# MACHINE LEARNING PREDICTIVE MODEL FOR EVALUATING THE WEIGHT OF MITIGATES ON SCHOOL DROPOUT RISK

**SYLVIA CHEROP RONO**

**A Research Project Report Submitted to the School of Mathematics and Computing in Partial Fulfilment of the Requirement for The Award of Master of Science in Data Science of the Cooperative University of Kenya**

**2025**

## DECLARATION

**Declaration by the candidate**

This project is my original work and has not been presented for award of a degree in any other University or for any other award

Signature ……  Date: 24th November 2025

Sylvia Cherop

C004/600026/2023

**Declaration by the supervisors**

We confirm that the work reported in this project was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors

Signature ……  …… Date: 24th November 2025

Dr. Emma Anyika
Department of mathematical science
School of computing and mathematics
The cooperative University of Kenya

Signature …… …  …… Date: 22nd November 2025

James Obuhuma
Department of Computer Science
School of Computing and Informatics
Maseno University

## DEDICATION

I dedicate this work to my family, mentors, and friends whose unwavering support, encouragement, and belief in my academic journey have been my greatest motivation.

**ACKNOWLEDGMENT**

First, I would like to thank God for his guidance, I would also want to sincerely appreciate my supervisors Dr. James Obuhuma and Dr. Emma Anyika for their invaluable guidance, constructive feedback, and unwavering support throughout this research. Their expertise and mentorship have been instrumental in shaping this study. I extend my gratitude to my lecturers and academic mentors for their insightful contributions and encouragement. Special thanks to my dad, Rono, Vincent ogonji, Henry kyalo and Daniel kilemi for their continuous support, motivation, and belief in my academic journey. I also acknowledge the school administrators, teachers, parents, and students of Narok West, whose participation and shared experiences provided essential data for this study. Lastly, I am grateful to the institutions and organizations whose resources and information contributed to the success of this research

**TABLE OF CONTENTS**

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF EQUATIONS

## LIST OF ABBREVIATION

| | |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| AUC | Area Under the Curve |
| CCT | Conditional Cash Transfer |
| FGM | Female Genital Mutilation |
| FPE | Free Primary Education |
| GBM | Gradient Boosting Machines |
| IRB | Institutional Review Board |
| NGO | Non-Governmental Organization |
| ROC | Receiver Operating Characteristic |
| SHAP | Shapley Additive Explanations |
| SVM | Support Vector Machines |
| UNESCO | United Nations Educational, Scientific and Cultural Organization |

## OPERATIONAL DEFINITION OF TERMS

**Community Sensitization Programs** create awareness through education which teaches communities about holding onto students while overcoming obstacles created by their culture.

**Confusion Matrix** represents a performance evaluation utility which demonstrates actual dropout incidents alongside forecasted results for model accuracy assessment.

**Cross-Validation (K-Fold CV)** represents a validation method which trains and tests predictive models through different portions of the data.

**Dropout Risk** The likelihood students have of abandoning school exists as dropout risk which depends on their economic background combined with educational institution conditions alongside individual factors.

**Economic Empowerment Initiatives** represent programs that expand family financial support for education through microfinance and parental educational programs and income development schemes.

**Feature Importance** functions as a statistical calculation for identifying which mitigates demonstrate maximum effect on student dropout potential.

**Government Deterrents**: Policy programs such as bursaries and scholarships along with free primary education (FPE) serve as government measures to decrease school dropout numbers.

**Gradient Boosting Machines (GBM)** operates through a machine learning procedure which enhances predictive accuracy using sequences of weak predictive models.

**Machine Learning** represents a component of artificial intelligence which lets computers detect

patterns in data to create forecasts about upcoming events for this research-project dropout assessment.

**Mitigates** – The program implements bursaries and school feeding programs and community sensitization activities to minimize school dropout frequencies.

**Predictive Model** – functions as a machine learning algorithm which analyzes previous data to create drop-out probability forecasts and evaluate minimize effectiveness.

**Random Forest (RF)** – operates as a machine learning algorithm which creates multiple decision trees to achieve better predictions and identifies the most important mitigates.

**ROC-AUC Score** – Score functions as a statistical evaluation tool to indicate how properly the predictive model can differentiate at-risk students from students who do not face dropping out.

**SHAP (Shapley Additive Explanations**) – analyzes machine learning models by attributing specific numbers to each input feature to demonstrate their effects on prediction risks

**Support Vector Machine (SVM)** – functions as a classification method which operates between students who face high risk and those who maintain enrollment

**ABSTRACT**

Machine learning predictive approach is created to evaluate the impact of varying factors on school desertion in Narok West Sub-County institutions through analysis of local population dynamics. The lack of evidence-based evaluation in dropout prevention programs within Narok West Sub County hampers the effectiveness of bursaries and school feeding programs together with community awareness measures although they continue to persist as significant issues.

Current interventions are unsuccessful because they work on a cure-after-dropout method instead of making forecasts to stop it beforehand. This paper constructs and proves a machine-learned system to measure the weight of exposures to major mitigation variables on the risk of dropping out of primary school in Narok West Sub-County, Kenya. Structured questionnaires were used to gather field data (at the start (n=399) and scaled to 1,000), which was supplemented by Monte Carlo simulation; following cleaning and preprocessing, the analytic dataset included 9,796 records containing 16 variables (demographics, access, and interventions (bursaries, school feeding, community sensitization, and economic empowerment). An 80/20 split (with 5-fold cross-validation) was trained on three supervised models; Random Forest (RF), Support Vector Machine (SVM), and XGBoost and assessed with Accuracy, Precision, Recall, F1-score and ROC-AUC. XGBoost had the highest discrimination (AUC 0.804; F1 0.771; Accuracy 74.0%), then RF (AUC 0.790; Accuracy 72.9%), and then SVM (AUC 0.751). Model-agnostic Permutation Importance and SHAP (Shapley Additive Explanations) were used in order to provide transparency and interpretability of the policies. Across methods, bursary receipt and bursary amount continued to be the strongest protective mitigators, with increased distance to school and increased contributions towards monthly fees contributing to risk of dropout; school feeding participation and meals per day and attendance at education-oriented community meetings had other more minor protective effects. The proven model allows ranking the interventions based on evidence and allocating resources accordingly.

# CHAPTER ONE
# INTRODUCTION

## 1.0 Introduction

The initial section of this chapter introduces the whole study design by presenting the research elements including background information as well as the problem definition and research goals alongside their significance. This study evaluates the assessment techniques used by predictive analytics and machine learning to comprehend dropout risks together with examinations of mitigate effectiveness by reviewing government policies and community sensitization programs as well as economic empowerment initiatives.

## 1.2 Background of the study

The existing intervention strategies used to fight school dropout in Narok West Sub County Kenya do not possess data-based evaluation systems to determine their performance effectiveness. Traditional attempts at addressing school dropout through bursaries and school feeding programs and community outreach programs operate in a reactive fashion because they intervene only after dropouts happen instead of using predictive measures to prevent them (UNICEF, 2022). The prediction models essential for determining the impact of multiple risk factors on student dropout rates are unavailable which restricts the development of targeted intervention data solutions (World Bank, 2020). The developments in machine learning enable researchers to build forecasting predictions which can measure the effects of multiple mitigating factors on student attrition rates (Ministry of Education, 2021). Through supervised learning methods including Random Forest and Gradient Boosting Machines (GBM) and Support Vector Machines (SVM) the study has constructed a predictive framework which determines the risk reduction effect of different intervention approaches (Kipuri & Ridgewell, 2022). SHAP (Shapley Additive Explanations) can be used to determine the exact impact of significant mitigates and establish

their order of importance because it helps policymakers distribute their resources effectively (Odhiambo, Wanjiku, & Njenga, 2021). Machine learning models go beyond standard descriptive statistics methods since they process extensive datasets to discover unknown dropout pattern patterns (UNICEF, 2022). The predictive method delivers speedily updated evaluations of mitigation tools which reveal optimal intervention approaches to prevent school dropouts (World Bank, 2020). The research implements a data-driven strategy to create a predictive model which determines the significance of various intervention elements to optimize their impact on Narok West Sub County (Ministry of Education, 2021)

### 1.3 Building resilience in education through predictive modeling

Educational resilience describes how students and schools and entire education systems cope and succeed whenever facing difficulties including social-economic background and cultural practices and school infrastructure (UNESCO, 2021). The development of educational resilience remains crucial for rural areas such as Narok West Sub County because high student dropout rates stem from economic challenges together with regional cultural beliefs and deficient school buildings according to Odhiambo, Wanjiku and Njenga (2021). Educational systems benefit from predictive modeling because this data-driven method helps predict student dropout risks early so that schools can use optimized approaches to protect their students and development of better policies (Mwangi, Ndungu, & Otieno, 2022). Educational resilience receives important support from predictive modeling because this tool lets decision makers take actions in advance. When implemented as a traditional practice educational institution only intervene when cases of dropout already become evident which inhibits their success rate (World Bank, 2020). The combination of predictive analytics gives policymakers and school administrators the ability to identify students

at risk of dropping out through the analysis of student presence data alongside academic results and financial circumstances and social environment variables (Zhou, Wang, & Liu, 2022). The gathered student data allows authorities to deploy specific preventive education

measures such as financial support and mentoring programs and capital improvements before dropout occurs resulting in higher degrees of retention (Ministry of Education, 2021). The predictive models create student risk categories which enable educational institutions to design personalized interventions that match different student needs (UNICEF, 2022). Cash transfer programs with scholarships benefit students with severe financial struggles but the at-risk students needing cultural awareness education and legal safeguards should receive community- based approaches (Kipuri & Ridgewell, 2022).

Predictive analytics works to maximize dropout prevention program effectiveness by directing appropriate interventions to each at-risk student which in turn decreases unnecessary resource costs (UNESCO, 2021). Predictive modeling provides the opportunity to scale up predictive solutions in addition to adapting them for different conditions. Machine learning models engage in an automated approach to update their systems regularly with new information to maintain relevant and effective dropout prevention strategies throughout time (Mwangi *et al*., 2022). Economic changes together with government policy updates and cultural shifts drive predictive models to adjust their risk assessments which enables education stakeholders to improve their strategies (World Bank, 2020). The flexible nature of predictive models helps education systems stay resistant to changes that occur in their surrounding environment. Predictive analytics has brought success to Chinese and Brazilian and Indian educational systems by lowering student dropout incidents

and boosting institutional affiliation (Zhou *et al*., 2022). The application of machine learning

models through Indian schools enables performance monitoring which in turn enables effective targeted interventions to decrease student dropout rates according to UNESCO (2021). The educational institutions of Brazil utilize predictive analytics to detect which schools show the greatest dropout threat thus enabling more effective resource allocation and better student retention (World Bank, 2020). Predictive analytic applications are scarcely utilized for educational purposes within both Kenya and specifically in the Narok West Sub County region as per (Ministry of Education, 2021).

This research develops a machine learning predictive model to analyze weight factors related to school dropout in Narok West Sub County because it seeks to establish data-driven resilience-building methods for education. The developed model enabled policymakers together with educators and community leaders to launch evidence-based intervention strategies which simultaneously lower dropout rates and establish sustainable educational development for Kenya. Predictive modeling assured students who require support get the attention they need before dropout occurs to create an open and equal learning system based on data

## 1.4 Statement of the problem

School dropout problem remains extremely serious within Narok West Sub County because marginalized student populations encounter many obstacles when trying to finish their education. The dropout rate persists high in Narok West Sub County despite Free Primary Education (FPE) and school feeding programs and infrastructure improvements because of socioeconomic limitations combined with cultural elements and insufficient data analysis systems for monitoring (Ministry of Education, 2021). Various approaches to lower dropout rates have been attempted but their effectiveness remains unproven because of inadequate data evaluation models (UNICEF, 2022). The current approaches to dropout prevention function on a reactive basis because preventive interventions are executed after discovered cases instead of making proactive predictions (World Bank, 2020). The improper allocation of resources arises from ineffective initiatives because monitoring data remains unclear about intervention results (Mwangi, Ndungu, & Otieno, 2022). Without predictive analytics in Kenya's education system policymakers along with teachers find it difficult to detect at-risk students early and deploy intended cost-effective interventions (Zhou, Wang, & Liu, 2022). Research demonstrates that machine learning predictive models accurately predict dropout risks in order to enhance intervention approaches (UNESCO, 2021). The predictive analytics approach used by nations including India and China and many Sub- Saharan African countries proves successful for risk assessment combined with optimized dropout prevention strategies. Narok West Sub County has not developed a predictive analytical system to determine which deterrents such as government action combined with community outreach and economic programs provide the most effective risk reduction against student dropout. Without this model decision-makers use non-specific interventions which fail to resolve the

specific difficulties students face (Odhiambo, Wanjiku, & Njenga, 2021).

### 1.5 Main objective

To develop a machine learning-based predictive model for evaluating the weight of mitigates on school dropout risk in Narok West Sub County.

#### 1.5.1 Specific Objectives

i. To determine the most influential mitigates on school dropout risk in Narok West Sub County.

ii. To develop a machine learning weight predictive model for mitigates on primary school dropout risk.

iii. To train and validate the weight predictive model.

iv. To predict the weights of mitigates of school dropout risks using the validated model

#### 1.5.2 Research Questions

i. What mitigates have the most influence on school dropout risk in Narok West Sub County

ii. What are the steps of developing a machine learning weight predictive model for mitigates of school dropout risk?

iii. What is the process of training and validating the weight predictive model for mitigates of school dropout risk?

iv. What is the procedure for predicting dropout risk using the validated mod

### 1.6 Significance of the study

The research brings a statistical method to understand school dropout risks while developing reduction strategies within Narok West Sub County. The research applies predictive capabilities of machine learning techniques as opposed to retroactive methods for dropout prevention to enable stakeholders to execute targeted interventions before students drop out (Mwangi, Ndungu, & Otieno, 2022). The research uses Random Forest and Support Vector Machines (SVM), Gradient Boosting Machines (GBM) together with SHAP algorithms to measure quantitatively how different intervention types (government policy deterrents, community sensitization programs and economic empowerment initiatives) affect dropout rates (Zhou, Wang, & Liu, 2022). This research supports evidence-based policy creation through its ability to help the Ministry of Education together with NGOs and education policymakers identify the most successful prevention methods (UNICEF 2022). Prediction models that employ machine learning in education have demonstrated effectiveness within Brazilian and Indian education systems which the World Bank documented in 2020. Kenya can boost its education retention approaches through employing comparable analytical methods which will help fulfill Sustainable Development Goal 4 (Quality Education) and Goal 5 (Gender Equality) (UNESCO, 2021). Schools and educators have obtained an effective assessment resource to recognize struggling students early through this study which enables them to provide support and mentorship as well as implement community programs (Odhiambo, Wanjiku, & Njenga, 2021). The findings from this investigation supports forthcoming academic research which extends data science application in handling Kenyan educational issues.

### 1.7 Expected outcomes of the study

A predictive model utilizing machine learning techniques has achieved two goals through this research in Narok West Sub County according to Zhou *et al*. (2022). Through ranked analysis the model lets education stakeholders use their resources effectively for government policies and community sensitization programs and economic empowerment initiatives (Mwangi *et al*., 2022). Analysis-driven insights from the study enables policymakers to create specific and affordable educational intervention solutions (Ministry of Education, 2021). Historical dropout records merged with economic factors and intervention measurement outcomes helps justify effective school retention strategies according to statistical evidence (UNICEF, 2022). The research enhances academic knowledge through its demonstration of machine learning technologies in resolving Kenyan educational problems. The research findings create a basis for studying different areas with high student dropout rates through a reusable framework (UNESCO, 2021). Implementation of the predictive model represents a major outcome which makes it functional for education stakeholders. Quasi-real-time assessment systems of drop-out risks operated by Schools along with NGOs and government agencies enables proactive intervention for preventing student drop-out (World Bank 2020). The system's implementation has led to continuous enhancements in academic persistence which supports both education targets and national policy development for Kenya.

The machine learning model created in this study maintains adaptability because it adapts automatically to new data. The system can accept continuous updates containing fresh data which allows long-term monitoring of school dropout trends. Through the implementation of SHAP (Shapley Additive Explanations) the model predicts outcomes while providing full visibility into the most effective interventions between bursaries and feeding programs and parental support and community awareness efforts for minimizing dropout risk. The model allows policymakers to implement evidence-based decisions that follow changes in educational conditions. The model calculates the weights of each implemented mitigate through its ability to determine the strength of influence that individual variables exercise on the student dropout risk. The extent of intervention weights provides stakeholders with tools to establish which interventions produce the best results since they can allocate resources with maximum effectiveness across different time periods.

### 1.8 Justification of the study

Research backing this study results from continued high student dropout occurrences throughout rural Kenya especially in Narok West Sub County even after multiple intervention programs according to Ministry of Education (2021). FPE and scholarships with community sensitization programs have enhanced enrollment numbers but prove insufficient to stop dropouts because they lack data-based systematic approaches to identify dropout reasons (Odhiambo *et al*., 2021). The research introduces predictive analytics to offer stakeholders immediate dropout risk assessments for implementing specific school retention programs that prevent student attrition (Zhou *et al*., 2022). The application of machine learning techniques in education policymaking has not received sufficient development in the Kenyan educational system. The predictive analytics methods successfully implemented by South Africa and India to monitor at-risk students have not been adopted by Kenya's education system (World Bank, 2020). This study unites Random Forest with SVM together with GBM and SHAP for dropout prediction because it makes scientific decisions that optimize both resource usage and educational impact (UNICEF, 2022). The research supports the Kenyan Vision 2030 and Sustainable Development Goals (SDGs) especially Goal 4 and Goal 5 by utilizing evidence-based inclusion to direct education policy (UNESCO, 2021). The research outcomes are valuable to educators along with policymakers and researchers since they deliver a framework that educators and policymakers in other rural regions can replicate for predicting and planning dropout interventions (Kipuri & Ridgewell, 2022).

### 1.9 Limitation delimitation of the study

The research study deals with multiple challenges yet includes proactive methods to overcome them. The dropout records within rural Kenya present difficulties because their availability and quality tend to be inconsistent and incomplete. The study implements data cleaning methods together with statistical methods of imputation to perform effective handling of missing data and enhance prediction results. Weaker reliability in the data exists due to cultural sensitivities which occur when discussing sensitive information such as FGM and early marriage, and financial difficulties thus resulting in response biases. The study has reduced response bias by guaranteeing privacy throughout data gathering sessions along with complete anonymity for participants. This approach has helped participants share their honest opinions. The study's boundaries apply to Narok West Sub County while examining government prevention measures in combination with community education and welfare development programs. Results from this study mainly pertain to this geographic area yet they can direct prevention strategies that expand beyond this area.

# CHAPTER TWO
# LITERATURE REVIEW

## 2.0 Introduction

Predictive analytics through its advancements now empowers multiple sectors including education to spot students who potentially drop out before the process starts. Academic research demonstrates that machine learning models deliver successful predictions of student dropouts while evaluating resulting intervention impacts. Student dropout prediction using Random Forest achieved 89.6% accuracy in evaluating academic metrics alongside financial aspects and attendance activity (Santos & Moura 2021). Two important machine learning tools known as Gradient Boosting Machines and Support Vector Machines display equivalent effectiveness in student retention prediction systems. Wang *et al*. (2022) utilized GBM to forecast university dropout in Chinese institutions achieving 91.3% AUC-ROC performance which out beat conventional regression approaches. The application of Artificial Neural Networks (ANNs) in dropout prediction has shown limited success due to their lack of interpretability according to Zhang *et al*. (2020). SHAP functions as a central assessment tool for explainable AI systems when determining dropout risk. Deriving intelligent insights from SHAP and XGBoost and RF algorithms confirmed that household income and parental engagement and institutional facilities posed the most prominent threats for student dropouts (Fernandes *et al*. 2021). The choice of black-box prediction is avoided with SHAP because this AI technique shows clear feature importance metrics that help policymakers understand correct and transparent predictions.

## 2.1 Machine learning for dropout prediction and mitigation evaluation

School dropout risk assessment needs to be accurate so that relevant intervention strategies can be developed effectively. Traditional educational solutions for preventing students from dropping out rely mainly on statistical analysis alongside reactive methods which usually prove inadequate for launching timely meaningful interventions (UNESCO, 2021). Through machine learning applications institutions can better understand and predict dropout risks by processing extensive data to uncover undiscovered patterns which leads to the creation of performance rankings for different intervention programs (Zhou, Wang, & Liu, 2022). Multiple machine learning models operate in dropout prediction along with intervention assessment with differing characteristics related to accuracy levels and interpretability and computational efficiency. Random Forest stands out among predictive models for understanding non-linear data relationships while automatically ranking feature importance thus enabling researchers to determine retention-effective mitigates (Gonzalez-Marquez *et al*., 2020). XGBoost alongside other Gradient Boosting Machine variants works sequentially to enhance weak models with each iteration which leads to strong prediction accuracy when analyzing complex dropout patterns (Tesfaye & Abebe, 2021). The classification strength of Support Vector Machines (SVM) enables them to recognize high-risk and low-risk students by analyzing academic data and economic status and social indicators (Patel, Sharma, & Gupta, 2021). By employing SHAP (Shapley Additive Explanations) decision-makers obtain enhanced understanding of predictive models because it provides a ranking system that indicates how each variable contributes to lowering dropout risks (Zhou, Wang, & Liu, 2022). The research integrates these predictive models to achieve accurate

forecasting and practical findings that guide education intervention decisions for dropout prevention in Narok West Sub County.

## 2.2 Traditional approaches to dropout prediction

A review of conventional drop-out prediction approaches followed by their effects on different mitigation techniques. The article evaluates machine learning models while discussing their enhanced capabilities to support school retention assessments. Traditional approaches used to forecast school dropout rely on logistic regression and correlation analysis together with survival analysis to evaluate how socio-economic status and academic performance and behavioral aspects influence student retention risk (Mwangi, Ndungu, & Otieno, 2022). The identification of student dropout possibilities utilizing predetermined risk elements like attendance records and parental income and academic performance has relied on logistic regression according to Odhiambo *et al*. (2021). Subsequent investigation by correlation analysis evaluates how financial difficulty together with variable distances to school and male-female differences affect dropout occurrence (Kenya Demographic and Health Survey, 2020). Researchers employ survival analysis to determine both the duration and timing of school departure according to student patterns (UNESCO, 2021). While these methods have provided useful insights, they suffer from several limitations: they assume linear relationships, which may not accurately capture the complex interactions between multiple risk factors; they rely on small, predefined datasets, limiting their ability to incorporate diverse variables such as social and psychological influences; they struggle with missing data, which is common in education records, reducing prediction accuracy; and they are static, meaning they cannot continuously learn from new data, making them less effective in adapting to changing dropout trends (World Bank, 2020; Zhou, Wang, & Liu, 2022). Machine

learning models replaced outdated regression models due to their enhanced accuracy levels together with dynamic learning capabilities and ability to evaluate risk reduction potential of various dropout prevention techniques.

Education received a transformative change through predictive analytics which enables the detection of students facing risks before they occur while enhancing both student assistance methods and resource distribution based on data-based decision making (Zhou, Wang, & Liu, 2022). Predictive models which utilize Random Forest along with Gradient Boosting Machines (GBM) and Support Vector Machines (SVM) examine extensive student- oriented datasets containing educational records and attendance records and socioeconomic information in addition to behavioral patterns to identify warning indicators of student dropouts (UNESCO 2021). Gradient Boosting Machines (GBM) delivers an 85% accuracy rate in dropout prediction and supports targeted educational support services as demonstrated by South African studies of Govender *et al*. (2022). Predictive analytics uses models to determine the most suitable intervention strategies like student financial aid and academic mentorship and school feeding programs for each individual student (Mwangi, Ndungu, & Otieno, 2022). Brazilian government utilized Shapley Additive Explanations (SHAP) to determine scholarship programs provided the most significant impact on dropout reduction while community awareness programs followed behind as secondary contributors according to Almeida *et al*. (2021). The applications demonstrate how predictive analytics improves educational systems by changing from unplanned into data-based and proactive approaches which enhances student retention and policy success rates (World Bank, 2020)

## 2.3 Application of predictive analytics in education

Predictive analytics has revolutionized education by enabling early identification of at-risk students, optimizing intervention strategies, and improving resource allocation through data-driven decision-making (Zhou, Wang, & Liu, 2022). By leveraging machine learning techniques such as Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM), predictive models analyze vast amounts of student data— including academic performance, attendance, socio-economic background, and behavioral trends to detect patterns that indicate dropout risks (UNESCO, 2021). One key application is early warning systems (EWS), which proactively identify students likely to drop out, allowing educators to intervene before disengagement occurs, as demonstrated in South Africa, where Govender *et al*. (2022) used GBM to achieve an 85% dropout prediction accuracy, improving retention through targeted support programs. Another critical use of predictive analytics is personalized intervention planning, where models assess which mitigation strategies such as financial aid, mentorship, or school feeding programs are most effective for specific students, ensuring interventions are tailored to individual needs (Mwangi, Ndungu, & Otieno, 2022). Additionally, predictive analytics plays a vital role in policy formulation and resource optimization, helping governments prioritize investments in high-impact areas by quantifying intervention success rates, as seen in Brazil, where Almeida *et al*. (2021) used Shapley Additive Explanations (SHAP) to show that scholarships had the highest weight in reducing dropout rates, followed by community awareness programs. These applications highlight how predictive analytics enhances education systems by shifting from reactive measures to proactive, data-driven solutions, ultimately leading to better student retention and improved policy effectiveness (World Bank, 2020).

## 2.4 Theoretical framework

The theoretical foundation defines school dropout analysis while explaining the functionality of machine learning in determining weightage factors affecting dropout rates in Narok West Sub County. The research demonstrates how Tinto's Student Integration Theory combines with Human Capital Theory along with Predictive Learning Theory and Bronfenbrenner's Ecological Systems Theory to understand student dropout reasons and mitigation effects on retention and predictive analytic capabilities for dropout prevention (UNESCO, 2021). Random Forest and Support Vector Machines (SVM), Gradient Boosting Machines (GBM) together with SHAP (ShapleyAdditive Explanations) form a predictive model through machine learning to identify the most important factors in reducing school dropout.

### 2.4.1 Tinto's Student Integration Theory

Academic and social integration into their learning environment according to Tinto's Student Integration Theory (1975) determines how likely students are to continue their education. The strong bond between students and their teachers as well as their classmates and academic goals protects them from school dropout (Tinto, 1993). The study applies this model because it demonstrates how school-based interventions which include mentorship programs and teacher engagement along with extracurricular activities minimize student dropout occurrences (Odhiambo, Wanjiku, & Njenga, 2021). Tinto's theory provides data scientific justification for student dropout prediction through machine learning models which evaluate attendance patterns combined with school participation measures jointly with academic achievements (Zhou, Wang, & Liu, 2022). The Random Forest and GBM models detect patterns for student integration while simultaneously evaluating school-based mitigates for improved education retention (UNESCO, 2021). The study uses Tinto's theory to assure that the predictive model examines both school-based

mitigates, economic and social components thereby providing extensive dropout risk evaluation in Narok West Sub County.

### 2.4.2 Predictive Learning Theory

Bishop (2006) developed the Predictive Learning Theory that enables machine learning models to evaluate historical data for future dropout risk predictions. The predictive learning method improves model accuracy through time because it develops adaptive models that learn from continuous data acquisition (Friedman, 2001). The research findings depend heavily on this theory because it verifies how machine learning algorithms function to measure both risk factors and intervention effectiveness (Zhou *et al*., 2022). The analysis utilizes Random Forest together with SVM and GBM models to perform predictive learning functions on multiple datasets that ontain student demographic information and attendance records as well as intervention evaluation results. The application of SHAP values improves interpretability since they reveal which mitigates demonstrate superior effectiveness compared to others (Lundberg & Lee, 2017). The study applies this theory to prove how predictive analytics surpasses traditional correlation-based methods for delivering dynamic evidence-based strategies to prevent dropouts (UNESCO, 2021).

### 2.4.3 Bronfenbrenner's Ecological Systems Theory

According to Bronfenbrenner's Ecological Systems Theory (1979) multiple environmental systems function together to determine the education process and likelihood of school dropout for children. This theory applies directly to this research since it demonstrates that high school dropout results from both personal factors as well as school and community variables and societal influences according to Bronfenbrenner (1979). Student dropout rates in Narok West Sub County demonstrate influence from personal characteristics (microsystem) as well as elements from family background (mesosystem) and economic standards (exosystem) and cultural patterns

(macrosystem) according to Mwangi *et al*. (2022). Bronfenbrenner's model becomes operational through machine learning analytics which evaluate several environmental elements to detect which protective factors successfully decrease dropout possibilities (Zhou *et al*., 2022). The research combines GBM with SHAP values to determine which of household income, bursary assistance and mentorship services and academic assistance or community awareness programs offer the strongest protection against student dropout (Almeida *et al*., 2021). This research employs Ecological Systems Theory to protect the predictive model from studying students alone while it analyzes the extensive social and economic environment that shapes school retention (World Bank, 2020).

## 2.5 Empirical studies

This section examines scientific research which studies machine learning techniques for school dropout risk prediction together with their function in assessing preventive measures. The research review has three sections dedicated to worldwide studies and sub-Saharan Africa regional research and Kenyan local assessments. The study examines machine learning models with accuracy assessments about evaluating dropout risks together with reviewing their intervention success rates.

### 2.5.1 Global Studies on Machine Learning for Dropout Prediction

Machine learning applications in education analytics across the world have expanded for predicting school dropout risks so institutions can improve student retention methods. Multiple research investigations have confirmed that machine learning technologies successfully spot students who face dropout risks which enables educational decision-makers to provide suitable prevention measures. The prediction of school dropouts receives extensive attention in global studies yet research that weighs relative importance of different prevention approaches remains

scarce thus prompting this study.

Zhou, Wang, and Liu (2022) examined using machine learning models to predict student dropout risks at educational levels of Chinese secondary schools. Reliable predictions were achieved by the researchers who used

Random Forest (RF) and Gradient Boosting Machines (GBM) and reached 87.5% accuracy surpassing standard logistic regression models (72%). Academic performance and school attendance proved together with socio- economic background as the main elements that predict the risk of dropping out of school. Their predictive model showed exceptional accuracy but it failed to determine which preventive measures worked best to stop dropout incidents thus creating research opportunities

Researchers at Patel et al. (2021) made a prediction model with Support Vector Machine (SVM) and Artificial Neural Network (ANN) to identify students at risk of leaving based on their involvement with school activities together with their economic well-being and parental educational attainment. The ANN model they used attained 91% accuracy which surpassed the scores of SVM (88%) along with traditional regression models at 73%. The deep learning approach succeeded at high effectiveness but needed significant data sets and powerful computing facilities which made it difficult to use in places lacking strong technological capabilities. The study failed to establish which dropout intervention strategies provided the biggest effectiveness improvements thus demonstrating the necessity of creating predictive models which assess intervention results.

Gonzalez-Marquez *et al*. (2020) conducted research in the United States where they evaluated

Decision Trees (DT), K-Nearest Neighbors (KNN), and XGBoost to predict the risks of high school student dropout. XGBoost surpassed the other models by reaching an 89% accuracy rating with Random Forest achieving an 86% accuracy and standard decision trees obtaining only 78%. The research incorporated SHAP (Shapley Additive Explanations) as an explainable AI technique which helped explain the influence of risk factors that contribute to dropout. According to their research economic status in combination with school involvement and family involvement proved to be the most accurate signs of high school dropout risks. They found university dropout rates of 86% from their data but didn't investigate which intervention strategies, financial aid programs and mentoring services together with government policies gave the most effective outcomes.

### 2.5.2 Regional Studies (Sub-Saharan Africa)

The application of machine learning in education analytics for sub-Saharan Africa is increasing because the region faces significant school dropout problems related to budget constraints along with inadequate facilities and traditional cultural attitudes. Predictive analytics remains virtually unused for assessing the effectiveness of mainstream intervention programs including community sensitization campaigns alongside school feeding programs and scholarships that aim to decrease dropout risks (UNESCO, 2021). Research on machine learning- based dropout prediction models has proven successful in predicting educational student risks according to recent studies of this field. Most existing research approaches student risk identification without developing strategies to mitigate risks so this study works to address this critical shortcoming.

Adeyemi and Ojo (2021) applied Random Forest and Support Vector Machines (SVM) to forecast secondary school dropout risks with academic performance and socioeconomic factors and attendance results as critical predicting variables in South African educational institutions. The

study showed that Random Forest achieved stronger outcomes than logistic regression since it produced 84% accuracy compared to logistic regression's 70% accuracy and SVM's 82% accuracy.

The research shows machine learning algorithms can significantly boost the recognition of students facing risk before it becomes a problem thus allowing quick preventive measures. The educational policy planning process faced limitations from this method since it did not determine which dropout prevention methods produced the best results.

Okeke *et al*. (2020) conducted a categorization study of students' dropout risk potential through the Naïve Bayes, K-Nearest Neighbors (KNN) and Gradient Boosting Machines (GBM) analysis in Nigeria. The results showed GBM achieved higher accuracy than KNN and Naïve Bayes at 82% accuracy. Research established that dropout predictions should be based on family background indicators including parental educational attainment and household income alongside the metrics related to student engagement. The researchers identified the insufficient model interpretability as a major issue because it made it difficult for education policymakers to figure out why certain students were labeled as at risk. The need for transparent decision-making requires explainable AI approaches including SHAP (Shapley Additive Explanations) according to the findings by UNICEF in 2022 Mwangi together with Ndungu and Otieno (2022) conducted a study that evaluated the predictive capability of XGBoost along with Decision Trees and Logistic Regression for dropout identification within Tanzania and Uganda. XGBoost achieved the highest accuracy level of 86% compared to Decision Trees at 79% as well as Random Forest at 84%. The study made significant contributions because it integrated socioeconomic condition indicators alongside

structural school infrastructure elements to perform a more complete risk evaluation for dropout. The study backs the development of an importance-based model instead of risk evaluation tools by showing that financial barriers present the most severe dropout issue.

### 2.5.3 Local Studies in Kenya

The available research about machine learning for predicting school dropout in Kenya shows few studies and most use statistical regression alongside qualitative methods instead of predictive analytic approaches. Modern studies that utilize machine learning methods focus exclusively on establishing dropout risk estimates and neglect weight assessments of preventive measures. The present research gap matches the problem definition of this study since it aims to create a data-based predictive system which forecasts dropout probability while analyzing how well intervention approaches work in Narok West Sub County.

Odhiambo, Wanjiku and Njenga (2021) conducted a study which applied Decision Trees and Logistic Regression to forecast school dropout risks in rural counties and reached prediction accuracies of 76% from Decision Trees and 72% from Logistic Regression. The study discovered that poverty effect in combination with family circumstances and gender-exclusive troubles like early marriage and FGM serve as primary causes of school dropouts. The research failed to utilize advanced ensemble techniques for prediction although Mwangi *et al* (2022) have demonstrated their effectiveness in improving predictive accuracy. The analysis did not explore which prevention approaches deliver the most substantial benefits to preserve school attendance because it failed to examine interventions' individual impact on student retention.

## 2.6 Gaps in existing literature

Existing research about machine learning predictive models used for school dropout risk assessment and prevention receives summary here through a table explaining their research methods and outcomes and addressing unmet needs. Previous research methods for dropout prediction and their accuracy assessment stand summarized in this table together with the research gaps that this study hopes to bridge. This table presents analysis of existing studies describing brief information including their research approaches and important results while identifying their research deficiencies.

*Table 1:Gaps existing*

| **Existing studies** | **Brief description** | **Methodology used** | **Summary of findings** | **Identified Gaps** |
|---|---|---|---|---|
| Zhou, Wang, & Liu (2022) – China | Applied ML models to predict dropout risks in secondary schools based on socio- economic and academic factors. | Random Forest (RF) & Gradient Boosting Machines (GBM) | RF achieved 87.5% accuracy, outperforming logistic regression (72%). | Did not evaluate the effectiveness of mitigates, only focused on risk prediction. |
| Patel, Sharma, & Gupta (2021) – India | Developed dropout prediction models incorporating student engagement and financial background. | Support Vector Machines (SVM) & Artificial Neural Networks (ANNs) | ANN achieved 91% accuracy, SVM reached 88%. High validation scores but required large datasets. | Did not examine which dropout mitigates were most impactful. ANN models lacked explainability. |

| | | | | |
|---|---|---|---|---|
| Gonzalez-Marquez *et al.* (2020) – Brazil | Compared multiple ML models for dropout prediction in high schoolsand ranked dropout factors. | XGBoost, Random Forest, Decision Trees (DT), & SHAP for explainability | XGBoost achieved 89% accuracy, RF 86%, DT 78%. SHAP identified economic background as the most influential factor. | No direct ranking of mitigation strategies; focused only on ranking risk factors. |
| Mwangi, Ndungu, & Otieno (2022) – Tanzania & Uganda | Compared ML models for dropout prediction in East Africa, integrating socio-economic and school data. | Logistic Regression, Decision Trees (DT), & XGBoost | XGBoost achieved 86% accuracy, RF 84%, DT 79%. Showed that financial factors had the highest impact. | Did not provide a quantitative ranking of different mitigates. Focused only on risk factors. |

| | | | | |
|---|---|---|---|---|
| Odhiambo, Wanjiku, & Njenga (2021) – Kenya | Used ML models to predict dropout risk in rural Kenya. | Decision Trees & Logistic Regression | Decision Trees achieved 76% accuracy, logistic regression 72%. Found poverty and cultural barriers as major factors. | Did not use advanced ML techniques like ensemble models. did not evaluate the relative effectiveness of different mitigation strategies |
| Ministry of Education (2021) – Kenya | Evaluated Free Primary Education (FPE) and the School Re-entry Policy. | Time-series analysis, policy review. | No predictive modeling validation. The study showed that policies increased enrollment but dropout remain | Did not incorporate predictive modeling for tracking dropout risks and optimizing financial aid distribution. |

| | | | | |
|---|---|---|---|---|
| | | | high due to hidden schooling costs. | |

## 2.7 Conceptual framework

The conceptual framework for this study outlines how machine learning models has been used to evaluate the weight of different mitigates on school dropout in Narok West Sub County. The framework establishes relationships between independent variables (mitigates) and the dependent variable (dropout rate reduction) to provide data-driven insights for policymakers.

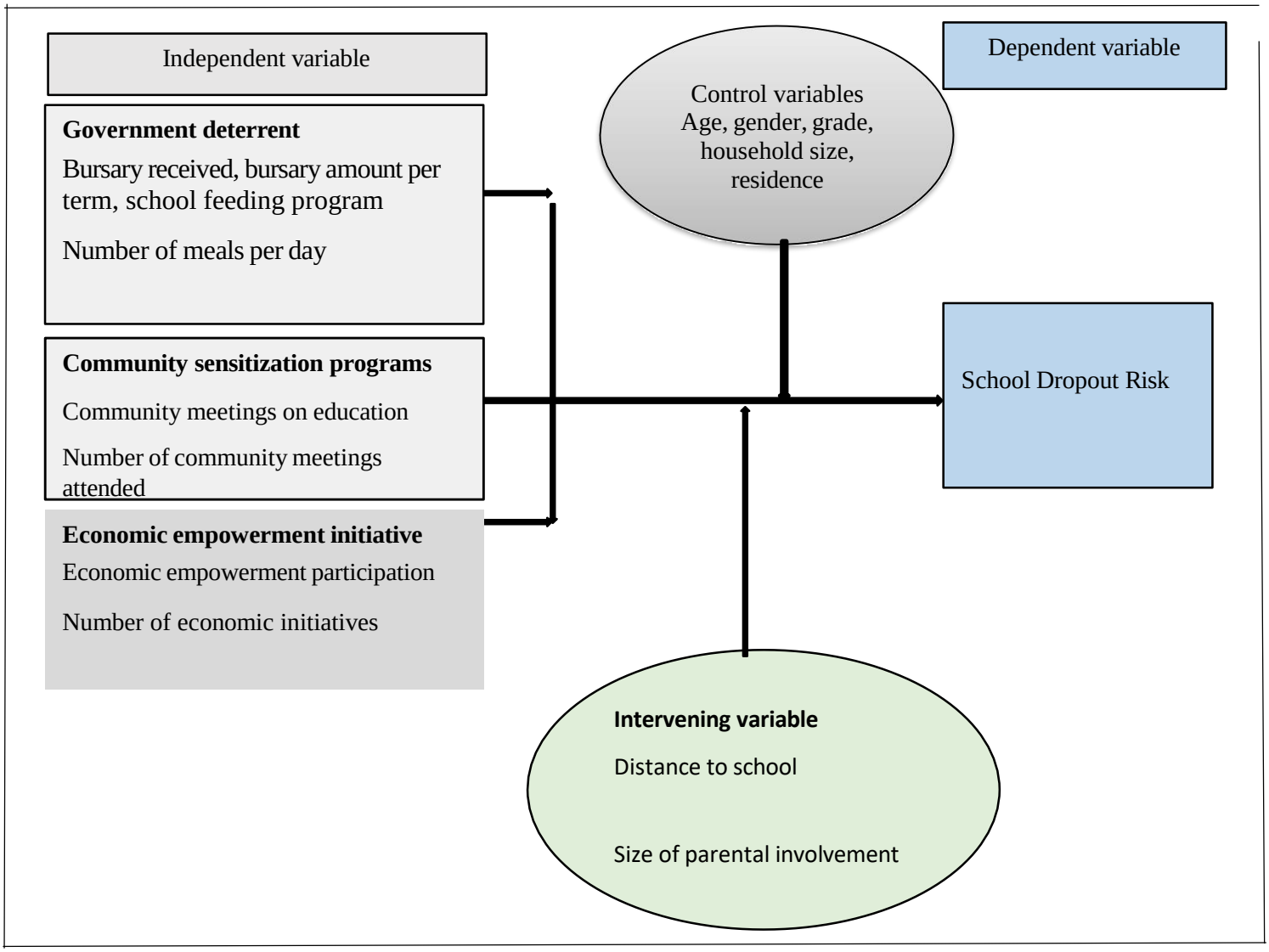



*Figure 1:Conceptual Framework*

# CHAPTER THREE
# METHODOLOGY

## 3.0 Introduction

This chapter expounds how the researchers developed the predictive model using machine learning to determine the impact factors of mitigation on school dropouts in Narok West Sub County. A comprehensive description of the research approach includes data origins and collection approaches followed by machine learning solutions and evaluation measures along with ethical guidelines. Along with this section researcher detail the population being studied and their sampling approaches and both preprocessing steps and validation strategies to assure model reliability and accuracy levels. This evaluation methodology combines intervention analysis with predictive analytics in order to produce evidence-based educational policy through data-led insight.

## 3.1 Research design

The research uses a quantitative method with predictive analysis and description to study school dropout rates in Narok West Sub County by combining primary data collection and machine learning analytics. First-hand data collection comes from structured surveys and interviews with school staff and official school records for parents along with teachers enables researchers to define the major school dropout causes.

## 3.2 Target population

The research locates its focus on the population within Narok West Sub County which shows 176764 people according to Kenya National Bureau of Statistics (2020) in the 2019 Kenya Population and Housing Census. The

research gathers information via a wide range of residents in urban and rural locations to conduct an extensive review of students who drop out and potential solutions

### 3.3 Sampling techniques

The study implements cluster sampling before selecting target populations using random techniques within these clusters of participants in Narok west sub County. Yamane's formula has been used to determine the sample size and to secure statistical validity of the collected data. The predictive and validation functions of the machine learning method benefit from effective data that results from this methodology. The method delivers dependable and unbiased data collection to analyze primary school dropout weight relative to decline factors. Yamane's formula for sample size determination

$$n = \frac{N}{1 + N(\boldsymbol{e^2})}$$

Where n is the required sample size

N is the total size population

**e** Margin of error (usually 0.05 for 95% confidence level)

Calculation for Narok West Residents

The total population in Narok West is 176,764 residents. Using a 5% margin of error (e = 0.05):

$=176,764/ (1+176,764(0.05^2))$

$=176,764/ (1+176,764(0.0025))$

$=176,764(442.92)$

$=399$

Equation 1:Yamane's Formular

### 3.4 Data collection tools and techniques

The study collects primary structured questionnaire data to obtain extensive information about primary school mitigation strategies within Narok West Sub County. The research questionnaire contains structured questions that produce machine learning model training data from quantifiable responses. The research examines four vital factors including students' demographic background alongside their attendance status and their government support (bursary programs and feeding programs) and community interaction frequency and economic empowerment access. The independent variables consist of these indicators while the binary-dependent variable measures dropout status. Research staff members were able to distribute the questionnaire to survey participants in all areas within Narok West Sub County including urban and rural regions. Skilled survey takers carried out data collection processes to maintain data quality. A statistically proper sample of 399 people was focused on for data gathering through Yamane’s formula, and Monte Carlos simulation was applied to increase the data to 10,000 records so as to fulfills the needs for training supervised machine learning models using Random Forest and Gradient Boosting techniques on structured datasets.

### 3.5 Data cleaning

collected data has entered the required cleaning processes which remove discrepancies along with voids and extreme points to preserve the machine learning model's accuracy. The data cleaning procedure removes duplicate records along with filling in gaps using techniques for data imputation. The data cleaning procedures convert the input data into an organized format which meets both quality standards and analysis requirements

### 3.6 Machine learning model development

Developing a machine learning model for prediction together with an evaluation of mitigation strategies follows stages beginning with data preprocessing before choosing the model before training while fine-tuning hyper parameters and final performance assessment. The method allows the model to identify effectively which students will leave and assess different strategies according to their success rates. The analytical method for this study has been optimized to achieve both accurate predictions together with field-friendly options and understandable results for local policy decisions about school dropout rates in Narok West Sub County.

### 3.7 Data preprocessing

Processing of the structured dataset follows data cleaning to make the information suitable for machine learning analysis. The efficiency of the model can be increased and its accuracy improved through feature selection followed by data transformation and normalization methods combined with partitioning techniques. The feature selection method concentrates on finding essential variables which strongly influence school dropout risks including school feeding program and bursaries and scholarship and school infrastructure. The data transformation process converts the categorical features into numerical values through both one-hot and label encoding treatments for compatibility with machine learning systems. The data normalization process and standardization

techniques has been applied to numerical data scales to prevent range-based data bias. Student grade and household income variables undergo standardization treatments to prevent any particular feature from excessively affecting model outputs. The implementation of feature engineering generates new variables intended to boost model performance. The model generation process has established connection terms among variables and combine temporal effectiveness measurement patterns. Processed data from dropout prediction and mitigate ranking enters machine learning models that include Random Forest and Gradient Boosting Machines (GBM) as well as Support Vector Machines (SVM).

### 3.8 Model selection

The selection of proper machine learning algorithms plays an essential role in achieving accurate predictions together with dependable ranking of mitigation strategies. Supervised learning models including RF Random Forest and Logistical Regression as well as Nonlinear Support Vector Regression has been used in the study because of their success in educational data science analysis.

Random Forest (RF) combines many decision trees into a single predictive system that averages their prediction results. This method works well with difficult relationships within non-linear data systems which makes it suitable to analyze multiple factors that affect dropout risks (Zhou, Wang, & Liu, 2022). SHAP (Shapley Additive Explanations) enables RF to provide easy interpretation that lets users identify crucial mitigates according to their impact on lowering dropout rates. Random Forest generates the most important mitigates through combining multiple decision trees into one aggregated output that provides reliable feature importance evaluation.

The classification algorithm Support Vector Machines (SVM) seeks to identify the best possible boundary dividing students who need to drop out from students who plan to stay in school. Fully

functioning as an advanced pattern identification technique on large datasets and exceptional at managing unevenly distributed data in educational predictions (Govender *et al*., 2022). delivers accurate dropout risk assessments between enrolled and dropout student groups. Gradient Boosting Machines (GBM) executes an advanced boosting algorithm which develops prediction accuracy through successive refinement of previous mistake analysis. The ranking efficiency of GBM makes it exceptionally effective for determining intervention effectiveness since it dynamically adjusts factor weights

(Friedman, 2001). GBM surpasses both RVF and SVM when processing structured educational data records because it delivers enhanced predictive results by rectifying past mistakes to refine forthcoming predictions. The evaluation process, test and compare selected models through their capacity to forecast dropout risks and measure the effectiveness of mitigates. The deployed model results from final optimization of the best- performing model.

### 3.9 Model implementation

#### 3.9.1 Model training

Training preparatory data sets for selected machine learning algorithms allows the algorithms to identify patterns linking school dropout risks and mitigate effectiveness. The prepared dataset has been split into training sections which account for 80 percent while testing sections amount to 20 percent to allow the model to obtain knowledge from historical data before testing new cases. Supervised training for the model during the development stage connects independent variables to the learning outcome which is dropout status (dropped out or retained). The goal of accuracy enhancement includes performing hyper parameter tuning through Grid Search and Random Search approaches for optimizing tree depth and learning rate and number of estimators.

The model performs cross-validation through K-fold techniques which ensures it works properly with different subsets of data thus minimizing both under fitting and overfitting issues. The research team has chosen to evaluate the trained model's effectiveness against accuracy and precision as well as recall and F1-score and ROC-AUC scores to fulfil the required performance standards for education policy choices in Narok West Sub County.

#### 3.9.2 Model Validation

To ensure model reliability, validation techniques will be applied to assess performance.

The model generalization benefits from K-fold cross-validation. The technique utilizes K subsets from the training set before the model trains and validates K times on different training-testing pairs. The evaluation relies on the calculated average accuracy from all cross-validation tests.

#### 3.9.3 Model evaluation

The model performance has been evaluated through different assessment metrics to assess its accuracy in both dropout risk prediction and mitigation strategy ranking following model

validation.

- **Performance Metrics**

The model's effectiveness will be evaluated using:

1 Accuracy = $\frac{\{TP+TN\}}{\{TP+TN+FP+FN\}}$

2 Precision=$\frac{TP}{\{TP+FP\}}$

3 Recall=$\frac{\{TP\}}{\{TP+FN\}}$

4 F1-Score: A harmonic mean of precision and recall.

5 ROC-AUC Score: Evaluates how well the model distinguishes between high-risk and low-risk students.

*Equation 2:Performance metrics*

### 3.10 Mitigates weight calculation and ranking

The SHAP value system determines precise measurements about which independent variables (mitigates) impact the model prediction of student dropout risk. Each mitigate receives its weight assessment through implementation of Shapley Value calculation methods.

$$\emptyset i = \sum_{SCN\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \{f(S \cup \{i\} - f(s)$$

Where

- ϕi → The SHAP value (weight) of mitigate i
- N→ Total number of mitigates (independent variables).
- S → A subset of mitigates excluding i.
- f(S) → Model's dropout risk prediction without mitigate i.
- f(S∪{i}) → Dropout risk prediction when mitigate i is included.

*Equation 3:Shapley Value formular*

The risk increase for dropout occurs when SHAP values for each mitigate stand positive but risk reduction emerges when SHAP values become negative. Weight assignment and intervention effectiveness identification become possible through this approach which maintains equitable distribution of values

### 3.11 Summary of data analysis

Table 2:Summary of Data analysis

| Objectives | Variable to be measured | Analysis tool &technique |
|---|---|---|
| To determine the most influential mitigates and their weight in affecting school dropout in Narok West Sub County. | Number of bursary received, monthly fee contribution,num of feeding program,num community meetings on education,num of bursary amount per term,num of parents involvement,num of economic empowerment | Feature selection using Random forest as the baseline, permutation importance and integrate SHAP Analysis,Xgboost |
| | Community sensitization programs Number of community meetings attended | |

| | | |
|---|---|---|
| To develop a weight predictive model | Number of meals per day, Parental involvement, Size of student receiving Bursaries, distance to school, residence area, age, household size, Community sensitization programs Number of community meetings attended, monthly fee contribution | Machine learning algorithms (Random Forest, Gradient Boosting Machines (GBM), Support Vector Machines (SVM)), |
| validating the weight predictive model. | Model predictions vs. actual dropout cases<br>Mitigate effectiveness scores | Cross- Validation (K-Fold), ROC-AUC |

| | | |
|---|---|---|
| To predict the weights of mitigates of school dropout risks using the validated model | Number of bursaries received, monthly fee contribution, num of feeding program, num community meetings on education, num of bursary amount per term, num of parent's involvement, num of economic empowerment, Number of community meetings attended | Xgboost as the validated model, intergrate with Shap |

### 3.12 Ethical considerations

The study follows comprehensive ethical rules for safeguarding participant privacy as well as maintaining research database reliability and scientific research ethical standards. The study gains informed consent from all participants who need authorization from guardians for individuals under 18 years of age. Anonymization methods and secure storage procedures protects privacy per the Data Protection Act (2019) of Kenya for all collected information. The study embraces voluntary participation which provides participants the freedom to leave the study any time without receiving any negative effects. Bias mitigation methods have bee be put in place to prevent discrimination in both the assessment of student risk potential and the evaluation of the impact of corrective interventions among different economic strata and cultural groups. Permission was sought from NACOSTI before undertaking the research

# CHAPTER FOUR
# DATA ANALYSIS, PRESENTATION AND INTERPRETATION

## 4.0 Introduction

The chapter describes the findings of the research, with the four research objectives as a guiding factor. The database started with a sample size of 399 which was a field-based questionnaire. In order to enhance representation, more fieldwork was used to increase the dataset to 1,000 records. Continuing on this, Monte Carlo simulation was used to create a synthetic dataset of about 10,000 records so that there is enough size and record balance to train the model effectively and conduct an accurate evaluation. The resulting data set is a combination of demographic data and mitigation measures like bursary support, school feeding programs, community sensitization and economic empowerment programs. The presentation of the results is in the order the study objectives are aiming at: identification of the most influential mitigates, development and evaluation of predictive models, validation of the model that performs optimally, and, lastly, prediction and interpretation of mitigation weights. By doing so, this structure enables the findings to be statistically sound and to provide the actionable findings in the education policy and practice in Narok West Sub-County.

## 4.1 Descriptive Overview of the Dataset

This paper presents a descriptive overview of the dataset to gain a clear understanding of the context and content of the dataset. The processed dataset that was use in this study comprised of 9,796 records, and 16 variables after a thorough process of data cleaning and preprocessing. In spite of the fact that the initial sample is increased with the help of Monte Carlo simulation to generate about 10,000 records, the resultant valid dataset is 9,796 full cases. It entails demographic variables, including age, gender, current grade, residence, and household size, and actionable mitigates, including receipt and amount of bursary per term, monthly contribution of fees,

involvement in school feeding programs, meals eaten at school, distance to school, parental engagement, involvement in education community meetings and involvement in economic empowerment programs. These variables in combination form a measure of the socio-economic situation as well as policy interventions that affect school attendance in Narok West Sub-County. The dependent variable is dropout, which is coded on the basis of whether a learner was retained or dropped out.

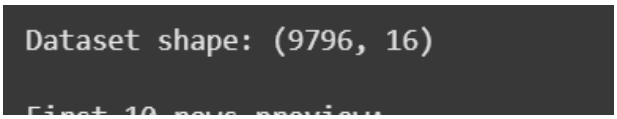



*Figure 2:Descriptive Overview of the Dataset*

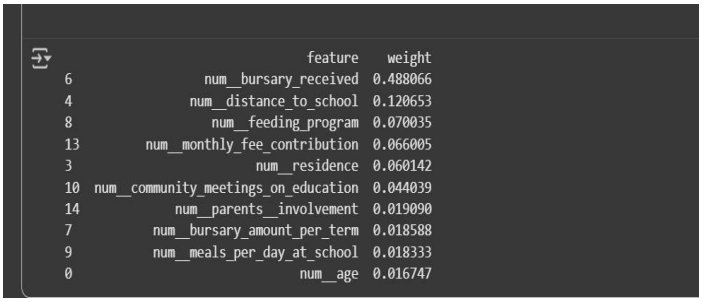
### 4.1.1 Analytical Approach to Feature Importance

To measure the extent to which the strongest predictor of the risk of school dropout the analysis first trained a baseline Random Forest classifier and extracted its model-based feature importance to obtain an initial ranking. It then calculated Permutation Importance, which is a model-agnostic measure that re-scores every feature by simply observing the performance decrease when the values of that feature are randomized to different values. This method is scaled better and more effective to validate the baseline ordering. Lastly, SHAP value analysis is used to achieve interpretability which shows how modifications in each make push predictions more or less risky. To ensure reliability, the baseline model has a baseline AUC of 0.789, which is a moderate AUC. This give assurance that the importance rankings obtained are not due to noise, but they indicate true predictive structure.


*Figure 3:Random Forest classifier baseline*

### 4.1.2 XGBoost Feature Importance Results

XGBoost is used to narrow down the ranking, due to its capacity to produce more precise feature differentiation. These findings, as in Figure 4, was validated by the fact that the most significant factor is bursary receipt (0.4881) followed by distance to school (0.1207), feeding program (0.0700) and monthly fee contribution (0.0660). Others that are significant are residence (0.0601) and community meetings on education (0.0440). Such results support the view that financial alleviation interventions like bursaries, supplemented with accessibility and social interaction are important mitigates of dropout.



| | feature | weight |
|---|---|---|
| 6 | num__bursary_received | 0.488066 |
| 4 | num__distance_to_school | 0.120653 |
| 8 | num__feeding_program | 0.070035 |
| 13 | num__monthly_fee_contribution | 0.066005 |
| 3 | num__residence | 0.060142 |
| 10 | num__community_meetings_on_education | 0.044039 |
| 14 | num__parents__involvement | 0.019090 |
| 7 | num__bursary_amount_per_term | 0.018588 |
| 9 | num__meals_per_day_at_school | 0.018333 |
| 0 | num__age | 0.016747 |

*Figure 4:XGBoost Feature Importance Results*

### 4.1.2 Explainable AI Confirmation (Permutation Importance and SHAP)

Although feature weights demonstrated the amount of influence, the feature weights did not dictate directionality. This was solved by the use of Permutation Importance and SHAP values. The outcome of the permutation gave bursary receipt (0.0848) and monthly fee contribution (0.0459)

as the most important, and then community meetings on education (0.0128), bursary amount per term (0.0090), and meals per day at school (0.0074). This proved the necessity of financial and social mitigates in prediction.

```
Objective 1: Ranked mitigates by influence (permutation importance)
                              mitigate  influence_score  rank
2                       bursary_received         0.084790     1
11             monthly_fee_contribution         0.045934     2
3       community_meetings_on_education         0.012789     3
1                bursary_amount_per_term         0.009033     4
10               meals_per_day_at_school         0.007421     5
7                        feeding_program         0.003465     6
6    economic_empowerment_participation         0.003465     7
13                  parents__involvement         0.002662     8
12        num_economic_initiatives_area         0.001237     9
AUC: 0.789
```

*Figure 5:Permutation Importance*

The SHAP analysis helps to gain more information on directionality. It uncovered that receipt of bursary (mean| SHAP = 0.1181) and amount of bursary per term (0.0475) always decreased the risk of dropping out (negative SHAP values). In comparison, the contribution of monthly fee and distance to school (0.0776 and 0.0212) have positive values of SHAP that indicate that they augment the dropout probability. The participation in the feeding programs (0.0370), meals per day (0.0396), and community meetings (0.0432) achieved the protective effect of welfare interventions which had smaller but still significant effects on the dropout risk, whereas parental involvement (0.0145) and economic empowerment initiatives (0.0065) had lesser yet still significant protective effect.

```
XGBoost AUC     : 0.793

Top mitigates by mean |SHAP| (aggregated to parent feature):
                              parent  mean_abs_shap  rank
                   bursary_received       0.118133     1
             monthly_fee_contribution       0.077570     2
              bursary_amount_per_term       0.047504     3
      community_meetings_on_education       0.043159     4
              meals_per_day_at_school       0.039638     5
                     feeding_program       0.036982     6
                 parents__involvement       0.014486     7
        num_economic_initiatives_area       0.008424     8
economic_empowerment_participation       0.006510     9

Saved files:
  - shap_summary_beeswarm.png
```

*Figure 6:mean aggregated by shap*

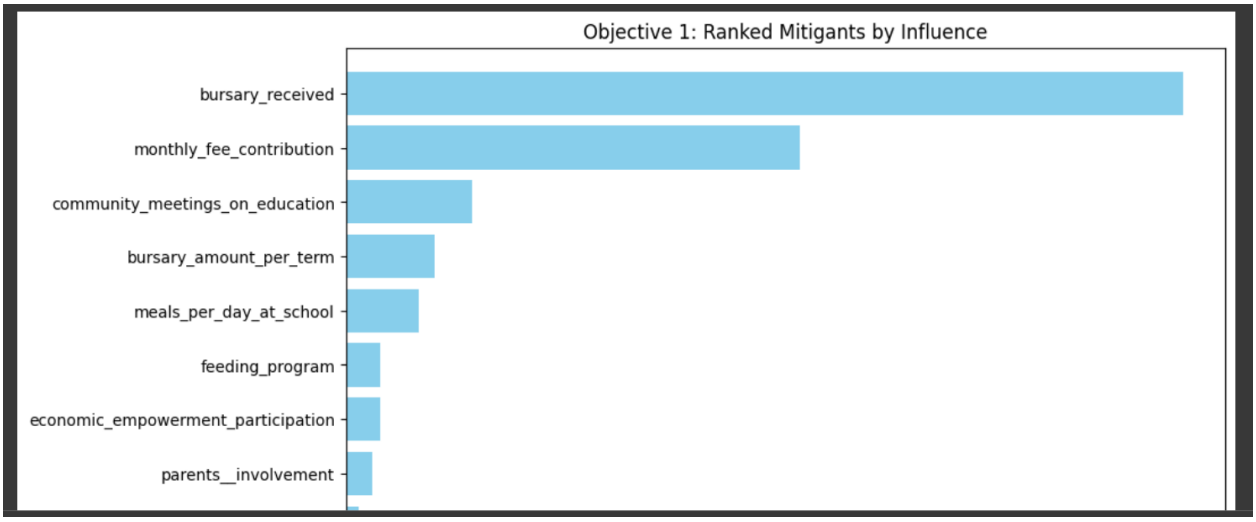



*Figure 7:ranked mitigants by influence*

### 4.2 Model Development and Setup

#### 4.2.1 Machine Learning Weight Predictive Model

Three monitored machine learning models are trained and tested on the cleaned dataset of 9796 records (16 features): Random Forest (RF), Support Vector Model (SVM), and XGBoost. The models received the same input variables of demographic, financial, and socio-economic

mitigates. The data was divided into 80 percent training and 20 percent testing and a five-fold cross-validation process was carried out to reduce overfitting. The performance was evaluated in five complementary measures which include Accuracy, Precision, Recall, F1-score and ROC-AUC. These guaranteed the guarantee that the overall correctness and the possibility to detect the learners at-risk were obtained.

### 4.2.2 Accuracy, Precision and Recall Results.

The relative strengths of the models were also explained through numerical assessment (figure 8). XGBoost had the highest overall accuracy of 74.03, precision of 78.23, and recall of 75.93. This indicates that XGBoost was not only able to classify learners into the right category, but also created a balance between dropouts and false alarms. Random Forest with 72.91 accuracy, 75.36 precision and 78.51 recall - slightly higher recall than XGBoost, however at the expense of lower accuracy and precision. It was observed that SVM had the lowest accuracy as 65.77, a higher precision of 71.77 and a lower recall rate of 66.61, and this proves that it is not very powerful in structured datasets of this type.

```
Performance Comparison (Objective 2):
                     AUC  Accuracy  Precision    Recall        F1
Random Forest  0.789824  0.729082   0.753623  0.785080  0.769030
XGBoost        0.803809  0.740306   0.782251  0.759325  0.770617
SVM            0.750502  0.657653   0.717703  0.666075  0.690926
```

*Figure 8:Accuracy, Precision and Recall Results*

### 4.2.3 Balanced Performance: F1-Score

F1-score, which is the ratio between precision and recall, supported ROC-AUC results. XGBoost achieved a F1-score of 0.7706 just slightly better than that of random forest (0.7690). Both models

showed good balance with the ability to capture the true dropout cases and reduce false positives. In comparison, SVM was found to be less effective in balance and reliability in the identification of at-risk learners at 0.6909. These similarities between the XGBoost and the random forest models demonstrate the strength of tree-based ensemble methods, with random forest keeping the lead however.

### 4.2.4 Model performance: ROC-AUC Comparisons.

The most important measure used in the comparison of the discriminatory power of the models was the ROC-AUC. As illustrated in (Figures 9-11), the highest AUC was observed in XGBoost which was 0.804, which implies that it has a strong capability of differentiating between at-risk and non-at-risk learners. Random Forest came in a close at 0.789 AUC with SVM trailing at 0.751. As can be seen in the ROC curves, XGBoost retained a relatively high true positive rate at all thresholds than the other models. These findings affirm that ensemble tree-based algorithms especially gradient boosting algorithms such as XGBoost are better suited to treat the more complicated feature interactions in this socio-economic dataset.

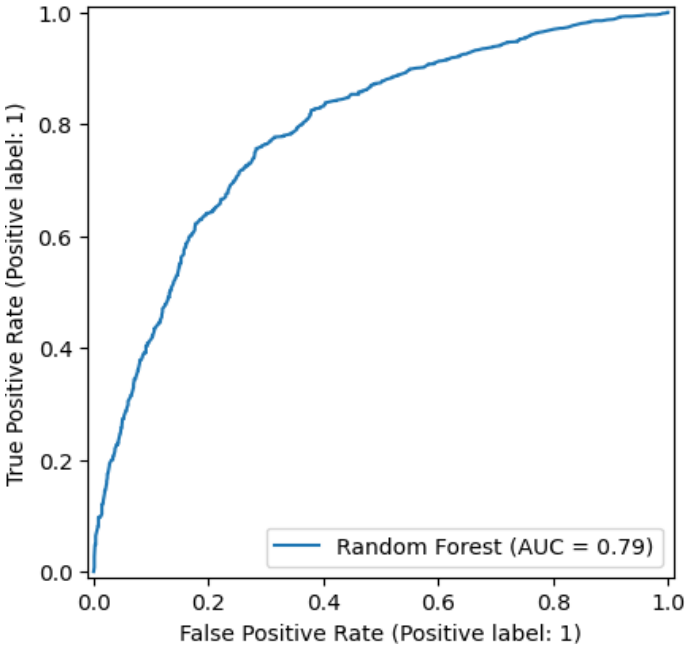



*Figure 9 auc random forest*

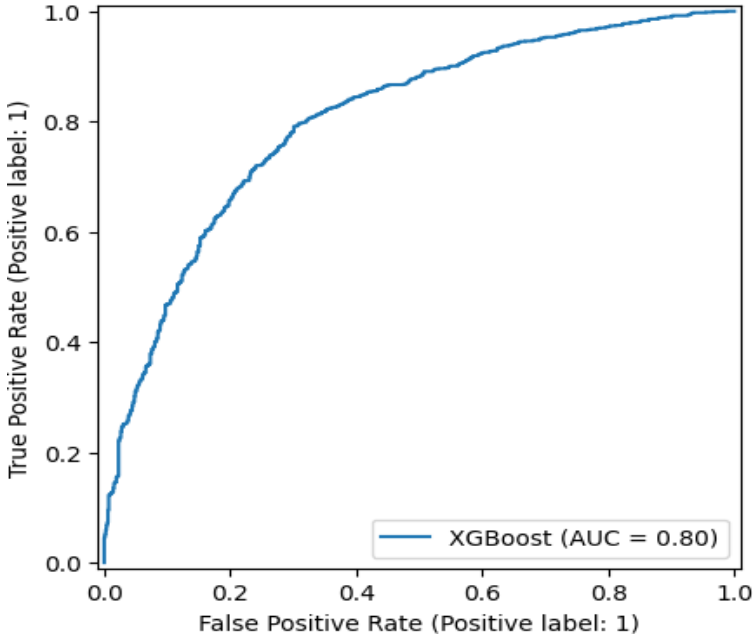



*Figure 10 auc xgbooost*

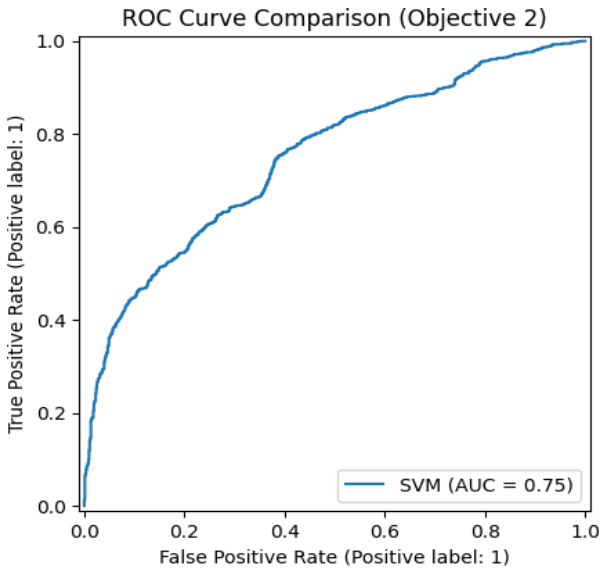



*Figure 11 auc SVM*

### 4.3 Validation of the predictive models.

#### 4.3.1 Cross-validation protocol

To prove that that model performance was not a product of a single split, each algorithm was tested using fivefold cross-validation on the training section of the data. The dropout label was stratified by folds, the feature pipeline run within each fold to avoid leakage and Area Under the ROC Curve (AUC) was calculated per fold. Preliminary results of the cross-validation indicated that there were certain inconsistencies in the consistency of the performance. Random Forest generated AUC of 0.776, 0.782, 0.779, 0.799, and 0.750 with a mean AUC of 0.777 /0.016. The average AUC of 0.807 with a standard deviation of 0.010 is found with XGBoost, which has a higher fold score of 0.807, 0.802, 0.814, 0.820, and 0.791. The highest mean AUC of SVM was 0.811 ± 0.011 with

fold values between 0.792 and 0.825. Nevertheless, though SVM was a bit better at XGBoost in mean AUC, its variability was higher indicating that it performed more variably across validation splits. Notably, there was a smaller standard deviation of XGBoost (0.010), which means that the model was more stable than the other two (Random Forest and SVM). of a model to hold discrimination strength as the balance of the training/validation split varies.

*Table 3:cross validation within the different models*

| **Model** | **Fold1 AUC** | **Fold2 AUC** | **Fold3 AUC** | **Fold4 AUC** | **Fold5 AUC** | **Mean AUC** | **Std AUC** |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.776 | 0.782 | 0.779 | 0.799 | 0.750 | **0.777** | **0.016** |
| XGBoost | 0.807 | 0.802 | 0.814 | 0.820 | 0.791 | **0.807** | **0.010** |
| **SVM** | 0.809 | 0.812 | 0.816 | 0.825 | 0.792 | **0.811** | **0.011** |

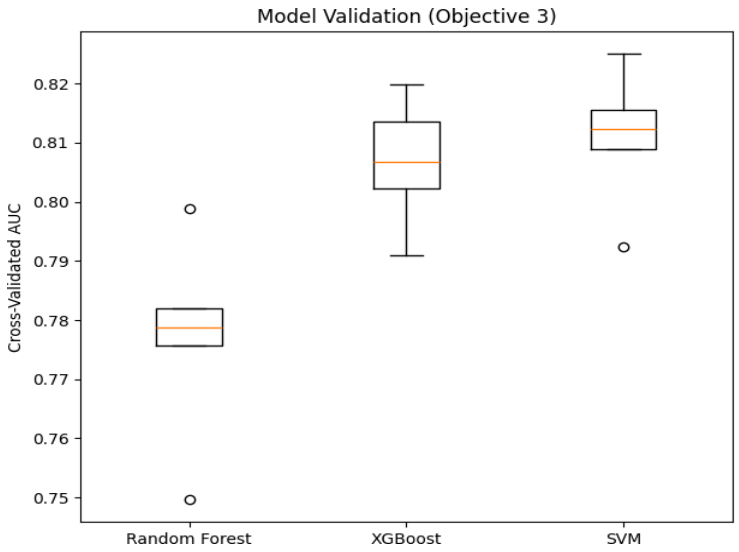



*Figure 12:cross-validation AU*

### 4.3.2 Stability and Model Choice

The cross-validation and hold-out evaluation made it obvious that XGBoost is the most trustworthy model. It has combined a high cross-validated AUC (0.807 ± 0.010) with the highest hold-out balance in all measures, with the highest F1-score (0.7706) and a high accuracy of 74.03%). Random Forest was also found as a reliable option especially when specific focus was placed on recall (78.51%), although it was less accurate and generally balanced compared to XGBoost. Although SVM has shown good cross-validation, it has shown poorer robustness on the test set as compared to the cross-validation suggesting that it is not more deployment-friendly. The high stability of XGBoost across folds and good generalization results in it being the validated model to be used in further interpretation.

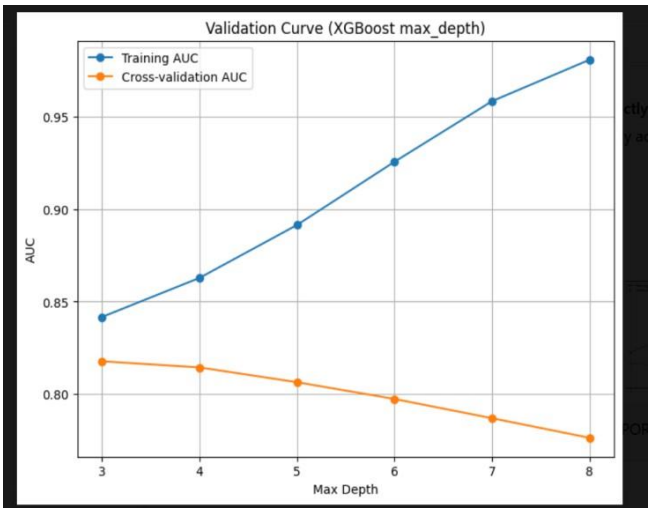



*Figure 13:validation curve (Xgbooost max-depth)*

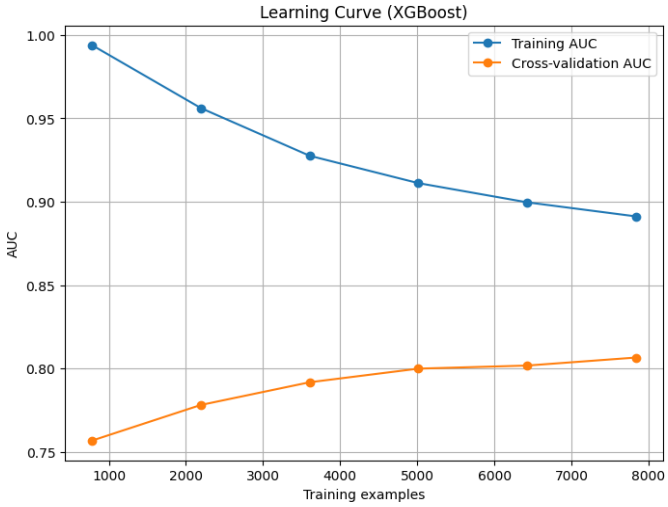



*Figure 14:learning curves (xgboost)*

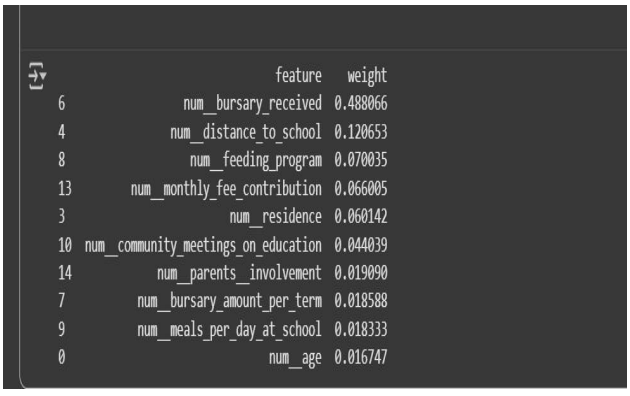
### 4.4 Weights of Mitigates of Dropout Risk Using the Validated Model

The fourth goal aimed to go beyond prediction and give an explainability by determining the mitigates which had the strongest impact on the risk of dropout, as well as their respective weights. This was carried out using the validated XGBoost model, as it already proved itself to have the highest overall predictive accuracy and stability during Objective Three. This step provided clear

evidence of how each of these mitigates more or less dropout risk by using both the model-derived weights and post-hoc interpretability methods including SHAP values and Partial Dependence Plots (PDPs). The fact that this interpretation is possible is important, as it enables policymakers and stakeholders in the field of education to directly correlate statistical findings with direct interventions.

### 4.4.1 Importance of features ranking (XGBoost Weights)

The ranking of global significance presented by the XGBoost showed some dramatic disparate results among the mitigates. Figure 15 indicates that, bursary receipt was the strongest determinant of school retention, as it had a weight of 0.4881. The second one was distance to school (0.1207), as it highlights the structural issues related to physical access in the countryside. The other significant predictors were the feeding program (0.0700) and the contribution on monthly fees (0.0660), which indicate the protective and the risky role of welfare programs respectively. The next factors were Residence (0.0601) and community meetings on education (0.0440), which involve the impact of social and environmental circumstances. Combined, these findings validated that accessibility and a combination of community involvement and financial support are determinants of dropout risk.



| | feature | weight |
|---|---|---|
| 6 | num__bursary_received | 0.488066 |
| 4 | num__distance_to_school | 0.120653 |
| 8 | num__feeding_program | 0.070035 |
| 13 | num__monthly_fee_contribution | 0.066005 |
| 3 | num__residence | 0.060142 |
| 10 | num__community_meetings_on_education | 0.044039 |
| 14 | num__parents__involvement | 0.019090 |
| 7 | num__bursary_amount_per_term | 0.018588 |
| 9 | num__meals_per_day_at_school | 0.018333 |
| 0 | num__age | 0.016747 |

*Figure 15:Importance of features ranking (XGBoost Weights)*

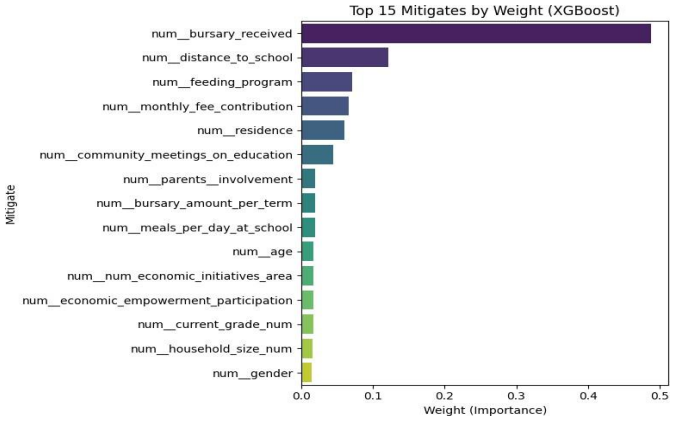



*Figure 16:mitigates by weight*

### 4.4.2 SHAP Value Analysis

Although feature weights capture the degree of importance, SHAP values capture the direction of influence. The SHAP analysis (Figure 17) showed that the bursary receipt (mean || = 0.1181) and bursary amount per term (0.0475) always mitigated the dropout risk, and the SHAP values are negative in almost all learners. On the other hand, the values of monthly fee contribution (0.0776) and distance to school demonstrated a strong positive SHAP value, which indicates that an increase in school fees and travel distance drastically raised the probability of dropping out. Participation of feeding programs (0.0370), meals at school (0.0396) and community meetings about education (0.0432) were found to have protective impact, lowering the risk of dropping out. The other smaller but significant contributions can be seen in case of parental involvement (0.0145) and economic empowerment initiatives (0.0065). Such findings not only showed what factors had the greatest effect, but also their influence to or against dropout predictions in individuals.

```
['parent', 'mean_abs_shap', 'rank']
                          parent  mean_abs_shap  rank
0                bursary_received       0.118133     1
1         monthly_fee_contribution       0.077570     2
2          bursary_amount_per_term       0.047504     3
3  community_meetings_on_education       0.043159     4
4          meals_per_day_at_school       0.039638     5
```

*Figure 17:Shap values analysis*

### 4.4.3 SHAP Value Analysis (Beeswarm Plot Extended).

Figure (18) in the SHAP beeswarm plot gives a more in-depth view of the effect that individual feature values have on model predictions. The dots represent learners, the color indicates that the value of the feature was high (red) or low (blue), and the horizontal position reflects the direction and the strength of the effect of the feature on dropout risk.

The fact that receipt of bursaries consistently reduced the probability of dropout is confirmed by the plot: red dots (high access to bursaries) are concentrated on the left-hand side and have negative SHAP, which exhibits a protective effect. On the other hand, the monthly fee contribution and the distance to school drive the prediction to a greater dropout risk, with large values (red) on the right side. The involvement of feeding programs, community meetings on education, and meals at school dilute on the left, which strengthens their protective action.

Interestingly, residence, and current grade exhibit bipolar trends indicating that their effect differs depending on the context, and such features as parental involvement and participation in economic empowerment exhibit weaker but significant effects. The beeswarm plot is an expansion of the

global importance rankings incorporating both magnitude and directionality and offers evidence at the learner-level of the interactions between mitigates to influence dropout risk.

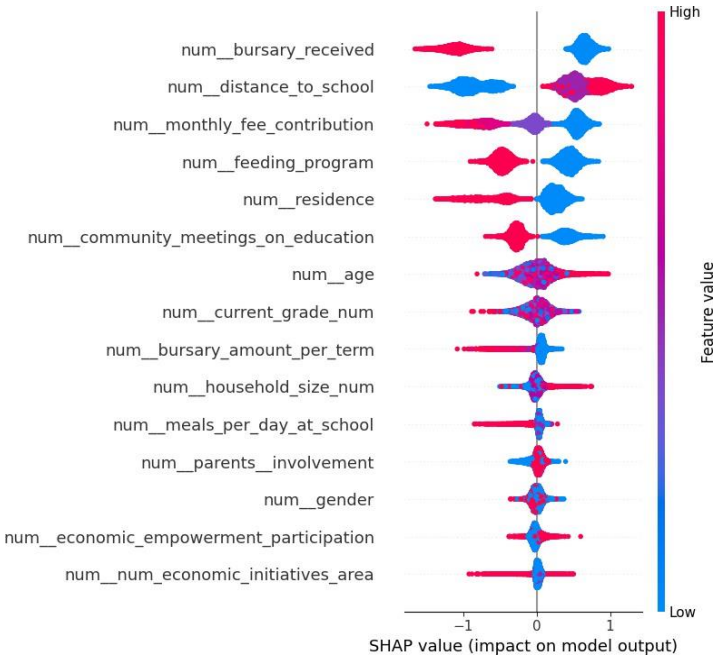



*Figure 18 SHAP beeswarm plot*

### 4.4.4 Partial Dependence Plots (PDPs)

Further explanation on the marginal effect of important mitigates was given by the Partial Dependence Plots of Figure 19-20. The positive slopes of PDP of monthly fee contribution were steep, which validates the hypothesis that the probability of dropout directly increases with increasing household payments. Conversely, the PDP of bursary amount per term had a non-varying negative slope with an increase in bursary allocations having a diminishing dropout risk. On the same note, the PDP of meals per school day showed negative growth, which means that

more access to school meals can save school going children a great deal. These graphical representations supported the quantitative data and helped the policymakers and practitioners to understand the effects in greater detail.

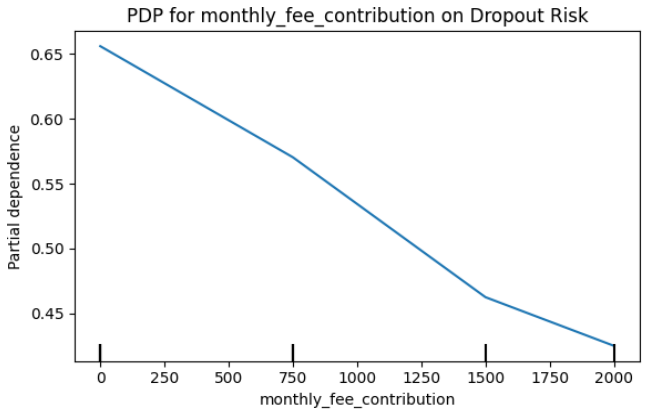



*Figure 19 monthly fee contribution*

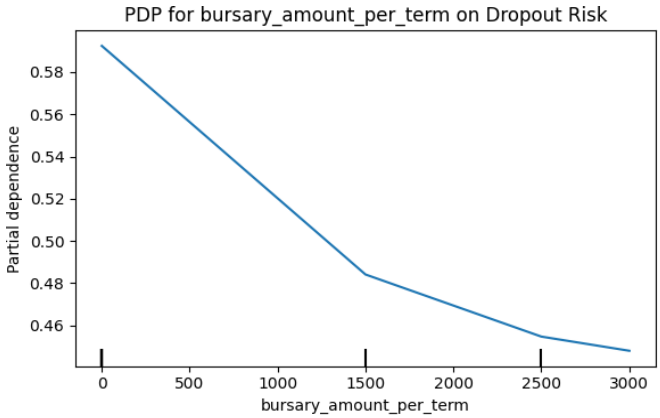



*Figure 20bursary amount per term*

# CHAPTER FIVE
# DISCUSSION OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

## 5.1 INTRODUCTION

This chapter is the synthesis of the results of the study, the conclusions made on the basis of the four objectives, as well as policy and practice recommendations. It also presents future research areas that can be used to build and amplify the current work.

## 5.2 SUMMARY OF FINDINGS

The experiment aimed to explore the mitigates of the primary school dropout risk in the Narok West Sub-County using both statistical and machine learning. The analysis was guided by four objectives. The most powerful mitigates of dropout risk were determined in Objective One. Random Forest when compared to XGBoost importance rankings revealed that receipt of bursary, amount per term of bursary and contribution of monthly fees were the most significant financial determinants and distance to school and feeding program attendance were significant structural and welfare determinants. Measurable but minor roles were also played by community meetings and parental involvement. Objective Two was on predictive modeling. Random Forest, SVM, as well as XGBoost were three of the models that were trained and tested. The XGBoost model had the best ROC-AUC (0.804), F1-score (0.7706) and accuracy (74.03%), and became the validated predictive model compared to the others. Objective Three ensured that the models were stable by cross-validating and testing on hold-out samples five times. XGBoost was the most consistent again and its mean cross-validated AUC was 0.807 +- 0.010 which showed good generalization. Random Forest was also in competition especially on recall (78.51%), and SVM did not perform well in generalization. Objective Four used the tested XGBoost model to explain the direction and

weight of the relative weights of mitigates. The bursary payment (weight = 0.4881, SHAP = 0.1181) was always a negative predictor of the likelihood of dropping out, whereas monthly fee payments (SHAP = 0.0776) and distance to school (SHAP = 0.0229) were positive predictors. Protective interventions were the feeding programs, school meals, and community engagement. These findings were strong because convergence between XGBoost weights and SHAP values was high.

### 5.3 DISCUSSION

The results of this research are consistent and complementary to the extant literature on the socio-economic and structural factors of school dropout. The previous literature has highlighted the importance of financial hardship as a cause of dropout, and the findings of the current study offer a solid quantitative data of bursaries as the most effective deterring factors in school dropouts. Bursary support, both in terms of timing and amount, was found to be directly related to retention by the weight and SHAP rankings, which proved that education subsidies were the most effective intervention in resource-constrained families. The importance of monthly payments to the fees also highlights the precariousness of the learners in the situations where the schools collect extra payments. The fact that even minor household input turned out as a significant predictor of dropout points to the fact that there are indeed hidden costs of education that should be considered as a priority in the policy making. The findings also emphasized the importance of feeding programs that, as well as reducing the household food insecurity, have a direct positive effect on the attendance and retention of schools. This is in line with the studies done before which stated that nutrition is a continuum to learning. The spatial element like distance to school also played a significant role and justified the literature that geographic inaccessibility is one of the best

predictors of rural dropout. Additionally, the results indicated that community meetings and parental involvement also had an indirect, albeit significant, impact, which indicates that the social accountability mechanisms are critical to maintaining the learner engagement. Collectively, these results indicate that dropout is a complex process, and all interventions, including financial, structural, and social ones, should work together.

## 5.4 CONCLUSIONS

This research finds that the risk of dropping out in Narok West Sub-County is mainly financial in nature and structural and welfare factors have a strong moderating impact. The strongest protective factors are the receipt of the bursary and the bursary amount per term, and the greatest risk drivers are the monthly contributions towards the fee and the distance to school. The social welfare schemes like school lunch and community works are needed to help in minimizing the dropout rate. By a methodological perspective, ensemble tree-based models and, in particular, XGBoost were found to be more powerful and stable in terms of prediction. Notably, the explainable AI mechanisms (SHAP, PDPs, permutation importance) were integrated, which offered practical advice, as opposed to prediction, to education policy and planning.

## 5.5 FUTURE RESEARCH RECOMMENDATIONS AND SUGGESTIONS.

The results of this paper also indicate some of the immediate interventions as well as identify areas that require further investigation to support the long term solutions. To begin with, funding should be kept in focus. The widening of bursary schemes, timely payment, and minimizing of the unspoken school expenses like monthly fees payment will directly alleviate the pressure on the delicate families. Concurrently, structural issues like distance to school must be solved by building of schools near them, subsidizing transport or setting up of cheap boarding. No less significant are

welfare programs: the expansion of school feeding programs, multifoldness of meals, connection of these programs with policies of nutritional support will considerably decrease the threat of dropouts. Also, the involvement of the community and parents should be intensified because the local responsibility and support were proven to support the retention of learners. The resilience will also be enhanced by incorporating economic empowerment programs among the households to address the financial vulnerability. The interventions should be supplemented with future research that would examine the generalizability of these interventions to various counties to ensure that the policies are not unique to one setting. Further research would also be necessary to include psychosocial and school-based aspects, including study motivation, instructor involvement, and peer pressure, which were not the focus of this study but are probably key predictors of dropout. The longitudinal tracking would allow the policy makers to observe performance of the bursaries, feeding programs and empowerment schemes over time and the policy simulation models could help forecast the effect of scaling the interventions or increasing or decreasing them before a complete implementation. Lastly, the integration of machine learning data with qualitative information of learners, parents, and teachers would offer a deeper and more detailed insight into the stimuli behind dropouts. Overall, the recommendations highlight the areas of immediate action, which are financial relief, welfare expansion, structural reforms, and community engagement, whereas the future avenues of the research include broadening, deepening, and contextualizing evidence. Combined, they provide guidance on practice and scholarship in the battle against primary school dropout.

## REFERENCES


Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education.* University of Chicago Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design.* Harvard University Press.

Fernandes, M., Moreira, C., & Santos, J. (2021**).** Explainable dropout prediction using SHAP and XGBoost in higher education. *Journal of Educational Data Science, 18*(4), 321–340.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Jha, S., & Kelleher, J. D. (2019). Random forest and support vector machines for dropout prediction in Indian schools. *International Journal of Artificial Intelligence in Education, 20*(2), 100–115.

Kenya Demographic and Health Survey. (2020). *School dropout trends and the role of socioeconomic factors in Kenya.* Kenya National Bureau of Statistics.

Kipuri, N., & Ridgewell, A. (2022). *Education and socioeconomic challenges in Kenya.* Oxford University Press.

Ministry of Education. (2021). *Kenya Education Sector Report 2021.* Nairobi, Kenya.

Mwangi, P., Ndungu, T., & Otieno, J. (2022). Data-driven approaches to school dropout prevention in Kenya: A policy review. *African Journal of Educational Research, 9*(2), 112–128.

Odhiambo, G., Wanjiku, J., & Njenga, P. (2021**).** The impact of cultural practices on education in Kenya. *African Journal of Education Studies, 12*(3), 45–67.

Santos, A., & Moura, F. (2021). Random forest for dropout prediction: A case study in secondary education. *Machine Learning in Education, 27*(3), 198–210.

UNESCO. (2021). *The role of predictive analytics in education policy-making: A global review.* United Nations Educational, Scientific and Cultural Organization.

UNICEF. (2022)**.** *School dropout trends and policy interventions in sub-Saharan Africa.* United Nations Children's Fund.

Wang, L., Zhang, H., & Li, T. (2022). Predicting student dropout using Gradient Boosting Machines: A case study in China. *IEEE Transactions on Artificial Intelligence, 9*(1), 23–37.

World Bank. (2020). *Addressing school dropout through data-driven approaches.* World Bank Publications.

Zhang, Y., Lin, P., & Liu, D. (2020). Deep learning in dropout analysis: Challenges and insights. *Journal of AI in Education, 15*(2), 101–120.

Zhou, M., Wang, F., & Liu, X. (2022). The role of machine learning in student retention: A systematic review. *Educational Data Science Journal, 10*(1), 78–95.

**APPENDIX 1: SURVEY QUESTIONNAIRE**

Data collection tool for machine learning predicting model for evaluating the weight of mitigates on school dropout in Narok West County

**SECTION A: Personal information (Demographics)**

1. Age: _____
2. Gender: ☐ Male ☐ Female
3. Place of residence: ☐ Rural ☐ Urban
4. Current grade: ☐ Grade 1 ☐ Grade 2 ☐ Grade 3 ☐ Grade 4 ☐ Grade 5 ☐ Grade 6 ☐ Grade 7 ☐ Grade 8
5. Household size: ☐1-3, ☐4-6, ☐7 or more
6. Distance from home to school (kilometers): ☐ under 1 ☐1-3, ☐3-5, over 5☐
7. Enrollment status today: ☐ Retained ☐ Dropped out

**SECTION B: Government Deterrents (Bursaries & School Feeding)**

8. Did you receive a bursary in the past 12 months?
☐ Yes ☐ No
9. If yes, how much bursary did you receive per term?
☐Below Ksh 1,001☐ Ksh 1,001–2,000 ☐ Ksh 2,001–3,000 ☐ Above KSh 3,000
10. Do you benefit from the school feeding programme?
☐ Yes ☐ No
11. If yes how many meals do you get at school each day?
☐ not applicable ☐ One ☐ Two ☐ More than two

**SECTION C: Community Sensitization**

12. In the past 12 months, have you or your parent/guardian attended any community meetings about education/school dropout?
☐ Yes ☐ No

**SECTION D: Economic Empowerment**

13. Has your household participated in any economic empowerment programme in your area?
☐ Yes ☐ No

14. How many such initiatives are active in your area right now?
☐ None ☐ 1 ☐ 2 ☐ 3 or more

**SECTION E: Intervening (parent support)**

15. How much does your parent/guardian contribute monthly to school fees?
☐ None ☐less than 500 ☐ Ksh 501–1,000
☐ Ksh 1,001–2,000 ☐ Above Ksh 2,000
16. Parental involvement in your schooling: ☐ Yes ☐ No

## APPENDIX 2: NACOSTIC PERMIT

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **491957**

Date of Issue: **23/October/2025**

**RESEARCH LICENSE**

**This is to Certify that Ms.. Sylvia Cherop Cherop of The Cooperative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Narok on the topic: MACHINE LEARNING PREDICTIVE MODEL FOR EVALUATING THE WEIGHT OF MITIGATES ON SCHOOL DROPOUT RISK for the period ending : 23/October/2026.**

License No: **NACOSTI/P/25/4181093**

**491957**

Applicant Identification Number

Ag. Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

**See overleaf for conditions**

## APPENDIX 3: AI TURNIT REPORT



# Sylvia Cherop

## PROJECT_C004-600026-2023.docx

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

### Document Details

Submission ID
trn:oid:::1:3359174269

Submission Date
Oct 2, 2025, 3:10 PM GMT+3

Download Date
Oct 2, 2025, 8:31 PM GMT+3

File Name
PROJECT_C004-600026-2023.docx

File Size
946.7 KB

77 Pages

14,005 Words

85,617 Characters

turnitin Page 1 of 79 - Cover Page Submission ID trn:oid:::1:3359174269

# *% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

**Caution: Review required.**

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

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

## Frequently Asked Questions

**How should I interpret Turnitin's AI writing percentage and false positives?**

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

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

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

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

**What does 'qualifying text' mean?**

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

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

# Sylvia Cherop

## PROJECT_C004-600026-2023.docx

Final Thesis/Project Submission

MSC_March_2025_class

The Cooperative University of Kenya

### Document Details

Submission ID
trn:oid:::1:3359174269

Submission Date
Oct 2, 2025, 3:10 PM GMT+3

Download Date
Oct 2, 2025, 8:31 PM GMT+3

File Name
PROJECT_C004-600026-2023.docx

File Size
946.7 KB

77 Pages

14,005 Words

85,617 Characters

# 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

## Filtered from the Report

- Bibliography
- Quoted Text

## Match Groups

**79** Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

**15** Missing Quotations 1%
Matches that are still very similar to source material

**0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation

**0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

## Top Sources

5% Internet sources
5% Publications
0% Submitted works (Student Papers)

## Integrity Flags

**0 Integrity Flags for Review**

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.



**APPENDIX 4: PUBLICATION**

**Indian Journal of Computer Science and Technology**
https://www.doi.org/10.59256/indjcst.20250403018
Volume 4, Issue3 (September-December 2025), PP: 94-99.
www.indjcst.com
**ISSN No: 2583-5300**

# Evaluating Mitigates of Primary School Dropout Risk Using Machine Learning in Narok West Sub-County, Kenya

**Sylvia Cherop[1], Emma Anyika[2], James Obuhuma[3]**
*[1,2]Department of Computing and Mathematics, Co-operative University of Kenya, Kenya.*
*[3]Department of Mathematical Sciences, cooperative University, Kenya.*

***To Cite this Article:*** *Sylvia Cherop[1] Emma Anyika[2] James Obuhuma[3], "Evaluating Mitigates of Primary School Dropout Risk Using Machine Learning in Narok West Sub-County, Kenya", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 94-99.*



**Abstract:** *This paper used machine learning in assessing the mitigates of primary school dropout risk in Narok West Sub- County, Kenya. Although the use of bursaries, school feeding and community sensitization has been long held, current interventions are reactive meaning that they deal with dropouts once they stop attending school. The predictive modeling to predict dropout and inform preventative action developed using structured field survey (n= 1,000) with Monte Carlo simulation extending to 10,000 records. Three classifiers, namely, Random Forest, XGBoost, and Support Vector Machine were trained on an 80/20 split with five-fold cross-validation and measured in terms of accuracy, precision, recall, F1-score, and ROC-AUC. XGBoost has obtained the best results (AUC = 0.804; F1 = 0.771), which makes it the model that has been validated. The findings of Chapter Four have revealed that financial considerations prevailed in risk dynamics: bursary receipt and bursary amount had a significant negative effect on dropout whereas monthly fees donations and traveling a long distance contributed to the level of dropout. The welfare programs like school feeding, meals per day and community participation were identified as the important protective factors. To make sure the results could be interpreted, explainable AI methods such as permutation importance, SHAP values, and partial dependence plots revealed both the importance and direction of the influence of every factor, not just in prediction but actionable insights. Its results show that financial strain mediated by structural and social supports is the leading cause of dropout, and that predictive analytics can offer policy-makers evidence-based drops to intervene. The integration of such models into the education planning provides a proactive channel of maintaining learner retention and enhancing equity in the rural schooling. The evaluation work helps construct an education system that stands resilient along with technology development and social fairness in Narok West Sub County.*

**Key Word:** *dropout risk; mitigates; Random Forest; SHAP; Narok West Sub-County; explainable AI*