A MACHINE LEARNING APPROACH FOR IDENTIFYING AT-RISK PUPILS AND RECOMMENDING SUPPORT STRATEGIES: A CASE STUDY OF PRIMARY SCHOOLS IN MUKONO DISTRICT, UGANDA

CHARLES JOVANS GALIWANGO

J24M19/002

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

May, 2026

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Abstract

Academic vulnerability and pupil dropout remain persistent challenges in Ugandan primary education, despite high enrollment rates. Current school support systems are often reactive, intervening only after academic failure has occurred. This study developed a predictive early warning system to proactively identify pupils at risk of academic failure in Mukono District, Uganda.

A mixed-methods approach was used, analysing structured records of pupils from Primary 4–6 and conducting interviews with teachers and administrators. The study first identified key behavioural and socioeconomic predictors of academic risk through statistical analysis. Four machine learning models were then evaluated and compared to determine the most effective approach for predicting vulnerability.

The analysis revealed that behavioural indicators, specifically disciplinary issues, incomplete homework, and poor attendance, were the strongest predictors of academic risk. Among the models tested, Logistic Regression proved most suitable, achieving a recall of 0.833 and ROC-AUC of 0.941 on unseen test data, while providing interpretable predictions crucial for educational settings. Based on these findings, a three-tiered intervention framework was developed, classifying pupils by risk level and linking specific risk factors to tailored support strategies.

The study concludes that a simple, interpretable predictive model using routinely collected school data can effectively identify vulnerable pupils early. The proposed framework offers Ugandan primary schools a practical, proactive tool for targeted intervention, shifting support from crisis management to prevention. This research contributes a feasible, evidence-based approach to enhancing educational equity and retention in resource-constrained settings.

Keywords: Academic vulnerability, Early warning systems, Predictive modeling, Primary education, Data-driven intervention.

Declaration

I, GALIWANGO CHARLES JOVANS, hereby declare that this is my original work and it has not been used anywhere else or presented by anyone for any academic reward/qualification in any institution of higher learning, and that references have been specified where I used other sources of information already published.

Signature:

Date: May 21st, 2026

Approval

I, the undersigned, acknowledge GALIWANGO CHARLES JOVANS of Registration Number J24M19/002 drafted this thesis under my supervision and he has been recommended for submission to the Department of Computing and Technology at Uganda Christian University.

ACADEMIC SUPERVISOR

Name: Dr. Daphne Nyachaki Bitalo.

Signature: 

Date: 28/05/2026

Acknowledgements

I am deeply grateful to my supervisor, Dr. Daphne Nyachaki Bitalo, for her guidance, critical feedback, and encouragement throughout this research. Her insights shaped this work excellently and her support was invaluable.

I thank the schools in Mukono District, their administration, teachers, and staff for welcoming this research and providing the data that made it possible. I am especially grateful to the pupils whose records formed the foundation of this study and to the educators who shared their experiences. I also appreciate the support of the Department of Computing and Technology and my colleagues for the stimulating discussions and shared learning.

Finally, I thank my family and friends for their patience and understanding during this demanding period, and I acknowledge God's grace in enabling me to complete this work.

Table of Contents

Abstract i
Declaration ii
Approval iii
Acknowledgements iv
1 Chapter One: Introduction 1
1.1 Introduction 1
1.2 Background Study 1
1.2.1 Global Trends in Educational Analytics and Local Relevance 2
1.2.2 Alignment with Sustainable Development Goals and National Goals 3
1.3 Problem Statement 4
1.4 Research Objectives 5
1.4.1 Main Objective 5
1.4.2 Specific Objectives 5
1.5 Research Questions 6
1.6 Justification of the Study 6
1.7 Scope of the Study 6
1.7.1 Geographical Scope 6
1.7.2 Target Population 7
1.7.3 Conceptual Scope 7
1.7.4 Methodological Scope 7
1.8 Conceptual Framework 7
2 Chapter Two: Literature Review 9
2.1 Introduction 9
2.2 Uganda's Primary Education Context 9
2.3 Factors Contributing to Academic Vulnerability 10
2.4 Limitations of Current Assessment and Intervention Methods 11
2.5 Predictive Models in Education 12
2.5.1 Logistic Regression in Dropout Prediction 12
2.5.2 Random Forest and Ensemble Methods 13
2.5.3 Gradient Boosting (XGBoost) 14
2.5.4 Support Vector Machines (SVM) 14
2.6 Early Warning Systems (EWS) in Schools 15

2.7 Summary of Literature Gaps 16
3 Chapter Three: Methodology 18
3.1 Research Design 18
3.2 Study Population and Sampling 18
3.2.1 Study Focus and Population 18
3.2.2 Study Area 19
3.2.3 Sampling Strategy 19
3.2.4 Sample Size 19
3.3 Data Collection Methods 20
3.3.1 Quantitative Data Collection 21
3.3.2 Qualitative Data Collection 21
3.4 Data Analysis Techniques 21
3.4.1 Qualitative Data Analysis 21
3.4.2 Analysis for Objective 1: Identifying Key Indicators and Risk Factors 22
3.4.3 Analysis for Objective 2: Model Implementation, Evaluation and Validation 23
3.4.4 Analysis for Objective 3: Evidence-based support strategies aligned with the model's outputs for timely and targeted intervention. 26
3.5 Ethical Considerations 27
4 Chapter Four: Results 28
4.1 Introduction 28
4.2 Objective One: Key Indicators of Academic Vulnerability 28
4.2.1 Academic and Engagement Profile 28
4.2.2 Behavioral and Socioeconomic Indicators 29
4.2.3 Non-Significant Factors 29
4.3 Objective Two: Performance and Validation of the Predictive Model 29
4.3.1 Comparative Model Performance 29
4.3.2 Validated Performance of the Optimized Model 30
4.3.3 Interpretation of Model Predictions 31
4.4 Objective 3: Evidence-Based Support Strategies from Model Outputs 33
4.4.1 Tiered Intervention Framework Output 33
4.5 Chapter Summary 34
5 Chapter Five: Discussion 35
5.1 Introduction 35

5.2 Interpretation of Key Risk Indicators 35
5.3 Interpretation of Model Performance and Selection 36
5.4 From Prediction to Recommended Support 37
5.5 Addressing Identified Research Gaps 38
5.6 Implications for Policy and Practice 39
5.7 Limitations of the Study 40
5.8 Recommendations for Future Research 40
6 Chapter Six: Conclusion 42
References 44
Appendix A: Summary of Key Variables 49
Appendix B: Interview Guide 50
Appendix C : Statistical Tests 52
Appendix D: Python Code and Jupyter Notebook for Model Development 53
Appendix E: Sample Informed Consent Form (Anonymized) 54

List of Figures

Figure 1:Conceptual Framework for Data-Driven Identification of At-Risk Pupils and Intervention Planning. 8
Figure 2: SHAP Summary Plot of Feature Effects on Risk Classification. 32
Figure 3: Feature Importance Rankings for Risk Prediction. 33

List of Tables

Table 4. 1: Key Factors Associated with Academic Vulnerability 29
Table 4. 2: Comparison of Model Performance for Identifying At-Risk Pupils 31
Table 4. 3: Final Model Performance on Test Data 31
Table 4. 4: Distribution of Pupils by Model-Derived Risk Tier 34

Table 5. 1: Comparison of Intervention Approaches 40

Table A: Summary of Key Variables collected for Predictive Modeling 51

Table C. 1: Cross-tabulation of Gender and Academic Risk Status 54
Table C. 2: Cross-tabulation of Class Level and Academic Risk Status 54
Table C. 3: Descriptive Statistics for Age by Risk Group 54

1 Chapter One: Introduction

1.1 Introduction

Education remains a key driver of individual progress and national development, and in Uganda, primary education plays a crucial role in shaping future opportunities for learners, especially those from vulnerable communities (The World Bank & UNESCO Institute for Statistics (UIS), 2024). While various efforts have been made to expand access to education, many pupils continue to face challenges that hinder their academic success and progression.

There has been growing interest in the use of data-driven approaches to improve learning outcomes. Predictive models using existing student data to identify those at risk of underperforming have emerged as a promising strategy in other education systems (Bowers & Sprott, 2012). However, application in Ugandan primary schools is still limited. Early identification of at-risk pupils could offer schools and policymakers a more proactive way to provide support before learning gaps widen (The Bantwana Initiative of World Education, 2021).

This study sought to explore the potential of predictive modeling to enhance early identification of academically at-risk pupils and guide the recommendation of targeted support strategies. Focusing on primary schools in Mukono District, this research aimed to identify key indicators, implement a predictive model and inform timely interventions.

1.2 Background Study

Universal Primary Education (UPE) implementation in 1997 increased access to education, with enrollment rising from 2.5 million (1996) to over 8.7 million (2022) (Uganda Bureau of Statistics, 2023). However, this quantitative expansion has been accompanied by persistent educational quality and retention challenges. Only 42.7% of P6 pupils demonstrate proficiency in literacy and 58% in numeracy (Uganda National Examinations Board, 2023).

According to the World Bank and UNESCO Institute of Statistics, the primary school completion rate was at 53.7% in 2017 for girls and 52.3% for boys (The World Bank & UNESCO Institute for Statistics (UIS), 2024), with significant disparities between urban (59%) and rural areas (53%) (National Planning Authority, 2018).

The traditional educational system mainly relies on end-of-term exams and limited implementation of continuous assessment techniques, despite policy recommendations (Mitana, 2018). This approach delays the identification of academic challenges, making early intervention difficult. The Ministry of Education and Sports has acknowledged this gap, citing a lack of systematic, data informed methods for identifying struggling pupils before performance declines significantly (Ministry of Education and Sports, 2021).

1.2.1 Global Trends in Educational Analytics and Local Relevance

Educational data science is revolutionizing student support worldwide, improving student outcomes. Predictive learning models have emerged as a powerful tool for identifying students at risk of failure and informing targeted interventions for students in that category (Herodotou et al., 2019). Research demonstrates predictive models can identify 78% of at-risk students 3-4 months before academic failure manifests,

giving schools a critical opportunity for timely support manifested (Adejo & Connolly, 2018).

Several established early warning systems demonstrate the effectiveness of this approach:

1. Early Warning Intervention and Monitoring System (EWIMS) - Implemented across multiple U.S. states, EWIMS uses attendance, behaviour, and course performance (the "ABCs") to identify at-risk students (Therriault et al., 2017). Schools implementing EWIMS experienced a 5% reduction in chronic absenteeism and improved course completion rates compared to control schools. (Faria et al., 2017)

2. Bridge International Academies' Teacher Guides - Operating in Kenya, Nigeria, Uganda, and India, Bridge uses structured teacher guides with embedded assessment tools that feed into a centralized monitoring system. Their approach includes weekly assessments that are analyzed to identify struggling pupils, resulting in reported learning gains of 0.89 additional years of learning compared to peers (Bridge International Academies, 2019).

Within the Ugandan context specifically, there are limited but emerging applications:

1. Kolibri Platform (Kampala/Wakiso) has demonstrated the potential of offline-capable analytics, supporting learning in areas with limited internet access (National Information Technology Authority-Uganda, 2022). However, it lacks the predictive capabilities to identify at-risk pupils early (NITA-U, 2022).

2. UNICEF's Primero System has shown cross-sector potential, supporting data collection and management in education, but it is not designed to predict academic vulnerabilities or provide targeted interventions (UNICEF Uganda, 2021).

While some efforts have been made to improve education through data-driven approaches, the use of predictive models in low-resource settings, especially in Sub-Saharan Africa, remains quite limited. In Uganda’s primary education system, such models are still rare. Most of the current tools focus more on basic monitoring rather than making predictions, and they often lack links to practical intervention strategies.

Some Local initiatives have started to bridge this gap. For instance, Alfred et al. (2023) analyzed Primary Leaving Examination (PLE) results from schools in Kiryandongo District using trend analysis. Their findings highlight that existing academic records can be effectively used to model and predict pupil performance, laying a foundation for early identification of those at risk of falling behind.

1.2.2 Alignment with Sustainable Development Goals and National Goals

This research aligns closely with both global and national education goals.

At the global level, it supports Sustainable Development Goal 4, particularly Target 4.1, which emphasises quality and inclusive primary education for all (United Nations, 2015). It also ties into indicator 4.1.2, which calls for stronger systems to identify learners at risk of falling behind or dropping out. (The World Bank & UNESCO Institute for Statistics (UIS), 2024)

Nationally, the study responds to Uganda's Fourth National Development Plan (NDP IV), which prioritizes improving access, equity, and quality of education at all levels (National Planning Authority, 2025). Similarly, the Education and Sports Sector Strategic Plan (ESSP) highlights the need for evidence-based interventions for vulnerable learners and the integration of technology in education (Ministry of Education and Sports, 2020).

The Education Response Plan for Refugees and Host Communities also recognizes early identification of struggling learners as key to addressing educational inequalities. These priorities all reinforce the relevance of this research, which aimed to implement a predictive model tailored to Uganda’s primary school context (Ministry of Education and Sports & UNHCR, 2022).

1.3 Problem Statement

Despite Uganda's 91% primary school enrolment rate (MoES, 2023), 60% of pupils fail to complete Primary 7 (National Planning Authority, 2025), with Primary 4 to 6 identified as the most vulnerable period (Uganda Bureau of Statistics, 2023). Over 50% of Primary 6 pupils fail to meet basic literacy and numeracy standards (Uganda National

Examinations Board, 2023), and like many districts, some schools in Mukono also face challenges in achieving consistent PLE performance.

Dropout is influenced by multiple factors, including socio-economic factors such as school fees and home challenges (Mike et al., 2008), but many cases begin with unnoticed academic-related disengagements such as chronic absenteeism, declining participation, incomplete homework, and learning difficulties (Gottfried, 2019). Unfortunately, schools only detect failure after it has already occurred, with no structured early warning system to identify struggling pupils early enough for effective intervention (Davis et al., 2019).

Various initiatives have introduced tools to improve learning outcomes in Uganda, yet they lack the predictive capabilities needed for early intervention. USAID's School Health and Reading Program (SHARP) tracks attendance but does not proactively identify struggling pupils. RTI's Early Grade Reading Assessment (EGRA) provides periodic literacy evaluations, but these are not linked to structured interventions. (RTI International, 2022). Kolibri, though offering offline-capable analytics, focuses on content delivery rather than academic risk prediction, while UNICEF's Primero System, designed for child protection case management, is not tailored for academic performance monitoring.

Machine learning approaches have shown promise in other contexts. Ahmed and Warsame (2024) predicted dropout in Somaliland; Mnyawami et al. (2022) achieved 99.8% accuracy in Tanzania; Krüger et al. (2023) applied explainable ML in Brazil. However, limited models exist for Ugandan primary schools, most prioritize accuracy over teacher interpretability, and predictions are rarely linked to actionable intervention strategies that schools can apply (Herodotou et al., 2019).

To address this gap, this study applied and evaluated machine learning techniques to build a predictive model tailored for primary schools in Mukono district, as the case study. By analyzing available data on attendance, academic records, classroom engagement, and related factors, the model aimed to identify pupils at risk of academic failure early enough to recommend appropriate support strategies. This proactive

approach is expected to improve retention, enhance learning outcomes, and contribute to national education goals.

1.4 Research Objectives

1.4.1 Main Objective

Apply machine learning to build a predictive model that identifies academically at-risk pupils and informs support strategies in primary schools within Mukono district, Uganda.

1.4.2 Specific Objectives

1. Identify key indicators and risk factors that predict academic vulnerability among pupils.
2. Evaluate, compare, and validate the performance of predictive modeling approaches for detecting at-risk pupils.
3. Propose evidence-based support strategies aligned with the model's outputs for timely and targeted intervention.

1.5 Research Questions

This Research sought to answer the following questions:

1. What are the key indicators and risk factors that can reliably predict academic vulnerability among pupils?
2. Which predictive modeling approach most accurately and reliably identifies at-risk pupils, and how does it perform upon validation?
3. How can the model's outputs be used to inform practical support strategies for at-risk pupils?

1.6 Justification of the Study

Many pupils in Ugandan primary schools struggle academically, but they are often only identified after poor results or when they drop out (Christine Mbabazi Mpyangu et al., 2014). This study aimed to change that by introducing a predictive model that helps schools detect at-risk pupils early, allowing timely support. The research was especially relevant in Mukono District, where dropout rates remain a concern. It supports national goals under Uganda's Fourth National Development Plan (NDP IV), which prioritizes

improving access, equity, and quality of education at all levels (National Planning Authority, 2025), and contributes to SDG 4, which targets inclusive and quality education for all. Improving learning at the primary level has lasting benefits. Studies show that more years in school can increase a person's future income and reduce poverty. By helping more pupils stay in school and succeed, this study also supports broader human development and economic growth.

1.7 Scope of the Study

1.7.1 Geographical Scope

The study was conducted in selected primary schools within Mukono District, representing a mix of urban and rural settings. This district was chosen due to its high dropout rates in some areas and performance in the Primary Leaving Examinations.

1.7.2 Target Population

The study focused specifically on pupils in Primary Four to Primary Six (P4–P6), as this stage has been identified as a critical risk period for dropout and academic decline (The World Bank & UNESCO Institute for Statistics (UIS), 2024).

1.7.3 Conceptual Scope

The study analyzed a combination of academic indicators (such as subject scores, attendance, and classroom performance), behavioral indicators (homework completion and class participation), and background factors (like socioeconomic status). These formed the basis for creating and testing the predictive model.

1.7.4 Methodological Scope

A mixed-methods design was used in this research. Quantitative methods were applied to analyze school data and create the predictive model. At the same time, qualitative data, such as interviews or teacher reports, allowed for an exploration of contextual challenges that pupils face and how schools can respond to the identified risks. This combined strategy ensured that both data-informed results and real-world applicability were attained (Herodotou, Rienties, Verdin, et al., 2019).

1.8 Conceptual Framework

To guide the prediction of academic risk and the following recommendation of support strategies within the Ugandan primary school context, this study adopted the data-informed conceptual framework illustrated in Figure 1 below. The framework was structured around the interaction of multidimensional factors, including academic history, attendance patterns, behavioral indicators, and socioeconomic background, that collectively influence a pupil's educational trajectory.

Central to this process is the Predictive Model, which ingests structured school data to generate insights. As detailed in the Data Variables section of Figure 1, the model processes key inputs ranging from subject scores and attendance rates to behavioral markers like homework completion and disciplinary history, alongside background characteristics such as family income level.

Following processing, the model generates individual Risk Probability scores. Based on established thresholds, these scores convert into Risk Classifications (for example high, moderate, or low likelihood of underperformance). These classifications serve as the trigger mechanism for context-appropriate Support Strategies, such as remedial instruction, parental engagement programs, mentoring, or psychosocial support.

Ultimately, this framework functions as a cyclical, data-driven system. Model predictions inform immediate intervention planning, while subsequent pupil outcomes feed back into the system for model refinement. This structure enables schools to transition from reactive measures to proactive support, leveraging existing data to improve academic retention in a sustainable manner.

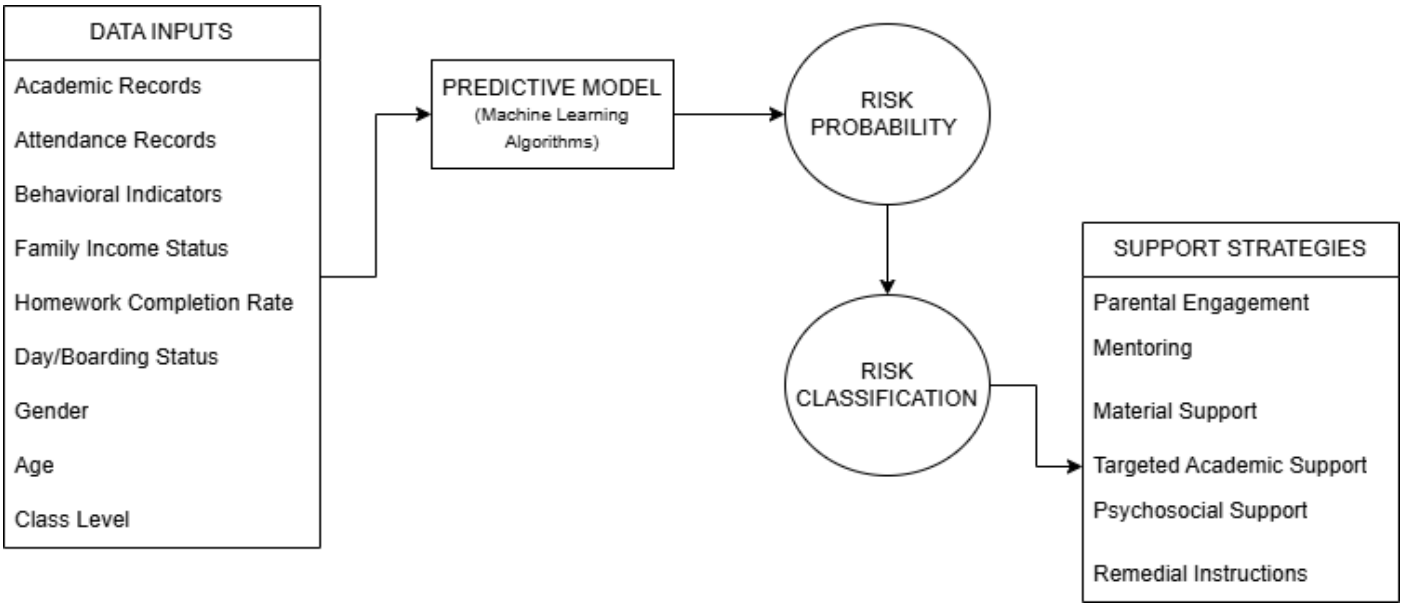

Figure 1:Conceptual Framework for Data-Driven Identification of At-Risk Pupils and Intervention Planning. The framework illustrates how data inputs are processed through a predictive model to generate risk probabilities, which are then classified into risk levels and linked to appropriate support strategies.

2 Chapter Two: Literature Review

2.1 Introduction

This chapter reviews existing literature relevant to the development of a predictive model for identifying and supporting academically at-risk pupils in primary schools. It discusses the state of Uganda's primary education, identifies the main factors influencing pupil dropout and poor academic performance, examines the limitations of current assessment and intervention methods, explores the application of predictive models and early warning systems in education, and highlights gaps in the current research that the study sought to address.

2.2 Uganda's Primary Education Context

Uganda introduced the Universal Primary Education (UPE) program in 1997 to eliminate financial barriers to education and boost enrolment rates across the country. The policy was a major milestone, increasing the number of children attending school from approximately 2.5 million in 1996 to over 8 million (Ministry of Education and Sports, 2022; Uganda Bureau of Statistics, 2023). Over the years, Uganda has maintained high enrolment levels, with a Net Enrolment Rate of 91% reported (Uganda Bureau of Statistics, 2020).

Despite the impressive gains in access, serious challenges remain regarding retention, progression, and completion. Data from the Uganda National Examinations Board (UNEB) and the Uganda Bureau of Statistics (UBOS) reveal that 60% of pupils who enroll in Primary One fail to complete Primary Seven (National Planning Authority, 2025). Dropout rates are particularly high between Primary Four and Primary Six, where learners are often affected by early marriage, child labour, financial hardship, long distances to schools, and loss of interest in learning (Ministry of Education and Sports, 2022; Oweyegha-Afunaduula, 2025; Uganda Bureau of Statistics, 2023).

Socio-economic factors persist, with poorer regions like Karamoja and West Nile experiencing even higher dropout and repetition rates compared to urban centers like Kampala (UNICEF UGANDA, 2020). Although several interventions have been made to address these disparities, their impact remains limited, and more sustainable, data-driven solutions are needed to support pupils at risk of academic failure.

2.3 Factors Contributing to Academic Vulnerability

Academic vulnerability among primary school pupils is rarely caused by a single issue; it usually results from a combination of academic, behavioral, socioeconomic, and environmental factors. Understanding these risk factors is essential for building effective early intervention strategies.

1. Academic Indicators

Low performance in core subjects such as literacy and numeracy is one of the strongest early signs of academic risk. Pupils who consistently perform below the expected level in class tests or standardized exams are more likely to fall behind their peers (The World Bank & UNESCO Institute for Statistics (UIS), 2024). Pupils who are frequently absent tend to miss important class activities and explanations, which affects their overall understanding and performance

2. Behavioral Indicators

Behaviours in the classroom, like staying quiet during lessons, not finishing homework, or causing disruptions, can reflect underlying academic or emotional issues. Research by Herodotou and others shows that tracking these behaviours alongside academic performance can help identify students who are likely to struggle in the future (Herodotou et al., 2019).

3. Socioeconomic Factors

Poverty remains a significant challenge to learning in Uganda. Pupils from low-income families often struggle with issues like not having school supplies, poor nutrition, or being expected to help with work at home (The World Bank & UNESCO Institute for Statistics (UIS), 2024). These hardships can also limit how much parents engage in their children's education, which plays a key role in school success.

4. Distance and School Infrastructure

In many rural areas, pupils walk long distances to get to school, which can leave them tired, often late, or even cause them to miss school altogether. On top of that,

overcrowded classrooms, a shortage of textbooks, and poor sanitation facilities can make it hard for pupils to learn (UNICEF UGANDA, 2020).

5. Family and Community Support

Research shows that when parents are involved in their children's education and provide a supportive home environment, children tend to do better in school (Epstein, 2018). Also, children from unstable households due to issues like divorce, domestic violence, or losing a parent are more likely to struggle.

While social and economic factors affect learning, signs like falling grades, missing school, or not participating in class show up more quickly, making them the most critical starting point for early intervention efforts.

2.4 Limitations of Current Assessment and Intervention Methods

In many primary schools, including those in Uganda, the education system heavily relies on summative assessments, such as end-of-term examinations, to evaluate pupils' academic progress. While these assessments provide a snapshot of achievement, they often come too late to offer timely support to struggling learners. By the time results reveal poor performance, pupils may have already fallen significantly behind, making interventions less effective and recovery more difficult (Herodotou et al., 2019).

Another major limitation is that summative assessments focus mainly on academic outcomes, ignoring important early warning signs like declining attendance, reduced classroom engagement, or behavioral changes. As a result, pupils facing challenges outside the classroom such as poverty, family instability, or health problems often go unnoticed until their academic performance drastically drops.

Additionally, intervention strategies in most schools are reactive rather than proactive. Support programs are often only initiated after a pupil has already failed or dropped out, rather than identifying risk early and preventing failure. This reactive approach not only limits the effectiveness of interventions but also places additional pressure on teachers and school administrators to "catch up" pupils who may have already disengaged from learning.

Research has shown that the absence of integrated data systems severely limits schools' ability to track pupil progress effectively. Without real-time access to indicators such as attendance, behavior, and academic performance, teachers and administrators struggle to recognize patterns of academic risk early enough for intervention (Elaine Allensworth & John Q. Easton, 2007).

The current assessment and intervention methods in primary education are too delayed, too narrow, and too reactive to address academic vulnerability effectively. There is a critical need for proactive, data-driven approaches that can detect risk factors early and support pupils before failure becomes inevitable.

2.5 Predictive Models in Education

While early educational research relied on simple correlation analysis, recent scholarship has shifted toward machine learning classifiers to handle the complexity of student data. A survey of machine learning approaches for dropout prediction in developing countries confirms that various algorithms have been applied, though the optimal choice often depends on local data characteristics and implementation context (Mduma, Kalegele & Machuve, 2019). The literature reveals a distinct trade-off between model interpretability and predictive accuracy. This section reviews four primary algorithms, Logistic Regression, Random Forest, XGBoost, and Support Vector Machines highlighting their application and performance in previous studies.

2.5.1 Logistic Regression in Dropout Prediction

Logistic Regression remains the foundational baseline in Educational Data Mining (EDM) due to its transparency. Unlike black-box models, Logistic Regression allows educators to quantify exactly how much a specific variable, such as absenteeism, increases the probability of risk. In a comparative study using data from a Korean distance university, Seo et al. (2024) evaluated multiple machine learning algorithms for dropout prediction. While gradient boosting methods achieved marginally higher accuracy, the authors specifically recommended Logistic Regression for deployment when interpretability is prioritized, noting its competitive performance and superior transparency for educational stakeholders.

In Africa, the algorithm has also demonstrated practical value. A study predicting student dropout in Somaliland using the 2022 National Education Accessibility Survey found that while ensemble methods achieved higher overall accuracy, Logistic Regression provided clearer insights into specific risk factors, highlighting age and student grade level as critical predictors through interpretable odds ratios (Ahmed & Warsame, 2024). This transparency is particularly valuable in resource-constrained settings where non-technical staff must understand and act upon model outputs. However, these studies consistently note that Logistic Regression struggles to capture complex interactions between behavioural and socioeconomic variables, often underperforming in diverse student populations. This suggests that while Logistic Regression is excellent for explaining risk, it may lack the sensitivity required for high-precision identification in some contexts.

2.5.2 Random Forest and Ensemble Methods

To address the limitations of linear models, researchers have increasingly adopted ensemble methods like Random Forest. It aggregates the predictions of multiple decision trees to reduce overfitting and handle imbalanced datasets, a common issue in dropout prediction where at-risk students are the minority. In East Africa, Mnyawami et al. (2022) used automated machine learning to predict secondary school dropouts in Tanzania, achieving 99.8% accuracy with Decision Trees. Their study achieved high prediction accuracies using automated feature selection and hyperparameter tuning. The researchers concluded that Random Forest's ability to model non-linear relationships made it effective for identifying struggling students, and that proper feature selection is critical for identifying root causes of dropout that can be addressed through early intervention.

Similarly, the Somaliland study by Ahmed and Warsame (2024) identified Random Forest as their most accurate model, achieving 95.00% accuracy in predicting dropout rates. The algorithm's feature importance rankings revealed that household income, housing type, and student grade level were among the most significant predictors. It also provides robust feature importance rankings, offering a compromise between the black-box nature of advanced AI and the need for educational interpretability.

2.5.3 Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) represents the current state-of-the-art in tabular data classification. It refines the boosting process by sequentially correcting errors from previous models. In large-scale educational studies, XGBoost has consistently demonstrated superior performance metrics. In a study from Bangladesh examining school dropout among students aged 6-24 years, Khatun et al. (2025) used a hybrid framework combining machine learning with explainable AI. The XGBoost model achieved the best performance with an accuracy of 94.4%, outperforming Random Forest. The study used SHAP and LIME to enhance interpretability, revealing that age, sex, completed grade, wealth index, and parental education were key factors influencing dropout.

A study in Brazilian schools by Krüger et al. (2023) applied explainable machine learning for dropout prediction, achieving an AUC-PR score of up to 89.5%. Their work demonstrated differences in prediction performance across educational stages and times of the academic year, highlighting the need for context-specific model development. However, the literature notes that XGBoost requires significant computational resources and careful hyperparameter tuning to prevent overfitting on smaller datasets (Chen & Guestrin, 2016), which may present challenges for schools with limited technical infrastructure.

2.5.4 Support Vector Machines (SVM)

Support Vector Machines are frequently cited for their effectiveness in high-dimensional spaces. Studies have found that while SVM can perform reasonably well with appropriate hyperparameter tuning, it typically underperforms compared to boosting algorithms on structured educational data (Villar & de Andrade, 2024). They often highlight a significant drawback: SVM models are difficult to interpret for non-technical stakeholders. While accurate in certain configurations, they do not easily translate into the if-then rules that teachers prefer for intervention planning. This limitation has led many recent studies to favor interpretable models or to supplement complex models with explainability techniques.

The literature reveals a clear trade-off between accuracy and interpretability. Complex models like XGBoost and Random Forest achieve higher accuracy, but simpler models like Logistic Regression offer the transparency educators need to understand and act on predictions. Most studies focus on secondary or higher education. Primary schools remain largely unexplored. Existing research also emphasises accuracy metrics over practical utility, with few studies providing actionable guidance for teachers.

This study addresses these gaps by comparing four machine learning algorithms in Ugandan primary schools within Mukono District, evaluating them on both performance and interpretability, and proposing a tiered intervention framework that links predictions to practical support strategies.

2.6 Early Warning Systems (EWS) in Schools

Early Warning Systems (EWS) are structured frameworks designed to identify students at risk of academic failure, dropping out, or disengaging from school. These systems use multiple data sources such as attendance, behaviour, and course performance to trigger timely interventions.

In the United States, the use of EWS has grown significantly over the last two decades. The American Institutes for Research (AIR) notes that schools implementing early warning systems can reduce dropout rates by as much as 20% (Faria et al., 2017). Key models, such as the ABC model (Attendance, Behaviour, and Course performance), have been widely adopted across many U.S. states to systematically monitor students and guide interventions.

Similarly, in South Africa, the Department of Basic Education piloted the Learner Early Warning System (LEWS) to track students' attendance and academic progress in rural schools. A report by Dr. Nic Spaull emphasized that using simple indicators like frequent absenteeism provided actionable insights that helped reduce dropout rates, especially among vulnerable learners in lower socio-economic communities (Spaull, 2019).

In Kenya, initiatives like the Tusome Early Grade Reading Assessment program have demonstrated the value of early monitoring (Global Schools Forum, 2022). Although initially focused on literacy, the tracking mechanisms set up under Tusome provided

templates for identifying struggling learners early and deploying targeted academic support (McCowan, 2018).

Although EWS models have proven effective internationally, implementing them in low-resource settings like Uganda requires customization. Limited infrastructure, inconsistent record-keeping, and heavy teacher workloads pose challenges. However, the experiences of South Africa and Kenya demonstrate that even low-cost, paper-based tracking tools can make a significant difference when well-managed.

In this study, lessons from established EWS models will guide the design of a predictive approach tailored to the realities of Ugandan primary schools, focusing on affordability, simplicity, and actionability.

2.7 Summary of Literature Gaps

While existing studies have advanced our understanding of academic vulnerability and the use of predictive models in education, several important gaps remain.

Most research on early warning systems has focused on high-income countries, though recent studies have begun to address African contexts such as East Africa (Mnyawami et al., 2022) and Somaliland (Ahmed & Warsame, 2024). However, these studies largely focus on secondary education, leaving primary schools underexplored. The critical Primary 4 to Primary 6 period, where dropout risk and academic failure sharply increase in Uganda, remains largely unexamined in machine learning-based research.

Many existing early warning systems rely heavily on structured academic records alone, such as grades and attendance, without fully integrating broader behavioural and contextual indicators like classroom participation, homework habits, parental involvement, and socioeconomic background. As highlighted by Herodotou et al. (2019) and Adejo and Connolly (2018), real-time and multi-dimensional data are essential for accurate predictions, yet practical implementations remain limited.

Few studies have translated predictive insights into actionable, school-level intervention strategies tailored to the realities of under-resourced schools. While models have been developed, guidance on how schools can effectively use predictions

to support at-risk pupils remains vague, especially in rural and semi-urban environments.

This study sought to address these gaps by:

- Implementing a predictive model grounded in the realities of primary schools in Mukono District, Uganda. Focusing specifically on pupils in P4–P6 levels where academic vulnerability is highest.
- Integrating academic, behavioural, and contextual factors to improve prediction accuracy.
- Linking model outputs directly to feasible, school level strategies, informed by local needs and capacities.

In summary, the reviewed literature highlights both the potential and the current limitations of existing approaches to identifying academically at-risk pupils. While predictive analytics and early warning systems have shown promise globally and increasingly in African contexts, there remains a significant gap in localized solutions tailored to the Ugandan primary school context. The next chapter presents the methodology that guided the design, implementation and validation of such a model.

3 Chapter Three: Methodology

This chapter outlines how the study was conducted. It covers the research design, study area, target population, sampling approach, data collection, and analysis methods.

3.1 Research Design

This study adopted an explanatory sequential mixed-methods design (Creswell & Plano Clark, 2018). This design involves collecting quantitative data first, followed by qualitative data to help explain and contextualize the quantitative findings.

In the first phase, quantitative data was collected from pupil academic records to implement and evaluate machine learning models for predicting at-risk pupils. In the second phase, qualitative data was gathered through semi-structured interviews with teachers and school administrators to validate the model's predictions, understand contextual factors influencing pupil performance, and inform the development of practical intervention strategies.

The explanatory sequential design was selected for three reasons. First, quantitative data alone could identify statistical patterns but not explain the underlying reasons behind pupil disengagement. Second, teacher insights were essential to ensure the predictive model aligned with classroom realities. Third, combining both methods strengthened the validity of findings through triangulation (Ivankova et al., 2006). This approach is recommended for educational research where both statistical patterns and human context matter (Creswell & Plano Clark, 2018).

3.2 Study Population and Sampling

3.2.1 Study Focus and Population

The study targeted pupils in Primary Four to Primary Six (P.4–P.6) from selected schools in Mukono District. These classes were chosen because dropout and academic decline tend to peak during this stage (Uganda Bureau of Statistics, 2023).

Teachers and school administrators also participated, as their firsthand observations helped explain the patterns found in pupil records.

3.2.2 Study Area

The study was conducted in Mukono District, located about 21 kilometers east of Kampala (Coordinates: 00°28′50″N 32°46′14″E). The district has both peri-urban areas around the Municipality and rural sub-counties, with over 200 primary schools serving a growing population (Uganda Bureau of Statistics, 2023).

Mukono was chosen because it reflects challenges seen across Uganda: high enrolment but persistent dropout, especially between P.4 and P.6 (Uganda Bureau of Statistics, 2023). The mix of urban and rural schools, along with varied income levels among families, made it a practical setting for this research.

3.2.3 Sampling Strategy

Sampling was done in stages:

Stage 1: School Selection

Two primary schools in Mukono District were purposively selected, to include both urban and rural settings for a balanced view. Schools were chosen based on accessibility, willingness to participate, and diversity in terms of pupil background.

Stage 2: Pupil Sampling

Within each school, pupils from P.4 to P.6 were selected using stratified random sampling to ensure representation across class levels and gender.

Stage 3: Teachers and administrators Sampling

Five teachers and administrators were purposively selected from each school based on their involvement in academic monitoring and pupil welfare. Their input provided context that the numerical data alone could not capture.

3.2.4 Sample Size

A Since the exact number of P.4 to P.6 pupils across the target schools was unknown beforehand, Cochran's formula was used to estimate the required sample size (Cochran, 1977). This formula is commonly applied when working with large or undefined populations.

The formula is expressed as:

$$n_0 = \frac{z^2.p(1-p)}{e^2}$$

Where:

n_0 = required sample size

z = z-score associated with the desired confidence level

p = estimated proportion of the population with the characteristic of interest

e = acceptable margin of error

For this study a 95% confidence level was selected, corresponding to a Z-score of 1.96. A proportion p, 0.5 was used, assuming maximum variability since the true prevalence of academic risk is unknown, A margin of error e, 0.05 was selected to maintain reasonable precision

Substituting into the formula

$$n_0 = \frac{(1.96)^2.0.5\ .(1-0.5)}{(0.05)^2}$$

$$n_0 = \frac{3.8416 * 0.25}{0.0025} = 384.16$$

This indicated that, under these assumptions, a minimum sample size of approximately 384 pupils was sufficient to produce generalizable estimates of academic risk with 95% confidence and a 5% margin of error.

In total, records for 409 pupils were collected from P.4 to P.6 across the participating schools. This exceeded the minimum of 384 required by Cochran's formula, giving the study adequate data for model training and validation. The sample was also balanced across class levels and gender.

3.3 Data Collection Methods

Data collection was guided by the first and three objectives of the study, which sought to identify key indicators of academic vulnerability and to propose evidence-based support strategies respectively.

3.3.1 Quantitative Data Collection

Quantitative data came from existing school records and covered attendance, academic scores, homework completion, disciplinary incidents, and basic demographic details. All records were anonymized before analysis.

A structured extraction template guided data collection across schools (see Table A: Summary of Key Variables collected for Predictive Modeling, Appendix A).

3.3.2 Qualitative Data Collection

Semi-structured interviews were conducted with class teachers (particularly those handling P.4 to P.6) and school administrators including headteachers and their deputies.

The interviews covered two areas. The first focused on what contributes to pupils struggling academically, family circumstances, parental involvement, motivation, and school conditions. The second looked at what support, schools currently offer struggling pupils and how well these measures work in practice.

The thematic interview guide kept conversations consistent while giving participants room to raise issues they considered important. (see, Appendix B :Thematic Interview Guide).

3.4 Data Analysis Techniques

Data analysis was carried out using Python and Microsoft Excel, following the CRISP-DM framework (Shearer, 2000), which structures the process into data preparation, modelling, evaluation, and interpretation.

3.4.1 Qualitative Data Analysis

Qualitative Data was analysed using Braun and Clarke's (2006) six-phase thematic analysis. The process began with data familiarization through repeated reading of interview transcripts to develop a deeper understanding of the data. Initial codes were then generated across the entire data, to identify meaningful patterns in participants' experiences and perspectives.

The coding process progressed to theme development, where related codes were grouped into potential themes representing significant aspects of the data. These preliminary themes underwent iterative review and refinement to ensure they accurately reflected participant perspectives while maintaining internal consistency and distinctiveness.

The final themes were named and documented, noting how they connected to each other and to the issue of academic vulnerability.

These qualitative findings informed the later quantitative work, especially the development of intervention strategies.

3.4.2 Analysis for Objective 1: Identifying Key Indicators and Risk Factors

This analysis addressed the first research objective of identifying key indicators and risk factors that predict academic vulnerability among pupils.

Data Preparation

The dataset contained 409 pupil records Variables included academic performance subject scores for English, Mathematics, Science, Social Studies (ENG, MTC, SCI, SST), along with derived metrics such as Average Score and Total aggregate (T/AGG). Behavioral and engagement metrics consisted of attendance and absence rates, homework completion frequency measure, and disciplinary incidents. Demographic and contextual factors included class level, gender, age, day/boarding status reflecting residential arrangements, and family income level as a socioeconomic factor. The target variable was academic risk status as a binary outcome measure.

Data cleaning involved removing student identifiers, checking for missing values, verifying data types, and removing duplicates. Values were also checked to ensure they fell within reasonable ranges.

Exploratory Data Analysis

Univariate analysis of each variable's distributional characteristics was conducted through measures of central tendency and dispersion metrics for all numerical variables, with histogram used for visualizations with kernel density estimation,

revealing skewness and modality patterns (Field, 2018). The dataset was inspected for outliers, making use of boxplot visualization supported by interquartile range statistical methods to identify extreme values (McKinney, 2017). For categorical variables, count plots visualized category distributions to highlight representation balance across demographic groups, while mode identification established the most frequent categories within each variable.

Bivariate analysis investigated relationships between predictor variables and academic risk status. Categorical predictors were analyzed through cross-tabulations with chi-square tests of independence (significance threshold $p < 0.05$), as recommended for examining associations between categorical variables (Field, 2018), supported by effect size measures and stacked bar chart visualizations. Numerical predictors were compared between risk groups using analysis of variance (ANOVA) to test for statistically significant differences between group means (Fisher, 1925), with boxplots visualizing distributional differences and effect sizes quantifying magnitude. Independent samples t-tests (Student, 1908) provided additional validation of mean differences.

Multivariate analysis examined relationships among predictor variables to identify potential redundancies and interaction effects. Pearson correlation matrices with heatmap visualization inspected linear relationships between continuous variables, while variance inflation factor analysis evaluated multicollinearity concerns for modeling later on.

Variables that showed both statistical significance and meaningful effect sizes were prioritized as potential risk indicators.

3.4.3 Analysis for Objective 2: Model Implementation, Evaluation and Validation

This analysis addressed the second research objective; to evaluate, compare, and validate the performance of predictive modeling approaches for detecting at-risk pupils, following a systematic pipeline from data preprocessing through final validation.

Data Preprocessing for Predictive Modeling

The predictive modeling phase started with careful data preparation to ensure optimal model performance. The target variable, academic risk status, was encoded into binary numerical values (0 for 'No', 1 for 'Yes') to facilitate classification modeling. Predictor variables were strategically selected by excluding academic outcome variables that would constitute data leakage, including subject scores and aggregates that directly define the risk status being predicted.

The final predictor set included class level, gender, age, attendance rate, absence rate, day/boarding status, homework completion, disciplinary issues, and family income level. These variables were categorized into numerical predictors and categorical predictors.

Scikit-learn's Pipeline library (Pedregosa et al., 2011) was used for systematic transformation alongside the Column Transformer library for data preprocessing. Numerical variables underwent standardization using the StandardScaler library to ensure uniform scaling across different measurement units. Categorical variables were encoded using one-hot encoding (OneHotEncoder) with the drop first parameter to avoid multicollinearity (Cabello-Solorzano et al., 2023).

Feature Selection and Importance

Recursive Feature Elimination (RFE) with Random Forest classifier as the estimator was used for feature selection, given the class imbalance observed in the target variable (91.4% not at-risk versus 8.6% at-risk), the RFE process was configured to select the most informative features through iterative elimination of the least important features (Guyon & Elisseeff, 2003).

Model Implementation

The dataset was partitioned using stratified sampling into training (70%), validation (15%), and test (15%) subsets to maintain proportional representation of academic risk categories. This ratio was selected to balance sufficient training data for model learning while retaining adequate samples for validation and testing. Given the naturally small at-risk population (8.6%), stratified sampling ensured proportional representation across all subsets, while five-fold stratified cross-validation during hyperparameter

tuning allowed every at-risk pupil to contribute to model evaluation across multiple iterations. The pronounced class imbalance was addressed through Synthetic Minority Oversampling Technique (SMOTE), a widely used approach for handling class imbalance in machine learning (Chawla et al., 2002). SMOTE was applied exclusively to training data, transforming it from 286 to 524 instances with balanced class distribution.

Four machine learning algorithms were implemented for comparative analysis:

Logistic Regression was configured with a liblinear solver and L2 regularization, with class weight balancing to address residual imbalance.

Random Forest Classifier used ensemble learning with several decision trees, configured with class weight balancing and random state consistency for reproducible results.

XGBoost Classifier used gradient boosting with optimized handling of imbalanced data, using log loss evaluation metric and reproducible random state configuration.

Support Vector Machine used a radial basis function kernel for non-linear classification boundaries, with probability estimation enabled for full performance evaluation and class weight balancing built in.

All models were trained on the SMOTE-balanced training data, with hyperparameters initially configured to default values to establish baseline performance prior to optimization.

Model Evaluation, Optimization and Validation

Model performance was evaluated on the validation set using comprehensive metrics with particular emphasis on recall for the minority class given the critical importance of correctly identifying at-risk students. Additional metrics included precision, F1-score, ROC-AUC analysis, and confusion matrix examination to provide holistic performance assessment across different classification aspects.

Following initial comparison, the most promising model underwent hyperparameter tuning using GridSearchCV with 5-fold stratified cross-validation, prioritizing F1-score as the optimization metric. The optimized model was then evaluated against the

completely unseen test set to assess generalizability, with special emphasis on recall as the critical benchmark for early warning system effectiveness.

3.4.4 Analysis for Objective 3: Evidence-based support strategies aligned with the model's outputs for timely and targeted intervention.

This section addresses the third research question by examining how model outputs can guide support strategies for academically at-risk pupils. The goal was to bridge the gap between quantitative predictions and actionable, qualitative interventions.

Intervention System Development

A tiered intervention system was developed to translate model probability scores into practical support strategies. Tiered approaches to risk classification are well-established in early warning system research, allowing schools to match intervention intensity to risk severity (Faria et al., 2017).

The system adopted a three-level risk classification framework:

High-risk: $P \geq 0.7$

Moderate-risk: $0.4 \leq P < 0.7$

Low-risk: $P < 0.4$

These thresholds were set based on how predicted probabilities were distributed and what would be realistic for schools to manage given limited resources.

Strategy Development and Personalization

Intervention strategies were matched to the specific risk factors flagged for each pupil, such as poor attendance, incomplete homework, or disciplinary issues. This allowed recommendations to address each pupil's particular situation rather than applying a one-size-fits-all approach.

Model Interpretation

Model interpretability was ensured through SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017). SHAP values quantify the contribution of each feature

to individual predictions by calculating how much that feature pushes the prediction toward or away from the at-risk classification. A positive SHAP value indicates the feature increases the predicted risk, while a negative value decreases it. Averaging the absolute SHAP values across all pupils reveals which features are most influential overall, enabling identification of the key risk drivers. Additionally, SHAP provides explanations for individual pupils, showing why a specific pupil was classified as high-risk based on their unique combination of features.

Implementation Framework and Output Generation

The system produced individual support plans for each pupil, along with summaries for school-level planning. Practical constraints like staffing, costs, and parent availability were considered when recommending interventions.

3.5 Ethical Considerations

The study was approved by the University's Research Ethics Committee before data collection began.

Teachers and administrators gave written consent after being informed about the study's purpose, procedures, and how their information would be protected. A sample of the consent form used for adult participants is provided in Appendix E. For pupil records, consent was obtained from school authorities, and all identifying information was removed during data extraction.

Electronic files were stored in password-protected folders, and physical documents were kept in locked cabinets. Participants were informed they could withdraw at any time without consequence.

4 Chapter Four: Results

4.1 Introduction

This chapter presents the study's findings, organized around the three research objectives: identifying key predictors of academic vulnerability, evaluating the predictive model, and developing an intervention framework.

4.2 Objective One: Key Indicators of Academic Vulnerability

The analysis revealed distinct differences between at-risk and not-at-risk pupils across behavioural, academic, and socioeconomic factors. Table 4.1 summarizes the key characteristics that distinguished the two groups.

Table 4. 1: Key Factors Associated with Academic Vulnerability

Factor	At-Risk Pupils (n=35)	Not-At-Risk Pupils (n=374)
Average Academic Score	41.2	69.1
Attendance Rate	62.3%	86.7%
Homework Completion (1-5 scale)	3.2	4.0
With Disciplinary Record	28.6%	0.8%
From Low-Income Households	48.6%	30.8%

4.2.1 Academic and Engagement Profile

The overall results on academic and engagement profiles as seen in Table 4.1, showed considerably lower values for at-risk pupils compared to their not-at-risk peers. The average academic score for at-risk pupils was 41.2, whereas not-at-risk pupils achieved an average score of 69.1, representing a difference of 27.9 points. This gap indicates that at-risk pupils were performing well below the expected academic standard.

Regarding school attendance, at-risk pupils recorded an attendance rate of 62.3%, compared to 86.7% for not-at-risk pupils. This difference of 24.4 percentage points suggests that at-risk pupils missed more school days than their peers. Homework completion rates followed a similar pattern. At-risk pupils scored 3.2 on a 5-point scale,

while not-at-risk pupils scored 4.0, indicating that at-risk pupils submitted homework less frequently.

4.2.2 Behavioral and Socioeconomic Indicators

As seen in Table 4.1, behavioural and socioeconomic factors also showed differences between the two groups. Chi-square tests confirmed significant associations between academic risk and both disciplinary issues ($x^2 = 71.427$, $p < 0.001$) and family income level ($x^2 = 10.214$, $p = 0.006$). Specifically, 28.6% of at-risk pupils had a disciplinary record, compared to only 0.8% of not-at-risk pupils. This means at-risk pupils were approximately 35 times more likely to have been involved in disciplinary incidents than the rest.

Regarding socioeconomic background, 48.6% of at-risk pupils came from low-income households, compared to 30.8% of not-at-risk pupils as shown in Table 4.1. This difference of 17.8 percentage points indicates that pupils from low-income families were disproportionately represented in the at-risk category, comprising nearly half of this group.

4.2.3 Non-Significant Factors

Several demographic variables showed no meaningful association with academic risk status. Chi-square tests revealed no significant relationship between academic risk and gender ($x^2 = 0.121$, $p = 0.728$) or class level ($x^2 = 0.242$, $p = 0.886$). Similarly, age did not differ significantly between at-risk and not-at-risk groups ($F = 0.091$, $p = 0.763$). Full statistical outputs are presented in Tables C.1, C.2 and C.3, Appendix C.

4.3 Objective Two: Performance and Validation of the Predictive Model

This section presents the findings on the evaluation, comparison, and validation of predictive models for identifying at-risk pupils.

4.3.1 Comparative Model Performance

Table 4.2 compares the four models tested. The main focus was recall—how well each model identified at-risk pupils—alongside ROC-AUC, which measures overall discriminatory ability.

Table 4. 2: Comparison of Model Performance for Identifying At-Risk Pupils

Model	Recall (At-Risk)	ROC-AUC	Overall Accuracy
Logistic Regression	0.60	0.821	0.79
Random Forest	0.00	0.671	0.84
XGBoost	0.00	0.596	0.85
Support Vector Machine	0.20	0.650	0.84

As shown in Table 4.2, Logistic Regression performed best at identifying vulnerable pupils. On the validation set, which contained five at-risk pupils reflecting the natural class distribution, the model correctly identified three (Recall = 0.60) and showed the best overall discriminatory ability (ROC-AUC = 0.821). Although the Random Forest and XGBoost models achieved higher overall accuracy, they failed to identify any at-risk pupils.

4.3.2 Validated Performance of the Optimized Model

The Logistic Regression model was selected for optimization and final validation. Its performance on completely unseen test data is presented in Table 4.3.

Table 4. 3: Final Model Performance on Test Data

Performance Metric	Score
Recall	0.833
ROC-AUC	0.941
Precision	0.385
F1-Score	0.526

As detailed in Table 4.3, the optimized model correctly identified five out of six at-risk pupils on the unseen test set (Recall = 0.833). The ROC-AUC score of 0.941 indicates

strong discriminatory ability between at-risk and not-at-risk pupils. The precision score of 0.385 indicates that when the model flagged a pupil as at-risk, it was correct approximately two out of five times. The F1-Score of 0.526 reflects the balance between identifying vulnerable pupils and managing false alerts.

4.3.3 Interpretation of Model Predictions

The relative influence of each predictor on the model's risk classifications revealed a clear hierarchy of factors. Figure 2 illustrates the impact of each feature on model output, while Figure 3 ranks them by importance magnitude.

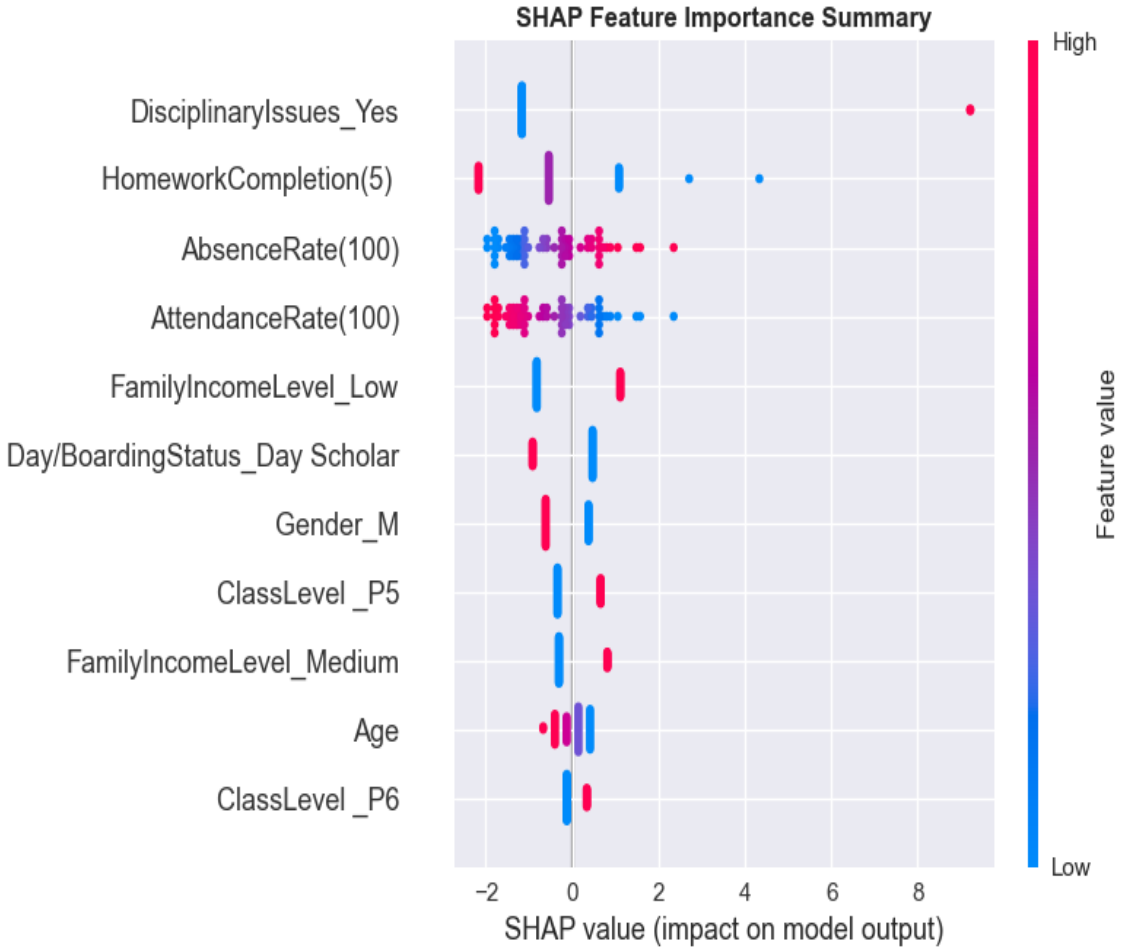

Figure 2: SHAP Summary Plot of Feature Effects on Risk Classification. Each dot represents one pupil. The vertical axis lists predictor variables ordered by importance. The horizontal axis shows the SHAP value, where positive values (right) indicate

increased predicted risk and negative values (left) indicate decreased predicted risk. Dot color represents the feature value for that pupil (pink = high, blue = low).

Figure 2 shows that disciplinary issues had the strongest effect on predictions, with pupils who had disciplinary records (pink dots) clustered on the right side, indicating substantially increased risk. Homework completion showed an inverse pattern, where lower completion rates (blue dots) pushed predictions toward at-risk. Similarly, higher absence rates (pink dots for AbsenceRate) increased predicted risk, while higher attendance rates decreased it. Demographic variables such as gender, age, and class level showed minimal spread, confirming their weak influence on predictions.

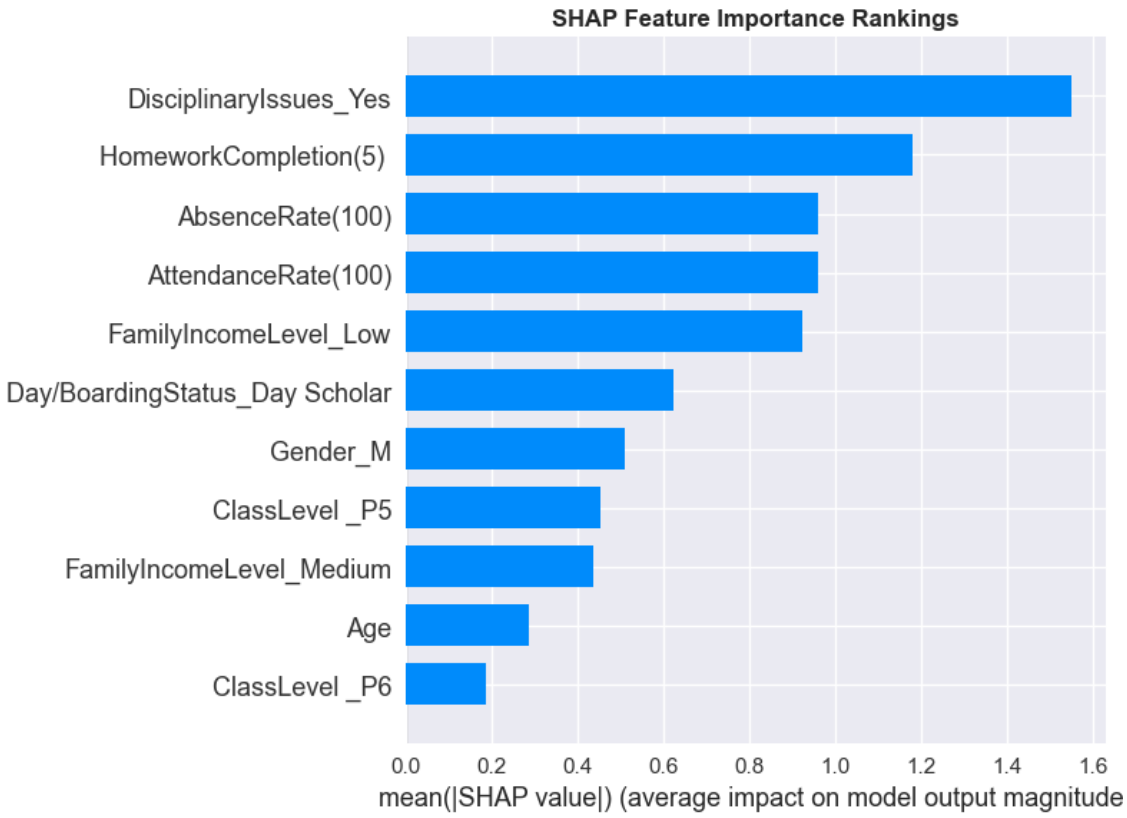

Figure 3: Feature Importance Rankings for Risk Prediction. The horizontal axis displays the mean absolute SHAP value, representing the average contribution of each feature to model predictions. The vertical axis lists features in descending order of importance.

Figure 3 confirms that disciplinary issues ranked as the most influential predictor, followed by homework completion and absence/attendance rates. These three behavioural factors dominated the top positions. Family income level also contributed notably. In contrast, gender, age, and class level appeared at the bottom with the smallest values, indicating minimal influence on risk classification.

4.4 Objective 3: Evidence-Based Support Strategies from Model Outputs

4.4.1 Tiered Intervention Framework Output

The model's probability scores were translated into a practical three-tiered intervention framework. The framework was applied to the 62 pupils in the test set, which represented 15% of the total sample held out for final model validation. The distribution of these pupils across risk tiers is shown in Table 4.4.

Table 4. 4: Distribution of Pupils by Model-Derived Risk Tier

Risk Tier	Number of Pupils	Percentage	Intervention Focus
High Risk	10	16.1%	Intensive, multi-domain support
Moderate Risk	3	4.8%	Targeted monitoring & support
Low Risk	49	79.0%	Preventive monitoring

As shown in Table 4.4, the majority of pupils (79.0%) were classified as low-risk, requiring only preventive monitoring. Ten pupils (16.1%) were identified as high-risk, warranting intensive, multi-domain support. A small proportion (4.8%) fell into the moderate-risk category, requiring targeted monitoring and support. This distribution suggests that intensive resources can be concentrated on a manageable subset of pupils rather than spread thinly across the entire population.

4.4.2 Personalized Intervention Strategies

The framework links each pupil's risk factors to specific support strategies drawn from the qualitative findings. For instance, a high-risk pupil flagged mainly for chronic

absenteeism would receive a plan including weekly check-ins with a teacher, agreements with parents, flexible attendance tracking, and home visits if needed. These target the attendance problem directly rather than applying generic interventions.

4.5 Chapter Summary

This chapter presented the findings corresponding to each research objective.

Key risk factors identified were disciplinary issues ($x^2 = 71.427$, $p < 0.001$), poor attendance, low homework completion, and low family income ($x^2 = 10.214$, $p = 0.006$). Demographic factors such as gender ($x^2 = 0.121$, $p = 0.728$), class level ($x^2 = 0.242$, $p = 0.886$), and age ($F = 0.091$, $p = 0.763$) were not significant predictors (see Tables C.1, C.2 and C.3, Appendix C).

Model Performance and Validation showed that Logistic Regression was the most effective algorithm, achieving a recall of 0.833 and an ROC-AUC of 0.941 on unseen data, correctly identifying five out of six at-risk pupils. The model's decisions were primarily driven by modifiable behavioural factors rather than fixed demographics.

A tiered intervention framework was proposed, classifying 16.1% of pupils as high-risk. The framework generated personalised support plans that would directly address the specific behavioural issues identified by the model, enabling a shift from reactive to proactive intervention.

5 Chapter Five: Discussion

5.1 Introduction

This chapter discusses the study's findings in relation to existing research. It explains why certain risk factors emerged as significant, why Logistic Regression was chosen over other models, and how the intervention framework can be applied in practice. The chapter also considers how these findings address gaps in the literature on academic vulnerability in Ugandan primary schools.

5.2 Interpretation of Key Risk Indicators

The first research question asked what indicators predict academic vulnerability. The findings showed that behavioural factors such as disciplinary issues, incomplete homework, and poor attendance, were the strongest predictors. Demographic variables like age, gender, and class level had no significant association with risk status.

This suggests that problems at home like poverty, family instability, show up first as changes in behaviour at school before leading to academic failure. A pupil struggling at home might start missing school or acting out, and grades follow. What makes these behavioural signs useful is that teachers can spot them daily and schools can actually do something about them, unlike poverty itself, which is beyond the school's control.

Similar findings have been reported elsewhere. Ahmed and Warsame (2024) found attendance and engagement to be strong predictors among secondary students in Somaliland. Mnyawami et al. (2022) observed comparable patterns in East African primary schools using Uwezo data. However, those studies focused on secondary education or regional-level assessments. This study shows that the same indicators matter at the primary level, particularly during the P4–P6 years when dropout risk is highest.

This aligns with research on the value of behavioural tracking for early identification (Herodotou et al., 2019). National data points to poverty as a main driver of dropout (UBOS, 2023), but this study shows that poverty's effect works through behaviour like attendance, homework, discipline. Schools cannot fix poverty, but they can monitor and respond to these behavioural signs. For teachers, this means the patterns they

already notice carry real predictive weight, giving them grounds to act rather than dismiss what they observe as just a hunch.

5.3 Interpretation of Model Performance and Selection

The second research question asked which modelling approach works best for identifying at-risk pupils. The results reveal an important gap between statistical accuracy and practical usefulness, a distinction that matters for schools with limited resources.

Random Forest and XGBoost had higher overall accuracy (84% and 85%), but they missed every at-risk pupil. Logistic Regression had lower accuracy but far better recall, the metric that matters most for an early warning system. The tuned model caught five of six at-risk pupils in the test set, with a recall of 0.833 and ROC-AUC of 0.941.

This choice comes down to what matters more: missing a struggling pupil or flagging one who turns out fine. Missing a pupil who then drops out is far worse than giving extra support to someone who did not strictly need it. Logistic Regression minimized missed cases, which is what an early warning system should do. A model with 85% accuracy that misses every at-risk pupil is useless for intervention, no matter how good the number looks.

The model compares well with similar work in developing countries. The ROC-AUC of 0.941 exceeds what Mduma et al. (2019) reported for dropout prediction in Tanzania and comes close to the 95% accuracy Ahmed and Warsame (2024) achieved in Somaliland with Random Forest, though their study was at secondary level. This study reached that performance with Logistic Regression, not a complex algorithm. This supports Seo et al.'s (2024) argument that simpler models are often better when interpretability matters more than squeezing out extra accuracy.

This directly engages with a well-documented trade-off in the literature between model complexity and interpretability. Studies such as Khatun et al. (2025) achieved high accuracy (94.4%) with XGBoost in predicting student dropout in Bangladesh, but required post-hoc explanation techniques such as SHAP and LIME to make model decisions comprehensible. Krüger et al. (2023) similarly employed explainable machine

learning in Brazil, achieving strong performance but with considerable technical overhead. In contrast, Logistic Regression provides inherent transparency through its coefficient structure, showing, for instance, that a high-risk classification was driven primarily by increased absenteeism or the presence of disciplinary issues. This interpretability is essential for adoption in under-resourced settings where teachers lack technical expertise and have limited time for training on complex systems.

Choosing Logistic Regression was not settling for less, it was picking the right tool for the job. It catches more at-risk pupils than the alternatives and teachers can actually understand why it flags someone. In this context, the interpretable model turned out to be the most effective one, which speaks to the tension Seo et al. (2024) and Krüger et al. (2023) have raised between accuracy and usability.

5.4 From Prediction to Recommended Support

The third research question asked how predictions could be turned into practical support. The answer was a tiered framework that converts probability scores into action plans for teachers.

The framework sorts pupils into three tiers: High Risk (probability ≥ 0.7), Moderate Risk (0.4 ≤ probability < 0.7), and Low Risk (probability < 0.4). This matters where resources are tight. Rather than allocating support uniformly across all pupils or waiting until failure occurs, schools can direct their most intensive efforts toward the subset of pupils with greatest need. In the test sample, 16.1% of pupils were classified as high-risk, a manageable proportion that allows for concentrated intervention without overwhelming available resources. This approach aligns with established early warning system research demonstrating the effectiveness of tiered interventions in matching support intensity to risk severity (Faria et al., 2017).

The framework also personalizes support. SHAP analysis shows which factors drove each pupil's risk score, so interventions can target the actual problem. A pupil flagged for attendance gets attendance-focused help like weekly check-ins, parent agreements, flexible tracking. A pupil flagged for behaviour gets a different plan aimed at conduct and engagement. This avoids generic responses that miss the real issue.

Comparative Advantage of the Predictive Approach

The developed framework represents a meaningful shift from traditional, reactive support methods commonly used in Ugandan primary schools. Table 5.1 outlines the key differences between the predictive system and conventional approaches.

Table 5. 1: Comparison of Intervention Approaches

Aspect	Traditional Approach	Predictive System
Timing of Identification	After academic failure	Before failure manifests
Basis for Support	Generic or cross driven	Individualized to specific risk factors
Resource Allocation	Reactive and often late	Strategic and preventive

Traditionally, schools only notice struggling pupils after they fail exams or accumulated absences, by then, options are limited. The predictive system spots pupils earlier, while behavioural patterns are still forming, and targets the specific issues before they snowball into failure.

This responds to a gap Herodotou et al. (2019) noted: many predictive models never get used because they need infrastructure, specialists, or systems that schools in developing countries do not have. This framework was built to work within existing constraints. It uses data schools already collect and does not require new technology or extra staff. By turning model outputs into simple categories, it gives schools something they can actually implement on their own.

5.5 Addressing Identified Research Gaps

This study was designed to address three specific gaps identified in the literature review, and the findings provide evidence that each gap has been meaningfully addressed.

The first gap was geographical. Predictive models for academic risk have mostly come from high-income countries, with recent work in African secondary schools (Ahmed & Warsame, 2024; Mnyawami et al., 2022). Ugandan primary education had not been

covered. This study applied a machine learning model for Ugandan primary schools, focused on P4–P6 where dropout risk peaks. The model's performance shows that approaches from elsewhere can work here when adapted to local data.

The second gap concerned practical utility. Much of the existing literature emphasises accuracy metrics with limited attention to whether models can actually be used by teachers and school administrators (Seo et al., 2024). This study prioritized usability from the outset by selecting Logistic Regression for its inherent interpretability and employing SHAP analysis to make individual predictions explainable. The result is a system that not only identifies at-risk pupils but explains why they are at risk in terms that educators can understand and act upon without specialized training.

The third gap concerned the translation from prediction to intervention. Few studies in the literature have moved beyond model development and implementation to address how schools should actually respond to identified risk (Krüger et al., 2023). The tiered intervention framework developed in this study bridges this gap by converting probability scores into actionable, personalised support plans that align with existing school capacities and require no additional resources beyond what schools already have. This represents a practical contribution that extends beyond the academic exercise of model building to address real-world implementation challenges.

5.6 Implications for Policy and Practice

The findings point to practical steps for different stakeholders.

For school leaders and teachers, the strong predictive power of routine behavioural data means that attendance registers, homework records, and conduct reports should function as diagnostic tools rather than merely administrative requirements. Regular review of these indicators, perhaps during weekly staff meetings or termly planning sessions, enables early identification without requiring new technology or data systems. The tiered framework provides a practical structure for organizing existing support, ensuring that intensive help reaches pupils with greatest need while avoiding the inefficiency of spreading limited resources too thin.

For district officials, the study shows that data-informed decisions are possible even with limited resources. Officials can help by encouraging consistent record-keeping across schools and training head teachers to interpret behavioural patterns and plan responses. This training should fit within existing professional development rather than adding to already full schedules.

For policymakers, the results support a shift from reactive to proactive support. Policies should encourage early warning systems in schools, with funding for consistent data collection and for training staff to use that data. This investment serves equity goals, early identification helps the most vulnerable pupils who would otherwise slip through unnoticed.

5.7 Limitations of the Study

While this study offers useful insights, its findings should be considered alongside certain limitations.

The model was implemented using data from schools in Mukono District. Its performance and the importance of specific predictors might differ in other regions of Uganda with different socioeconomic or cultural conditions.

The analysis relied on data from one point in time. Tracking the same pupils over several terms or years would provide stronger evidence of how well these indicators predict long-term outcomes and how risk develops over time.

The study used available school records, which, while practical, limited the depth of information available. Richer data, such as the specific reasons behind absences, the nature and quality of parental involvement, or pupils' own perspectives on their difficulties, could potentially improve model accuracy and help tailor more relevant support strategies. The qualitative component of this study provided some contextual insight, but more systematic collection of such information could strengthen future iterations of the model.

5.8 Recommendations for Future Research

Based on the findings and limitations of this study, several avenues for further research are recommended.

Validation across contexts: Testing the model in other Ugandan districts with contrasting socioeconomic profiles would establish generalizability and reveal any necessary context-specific adaptations.

Longitudinal investigation: Tracking pupil cohorts across multiple academic years would strengthen evidence for the predictive indicators and illuminate how risk trajectories develop and respond to early intervention.

Qualitative depth: Integrating qualitative data, such as detailed absence reasons, teacher observations of behavioural change, or pupil perspectives would enrich understanding of vulnerability's underlying causes and potentially improve model accuracy.

6 Chapter Six: Conclusion

This study set out to determine whether a data-driven, proactive approach could effectively identify academically vulnerable pupils and recommend appropriate support strategies in Ugandan primary schools. Three specific objectives guided the research: identifying key risk indicators, evaluating predictive modelling approaches, and developing an evidence-based intervention framework.

Regarding the first objective, the research confirmed that observable behavioural indicators are the most reliable early warning signals. Disciplinary issues, incomplete homework, and poor attendance proved to be stronger immediate predictors of academic risk than demographic factors such as age and gender. This finding is significant because these behavioural indicators are both visible to teachers in their daily practice and potentially modifiable through school-based support, offering a practical entry point for intervention.

For the second objective, the study demonstrated that model interpretability is as crucial as predictive accuracy in an educational early warning system. The Logistic Regression model achieved strong performance (Recall = 0.833, ROC-AUC = 0.941) while providing transparent reasoning through SHAP analysis. This combination builds the necessary bridge of trust between algorithmic output and classroom action, enabling teachers to understand and act upon model recommendations without requiring technical expertise.

The third objective was addressed by developing a tiered intervention framework to translate model outputs into actionable support strategies. By categorizing pupils into high, moderate, and low-risk tiers, the framework enables schools to allocate intensive resources to those with greatest need while maintaining preventive monitoring for others. The framework is designed to work within existing school capacities, requiring no complex technology or additional staffing.

These findings respond to the three gaps identified in the literature review. First, the geographical gap: a predictive model was implemented specifically for Ugandan primary schools, targeting the P4–P6 years when dropout risk peaks. Second, the

practical utility gap: interpretability was prioritized alongside accuracy so that teachers can use the model without specialist training. Third, the implementation gap: predictions were translated into concrete, personalised strategies rather than leaving schools to figure out responses on their own.

While acknowledging the limitations of sample size and geographical scope, this study provides a feasible, evidence-based foundation for transforming pupil support from a reactive to a proactive endeavor. It demonstrates that through systematic use of existing school records, educators in resource-constrained settings can begin to identify vulnerable pupils early, based on modifiable behaviours, before disengagement leads to dropout. This represents a meaningful step towards more equitable educational practice, extending timely support to the most vulnerable pupils before they fall behind.

References

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61–75. https://doi.org/10.1108/JARHE-09-2017-0113

Ahmed, A. M., & Warsame, M. H. (2024). Predicting Student Dropout Rates Using Supervised Machine Learning: Insights from the 2022 National Education Accessibility Survey in Somaliland. Applied Sciences, 14(17), 7593. https://doi.org/10.3390/app14177593

Alfred, W., Oriangi, G., Odama, S., & Ologe, D. O. (2023). Modelling Academic Performance in Science-Based Subjects in Primary Schools in Uganda. East African Journal of Education Studies, 6(1), 311–319. https://doi.org/10.37284/eajes.6.1.1150

Bowers, A. J., & Sprott, R. (2012). Why Tenth Graders Fail to Finish High School: A Dropout Typology Latent Class Analysis. Journal of Education for Students Placed at Risk (JESPAR), 17(3), 129–148. https://doi.org/10.1080/10824669.2012.692071

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cabello-Solorzano, K. et al. (2023). The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023)

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Christine Mbabazi Mpyangu, B., Awich Ochen, E., & Olowo Onyango Yovani Moses Lubaale, E. A. (2014). OUT OF SCHOOL CHILDREN STUDY IN UGANDA THE REPUBLIC OF UGANDA.

Cochran, W. G. (1977). Sampling techniques (3rd Edition). John Wiley & Sons

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and Conducting Mixed Methods Research (3rd ed.). SAGE Publications.

Davis, M. H., Mac Iver, M. A., Balfanz, R. W., Stein, M. L., & Fox, J. H. (2019). Implementation of an early warning indicator and intervention system. Preventing School Failure: Alternative Education for Children and Youth, 63(1), 77–88. https://doi.org/10.1080/1045988X.2018.1506977

Elaine Allensworth, & John Q. Easton. (2007). What matters for staying on-track and graduating in Chicago public high schools. Chicago: Consortium on Chicago School Research. Https://Consortium.Uchicago.Edu/Publications/What-Matters-Staying-Track-and-Graduating-Chicago-Public-Schools.

Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: A case of Tanzanian's secondary schools. Applied Artificial Intelligence, 36(1), 2071406. https://doi.org/10.1080/08839514.2022.2071406

Epstein, J. L. (2018). School, Family, and Community Partnerships. Routledge. https://doi.org/10.4324/9780429494673

Faria, A.-M., Sorensen, N., Heppen, J., Bowdon, J., Taylor, S., Eisner, R., & Foster, S. (2017). Getting students on track for graduation: Impacts of the Early Warning Intervention and Monitoring System after one year At American Institutes for Research. http://ies.

Field, A.P. (2018) Discovering Statistics Using IBM SPSS Statistics. 5th Edition, Sage, Newbury Park.

Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd.

Global Schools Forum. (2022). Tusome Early Grade Reading, Kenya.

Gottfried, M. A. (2019). Chronic Absenteeism in the Classroom Context: Effects on Achievement. Urban Education, 54(1), 3–34. https://doi.org/10.1177/0042085915618709

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

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. https://doi.org/10.1007/s11423-019-09685-0

Herodotou, C., Rienties, B., Verdin, B., & Boroowa, A. (2019). Predictive learning analytics "at scale": Towards guidelines to successful implementation in higher education based on the case of the open university UK. Journal of Learning Analytics, 6(1), 85–95. https://doi.org/10.18608/jla.2019.61.5

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice. Field Methods, 18(1), 3–20. https://doi.org/10.1177/1525822X05282260

Khatun, M. R., Mim, M. A., Tasin, M. M., & Hossain, M. M. (2025). A hybrid framework of statistical, machine learning, and explainable AI methods for school dropout prediction.PLOS ONE, 20(9), e0331917. https://doi.org/10.1371/journal.pone.0331917

Krüger, J., Britto, A., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. Expert Systems with Applications, 233, 120933. https://doi.org/10.1016/j.eswa.2023.120933

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

McCowan, T. (2018). Quality of higher education in Kenya: Addressing the conundrum. International Journal of Educational Development, 60, 128–137. https://doi.org/10.1016/j.ijedudev.2017.11.002

McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.

Mduma, N., Kalegele, K., & Machuve, D. (2019). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. Data Science Journal, 18(14), 1-10. https://doi.org/10.5334/dsj-2019-014

Mike, I. O., Nakajjo, A., & Isoke, D. (2008). Socioeconomic Determinants of Primary School Dropout: The Logistic Model Analysis. In Research Series (Issue 54).

Ministry of Education and Sports & UNHCR. (2022). Education Response Plan for Refugees and Host Communities (ERP II) 2023-2025. Government of Uganda.

Ministry of Education and Sports. (2020). Education and Sports Sector Strategic Plan (ESSP) 2020/21-2024/25. Government of Uganda.

Ministry of Education and Sports. (2021). COVID-19 Education Response Plan. Available at: Https://Www.Education.Go.Ug/Covid-19-Sector-Response/.

Ministry of Education and Sports. (2023). Education Sector Performance Report 2021/2022. Government of Uganda.

Mitana, J. (2018). Towards a holistic and relevant educational assessment in primary schools in Uganda. African Educational Research Journal, 6(2), 58–68. https://doi.org/10.30918/AERJ.62.18.018

National Planning Authority. (2018). EDUCATION MODELLING AND FORECASTING COMPREHENSIVE EVALUATION OF THE UNIVERSAL PRIMARY EDUCATION (UPE) POLICY.

National Planning Authority. (2025). Fourth National Development Plan (NDP IV) 2025/26–2029/30. Government of Uganda.

Oweyegha Afunaduula. (2025, February 6). CONFRONTING THE SCHOOL DROPOUT DILEMMA IN UGANDA. Https://Muwado.Com/Confronting-the-School-Dropout-Dilemma-in-Uganda/?V=2a0617accf8b.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Seo, E. Y., Yang, J., Lee, J. E., & So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. Heliyon, 10(11), e32144. https://doi.org/10.1016/j.heliyon.2024.e32144

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

Spaull, N. (2019). Priorities for education reform in South Africa: A report to President Ramaphosa and Minister Mboweni

Student. (1908). The probable error of a mean. Biometrika, 6(1), 1–25.

The Bantwana Initiative of World Education. (2021). Dropout Early Warning System Technical Paper.

The World Bank, & UNESCO Institute for Statistics (UIS). (2024, September 30). Primary completion rate (% of relevant age group). Https://Genderdata.Worldbank.Org/En/Indicator/Se-Prm-Cmpt-

Zs?Gender=female&gender=male&geos=ERI_SWZ_ETH_GMB_GHA_KEN_LSO_LBR_MWI_MOZ_NGA_RWA_SLE_SOM_SSD_TZA_UGA_ZMB_ZWE&groups=incomeGroup&view=bar.

Therriault, S. B., O'cummings, M., Heppen, J., Yerhot, L., & Scala, J. (2017). Early Warning Intervention and Monitoring System (EWIMS) Implementation Guide.

Uganda Bureau of Statistics Statistical Abstract 2020. Government of Uganda.

Uganda Bureau of Statistics. (2023). Statistical Abstract 2022. Government of Uganda.

Uganda National Examinations Board. (2023). The Achievement of Primary School Learners in Numeracy and Literacy in English in Uganda: 2023 NAPE Report. UNEB.

UNESCO. (2022). Global Education Monitoring Report 2022: Non-state actors in education: Who chooses? Who loses? UNESCO.

UNICEF UGANDA. (2020). UNICEF Uganda Annual Report 2020.

United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development. United Nations General Assembly. Retrieved from https://sdgs.un.org/goals.

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence, 4, 2. https://doi.org/10.1007/s44163-023-00079-z

Appendix A: Summary of Key Variables

Table A: Summary of Key Variables collected for Predictive Modeling

Variable Name	**Description**
Pupil ID (anonymized)	Unique code for each learner
School Type	Whether it's a government aided or private school
Class / Grade level	Class level of pupil e.g. P.4, P.5, or P.6
Gender	Male / Female
Age	Pupil's Age
Academic Records	Average exam/ test scores per subject
Attendance Rate	Percentage of school days attended
Day/ Boarding Status	Day or Boarding pupil
Homework Completion rate	Percentage of submitted homework
Disciplinary Issues	Number of recorded misbehavior events
Family Income Level	Categorized socioeconomic status
Academic Risk Status	Academic risk status of the student

Appendix B: Interview Guide

THEMATIC INTERVIEW GUIDE

Study Title: A Predictive Model for Identifying At-Risk Pupils and Recommending Support Strategies: A Case Study of Primary Schools in Mukono District, Uganda

Participants: Class teachers and School administrators (P.4–P.6)

Purpose: To gather qualitative insights aligned with the study's analytical themes: academic vulnerability, social challenges, and school-level intervention strategies.

Section A: General Information

1. What is your current role at this school?
2. How long have you worked with learners in Primary 4 to Primary 6?
3. What is the average number of pupils in the class(es) you handle?

Section B: Academic Vulnerability

1. What early signs do you typically observe that indicate a pupil may be academically at risk?
2. In your view, how do poor academic results usually develop, is it sudden or gradual?
3. Which academic indicators (e.g. test scores, incomplete work, class participation) do you consider most important for identifying struggling pupils?
4. How do you monitor academic progress in your class?

Section C: Social and Environmental Challenges

1. What external (non-academic) challenges do you think contribute to academic decline in pupils?
2. How do socioeconomic factors influence pupil performance here?
3. Do pupils from certain backgrounds face more challenges than others?
4. In your experience, how does absenteeism relate to family or community issues?

Section D: School-Level Practices and Interventions

1. How do you currently identify pupils who are struggling or at risk of dropping out?
2. Are there any systems or tools that help you track pupil well-being?
3. Once a pupil is identified as struggling, what actions are typically taken to support them?
4. What kind of support is most effective?
5. How involved are parents or guardians in interventions?
6. What challenges do you face when supporting struggling pupils?

Section E: Use of Predictive Tools

1. Would you find a tool that predicts academic risk based on pupil data helpful?
2. What pupil data do you think would be most helpful in identifying risk early?
3. Are there any concerns about using such predictive models in your school?
4. Is there anything else you'd like to share that would help this research?

Appendix C : Statistical Tests

Statistical Tests for Non-Significant Demographic Variables

Table C. 1: Cross-tabulation of Gender and Academic Risk Status

Gender	Not At-Risk	At-Risk	Total
Female	175	18	193
Male	199	17	216
Total	374	35	409

Chi-square test: p = 0.728

No statistically significant association was found between gender and academic risk status.

Table C. 2: Cross-tabulation of Class Level and Academic Risk Status

Class Level	Not At-Risk	At-Risk	Total
P4	138	12	150
P5	134	14	148
P6	102	9	111
Total	374	35	409

Chi-square test: p = 0.886

No statistically significant association was found between class level and academic risk status.

Table C. 3: Descriptive Statistics for Age by Risk Group

Risk Status	n	Mean Age	Std. Dev.
Not At-Risk	374	10.48	1.25
At-Risk	35	10.54	1.27

ANOVA test: p = 0.763

No statistically significant difference in mean age was found between at-risk and not-at-risk groups.

Appendix D: Python Code and Jupyter Notebook for Model Development

The complete Python code used for data analysis, model development, and evaluation is available in the accompanying Jupyter Notebook. The notebook includes:

- Data preprocessing and cleaning procedures
- Exploratory data analysis
- Statistical tests
- Model training and evaluation
- SHAP analysis for model interpretation
- Intervention framework implementation

The repository below also contains the raw dataset and pre-processed dataset used in this study.

Repository: https://github.com/ChuckJovans/A-Predictive-Model-for-Identifying-At-Risk-Pupils

Appendix E: Sample Informed Consent Form (Anonymized)

This is a blank copy of the consent form used in the study. Any personal details, such as names, signatures, and dates have been removed. The original signed forms are kept securely.

UCU REC INFORMED CONSENT FORM

A Predictive Model for Identifying At-Risk Pupils and Recommending Support Strategies: A Case Study of Primary Schools in Mukono District, Uganda

No.	Name of Investigator	Designation	Address/Telephone/Email	Institution of Affiliation
1.	Galiwango Charles Jovans	Principal Investigator	galiwangocharlesjovans@gmail.com	UCU
2.	Dr Bitalo Daphne	Co-investigator	dbitalo@ucu.ac.ug	UCU

Many primary school pupils in Uganda continue to struggle academically, with high dropout rates particularly between Primary Four and Six. This study aims to develop a data-driven model that will help schools identify pupils who are at risk of academic failure early enough to provide appropriate support. Mukono District was selected due to its mix of urban and rural schools and documented challenges in learning outcomes

This is a mixed-methods study combining school data analysis with stakeholder interviews. Participants will be involved in a brief interview lasting approximately 30–45 minutes. These interviews will gather insights on how schools currently monitor pupil performance, the challenges faced, and existing support practices.

Participants will include school administrators, class teachers (P.4–P.6), and relevant staff involved in pupil support within selected primary schools in Mukono District. No pupils or minors will be interviewed.

Participation in this research is entirely voluntary. You may choose not to answer specific questions or withdraw from the study at any point, with no consequences whatsoever. There is no obligation to continue if you no longer wish to participate.
All information shared will be kept strictly confidential. Your name, school, or any identifying details will not appear in any reports. Data collected will be anonymized and securely stored, and it will be used solely for academic purposes related to this study.

- Better understanding of how predictive models can support early identification of academically at-risk pupils
- Development of practical, school-based support strategies to improve learning outcomes
- Contribution to evidence-based policymaking and educational interventions in low-resource settings

UCU REC INFORMED CONSENT FORM

The information you provide will be kept strictly confidential and will only be used for the purposes of this research study. During the process of writing the report, your name will never be used, and everything you share with us will remain anonymous. We will be asking questions about the availability and functionality of childhood vaccination services in your community. If there are any questions you do not wish to answer, you can simply let us know, and we will not insist. Before participating in the study, each participant will be asked to sign a written informed consent form, which ensures that your participation is voluntary and that you agree to take part in the study.

11. Whom to contact in case of ethical related concerns

This study was Approved by Uganda Christian University Research Ethics Committee (UCU-REC) and cleared by Uganda national Council for Science and Technology (UNCST), In case of any Ethical or your rights related concerns or inquiries, please contact UCUREC Chairperson; Prof. Peter Waiswa, 0772405357, pwaiswa@musph.ac.ug or UCUREC Manager, Mr. Osborn Ahimbisibwe, 0775737627 or oahimbisibwe@ucu.ac.ug, UNCST: Tel; +256 414 705500, info@uncst.go.ug

STATEMENT OF CONSENT

Do you accept to be recorded?

Yes ☐ **No** ☐

I voluntarily agree to participate in this research program; to tick appropriately

Yes ☐ **No.** ☐

I understand that I will be given a copy of this signed Consent Form.

Name of Participant: .....

Signature: ..... Date: .....

Name of Researcher/designee: .....

Signature: ..... Date: .....