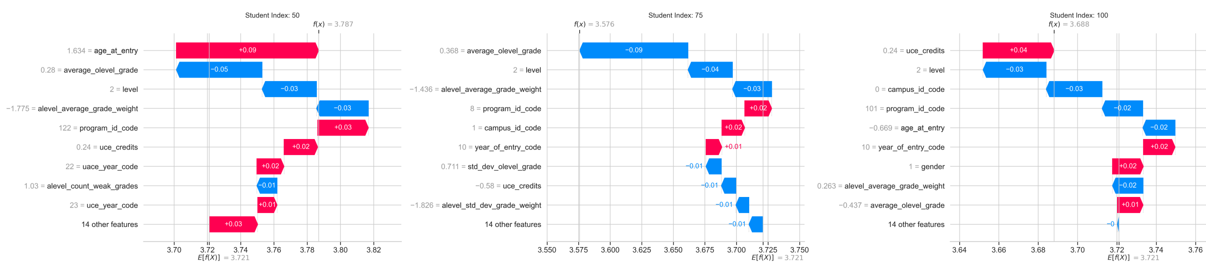

Figure 4.8: SHAP waterfall plots for Students 50, 75, and 100.

Table 4.2: Top influences on CGPA for sample students (SHAP-based).

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

4.4.4 Feature Interactions

Partial dependence relationships illuminate important interactions (Figure 4.9). For example, strong O-Level and A-Level profiles jointly amplify predicted CGPA, while interactions with programme and campus codes suggest structural differences across units.

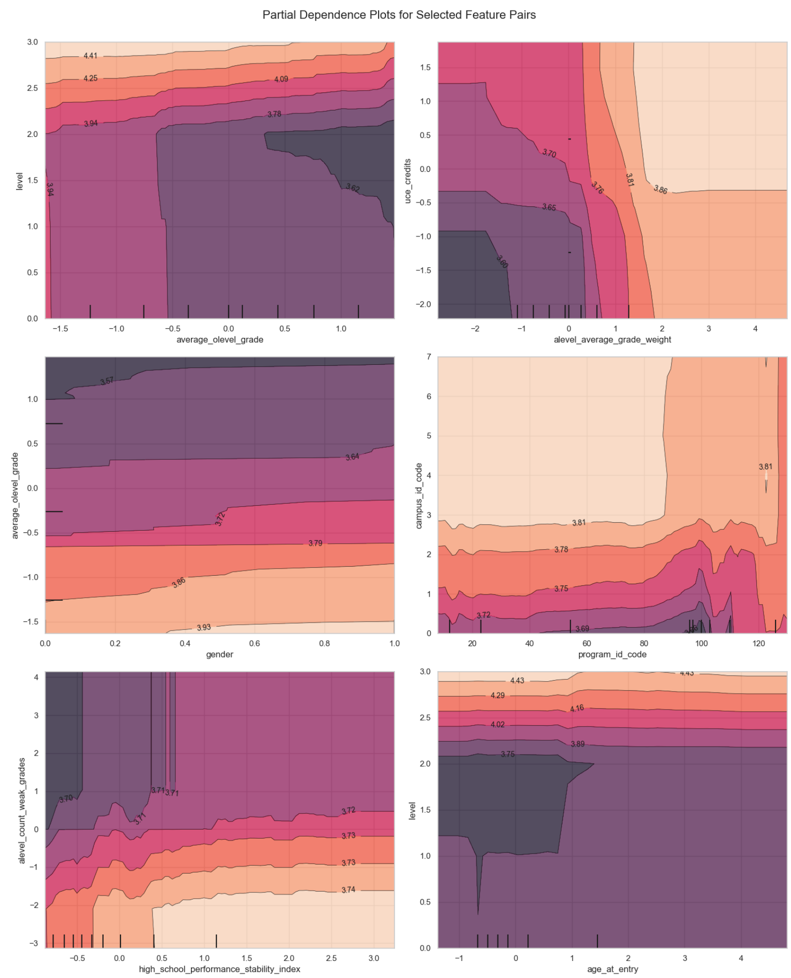

Figure 4.9: Partial dependence plots for key feature interactions.

4.5 Band-Level Predictions and Institutional Use

4.5.1 CGPA Band Distribution

To support advising workflows, continuous predictions were mapped to institutional bands: **High** (CGPA $\geq$ 3.60), **Moderate** ($2.80 \leq$ CGPA < 3.60), and **Low** (CGPA < 2.80). These thresholds align with UCU's academic policy conventions. Table 4.3 and Figure 4.10 summarise the distribution and suggested actions.

Table 4.3: Predicted CGPA band distribution with suggested actions.

Band	Count	Avg. CGPA	Suggested institutional action
High (≥ 3.60)	653	3.82	Recognition: scholarships, honours tracks, leadership opportunities
Moderate (2.80–3.59)	343	3.47	Guided support: advising, structured study plans, mentorship
Low (< 2.80)	0	–	None detected; recalibrate thresholds in future intakes if needed

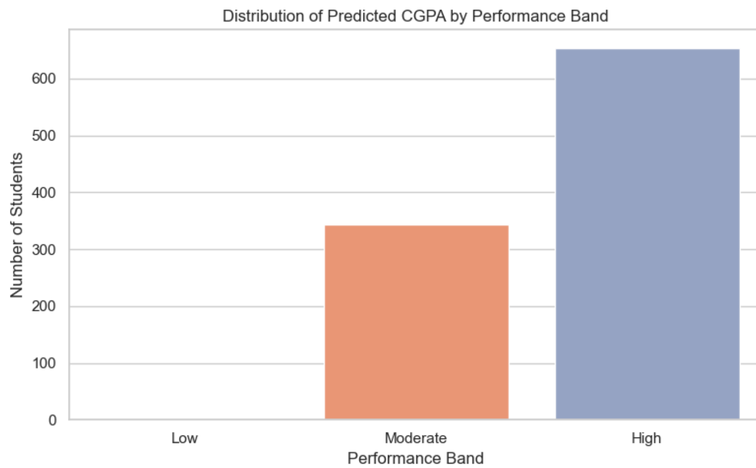

Figure 4.10: Distribution of predicted CGPA by performance band.

4.6 Fairness and Subgroup Evaluation

4.6.1 Gender-Based Error Comparison

Mean Absolute Error (MAE) by gender is shown in Table 4.4.

Table 4.4: MAE by gender.

Gender	MAE	Sample size
Male	0.358	425
Female	0.279	571

Errors are modestly higher for male students; this warrants continued monitoring as additional intakes are incorporated.

4.6.2 Institutional Subgroups

Figure 4.11 summarises predicted CGPA by campus and study level. Larger campuses present wider predictive dispersion, consistent with greater heterogeneity; smaller campuses show tighter bands, though some effects may reflect limited sample sizes.

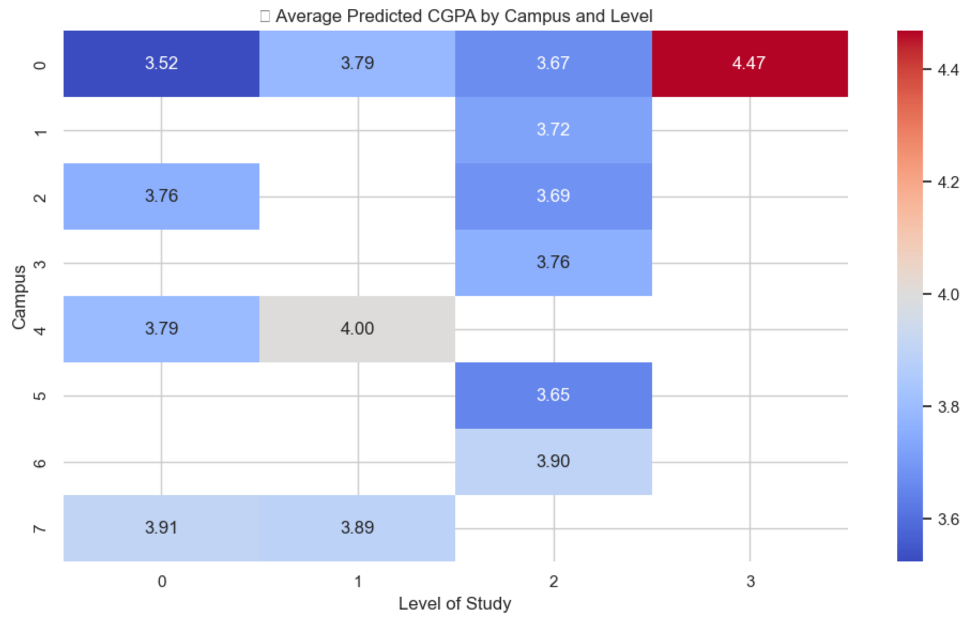

Figure 4.11: Average predicted CGPA by campus and level.

Tables 4.5 and 4.6 contrast predicted and observed CGPA by gender and level, while Table 4.7 summarises MAE by campus.

Table 4.5: Predicted vs. actual CGPA by gender.

Gender	Students	Pred. CGPA	Actual CGPA
Male	425	3.71	3.67
Female	571	3.70	3.75

Table 4.6: Predicted vs. actual CGPA by level of study.

Level	Students	Pred. CGPA	Actual CGPA
Diploma	21	3.84	3.91
Certificate	13	3.94	4.11
Bachelor	940	3.68	3.69
Master	22	4.47	4.35

Table 4.7: MAE by campus.

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

Overall, subgroup performance is acceptable, with some campus-level variation meriting follow-up as sample sizes grow.

4.7 Sensitivity and What-If Simulations

4.7.1 Single-Feature Sensitivity

Table 4.8 illustrates how targeted adjustments to individual inputs affect a representative student's predicted CGPA.

Table 4.8: Sensitivity analysis for selected single-feature changes.

Scenario	Pred. CGPA	Δ	Impact
Add 2 UCE credits	3.985	+0.022	Slight increase
Reduce average O-Level grade by 1	3.754	-0.210	Moderate drop
Set academic level to Masters (3)	4.503	+0.540	Major increase

4.7.2 Programme-Fit Simulation

To explore programme alignment, the programme identifier was varied while holding other features constant. Figure 4.12 shows the top and bottom simulated outcomes for a sample student (ID 10), and Table 4.9 lists the top ten predicted programmes.

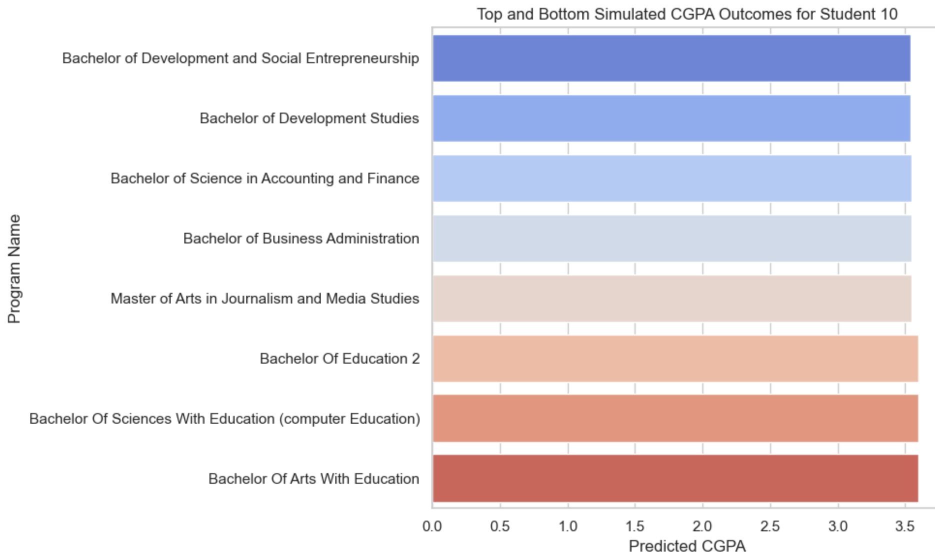

Figure 4.12: Top and bottom simulated CGPA outcomes for Student 10

Table 4.9: Top ten programmes by predicted CGPA: Student 10.

Programme	Code	Pred. 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

These simulations are advisory rather than prescriptive and should be interpreted alongside student preferences and programme goals.

4.8 Bias and Risk Considerations

4.8.1 Data Coverage and Small-Group Risk

Some programmes and campuses are sparsely represented in the test set, increasing the risk of unstable subgroup estimates. Table 4.10 lists groups with fewer than ten students; predictions for these segments should be interpreted with caution.

Table 4.10: Subgroups with fewer than ten students

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

4.8.2 Sources of Bias and Mitigation

Admission-time academic indicators are strong signals but may under-represent late-blooming potential. Contextual variables can proxy wider inequalities if not monitored. Mitigation includes periodic fairness audits, retraining with rebalanced intakes, cautious use of contextual features, and routine transparency through SHAP-based explanations to maintain accountability.

4.9 Illustrative Student Cases

Two anonymised profiles demonstrate how predictions and explanations translate into advising. For each, a partial JSON profile is shown, followed by predicted CGPA and leading SHAP contributors.

4.9.1 Student A: High Predicted Performance

Profile (partial):


```
{"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
Dominant drivers (SHAP): strong A-Level performance; consistent O-Level grades; small positive effect of age at entry.

4.9.2 Student B: Moderate Predicted Performance

Profile (partial):


```
{"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

Dominant drivers (SHAP): solid O-Level performance; negative influence from grade variability; minor level/demographic effects.

These cases illustrate how transparent explanations convert numeric predictions into targeted guidance (e.g., mentoring for stability, early recognition for excellence), supporting ethical, student-centred advising.

4.10 Conclusion

The results demonstrate that admission-only predictors yield actionable forecasts of final CGPA. The tuned Random Forest provides consistent accuracy in the mid-range, with limited bias and interpretable contributions driven primarily by prior academic attainment. Global and local explainability (feature importance, PCA compactness, SHAP, and partial dependence) supports transparency and trust. Band-level summaries translate continuous predictions into institutional actions, while sensitivity and programme-fit simulations illustrate how outputs can inform individual advising and resource planning. Fairness analyses identify minor subgroup differences and motivate ongoing monitoring, retraining, and governance.

Collectively, the findings address the study's research questions by establishing the influence of key pre-admission features on final CGPA prediction (RQ1), demonstrating the predictive performance of admission-time machine learning models (RQ2), showing how SHAP explanations and subgroup fairness diagnostics support ethical and transparent use of predictions (RQ3), and translating model outputs into advising-ready artefacts such as CGPA bands, early-support indicators, and programme-fit simulations (RQ4).

Chapter 5

Conclusion and Recommendations

5.1 Introduction

This chapter concludes the study by summarising its purpose, methodology, key findings, limitations, and implications for both institutional practice and future scholarship. The central aim of the research was to determine whether pre-admission data alone could provide reliable predictions of university academic performance, expressed as the final Cumulative Grade Point Average (CGPA). The study sought to balance predictive accuracy, interpretability, and fairness in developing a model suitable for practical use in higher education.

A quantitative research design was applied, combining rigorous data preprocessing and feature engineering with a tuned Random Forest algorithm. The model achieved consistent performance and was further enhanced with explainable artificial intelligence techniques, particularly SHapley Additive exPlanations (SHAP) and sensitivity analyses. These tools translated complex model outputs into transparent, student-specific insights, enabling practical application by academic advisors and institutional leaders.

The chapter first summarises the study's principal findings, then outlines limitations and challenges, evaluates deployment readiness, proposes an institutional adoption strategy, highlights directions for future research, and concludes with broader reflections on the contribution of predictive analytics to higher education.

5.2 Summary of Findings

This study demonstrates that pre-admission data can yield actionable predictions of university CGPA. Using historical records from Uganda Christian University (UCU), an optimised Random Forest model achieved a moderate coefficient of determination ($R^2 \approx 0.24$) and a mean absolute error (MAE) of approximately 0.31 on the test set. Predictions were most accurate in the mid-range of the CGPA distribution, which covers the majority of students, and residuals were approximately normal and centred near zero. The model slightly underestimated the highest performers, reflecting the absence of in-programme data, but remained reliable overall.

Feature analysis showed that academic indicators were the dominant predictors. Average

O-Level grade, weighted A-Level performance, and UCE credits consistently emerged as the strongest features. These findings confirm the long-established principle that prior attainment is the best predictor of future academic success. Demographic variables such as gender and age of entry, and institutional placement variables such as programme and campus, had smaller but contextually meaningful contributions without overshadowing core academic measures. Importantly, these results suggest that entry scores should not only support admission ranking, but also serve as early advising signals for enrichment, mentoring, bridging support, and first-year monitoring.

Interpretability was addressed through SHAP and sensitivity simulations. Global SHAP values revealed the consistent influence of prior academic performance, while individual-level breakdowns demonstrated how the same feature could have varying impacts across different students. Sensitivity analyses, including "what-if" scenarios, highlighted how improvements in O-Level or A-Level performance or programme changes could alter outcomes. These insights moved the model from being merely predictive to becoming more interpretable and advisory.

Fairness was assessed through subgroup analyses by gender, campus, and academic level. Results were generally consistent across groups, though male students exhibited slightly higher prediction errors and smaller campuses showed greater variability due to under-representation. These differences were modest but emphasise the importance of continuous monitoring to ensure equity.

The model's outputs were made actionable through advising artefacts such as CGPA performance bands and programme-fit simulations. Banding predictions into high, moderate, and low categories allowed for targeted interventions, from scholarships and enrichment opportunities to structured advising and mentorship. Programme-fit simulations offered personalised insights into how students might perform across different study paths, providing valuable tools for academic planning.

In sum, the study established that admission-only models can deliver meaningful, interpretable, and fair predictions of CGPA. These findings answer the research questions by demonstrating predictive viability, identifying salient admission-time predictors, providing transparent explanations, and translating outputs into advising-ready tools.

5.3 Limitations and Challenges

Despite encouraging results, several limitations frame the interpretation of this study. First, the dataset was restricted to students with recorded final CGPA, excluding dropouts and transfers. This introduces survivorship bias and limits insights into broader student trajectories. Beyond this, while preprocessing addressed missing data through imputation, residual inconsistencies in historical records may have affected robustness.

Second, the reliance on pre-admission data excluded other influential predictors such as first-year performance, attendance, socioeconomic background, personal status, and learner competences. Although intentional, this limited the model's overall explanatory power.

Third, certain subgroups were under-represented. Some programmes and campuses had fewer than ten students in the test set, producing volatile predictions and constraining generalisability. These gaps require cautious interpretation when applying outputs to smaller or remote groups.

Fourth, the choice of Random Forest reflected a deliberate trade-off between predictive accuracy and interpretability. While SHAP explanations mitigated opacity, the ensemble model remains less transparent than simpler baselines such as linear regression. Non-technical stakeholders may still find interpretation challenging.

Fifth, the dataset did not include school-source categorisation variables, such as urban or rural school background, school ownership type, or traditional versus international curriculum pathway. This limited the study's ability to examine how prior school context may influence CGPA predictions.

Finally, equity considerations remain. Although subgroup audits indicated broadly consistent performance, modest differences, such as slightly higher MAE for male students, highlight the need for ongoing fairness monitoring. Without continuous evaluation and retraining, small discrepancies could accumulate into systemic bias.

These limitations do not diminish the contribution of the study but provide essential context. The findings are robust at an institutional level, but further refinements and monitoring are necessary for long-term, equitable use.

5.4 Deployment Readiness

Beyond accuracy, the study examined whether the model is practically deployable within UCU's institutional context. The Random Forest was implemented as a deployable service, serialised with `joblib` for persistence and wrapped in a Flask RESTful API. This architecture enables flexible deployment either on-premises or in the cloud, depending on infrastructure and governance requirements. The API supports real-time inferences from structured JSON inputs and provides outputs including predicted CGPA and feature contributions.

Two integration modes were identified. Batch integration allows predictions to be generated periodically from exported admissions data and imported into the MIS for advisor review. This option requires minimal system modification and is suitable for pilot deployment. Real-time integration connects the model directly to the MIS during registration or advising, providing immediate predictions and greater intervention potential, though at a higher technical cost.

A feasibility assessment rated the model as technically mature, ethically compliant, and institutionally relevant. Performance metrics are stable, privacy safeguards are in place, and fairness audits have been conducted. However, successful deployment requires developing an advisor-facing interface, staff training, and institutional policies to ensure responsible use. Governance frameworks will be critical to address transparency, accountability, and data privacy.

Overall, the model is ready for deployment as a decision-support tool, provided it is accompanied by supportive infrastructure, clear policies, and continued monitoring.

5.5 Institutional Adoption Strategy

The effective use of predictive analytics requires not only technical integration but also thoughtful adoption aligned with institutional culture and priorities. At UCU, this model can enhance decision-making across several domains.

In academic advising, integration into faculty dashboards would provide advisors with early risk profiles during admission or routine counselling. This would enable tailored support plans, including structured study schedules, mentorship, or enrichment opportunities. High-performing students could be channelled into honours tracks, scholarships, and leadership roles.

In student onboarding, providing personalised performance estimates at orientation would help learners understand their strengths and weaknesses from the outset. When combined with bridging courses, peer mentoring, and study skills workshops, such insights can foster preparedness and motivation.

The model also strengthens early warning and intervention systems. By comparing real-time performance to baseline admission predictions, deviations can trigger timely interventions before problems escalate. This positions the model as a proactive component of institutional student support.

At the policy level, aggregated predictions can inform curriculum reforms, highlight programmes with systemic challenges, and guide equitable resource allocation. Disaggregated outputs allow leadership to monitor fairness across campuses, gender, and other demographic groups.

UCU should also review the policy framework used to rate and combine admission scores into final entry points. Since O-Level performance, weighted A-Level scores, UCE credits, and related academic indicators showed different levels of predictive value, admission policy should examine whether current score-weighting practices reflect their contribution to academic success. Such a review should remain fair, transparent, and programme-sensitive, supporting better admission guidance and early academic support rather than exclusion.

UCU should further enrich its admissions data collection framework by capturing structured information on students' prior school context, including school location category, curriculum pathway, and other relevant background indicators. This would allow future models to distinguish more clearly between academic performance and educational opportunity, while guiding bridging, mentoring, or early support where needed.

A stepwise adoption plan is recommended: pilot testing in selected faculties, followed by staff training and incremental scaling once governance frameworks are in place. Such a phased approach will build confidence, incorporate feedback, and ensure responsible integration into UCU's MIS.

5.6 Future Work

The study opens multiple avenues for further research and development. One promising direction is real-time prediction, integrating the model with admissions and advising workflows to provide instantaneous forecasts. This would enhance responsiveness and allow for continuous model updates as new data becomes available.

Another extension is trajectory modelling. Rather than predicting only final CGPA, future models could forecast semester-by-semester outcomes, providing more granular insights into student development and enabling earlier interventions.

Enhanced explainability also warrants attention. While SHAP was effective, more user-friendly explanatory tools could increase accessibility for non-technical stakeholders. Faculty-led summaries of model logic in plain language and fairness-aware interpretability methods would further strengthen transparency.

Interactive dashboards and visual analytics represent another priority. Beyond static tables and figures, dashboards would allow advisors and administrators to explore predictions dynamically, filter by subgroup, and visualise trends for strategic planning.

Finally, expanding the feature set to include in-programme variables such as coursework grades, attendance, and LMS engagement would improve predictive accuracy and contextual depth. This would extend the model from admission-only forecasting to a more comprehensive academic performance monitoring system.

Future studies should broaden the model beyond grade-based predictors by incorporating personal status and competence-based indicators where they can be ethically collected. These may include socioeconomic background, employment status, family responsibilities, accommodation status, motivation, study habits, digital literacy, communication skills, resilience, and self-regulated learning. Such variables would provide a more holistic view of academic success and support more personalised student interventions.

Together, these directions would transform the current prototype into a fully adaptive, transparent, and integrated decision-support framework.

5.7 Conclusion

This study has demonstrated that machine learning, applied responsibly to pre-admission data, can generate actionable predictions of university academic performance. The Random Forest model achieved consistent accuracy while preserving interpretability and fairness through SHAP explanations, sensitivity analyses, and subgroup audits. By translating outputs into advising artefacts such as performance bands and programme-fit simulations, the model provided practical tools for academic support and planning.

The findings align directly with the revised research objectives and questions. Key pre-admission predictors of final CGPA were identified (RQ1); the predictive viability of admission-time machine learning models was established (RQ2); interpretability and fairness were demonstrated through SHAP explanations and subgroup analyses (RQ3); and practical applications for proactive advising, early identification of students requiring support, and evidence-informed programme guidance were developed (RQ4).

The implications extend beyond UCU. This work contributes to the broader field of educational data mining by showing how predictive models can be designed to align with institutional priorities, fairness standards, and transparency requirements. The results highlight both the potential and the responsibility inherent in applying predictive analytics in education: to empower students and institutions without undermining equity or trust.

In conclusion, the study illustrates how machine learning, when transparent, fair, and responsibly governed, can enhance academic advising, institutional planning, and student success. It sets the stage for scalable and ethical predictive systems in higher education that complement, rather than replace, human judgement.

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 sub-saharan 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.

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A. E., Isayev, O., Curtalolo, S., Fourches, D., & Tropsha, A. (2017). Application of random forest models to predict academic success based on pre-admission variables in pharmd programs. *American Journal of Pharmaceutical Education*, *81*(7), Article 118. https://doi.org/10.5688/ajpe817118

Myburgh, C. (2019). *Predicting academic success using pre-admission data: A machine learning approach in south african higher education* [PhD thesis]. North-West University.

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.

Puri, N., McCarthy, M., & Miller, B. (2022). Validity and reliability of pre-matriculation and institutional assessments in predicting usmle step 1 success: Lessons from a traditional 2 x 2 curricular model. *Frontiers in Medicine*, *8*, 798876.

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access*, *10*, 72480–72503.

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 highlights global studies on predictive analytics, with emphasis on CGPA prediction, institutional analytics in African higher education, ethical considerations, and practical applications such as programme recommendation systems. A distinct category is included for *admission-only models*, underscoring their importance and limited representation in existing research. This further justifies the contribution of the present study.

Table A.1: Classification of Literature Used in Chapter 2

Citation Key	**Focus Area**	**Region/Scope**
	1. CGPA and Academic Performance Prediction	
`nur2021developing`	Temporal ML models for CGPA prediction, interpretability, and stakeholder usability	USA
`fahd2022application`	Meta-analysis of ML for academic performance, at-risk detection, and attrition	Global
`zhang2021educational`	Systematic review of EDM techniques for predicting academic outcomes	China/Global
`ismanto2022recent`	Review of deep learning methods for academic performance prediction	Global
`alnasyan2024power`	Deep learning for predicting student success in online/virtual settings	Global
	2. Admission-Only Prediction Models	
`muratov2017application`	Application of Random Forest models on pre-admission variables to predict PharmD outcomes	USA
`myburgh2019predicting`	Machine learning with pre-admission (school-leaving) data to predict academic success and fairness impacts	South Africa
`puri2022validity`	Pre-matriculation assessments and their link to academic performance	USA
`walker2024can`	Validity of admission metrics in medical education, with caution on predictive rigidity	USA
	3. Institutional Analytics in African HEIs	
`mpofu2025data`	Strategic use of analytics in Zimbabwean universities for planning and innovation	Africa
`ajani2024leveraging`	Leveraging AI for education quality with equity and ethics concerns	Africa
`kunjumuhammed2024artificial`	Risk of AI reinforcing inequality in African HEIs	Africa

Continued on next page

Table A.1 – continued from previous page

Citation Key	Focus Area	Region/Scope
patel2024implementation	AI implementation challenges and governance frameworks in South African HEIs	South Africa
4. Programme Recommendation and Course Selection		
kamal2024recommender	Systematic review of recommender systems in academic advising	Global
algarni2023systematic	Recommender methodologies and programme matching strategies	Global
simionintelligent	Intelligent digital platform for programme choice and student guidance	Europe
5. Ethical and Responsible AI in Education		
memarian2023fairness	Systematic review of Fairness, Accountability, Transparency, and Ethics (FATE) in predictive systems	Global
walker2024can	Ethical implications of relying solely on pre-entry data in predictions	USA
6. Student Retention and Dropout Prediction		
shafiq2022student	Review of ML approaches for student retention across modalities	Global
sosa2025predicting	Dropout risk modelling using SEM and data mining approaches	Europe
colpo2024educational	EDM in dropout prediction with focus on implementation gaps	Global
7. Broader Applications of AI in Education		
forero2024techniques	Review of ML/AI techniques applied across education levels	Global
moussa2024predictive	Predictive assessment using AI with ethical/policy considerations	Global/Arabic contexts
adewale2024impact	Application of AI in Open and Distance Learning; process-based predictive frameworks	Africa