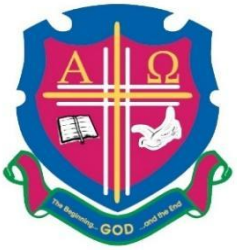

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

NAME: ATUHAIRE PAULINE

ACESS NUMBER: B35093

REG.NO: S25M19/016

COURSE CODE: RSM8101

COURSE NAME: RESEARCH METHODS AND PUBLICATIONS

EXAMINATION TYPE: PROJECT BASED -EXAM

Question one: Milestone One:

Project Title**: ''Revolutionizing Education in Uganda: Leveraging ML to Tackle Secondary School Dropout and Enrollment Rates.''**

Executive Summary

Education is a primary human right for all children. Failure to access and complete a basic cycle of quality inclusive education gravely limits future opportunities, denies fundamental rights, and deepens cycles of poverty. The Convention on the Rights of the Child (CRC), which Uganda has validated, instates a legal obligation for the government to secure compulsory primary education free of costs. This legal and ethical commitment must stretch to the relentless effort to maximize educational attainment at the secondary level.

The purpose of this project is to develop a data-driven Early Warning System (EWS) to directly address the persistent challenge of alarming student attrition rates and erratic enrollment in Ugandan secondary schools. In spite of notable significant policy milestones, namely the Universal Secondary Education (USE) program, the secondary net enrollment rate stays low (estimated at 27%), and up to 43 out of every 100 pupils drop out ahead of completing the primary level, greatly hindering progression to secondary school. This situation symbolizes a critical impediment to Human Capital Accumulation, and a significant opportunity for innovative policy intervention [2].

Traditional, reactive intervention methods are historically reactive or inefficiently timed, failing to capitalize on the rich student data already at hand within the education system. The core solution is the design and implementation of a scalable Machine Learning (ML) model built on historical academic, demographic, and socio-economic data. This prognostic modelling will with precision classify students into severe, moderate, or minimal risk of dropping out before the risk exhibits in poor performance or physical absence. This approach offers an instant unique competitive advantage by converting intervention from being reactive to being preventative and evidence-based. The gains of this model are strategic: it permits school administrators and the Ministry of Education to optimize resource

deployment (e.g., mentorship, financial aid, counseling) to the precise students who need them most. Therefore, this contributes to improved secondary school completion rates, and reduced educational inequity (particularly for girls and rural students), and highly skilled and educated youth population prepared for the workforce.

The project is closely aligned with both national and global development priorities. It strongly advocates for Sustainable Development Goal (SDG) 4: Quality Education, particularly Target 4.1 (ensuring all complete equitable and quality secondary education). Nationally, it is a key driver for Uganda's Fourth National Development Plan (NDP IV: 2025/26–2029/30), adding to the core theme of Sustainable Industrialization by fortifying the Human Capital Development pillar and leveraging the Digital Transformation agenda. The technical feasibility of the system is high, depending mainly on open-source tools and prevailing data infrastructure, making it cost-effective and scalable across the country. The project team features the core competence in Data Science, Predictive Modeling, and Educational Policy to provide a strong and actionable system. We encourage the Ministry of Education and Sports and the National Planning Authority (NPA) to sanction this project for pilot deployment to secure these tactical national benefits.

Milestone two (Chapter One)

1.0 Introduction

In the contemporary world, education remains one of the most powerful tools for progress. As Nelson Mandela said, "Education is the most powerful weapon which you can use to change the world." However, many emerging economies, including Uganda, face high dropout rates, undermining national development. This study intends to address this issue by developing a data-driven Early Warning System using Machine Learning to enhance student retention in Ugandan secondary schools and support equitable education outcomes.

Background to the Study

Within Sub-Saharan Africa, the issue of Dropout rates from secondary education is considerably a huge hurdle, with an estimation of close to,'' 89 million children

out of school,'' thus negatively hindering efforts to attain quality education worldwide. For instance; following the implementation of SDG 4 (4.1) which aims at promoting inclusivity and equitable quality education and additionally the promotion of lifelong learning opportunities for all regardless of social status (rich or poor). Furthermore, based on the high attrition rates presently in Uganda, there has been a great block on human capital development and fulfillment of the National Development Plan has been greatly impacted in a negative way [1]. That is to say; the NDP IV is founded on the principle of attaining sustainable industrialization for inclusive growth, employment and wealth creation by changing the economy from a livelihood to a fully capitalized one. In spite of government investment in education, throughput rates remain low, which sets off loud alarms in the education sector, with many students dropping out as a result of complex factors.

There's an urgent need for a more tailored approach to tackle these issues. It is known that current research identifies causative factors, that often lack data-driven insights which can limit the effectiveness of policy interventions. Therefore, a methodology that ranks key determinants and predicts individual student risk could allow personalized interventions, improving student retention and education outcomes in Uganda [2].

Problem Statement

The high rate of student dropout in Uganda, more specifically in the Central Region, forms a major educational pipeline failure that critically constrains national human capital development and fuels poverty cycles (hinders job creation and wealth accumulation). Current interventions are being limited by a notable knowledge gap: that is to say; interventions are based on wide-ranging assumptions rather than an evidence-based hierarchy of the most impactful determinants (more subjective). Furthermore, the existing descriptive statistical models are lacking (can't be trusted or accurate), individual-level prediction of dropout risk [14]. This absence of a high-precision predictive model and a systematic ranking of determinants prevents educational stakeholders from optimally allocating limited resources. Thus, policies remain reactive, experimental, and generalized, failing to address the highly localized and dynamic nature of student attrition risk. The

purpose of this study is to examine more accurate and reliable tools in order to address gaps in the education system and also mitigate the high levels of students not completing school. Additionally, this study aims to develop and validate a high-performance machine learning system capable of providing accurate prediction and systematic ranking of risk factors leading to drop out of the students.

Research Objectives

General Objective

To develop a high-performance machine learning system for predicting school dropout risk and systematically ranking the key causative factors among secondary school students in the Central Region of Uganda.

1.1.1 Specific Objectives

To identify and rank the most significant sociodemographic, institutional, and behavioral determinants of school dropout in secondary schools in the Central Region of Uganda.

To develop, train, and validate a high-precision machine learning model (Logistic Regression, Random Forest, and XGBoost) to predict individual student dropout risk.

To propose a framework for evaluating the effectiveness of data-driven, targeted interventions based on the predictive model's output using an econometric evaluation method (RCT framework).

Research Questions

What is the empirical ranking of sociodemographic, institutional, and behavioral determinants that most significantly influence school dropout in the Central Region of Uganda?

Which machine learning model (Logistic Regression, Random Forest, or XGBoost) achieves the highest precision and recall in predicting individual student dropout risk?

How can the results of the predictive model be used to design and evaluate targeted interventions using an econometric method (RCT framework)?

Research Hypotheses

H_0: There is no numerically significant difference between the predictive performance of Logistic Regression, Random Forest, and XGBoost models in predicting school dropout risk.

H_1: The XGBoost model will demonstrate statistically superior predictive performance (higher F1-Score) against Logistic Regression and Random Forest in predicting school dropout risk.

Justification of the Study

This study is warranted by the need to fill key gaps in educational management and research. It yields quantitative evidence to prioritize impactful interventions, optimizes resource allocation, and constructs a high-precision predictive model to identify high-risk students (evidence -based). By implementing advanced machine learning techniques, the study establishes a new methodological standard for research on attrition in low- and middle-income countries, in particular Uganda.

Chapter Two: Literature Review

Introduction

Previous research has shown that student retention, non-enrolment, and early school leaving (dropout), has been due to a number of factors that are internal and external in nature, though results remain inconsistent [5]. The majority of existing literature has zeroed in on particularly the sociodemographic, institutional, and economic determinants of educational attrition in low- and middle-income countries, with a primary emphasis on the Ugandan context. The review is structured to first define the global and national scale of the dropout crisis using key statistics, then scrutinize the theoretical frameworks and empirical studies identifying the contributing factors. Lastly, recent studies have highlighted the need for further investigation into employment of machine learning to mitigate the problem of student dropouts through early intervention curbing down on the

number of children who leave school without completion and fills the existing knowledge gaps. In conclusion, this review sums up key findings and identifies the main research gap addressed in this study.

Global and National Context of Educational Attrition

The challenge of early school leaving is widely acknowledged as a significant barrier to attaining Sustainable Development Goal 4 (SDG 4) for quality education [3]. The issue is severe across Africa, raising an existential challenge to the continent's socioeconomic development, with nations like Nigeria (16.9%) and Ethiopia (13%) reporting high rates of attrition [4].

Within this continental context, Uganda is battling a major pipeline failure, with dropout rates spanning from 4.3% to 9.7% depending on the educational level. Despite commitments to Universal Primary Education (UPE) and Universal Secondary Education (USE), the national primary school completion rate hovers around 52%, and the survival rate to the final primary grade stands at a staggering 33% [3], [4]. This systemic failure sabotages Uganda's National Development Plan (NDP III, 2020), which champions human capital development, indicating that the current education system is falling short of national strategic priorities. Tackling this problem is, therefore, not merely an academic exercise but a strategic requirement for long-term national development and the mitigation of poverty cycles.

Theoretical Framework

The Push and Pull Theory: This study is rooted in the Push and Pull Theory of Educational Attrition. The theory suggests that dropout results from the interaction of two force types: Push Factors: These pushes dwell within the school system (e.g., poor teacher quality, institutional culture, fees). They push students out. Pull Factors: These forces are situated in the external environment (e.g., socioeconomic status, early marriage, income-generating activities). They pull students away from school [6]. This structure is critical for categorizing the study's determinants into institutional (Push) and sociodemographic/behavioral (Pull) categories, allowing for targeted policy analysis.

Human Capital Theory

The Human Capital Theory asserts that education is an investment. Individuals invest time and resources expecting future yields (e.g., higher wages, better employment). Dropping out happens when the perceived costs of education surpass the expected lifetime benefits [7]. This theory validates analyzing socioeconomic and household-level factors as key determinants of educational investment choices.

2.3 Conceptual Framework

The conceptual framework is derived from the Push-Pull Theory. It demonstrates that the independent variables (determinants) sway the dependent variable (dropout) [6]. Independent Variables (Determinants): Sociodemographic Factors: Household income, parental status, student gender, etc. Institutional Factors: School type, student-to-teacher ratio, school distance, etc. Behavioral Factors: Attendance, academic performance, discipline records. The analysis process serves as the Intervening Variable, where data is converted through machine learning to produce an output: Predictive Model and graded Determinants.

2.4 Synthesis of the Literature and Research Gap

Available data clearly pinpoints many factors influencing dropout but suffers from two main deficiencies: Methodological Deficiency: Most local studies rely on traditional linear statistical models (e.g., OLS or simple Logistic Regression). These models often fail to capture the complex, non-linear interactions between numerous determinants. They lack the predictive power necessary for high-precision, individual-level risk identification. Lack of Empirical Prioritization: Current studies fail to provide a ranked hierarchy of determinants derived from their quantitative impact [8]. This is crucial for resource allocation, as policy makers do not know which factor (e.g., poverty vs. teacher absenteeism) is most efficient to address first.

The Research Gap: This study focuses on the need for a robust, high-precision machine learning framework (through the use of XGBoost) to empirically rank the primary drivers of secondary school dropout in Uganda's Central Region. This approach will change generalized policy into targeted, evidence-based intervention.

Milestone 4 (Chapter Three)

Research Paradigm and Design

Pragmatic paradigm is employed for the project, thus integrating quantitative (machine learning) and qualitative (policy evaluation) factors. The core method is a quantitative, predictive modeling design. This is a non-experimental, cross-sectional design that examines secondary data to predict a binary outcome (dropout) [13].

3.2 Scope of the Study

3.2.1 Geographical and Contextual Scope

- Central region (Uganda) focus due to high population distribution and varied institutional factors.
- Secondary students

3.2.2 Thematic Scope

- The research centers on three groups of determinants: sociodemographic, institutional, and behavioral factors. The outcome variable is student dropout.

3.2.3 Data Source and Target Population

- uses secondary data from the current Uganda Bureau of Statistics (UBOS) National Household Survey.
- This dataset offers a longitudinal perspective over one academic year. The target population is all secondary school students embedded in the survey.
- Sampling is not mandatory as the study utilizes the full population from the secondary dataset.

3.3 Data Processing

Data processing comes after a standard measure in the field:

Data Cleaning: Handling missing values and removing outliers.

Feature Engineering: Creating new variables from existing raw features (e.g., socioeconomic index).

Encoding: Changing categorical variables (e.g., school type) into numeric expressions (One-Hot Encoding).

Normalization: Scaling all features to ensure equal contribution during model training.

3.4 Analytical Plan and Modeling

3.4.1 Machine Learning Models

Three binary classification models were compared: Logistic Regression (LR): A baseline, linear model for probability estimation. Random Forest (RF): An ensemble method using decision trees to mitigate overfitting. Extreme Gradient Boosting (XGBoost): A high-performance boosting technique expected to yield the highest precision [11].

3.4.2 Model Training and Validation

The dataset was segmented into 70\% for training and 30\% for testing. Training used 10-fold cross-validation (CV) for robust parameter tuning.

3.4.3 Performance Metrics

Model performance was evaluated based on: Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). The F1-Score was the primary metric because of potential class imbalance (dropout vs. retention).

3.4.4 Determinant Ranking

The top-performing model's results were used to rank determinants. The Feature Importance Score (Gini Importance for RF/XGBoost, or coefficients for LR) identified the most impactful factors.

3.4.5 Proposed Econometric Evaluation (RCT Framework)

The predictive model's output is used to advocate a Randomized Control Trial (RCT) framework for future policy evaluation. This framework targets high-risk students identified by the ML model. The RCT divides students into:

Treatment Group: Receives the new, targeted intervention.

Control Group: Receives the standard intervention.

The shift in dropout rate between groups is quantified using the Difference-in-Differences (DiD) method. This offers rigorous, causal evidence on intervention effectiveness.

3.5 Quality Assurance (Validity and Reliability)

3.5.1 Validity

External Validity is guaranteed by using large-scale UBOS data, supporting generalizability across Uganda. Construct Validity is confirmed through mapping all variables directly to the established Push/Pull factor conceptual framework [9].

3.5.2 Reliability

Reliability is sustained by using established, open-source programming libraries (Python, Scikit-learn). All data processing and modeling procedures are recorded and fully replicable [10].

3.6 Ethical Considerations

All data utilized is secondary and anonymized by the provider (UBOS), voluntary participation, no personal identifying information (PII) was used. Data access was granted through formal permission from the Ministry of Education and Sports (MoES),

3.7 Assumptions and Limitations

3.7.1 Assumptions

The UBOS dataset is faultless and representative of the overall secondary school student population in the Central Region. The features selected in the study symbolize the true independent determinants of the dropout outcome [12].

3.7.2 Limitations

Small sample size

Limited generalizability

Potential for response bias.

5. Milestone 5

Conclusion

This study established a specific predictive framework for student dropout in Uganda's Central Region, using advanced machine learning models. The research confirmed demographic and educational parameters, like age and grade, as crucial determinants of student risk levels. The study's findings are anticipated to significantly reduce secondary school dropout rates, enhance educational outcomes, and support Uganda's National Development Plan IV and Sustainable Development Goal 4: Quality Education. The predictive framework can be used to target interventions, allot resources efficiently, and foster inclusive and equitable education.

Workplan and Budget

A detailed work plan outlining research activities, timelines, and budget is displayed as follows;

Milestones	**Task/section**	**Estimated Duration (weeks)**	**Budget (Cost)**
Milestone 1	a) Topic b) Executive summary c) SDG and NDP (IV)	3 days	$0
Milestone 2 (Introduction)	a) Background of study b) Problem statement c) Objectives d) Research questions	2 weeks	$ 50
Milestone 3 (Literature review)	Literature review	4 weeks	$150

Milestone 4 (Methodology)	a) Research Design b) Population and study area c) Data Collection& sources d) Data preprocessing e) Predictive modeling and evaluation	6 weeks	$ 200

Total Duration 12 weeks and 3 days and total cost is $400.

Qn 2: For my proposed project, I would submit to the following two reputable publication outlets that is; International Journal of machine learning and computing (IJMLC) and Nature Machine Intelligence (peer-reviewed journal).

Reputable publication Outlet	International Journal of machine learning and computing (IJMLC)	Nature Machine Intelligence
Justification		
Scope and relevance of the outlet	it focuses on the application of technology, such as machine learning (ML) which easily detects students prone to dropping out of secondary school within the education sector and helps the school administrators, stakeholders and policymakers mitigate the problem at hand before it gets out of control ie the severe dropout rates.	Highly relevant to my research area, as it publishes ground breaking work in machine learning and artificial intelligence, corresponding closely with my project's focus.
Target audience	policymakers, researchers, educators, ministry of education, this is because they would aim to provide a solution to the real-world education problem (dropout rates).	Targets a broad audience of researchers, engineers and practitioners in machine learning, making it a suitable outlet for sharing my work to those

		who can benefit from and build upon it.
Impact factor	IJMLC being a well-recognized journal in the field of machine learning, aims at validating its technical contribution to the project through the recognition it gives in detecting early warning signs of dropouts and possible solutions to mitigating the high number.	With an impact factor close to 23.9, Nature Machine Intelligence is respected for publishing influential and extensively viewed research, guaranteeing my work will reach an influential readership.
Regional or global visibility	the journal has a global viewership, thus the Ml methodology employed in the project is subjected to the international community of computer scientists.	As a Nature-branded journal, Nature Machine Intelligence has worldwide visibility, with a strong online presence and distribution network that ensures my research a worldwide audience.
Open access or indexing status	IJMLC is an open access journal which means it can be accessed freely by anyone and is indexed in couple of academic databases, increasing its visibility.	Nature Machine Intelligence provides open-access options, allowing my research to be freely available to anyone, anywhere, and at any time amplifying its potential reach and impact.

References

1. Haimovich, F.; Vázquez, E.; Adelman, M. Scalable Early Warning Systems for School Dropout Prevention: Evidence from a 4.000-SchoolRandomized Controlled Trial; Universidad Nacional de La Plata, Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS): La Plata, Argentina, 2021.
2. Adelman, M. A., & Szekely, M. (2016). School dropout in Central America: An overview of trends, causes, consequences, and promising interventions. World Bank Policy Research Working Paper, (7561).
3. Uganda Bureau of Statistics. (2021). Uganda National Household Survey 2019-2020 report.https://www.ubos.org/wp-content/uploads/publications/09_2021Uganda-National-Survey-Report-2019-2020.pdf
4. Villano, R., Harrison, S., Lynch, G., & Chen, G. (2018). Linking early alert systems and student retention: a survival analysis approach. Higher Education, 76, 903-920.
5. Williamson, B. (2016). Digital education governance: data visualisation, predictive analytics, and real-time policy instruments. Journal of education policy, 31(2), 123-141.

6.Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

7. Elliot, A. J. (2006). The hierarchical model of approach-avoidance motivation. Motivation and Emotion, 30(2), 111-126. https://doi.org/10.1007/s11031-006-9028-7

8. Becker, G. S. (1993). Human capital: A theoretical and empirical analysis, with special reference to education (3rd ed.). University of Chicago Press.

9. Freeman, J., & Simonsen, B. (2015). Examining the impact of policy and practice interventions on high school dropout and school completion rates: A systematic review of the literature. Review of educational research, 85(2), 205-248.

10. Lubaale Y.A.M (2010) Orphans and Vulnerable children in Uganda; Is it a homogenous group PhD Thesis Makerere University, Unpublished

11. Kim,S.; Choi, E.; Jun, Y.-K.; Lee, S. Student Dropout Prediction for University with High Precision and Recall. Appl. Sci. 2023, 13, 6275. [CrossRef]

12.[Fluke Corporation, 2005] Fluke Corporation (2005). The basics of predictive / preventive maintenance. *none*, pages 1–6.

13.Almalki, S. (2016). Integrating Quantitative and Qualitative Data in Mixed Methods Research--Challenges and Benefits. Journal of education and learning, 5(3), 288-296.

14. UNESCO (2018). Uganda: Education and literacy. UNESCO Institute for Statistics.