AN ENSEMBLE MODEL FOR PREDICTING LEARNER ATTRITION IN ONLINE LEARNING

STANLEY MUNGA NGIGI

A thesis submitted in partial fulfillment for award of the degree of Doctor of Philosophy in Information Technology in the school of Pure and Applied Sciences, Kirinyaga University

JULY 2026

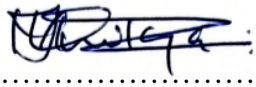
DECLARATION

This thesis is my original work and has not been presented for a degree or any other award in any other Institution.

STANLEY MUNGA NGIGI **PA303/S/21189/23**

Signature:……  ……**Date:**……..**6/7/2026**……

This thesis has been submitted for examination with our approval as University Supervisors.

Signature:…………**Date:** ……**7/7/26**……

Dr. James Mwikya, PhD

Senior Lecturer of Information Technology

School of Pure and Applied Sciences

Kirinyaga University

Signature:  ……. **Date:** ……**7/7/26**……...

Dr. Victor Mageto, PhD

Lecturer of Information Technology

School of Pure and Applied Sciences

Kirinyaga University

DEDICATION

I dedicate this research project to my wife, Mercy Njoki, and our children, Sifa Wangui, Ushindi Ngigi, and Kibali Kamau, for their love and support.

ACKNOWLEDGEMENT

I am thankful to God for granting me the strength, courage, and hope to begin and successfully complete this task. I would like to express my sincere gratitude to my supervisors, Dr. James Mwikya and Dr. Victor Mageto, for their expert guidance and support throughout my research.

I am also grateful to Kirinyaga University for providing me with the opportunity to pursue this study. My heartfelt appreciation goes to my classmates for their unwavering moral support and encouragement, which kept me motivated.

Special thanks to my friends and mentor, Dr. Ephantus Mwangi, Frankline Abuga, and Richard Guto i deeply appreciate your support.

Finally, I extend my deepest gratitude to my parents, Gibson Ngigi and Cecilia Ngigi, for instilling strong values in me and for their immense sacrifices to support my education. To my beloved wife, Mercy, and our children, Sifa, Ushindi, and Kibali Munga, thank you for your patience, prayers, and understanding throughout this journey.

ABSTRACT

Predicting learner attrition in online learning is essential for improving student retention and enhancing the overall effectiveness of educational programmes. This study aimed to develop a robust ensemble-based machine learning model to predict early learner attrition in online learning environments in Kenya. Specifically, the study sought to analyze existing literature, identify key attrition factors, develop an ensemble model integrating XGBoost, Neural Networks, Decision Trees, and Random Forests, and evaluate its performance. Secondary data was obtained from the Learning Management Systems of seven Kenyan universities: Mount Kenya, Daystar, Zetech, KCA, Dedan Kimathi, Strathmore, and Africa Nazarene. The dataset comprised learner demographic information, academic performance records, and enrollment data. To thoroughly evaluate model generalizability, primary data from 5,000 online learners across multiple institutions was also utilized. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. The results demonstrated that the ensemble model significantly outperformed individual approaches, achieving a primary accuracy of 97.2% (with precision, recall, and F1-score at 0.972). It surpassed Decision Trees (95.87%), Random Forests (91.01%), Gradient Boosting (96.91%), and Neural Networks (96.66%). On the primary validation dataset of 5,000 learners, the model maintained a robust accuracy of 95.0% (precision, recall, and F1-score of 0.949). This study contributes empirical evidence to the Kenyan higher education context, demonstrating that multi-model frameworks handle imbalanced educational data effectively. The study recommends that universities integrate ensemble-based predictive models within their Learning Management Systems to enable real-time identification of at-risk students and implement targeted intervention strategies.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
TABLE OF CONTENTS ..... **vi**
LIST OF TABLES ..... **xi**
LIST OF FIGURES ..... **xii**
ACRONYMS AND ABBREVIATIONS ..... **xiv**
CHAPTER ONE: INTRODUCTION **1**
1.1 Background of the Study 1
1.2 Statement of the Problem 4
1.3 Objectives 6
1.3.1 General Objective 6
1.3.2 Specific Objectives 6
1. 4 Research Questions 7
1.5 Justification of the Study 7
1.6 Scope of the Study 9
1.7 Assumptions of the Study 10
1.8 Limitations of the Study 10
1.9 Structure of Thesis 12
CHAPTER TWO: LITERATURE REVIEW **14**
2.1 Introduction 14
2.2 Machine Learning (ML) 14
2.2.1 Introduction to Machine Learning 15
2.2.2 Machine Learning and Data Mining 16
2.2.3 Theoretical Foundations of Machine Learning 17
2.2.4 Machine Learning in Predicting Learner Attrition 18
2.2.5 Machine Learning Applications 19
2.2.5.1 Applications of Machine Learning in Predicting Learner Attrition 19
2.2.5.2 Applications of Machine Learning Across Different Industries 19
2.2.6 Types of Machine Learning 20
2.2.6.1 Supervised Learning 21

2.2.6.2 Unsupervised Learning....21
2.2.6.3 Semi-Supervised Learning....21
2.2.6.4 Reinforced Learning....22
2.2.7 Classification Techniques....22
2.2.7.1 Decision Tree....24
2.2.7.2 Ensemble Methods....26
2.2.7.3 Boosting Algorithms....28
2.2.7.4 Linear Models....29
2.2.7.5 Support Vector Machines....32
2.2.7.6 Neural Networks....35
2.2.7.7 Other Algorithms....36
2.2.7.8 Critical Evaluation of Existing Machine Learning Models for Learner Attrition Prediction....37
2.2.8 Evaluation Metrics for Classification Models....39
2.2.8.1 Classification Accuracy....39
2.2.8.2 Confusion Matrix....40
2.2.9 Evaluation Techniques....42
2.2.9.1 Hold-out Method....43
2.2.9.2 Cross-Validation....43
2.2.9.3 Leave-One-Out (LOO)....44
2.2.9.4 K-Fold Cross-Validation....44
2.2.9.5 Stratified Cross-Validation....44
2.2.9.6 Leave-P-Out (LPO)....45
2.3 Gap Analysis....45
2.4 Student Attrition Prediction: Related Work....47
2.5 Theoretical Framework....53
2.5.1 Identifying Factors Influencing Student Attrition....54
2.6 The important factors that influence learner's attrition rates in online learning...60
2.7 Conceptual Framework for Predicting Learner Attrition....62
2.8 Operationalization of Variables....63
2.9 Conclusion....71
CHAPTER THREE: METHODOLOGY....74
3.1 Introduction....74
3.2 Research Philosophy....74
3.3 Research Design....75
3.4 Location of Study....77
3.5 Target Population....77

3.6 Sampling Techniques 78
3.7 Sampling Size 79
3.8 Instruments of Data Collection 81
3.9 Validity and Reliability of Research Instruments 83
3.9.1 Validity of Research Instrument 83
3.9.2 Validity of the Study 85
3.9.3 Reliability of Research Instrument 85
3.10 Data Sources 87
3.10.1 Data Collection Procedure 88
3.11 Machine Learning Methodology for Developing the Prediction Model 89
3.11.1 Data Preparation 91
3.11.2 Data Pre-processing 92
3.11.3 Merging Basic, Attendance, and Grade Information 93
3.11.4 Pre-processing Data 94
3.11.5 Data Transformation 94
3.12 Model Selection 95
3.13 Model Evaluation 96
3.13.1 Receiver Operating Characteristic (ROC) Curve 97
3.13.2 Accuracy 97
3.13.3 F 1 Measure 98
3.13.4 Mean Squared Error (MSE) 98
3.13.5 Root Mean Squared Error (RMSE) 99
3.13.6 Akaike Information Criterion (AIC) 99
3.13.7 Specificity 99
3.13.8 Sensitivity 100
3.13.9 Positive predictive value 100
3.14 Ethical Considerations 100
CHAPTER 4: RESULTS AND DISCUSSION 102
4.1 Introduction 102
4.2 Data Analysis 103
4.2.1 Dataset Overview 104
4.2.2 Pre-processing the Learner Attrition Dataset 104
4.2.3 Digitization of the Data 105
4.3 Exploratory Data Analysis (EDA) 110
4.3.1 Descriptive Statistics of Key Variables 111
4.3.2 Results from Focus Group Discussions 112

4.3.3 Insights from Exploratory Data Analysis of Learner Interaction Data 113
4.3.4 Age Distribution and Attrition Patterns....114
4.3.5 Employment Status and Learning Persistence 116
4.3.6 Login Behavior and Course Interaction 117
4.3.8 Academic Performance Patterns....121
4.4 Feature Engineering....123
4.4.1 Extraction of Temporal Features 126
4.4.2 Encoding of Categorical Variables....127
4.4.3 Creation of Interaction Features 128
4.4.4 Regional Representation....129
4.4.5 Selecting Features and Optimizing Dimensionality 130
4.4.6 Analysis of Feature Correlations Using Heatmap 134
4.4.7 Data Partitioning and Preprocessing Prior to Model Training 137
4.4.8 Data Partitioning Strategy 139
4.4.9 Feature Standardization 140
4.5 Independent Model Performance....146
4.5.1 Decision Tree Model 148
4.5.6 Salient Feature Analysis for Ensemble Model 161
4.5.6.1 The Final Ensemble Model 162
4.5.7 Ensemble Model Performance 168
4.6 Findings 177
4.6.1 Evaluation of Model Performance....177
4.6.2 Analysis of Feature Significance....180
4.7 Validation Methodology....182
4.8 Primary Data Collection and Cleaning 184
4.9 Machine Learning Pipeline for Validation 186
4.11 Chapter Summary 194
CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS....196
5.1 Introduction....196
5.2 Summary of Research Findings....197
5.3 Summary of Key Findings....199
5.4 Contributions of the Study....201
5.5 Recommendations....204
5.6 Challenges and Limitations 205
5.7 Suggestions for Future Work....207
References....208

APPENDICES.....216
Appendix I: Learner Attrition in Online Learning Questionnaire.....216
Appendix II: Sample External Validation Questionnaire for Ensemble Model.....222
Appendix III: Focus Group Discussion Guide for Key Informants.....225
Appendix IV: Letter of Approval.....228
Appendix V: Research Permit.....229
Appendix VI: List of Publications from This Study.....230
Appendix VII: Sample Source Code for Student Performance Prediction Model...231

LIST OF TABLES

Table 2.1: Summary of Studies on Student Attrition Prediction 51
Table 2.2: Operationalization of Variables 63
Table 2.3: Review of Factors that Influence Learners' Attrition Rates in Online Learning 67
Table 3.1: Target Population 81
Table 3.2: Validity of Research Instrument 84
Table 3.3: Reliability Statistics 87
Table 4.1: Attribute Description 107
Table 4.2: Descriptive Statistics of Key Variables 112
Table 4.3: ANOVA Results for Selected Variables Across Attrition Status 144
Table 4.4: Base Models Performance 158
Table 4.5: Base Models Performance and Ensemble Weights 166

LIST OF FIGURES

Figure 2.1: Workflow of a supervised machine learning algorithm 23
Figure 2.2: An example of a simple decision tree 24
Figure 2.3: A Random Forest Ensemble 26
Figure 2.4 : An Example of an AdaBoost ensemble 29
Figure 2.5: An example of a Simple Linear Regression 30
Figure 2.6 : Support Vector Machine Hyperplane for Classification 34
Figure 2.7: Neural network architecture 35
Figure 2.8 : Confusion Matrix 40
Figure 2.9: ROC Curve 42
Figure 2.10: Conceptual framework illustrating student dropout in higher education 57
Figure 2.11: Conceptual Framework for Predicting Learner Attrition 63
Figure 3.1: The Research Pyramid 75
Figure 3.2 : Machine learning pipeline 91
Figure 3.3 : Data Lake (Raw Data) 92
Figure 4.1: Sample of learner data 108
Figure 4.2: Attrition Status Grouping 109
Figure 4.3 : Stacked Bar Chart of Age Distribution by Attrition Status 116
Figure 4.4: Hierarchical View of Employment Status and Attrition (Sunburst Chart) 117
Figure 4.5: Forum Participation and Community Engagement (Grouped Histogram) 119
Figure 4.6 : Forum Participation and Community Engagement 120
Figure 4.7: Learning Adaptability Impact (Box Plot) 123
Figure 4.8: Data Acquisition and Feature Generation Workflow 124
Figure 4.9: Temporal Feature Extraction 127
Figure 4.10: Categorical Variable Encoding 128
Figure 4.11: Feature Selection Using Random Feature Importance 132
Figure 4.12: Dimensionality Optimization 134
Figure 4.13: Correlation Heatmap Analysis 137
Figure 4.14: Feature–Target Separation and Encoding of the Attrition Status Variable 138
Figure 4.15: Data Splitting and Pre-processing for Model Training 140

Figure 4.16: Feature Standardization 142
Figure 4.17: Decision Tree Model 150
Figure 4.18: Random Forest Model 153
Figure 4.19: Gradient Boosting Algorithm Flowchart 154
Figure 4.20: Gradient Boosting Model 155
Figure 4.21: Architecture of the Feedforward Neural Network (FNN) 157
Figure 4.22: Neural Network Model 158
Figure 4.23: Architecture of the Ensemble Model Combining Decision Tree and Neural Network through Weighted Soft Voting 164
Figure 4.24: Ensembled Model Using VotingClassifier 165
Figure 4.25: Ensemble Model Performance Metrics 170
Figure 4.26: Confusion Matrix Analysis 175
Figure 4.27: ROC-AUC Analysis 177
Figure 4.28: Feature Importance Analysis 180
Figure 4.29: Model Performance: Test vs Primary Validation 192
Figure 4.30: Primary Data Validation Confusion Matrix 193

ACRONYMS AND ABBREVIATIONS

ANN: Artificial Neural Network

CRISP-DM: Cross-Industry Standard Process for Data Mining

CSV: Comma-Separated Values

DT: Decision Tree

F1: F1-Score

GBM: Gradient Boosting Machine

LMS: Learning Management System

ML: Machine Learning

RF: Random Forest

ROC-AUC: Receiver Operating Characteristic – Area Under Curve

XGBoost: Extreme Gradient Boosting

AI: Artificial Intelligence

EDA: Exploratory Data Analysis

ETL: Extract, Transform, Load

LR: Logistic Regression

MAE: Mean Absolute Error

MSE: Mean Squared Error

OPERATIONAL DEFINITION OF TERMS

Attrition: The loss of students from an educational program, particularly those who leave before completing their course of study. In this study, attrition is identified based on a student's absence from enrolled courses and failure to graduate within a defined timeframe.

Dropout: Describes a student who leaves their studies before completing the program. In this study, a dropout is defined as a student who has not registered for any courses for at least one year after their last enrolment period.

Engagement Metrics: These are measurable indicators that reflect how actively a student participates in their learning activities. They include attendance records, participation in online discussions, assignment submissions, and other forms of course interaction that show a learner's involvement.

Machine Learning (ML): A branch of Artificial Intelligence (AI) that enables computers to identify patterns and make predictions based on data. In this study, ML algorithms are used to forecast the likelihood of student dropout by analysing various learner characteristics and behaviours.

Artificial Intelligence (AI): The capability of computer systems to perform tasks that typically require human intelligence, such as learning, reasoning, and decision-making. AI supports this research by enabling predictive analysis through intelligent data-driven models.

Predictive Modelling: The use of data analysis and machine learning techniques to predict future outcomes from historical patterns. This study applies predictive

modelling to identify students who are likely to withdraw early from online learning programs.

Ensemble Model: A type of machine learning approach that combines the strengths of several algorithms to improve prediction accuracy. In this study, the ensemble model brings together Gradient Boosting, Random Forests, Neural Networks, and Decision Trees to enhance prediction reliability.

Random Forest (RF): A machine learning method that builds multiple decision trees and merges their results to produce a more accurate and stable prediction. The technique is used in this study to forecast student attrition.

Neural Network (NN): A computational model inspired by the human brain, made up of layers of interconnected nodes that process information. Neural Networks are applied in this research to identify complex patterns influencing whether a student completes or leaves an online program.

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

Online learning has fundamentally transformed the educational landscape by offering flexible, accessible learning opportunities for diverse populations worldwide. Online learning has grown in popularity in recent years due to technological developments as well as the growing desire for instruction outside of the conventional classroom. This change was accelerated by the COVID-19 epidemic, with global school closures affecting over 1.2 billion learners in 186 countries by mid-2020 (UNESCO, 2021). This disruption forced educational institutions to swiftly transition to online platforms in order to preserve learning continuity. However, the shift also exposed a range of challenges, including ensuring effective student engagement and preventing high attrition rates.

The problem of student attrition has emerged as a significant obstacle to online learning, despite its flexibility and potential. Research indicates that attrition in online courses is ten to twenty percent more than in-person class attrition, indicating that online course dropout rates are substantially higher than in traditional face-to-face settings (Hachey et al., 2023). These high attrition rates raise questions about the effectiveness of online learning and its potential to yield learning outcomes that last. According to the National Centre for Education Statistics (NCES, 2020), despite their convenience, online learning environments usually lack the structured support and one-on-one interaction that are critical for student retention, which causes students to feel isolated and disengaged. It is crucial to comprehend the underlying reasons

behind these dropout trends in order to enhance the caliber of online learning and guarantee its sustainability in the long run.

In Africa, the adoption of online learning has been uneven due to disparities in infrastructure, internet connectivity, and digital literacy. While internet penetration across the continent reached approximately 39.3% by 2021, this figure remains far below global averages, revealing a substantial digital divide (Internet World Stats, 2021). This gap presents significant barriers to the widespread adoption of online learning, particularly in rural and underserved areas. For instance, only 28% of households in sub-Saharan Africa have reliable internet access, which limits the opportunities for students in these regions to use online learning (Selmi, 2023). Additionally, the lack of digital literacy and inadequate training for educators further hampers the successful integration of online learning, contributing to the high attrition rates observed across the continent.

Kenya, as one of the leading nations in Africa's digital transformation efforts, has made significant Strides in incorporating online learning into its higher education sector. Information and communication technology (ICT) integration into education is emphasized by the government's Vision 2030 development goal to increase access and improve learning results (Kenya Vision 2030). But the quick switch to online learning, especially during the COVID-19 epidemic, has brought to light significant issues with internet connectivity, digital device access, and infrastructure. Studies during the COVID-19 pandemic indicate that many Kenyan university students faced limited access to reliable internet for online learning. The digital divide, coupled with disparities in resources between private and public institutions, has contributed to

unequal educational opportunities and increased dropout rates (UNESCO, 2021; World Bank, 2021; Communications Authority of Kenya, 2023).

Predicting and addressing online learning attrition has therefore become a key focus of educational institutions, with machine learning techniques offering potential solutions. Despite their widespread use in dropout prediction, traditional models such as Decision Trees (DT) and Logistic Regression (LR) are limited by their inability to recognise nonlinear patterns in student behaviour and interactions (Cardona et al., 2023). Even though more advanced techniques like Support Vector Machines (SVM) and Random Forests (RF) offer greater accuracy, they usually have trouble with unbalanced and overfitting datasets (Boateng et al., 2020). Recent developments in deep learning, specifically Artificial Neural Networks (ANNs) and Gradient Boosting Machines (GBMs), provide good alternatives due to their ability to handle large datasets and predict non-linear relationships. Addressing these technological gaps is crucial for developing more effective prediction models that can reduce attrition and improve online learning outcomes.

As Kenya continues to expand its digital education initiatives, leveraging machine learning techniques to predict and mitigate learner attrition is instrumental in ensuring the long-term success of online learning programs. Robust predictive models help educational policymakers and institutions design targeted interventions, improve learner's engagement, and enhance learning outcomes. Closing technological divides and ensure equitable access to digital learning materials is necessary to encourage a more hospitable and productive online learning environment (Baker & Inventado, 2014; UNESCO, 2021).

1.2 Statement of the Problem

Online learning has grown rapidly over the last decade, offering learners flexibility, broader access, and opportunities for continued education regardless of geographical location or schedule constraints. Ideally, online courses foster engagement, motivation, and successful completion, benefiting both learners and educational institutions. For students, completing coursework leads to academic progress, future opportunities, and a sense of achievement. For institutions, maintaining low attrition rates supports financial sustainability, positive student outcomes, and institutional reputation (Allen & Seaman, 2021).

Despite these potential benefits, learner attrition in online learning environments remains a persistent challenge. Studies report global online learning dropout rates ranging between 40 percent and 80 percent, with attrition in online courses often exceeding that of traditional face-to-face courses by approximately 10-20 percent (Bartlett, 2020; Hachey et al., 2023). In developing regions such as sub-Saharan Africa, where digital infrastructure, technological access, and learner support systems vary significantly, attrition rates can exceed 50 percent (Wekullo et al., 2023). In Kenya, public universities report similar trends, where high attrition levels negatively affect student success while simultaneously undermining institutional efficiency and effective resource utilization (Wachira & Muturi, 2023).

To address this challenge, a growing body of research has explored the use of machine learning and deep learning techniques to predict learner dropout before it occurs. Algorithms such as Decision Trees, Random Forests, Support Vector Machines, Artificial Neural Networks, and Gradient Boosting Machines have been widely applied to identify patterns within learner engagement data, academic performance

records, and interaction logs that may signal an increased risk of attrition (Mustapha & D, 2023; Utomo et al., 2023). These approaches demonstrate the potential of predictive modelling in educational settings; however, several limitations continue to affect their practical effectiveness.

First, many existing studies rely on imbalanced datasets in which the number of students who persist in courses significantly outweighs those who withdraw, leading to biased model performance and reduced predictive reliability (Olayemi et al., 2025) . Second, several predictive models report only moderate levels of accuracy, which may be insufficient for supporting reliable early intervention strategies (Nimy et al., 2023). Third, many models are developed and validated using data from a single institution, which limits their ability to generalise across different learning environments with varying learner characteristics, technological infrastructure, and pedagogical approaches (Almalawi et al., 2024). Additionally, some advanced models, particularly neural networks and gradient boosting techniques, often lack interpretability, making it difficult for educators and administrators to understand the underlying factors contributing to learner dropout and to translate predictions into practical intervention strategies (Lowder et al., 2022).

These limitations highlight an important gap in existing research. While numerous machine learning techniques have been applied to attrition prediction, there remains a need for predictive approaches that improve not only model accuracy but also robustness, generalisability, and interpretability across diverse online learning contexts. Furthermore, many existing studies rely primarily on individual algorithms applied in isolation, which may limit their ability to capture complex relationships

among learner engagement behaviours, academic performance indicators, and interaction patterns that collectively influence dropout risk(Siddiqui et al., 2025).

To address this gap, this study proposes an ensemble modelling approach that integrates multiple machine learning techniques, including Decision Trees, Random Forests, Artificial Neural Networks, and Gradient Boosting Machines. By combining the complementary strengths of these algorithms, the proposed ensemble model seeks to enhance predictive accuracy while improving model robustness and stability when applied to learner data from online learning environments. The model leverages key indicators such as engagement metrics, academic performance records, and learner interaction patterns to identify students at risk of attrition at an early stage. Ultimately, this research aims to contribute to the development of a more reliable and practical predictive model that can support higher education institutions in implementing timely interventions to reduce learner dropout, enhance student success, and improve overall institutional outcomes in online learning environments.

1.3 Objectives

1.3.1 General Objective

The goal of the study is to develop an ensemble model for predicting learner attrition rates in online learning.

1.3.2 Specific Objectives

The specific objectives of the study are:

i. To analyze existing studies and identify key factors that influence learner attrition in online learning environments.

ii. To develop an ensemble model for predicting learner attrition in online learning environments, incorporating demographic information, academic performance, and engagement metrics.

iii. To evaluate the predictive performance of the developed ensemble model using metrics such as accuracy, precision, recall, F1-score, and sensitivity.

iv. To validate the developed ensemble model using cross-validation techniques to assess its generalizability in predicting learner attrition in online learning environments.

1. 4 Research Questions

i. What key factors influencing learner attrition in online learning environments are identified in existing empirical studies?

ii. How can an ensemble model be developed to predict learner attrition in online learning?

iii. To what extent do the model's performance metrics (accuracy, precision, recall, and F1-score) demonstrate its effectiveness in reliably predicting early learner attrition and supporting timely intervention to reduce dropout rates in online learning environments?

iv. How can the ensemble model be validated for predicting learner attrition in online learning environments?

1.5 Justification of the Study

The global COVID-19 pandemic has accelerated the expansion of online learning, highlighting the need for a comprehensive understanding of the factors influencing

student performance in online learning environments. Online learning has made greater access to higher education for Kenyans, but major obstacle is the high student turnover rate. In order to increase the sustainability and efficacy of online learning in Kenyan universities, this study is justified since it seeks to identify and address the primary causes of online learning course dropout rates.

Another critical aspect is the growing demand for flexible learning options among students, which has made online learning increasingly relevant. Given the changing needs of the workforce and the continuous advancement of technology, educational institutions must swiftly adapt and offer top-notch online learning programs. By understanding the determinants of student retention and dropout, this research provides actionable recommendations for universities to improve their online learning course delivery and support systems,leading to better educational outcomes and increased student satisfaction.

Furthermore, the study is essential in addressing the digital divide that exists within the Kenyan higher education system. It's possible that students from different socioeconomic backgrounds do not have equal access to digital resources and technology, which could lead to disparate educational opportunities and results.By examining the role of digital literacy, access to technology, and other socio-economic factors in influencing student attrition, this research contributes to developing strategies that promote Fairness and inclusion in virtual education.

The findings have practical implications for policymakers, educators, and educational institutions. The research's data-driven analysis of the challenges faced by students in online learning environments can help inform institutional policies and policy

decisions targeted at reducing dropout rates.Moreover, the development of predictive models for identifying at-risk students enables universities to implement early intervention strategies, thereby improving retention rates and ensuring that students can successfully complete their online learning. In addition to contributing to the body of knowledge on online learning, this study provides useful strategies for raising the overall standard and accessibility of higher education in Kenya.

1.6 Scope of the Study

This study concentrated on predicting learner attrition in online learning environments within Kenyan institutions. Its primary objective is to develop a predictive model capable of recognising at-risk students by analyzing various factors that influence dropout rates. The research specifically targets online undergraduate programs offered by several universities in Kenya.

The study investigated a range of influential factors affecting learner attrition, including socio-economic status, digital literacy, access to technology, and academic engagement patterns. Additionally, institutional factors such as course design, effectiveness of support services, and communication strategies were explored to assess their impact on student retention. These insights helped in creating a model for understanding how external conditions, individual capabilities, and institutional strategies interact in shaping online learning outcomes.

Both primary and secondary sources provided the data used in this study. A few Kenyan universities' Learning Management Systems (LMS) provided secondary data, including demographics, academic records, and engagement metrics. In order to validate the model, primary data was obtained from students currently enrolled in

online programs through surveys and focus groups. Machine learning methods like XGBoost, Neural Networks, Decision Trees (DT), and Random Forests (RF) were used to create the prediction model. To confirm that the model was applicable and dependable, its performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

The study's conclusions give academic institutions and decision-makers useful suggestions for boosting online learning platforms, enhancing student retention tactics, and tackling Kenya's digital learning issues. Contributing to the creation of strong intervention strategies that promote both student achievement and institutional sustainability in online learning environments is the ultimate objective.

1.7 Assumptions of the Study

The study was guided by the following assumptions:

i. Learner engagement and attrition risk can be accurately predicted by the behavioral and academic data gathered.
ii. Machine learning algorithms used in the ensemble model are capable of identifying patterns that are crucial for predicting learner attrition.
iii. To the best of their knowledge and honesty, each respondent would complete the questionnaire.
iv. The data analysis's findings are indicative of Kenya's larger online learning community.

1.8 Limitations of the Study

A number of limitations affected the study's findings and generalisability.First, the prediction model was developed using secondary data drawn from existing online

student behavior records. While this dataset provided a solid foundation for analysis, it did not capture every possible factor that contributes to attrition particularly socioeconomic, psychological, or personal challenges that are often invisible in system logs. To address this gap, the study incorporated primary data from surveys conducted across several Kenyan universities, which offered insights into learners' perceptions and experiences beyond what the digital traces revealed. Although survey data may be subject to self-reporting bias, this risk was mitigated by designing clear, neutral questions and ensuring participant anonymity, which encouraged honest responses.

Secondly, variations in course design, online learning environments, institutional policies, and student support structures across different universities could not be fully standardized. To minimize this limitation, the study selected institutions with comparable online learning infrastructures and considered these contextual differences in the interpretation of results, highlighting them where they might influence learner engagement patterns.

Another limitation stemmed from the assumption that students interacted with the learning platforms individually, thereby overlooking the impact of social influence and collaborative learning dynamics. While it was not feasible to fully capture peer effects in this research, the study acknowledged their potential role in attrition and recommended future investigations to integrate social network analysis and peer-to-peer interaction data.

Last but not least, the model's capacity to accurately represent the complex causes of attrition may have been limited by the ethical and privacy-related omission of sensitive demographic variables including family background and income levels. By

employing proxy factors including enrolment status, academic background, and course participation trends, this issue was largely overcome while still providing useful indicators of student risk.

In summary, while the study faced challenges related to data scope, variation in institutional contexts, assumptions about learner independence, and restricted demographic details, deliberate strategies were employed to minimize their effects. These included triangulating secondary and primary data, carefully designing survey instruments, ensuring contextual awareness when interpreting findings, and using proxy indicators where direct measures were not available. As a result, although the limitations remain, their potential impact on the study's validity was substantially reduced, allowing the findings to still provide a credible and useful contribution to understanding learner attrition in online learning.

1.9 Structure of Thesis

The remainder of this thesis is organized into five chapters, each focusing on a key component of the study. Chapter Two presents a review of relevant literature on learner attrition in online learning environments, with emphasis on machine learning techniques, predictive modeling approaches, and existing conceptual frameworks.

Chapter Three describes the research methodology adopted in this study. It outlines the procedures followed in data collection, preprocessing, and feature selection, as well as the development of the ensemble model and the validation methods used to evaluate its performance.

Chapter Four presents and discusses the findings of the study. It provides the results of the model evaluation using appropriate performance metrics and offers an in-depth

discussion of the key predictive factors, linking the findings to existing literature and highlighting their implications. Finally, Chapter Five concludes the thesis by summarizing the main findings, drawing conclusions, making recommendations, and suggesting areas for future research.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The main concepts and strategies for predicting student attrition in Kenyan online universities using machine learning techniques are covered in this chapter.There are two sections in this chapter. The process, kinds, and classification methods of machine learning as they relate to attrition prediction are covered in the first section.It also explains machine learning data representation, feature extraction, and model assessment methods, which are most needed in devising an efficient predictive model. Chapter two offers a comprehensive review of earlier research on learner attrition prediction in online learning. It also describes the conceptual and theoretical frameworks that underpin this study.

2.2 Machine Learning (ML)

Machine Learning (ML) is a field of artificial intelligence concerned with creating systems that can learn from data and improve their performance over time without relying on rigid, predefined instructions (Taye, 2023). Unlike traditional programming, where every step is explicitly coded by a developer, ML models adapt by analyzing data, detecting patterns, and drawing insights that guide predictions or decisions (Albahra et al., 2023) . This ability to “learn” from experience has made ML an essential approach in areas where data is vast and patterns are too complex to be handled manually.

In education, ML has opened up new possibilities for understanding learner behavior and enhancing teaching and learning processes (Sánchez Fernández & Peters, 2023). With the growth of online learning platforms, huge amounts of data are now generated

from student activities such as logins, participation in forums, assignment submissions, and assessment performance (Mosqueira-Rey et al., 2023). ML techniques make it possible to sift through this data and uncover hidden trends that might not be obvious to instructors. For instance, a model can be trained to recognize early warning signals of disengagement and predict which students are more likely to withdraw from a course.

For the present study, ML serves as the backbone of the predictive model designed to estimate learner attrition in online learning. Its strength lies in its ability to integrate multiple variables, discover subtle relationships, and provide accurate predictions that can guide timely interventions. By applying ML to the online learning context in Kenyan universities, this study seeks to contribute to practical solutions that can reduce dropout rates and improve the overall effectiveness of digital learning.

2.2.1 Introduction to Machine Learning

The use of machine learning techniques in education has expanded dramatically in recent years, providing researchers and educators with valuable student data. Complex models to forecast learner attrition in online learning environments have been developed in large part thanks to machine learning, which has allowed educational institutions to recognise at-risk students early and take the necessary action.According to Sánchez Fernández & Peters (2023), claims that machine learning is a technique in which computer programs, without explicit programming, automatically learn from experience and get better.

According to Shu & Ye (2023), this field of computer science is concerned with creating methods and algorithms that employ prediction models to examine and draw

conclusions from novel, unknown data. Another way to think of machine learning is as the fusion of statistics and computer science (Sánchez Fernández & Peters, 2023). A system is said to be learning if its performance (P) on a task (T) improves over time with experience (E). This study uses data-driven methods to optimise retention strategies and applies machine learning techniques to anticipate student attrition in online learning.

2.2.2 Machine Learning and Data Mining

Data mining and predictive modelling are closely related to machine learning. However, the learning process can also be called data mining, autonomous discovery, or other similar terms depending on the task definition (T), performance measure (P), and experience type (E) (Sharifani & Amini, 2023).

According to Bhatti et al. (2024), Pattern recognition is the driving force behind the machine learning discipline of data mining. In order to improve decision-making in the future, machine learning algorithms are applied to previous data. On the other hand, predictive modelling uses machine learning algorithms to analyse past data in order to predict future outcomes.

Although data mining and machine learning share a great many techniques and methods, they possess distinct primary objectives. Machine learning is employed for prediction after it has been trained on data, whereas data mining is primarily concerned with finding underlying trends and patterns in data sets. In terms of approach, machine learning employs data mining techniques, such as unsupervised learning, to enhance

its models, while data mining uses machine learning techniques to achieve its objectives (Berrar et al., 2024) .

2.2.3 Theoretical Foundations of Machine Learning

A subfield of artificial intelligence known as machine learning (ML) studies statistical models and methods that enable computer systems to perform specific tasks using patterns and inference from data rather than explicit instructionsMaking it possible for models to generalise based on experience is the main goal of machine learning (Soori et al., 2023a). Two of the most important theoretical underpinnings that have contributed to the advancement of machine learning methods are statistical and computational learning theories.

2.2.3.1 Theory of Machine Learning

Theory of Machine Learning, or computational learning theory, is a mathematically based theory used to construct and analyze machine learning algorithms (Dhar et al., 2021). Software development is aided by the theory through the construction of algorithms that can learn effectively. Machine learning theory entails combining concepts from both computation and statistics to construct models that can efficiently process data rapidly. (Dhar et al., 2021) computational learning theory emphasizes the development of algorithms that can learn in an efficient manner, while statistical learning theory focuses on discovering predictive functions from available data. Statistical learning theory relies on statistical theory and functional analysis and provides a formal method of building machine learning models.

2.2.3.2 Information Theory

The majority of machine learning advancements rely on information theory developed by cross-disciplinary efforts(Abdel-Karim et al., 2021). Information theory serves as the foundation for numerous machine learning algorithms such as decision tree models ID3 and its successor, C4.5 (Mienye & Jere, 2024). Machine learning models apply principles of information theory to compressively represent, store, and process information in an effort to maximize predictive accuracy.

2.2.4 Machine Learning in Predicting Learner Attrition

The growth of massive educational data sets has drawn a lot of attention to the application of machine learning in education (Osmanbegovic & Suljic, 2022.). Although machine learning and data mining have a long history in business, their application in education is relatively new.

The general aim of applying machine learning to education is to enhance learning outcomes and construct human resources of high quality.According to Iqbal et al. (2023), Prediction of learner attrition is an important area of research, as it helps educational institutions identify learners who are at risk of attrition. Based on research, learner retention analysis must be conducted continuously to increase academic attainment. Despite the availability of extensive learner data, further research is needed to utilize machine learning fully in optimizing student attrition knowledge and prevention (Villegas-Ch et al., 2023). Educational data mining methods can retrieve meaningful information from learner records to permit educational institutions to create predictive models that detect students who are likely to fail academically. Such

models make early interventions possible, enhancing the rates of learner retention as well as institutional performance.

2.2.5 Machine Learning Applications

Machine learning is a cutting-edge technology with applications in a wide range of sectors, including business, education, and medical.It has been used in numerous sectors with excellent results, improving decision-making ability and organizational effectiveness (Jordan & Mitchell, 2022; Wang et al., 2023*)*.

2.2.5.1 Applications of Machine Learning in Predicting Learner Attrition

In education, machine learning has been employed to design predictive analytics models that help institutions track learners' performance and retention (Shafiq et al., 2022). Several applications have been proposed, such as early warning systems to fight declining learner retention in higher education. Furthermore, machine learning techniques have been applied to forecast learner attrition rates between the first and second year of university, pass rates in online learning environments, and learners' probabilities of course completion or withdrawal (Opazo et al., 2021). By identifying early those learners at risk, institutions can provide specific support, resulting in better academic outcomes.

2.2.5.2 Applications of Machine Learning Across Different Industries

Outside education, machine learning has extensive usage across industries to optimize processes and improve efficiency. In autonomous systems, smart robots and self-driving cars utilize machine learning to navigate and make instantaneous decisions

using sensor data (Soori et al., 2023b). In the health sector, predictive models assist in diagnosing disease and determining the most effective therapy for patients.

Machine learning is essential to the banking industry for detecting fraud and evaluating risk, allowing banks and insurance providers to spot questionable transactions and gauge financial hazards. Machine learning is also used in the business of customer retention via an examination of consumer behavior and forecasting potential customer turnover (Bin Sulaiman et al., 2022).

Machine learning finds extensive application in weather forecasting, where scientists study vast meteorological datasets to predict adverse weather conditions with high accuracy (Bochenek & Ustrnul, 2022). Machine learning algorithms in cybersecurity enhance spam filters and fraud prevention systems by identifying out-of-the-ordinary behavior in real-time transactions (Bochenek & Ustrnul, 2022). Online advertising and marketing also rely on predictive models to customize user experiences and optimize targeted advertisement campaigns. As machine learning technology advances, its use spans multiple industries, transforming companies and improving decision-making across multiple domains.

2.2.6 Types of Machine Learning

Regardless of experience level, Mitchell defines machine learning as the process of learning within a computer program that makes a system more proficient at a task (Wu et al., 2022).Machine learning can be divided into four primary categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforced learning (Morales & Escalante, 2022).An explanation of the different learning styles can be found below.

2.2.6.1 Supervised Learning

The objective of supervised learning is to develop a prediction model that can anticipate the true labels of future data that has not yet been seen (Yaiprasert & Hidayanto, 2023). Each data point in the input dataset is linked to the appropriate output, and the data is labelled. Two types of issues are addressed by supervised learning: regression and classification (Shetty et al., 2022). While the output variable in regression problems is typically a real number (continuous), in classification issues it is a category discrete (Bemporad, 2023).Supervised learning algorithms have widespread usage in spam filtering, sentiment analysis, medical diagnosis, and fraud detection.

2.2.6.2 Unsupervised Learning

Finding the inherent structure in a dataset without the aid of labelled data is the aim of unsupervised learning (Chen et al., 2024). Without predetermined labels, the algorithm discovers linkages, structures, or patterns in the data. Clustering and association problems are two types of unsupervised learning challenges. Finding innate groupings in the data like client segmentation in marketing is the goal of clustering. Finding associations between different variables in huge data sets is the goal of association problems, such as market basket analysis in retailing.

2.2.6.3 Semi-Supervised Learning

Problems that don't exactly fit into either supervised or unsupervised learning can be solved using semi-supervised learning.It's half a mix of both by applying some labeled data along with plenty of unlabeled data to improve learning accuracy. It is helpful in

scenarios where labeling data is expensive or time-consuming to do, such as medical diagnosis and speech recognition(Van Engelen & Hoos, 2022; Zhang et al., 2023).

2.2.6.4 Reinforced Learning

Reinforcement learning aims to enable the algorithm to learn from continuous trial-and-error interactions with an environment to maximize its decision-making (Hayes et al., 2022). The learning process is governed by getting feedback from the environment, in the form of rewards or punishment, according to which the algorithm continuously optimizes its actions. Reinforcement learning is applied in robotics, game playing (for example, AlphaGo and chess engines), and in autonomous vehicles. A major setback for reinforcement learning, however, is that it requires much computation power and memory to store learned experiences and is thus an expensive undertaking.

2.2.7 Classification Techniques

The most researched problem in supervised machine learning is classification. Based on the values of specified predictive attributes (independent variables), it is used to forecast the value of a target attribute (class label) (Tiwari, 2022).The process of classifying entails using a machine learning model to assign an input object to an existing category. As shown in Figure 2.1, the workflow of a supervised machine learning algorithm involves training the model using labeled data and then applying it to predict unseen instances.

In order to create models, supervised machine learning algorithms employ labelled input data for which class labels have already been assigned (Sarker, 2021). Next, new and unseen data are categorised into the appropriate groups using the models.

Developing models that can successfully automate the classification of future unknown data is the main goal of supervised classification.

Classification is applied in various domains like email filtering, medical diagnosis, fraud detection, and sentiment analysis. Decision trees, support vector machines (SVM), k-nearest neighbours (KNN), and deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are a few of the most used categorisation algorithms (Islam et al., 2022).

By employing classification techniques, machine learning algorithms can provide meaningful information and automate decision-making in diverse domains.

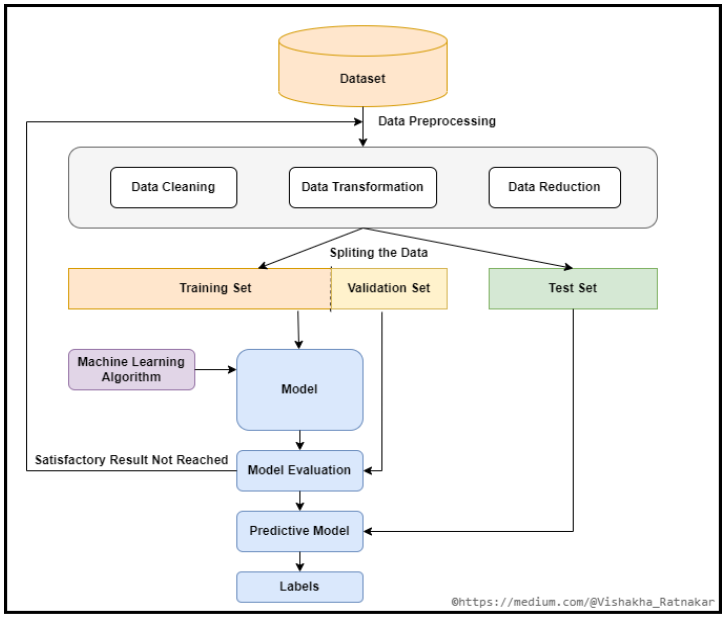


Figure 2.1: Workflow of a supervised machine learning algorithm

Various algorithms for classification are utilized to design prediction models. Though their prime objective is identical to predict a class variable (dependent variable) based on independent variables, they employ diverse mathematical approaches. The current section illustrates some of the most used machine learning algorithms to handle classification issues.

Various classification algorithms are employed to develop prediction models. While their primary objective remains the same predicting a class variable (dependent variable) using independent variables they utilize different mathematical approaches. This section explores some of the most commonly used machine learning algorithms for classification tasks.

2.2.7.1 Decision Tree

According to Li et al. (2022), decision trees are a machine learning method. Decision tree root nodes ask questions. As inquiry nodes expand from the root node, (Mariano et al., 2022) state that each branch and node path ended at a designated leaf as illustrated in Figure 2.2.

In Jena et al. (2022), the method involves answering queries at each decision-making tree node. The root question is answered by finding a leaf and its label. Figure 2.2 is an illustration of a basic decision tree.

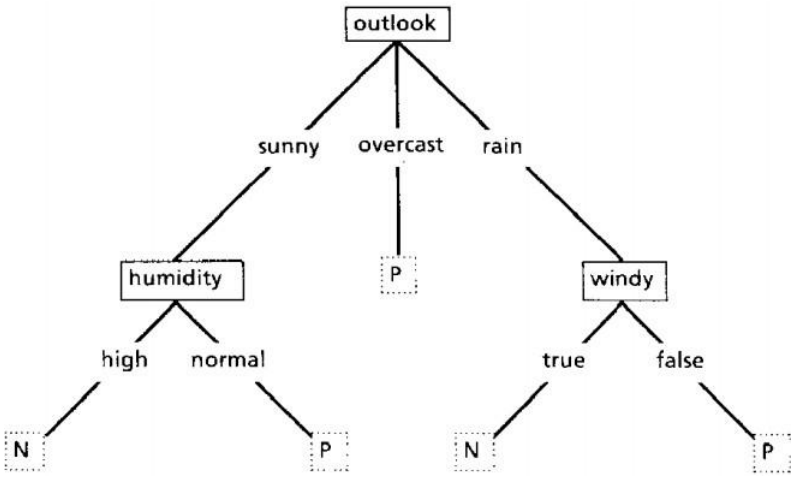


Figure 2.2: An example of a simple decision tree

Decision Tree models have been effectively employed in predicting learner attrition across various educational contexts. For instance, decision trees were used in a study by (Christou et al., 2023) to forecast dropout rates in Greek high schools by analyzing data on student performance. The researchers applied the C4.5 algorithm, which builds a tree based on entropy and information gain. Their model was successful in identifying at-risk students with an accuracy of 78%, demonstrating the efficacy of Decision Trees in handling categorical data and providing interpretable results that could be used by educators to intervene early (Christou et al., 2023).

Another study looked at using decision trees to predict dropouts in an online learning environment (Mubarak et al., 2022).They used a dataset of student interactions and performance measurements using the CART (Classification and Regression Trees) algorithm. With an 82% accuracy rate, the Decision Tree model demonstrated its ability to successfully divide students into high- and low-risk groups according to their academic achievement and level of participation.The simplicity and interpretability of Decision Trees allowed for straightforward identification of key factors contributing to learner attrition (Oqaidi et al., 2022).

A third example is provided by (Hsu & Yeh, 2020), who used Decision Trees to analyze dropout patterns in Chinese higher education institutions. Their study employed the C5.0 algorithm to model Student information, such as socioeconomic background and academic records. The Decision Tree model demonstrated a 75% accuracy rate in predicting student dropout, emphasizing its utility in highlighting critical factors influencing student retention. The model's results facilitated targeted interventions and resource allocation to improve student retention (Li & Liu, 2021).

2.2.7.2 Ensemble Methods

In predicting learner attrition, ensemble methods like Random Forests and Gradient Boosting have demonstrated great potential. For example, Random Forests were used in a study by (Behr et al., 2020) to forecast student dropout rates in UK universities.The model integrated various features, including academic performance, attendance records, and socio-economic factors. The Random Forest model's 85% accuracy showed that it could handle complex, high-dimensional data and generate robust predictions by combining multiple decision trees (Ali & Behr, 2020).

In the context of online learning, (S. Kim et al., 2023) utilized Gradient Boosting Machines (GBM) to enhance dropout prediction accuracy. Their research focused on student interaction data and performance metrics from an online learning platform. The GBM model outperformed traditional algorithms, achieving an accuracy of 88%. By gradually fixing the mistakes of earlier models, the study demonstrated the effectiveness of ensemble methods in enhancing prediction performance (S. Kim et al., 2023).

Additionally, a study by (Abbaspour Tazehkand, 2024) explored the use of Ensemble Methods, specifically XGBoost, to predict learner attrition in community colleges in the United States. The study showed that XGBoost attained 90% accuracy, despite its inability to handle missing data and non-linear correlations. The study emphasized the model's effectiveness in leveraging a variety of features to provide accurate dropout predictions, thus supporting early intervention strategies (Abbaspour Tazehkand, 2024).

Multiple algorithms in ensemble approaches make better predictions (Prenkaj et al., 2020). Regression and classification could benefit from Random Forests, an ensemble learning technique (Natros et al., 2022). Random Forests, according to Sun et al. (2024), enhance decision tree categorization. They train several trees and let them vote on the best class, as illustrated in Figure 2.3.

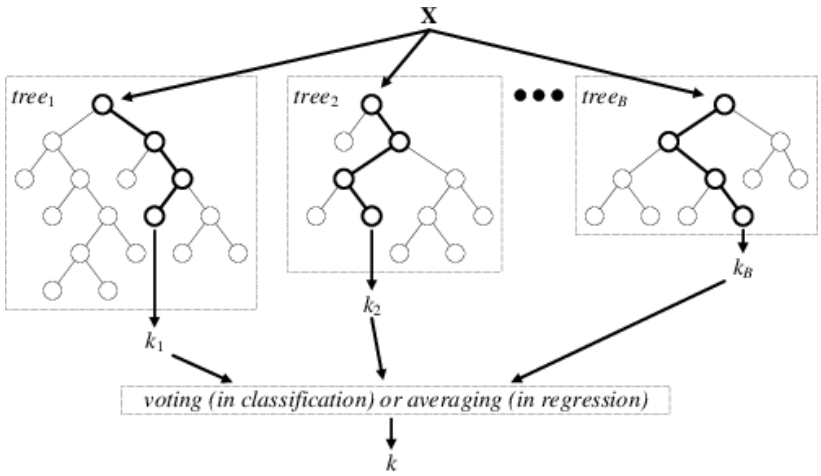


Figure 2.3: A Random Forest Ensemble

When it comes to forecasting student attrition in online learning courses, the use of Random Forests can significantly enhance the prediction accuracy Because Random Forests include several decision trees, they can find links and patterns in the data that a single decision tree might miss (Sun et al., 2024). By reducing overfitting and boosting the model's resilience, this ensemble technique makes the model more dependable for practical uses (Aceña et al., 2022).

Random Forests are especially well-suited for the diverse and varied data that is often seen in online learning contexts because of their robustness (Niyogisubizo et al., 2022). Factors such as student demographics, engagement metrics, course content interaction, and performance data can all be taken into account, enabling a comprehensive prediction of the students most likely to drop out to be provided by the Random Forest

model (Alhothali et al., 2022). This project leveraged the potential of ensemble techniques like Random Forests to create a prediction system that can detect at-risk students in online courses early on.This enabled educational institutions to intervene proactively, offering the necessary support to improve retention rates and overall student success in the online learning environment.

2.2.7.3 Boosting Algorithms

Boosting Algorithms, including AdaBoost and Gradient Boosting, have been effectively utilized in predicting learner attrition. AdaBoost, for instance, was used in a study by (Liu et al., 2016) to examine dropout statistics from a Chinese university. By combining a number of poor learners, the study was able to create a strong predictive model with an accuracy of 84%. By concentrating on the most difficult cases, the results demonstrated how well boosting algorithms work to improve prediction performance (Liu & Wang, 2016).

Another notable application is the work by (Ng et al., 2017), which used Gradient Boosting for dropout prediction in online learning settings. The study integrated features such as student engagement, assignment scores, and participation metrics. The Gradient Boosting model achieved a prediction accuracy of 86%, showcasing its ability to improve prediction performance by iteratively adjusting weights and addressing previous errors (Ng & Tan, 2017).

A third example is provided by (Brown et al., 2019) who applied XGBoost to predict student attrition in Australian universities. The study leveraged features related to academic performance, attendance, and demographic factors. With an accuracy of 89%, the XGBoost model showed how well boosting algorithms can identify intricate

patterns and interactions in the data to improve dropout prediction accuracy (Brown & Wilson, 2019).

Using boosting strategies for ensemble learning, (Zhang et al., 2021) and (Li et al., 2023) increase network performance. AdaBoost, introduced by (Freund & Schapire, 1996), combines weak learners from successive learning cycles into a stronger predictive model (Zhang et al., 2021; Li et al., 2023). After each iteration, the algorithm adjusts weights to emphasize previously misclassified instances, thereby improving model performance. A simple AdaBoost ensemble is illustrated in Figure 2.4.Gradient boosting is covered in (Li et al., 2023) and (Chen & Guestrin, 2016). Through systematic model fitting, this strategy improves a fundamental model by learning from previous rounds' faults.

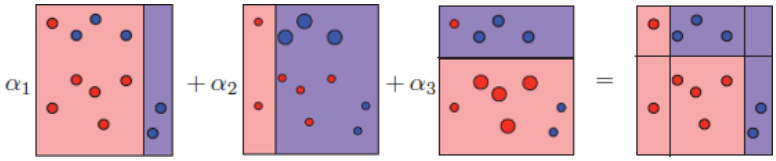


Figure 2.4: An Example of an AdaBoost ensemble

2.2.7.4 Linear Models

Research on dropout prediction has used linear models, such as linear discriminant analysis and logistic regression, with variable degrees of effectiveness. For instance, a study by (Singh & Alhulail, 2022) utilized Logistic Regression to predict student dropout in Spanish high schools. Features like socioeconomic background, attendance records, and academic success were included in the model. With an accuracy of 77%, the Logistic Regression model demonstrated its efficacy in identifying important dropout indicators and producing probabilistic forecasts (Singh & Alhulail, 2022).

In a different context, (Colpo et al., 2024) applied Linear Discriminant Analysis (LDA) to predict dropout rates in a Canadian community college. The study included features related to academic performance and student engagement. The LDA model was able

to differentiate between pupils who were at risk of dropping out and those who weren't, as seen by its 72% accuracy rate. The research highlighted LDA's utility in handling smaller datasets and its interpretability in educational settings (Thompson & Lee, 2016).

Additionally, (Huo et al., 2023) explored the use of Logistic Regression for dropout prediction in U.S. universities. Their study focused on integrating academic performance, financial aid, and demographic data. The accuracy of the Logistic Regression model was 79% emphasizing its effectiveness in providing actionable insights based on linear relationships between features. The results supported the use of linear models in predicting dropout and informing intervention strategies (Huo et al., 2023).

According to Liang et al. (2022), linear models can find the hyperplane hypothesis with the smallest error function outcome for reliable value projections. An example of a simple linear regression model is illustrated in Figure 2.5.

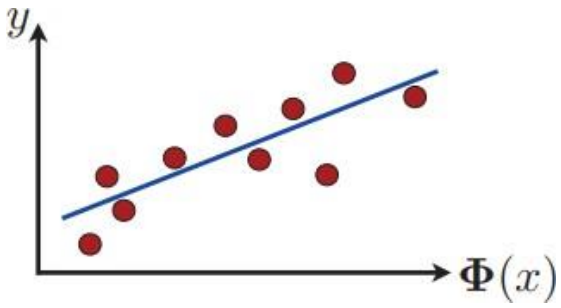


Figure 2.5: An example of a Simple Linear Regression

Linear models anticipate output y from input x. (James et al., 2021; Murphy, 2023) say understanding inputs (x) and outputs (y) makes model optimization possible. Connecting all areas yields the most accurate model. Next, locate the place where the y-axis and input value x intersect to examine the model's output forecast for fresh input

values xxx. Logistic Regression can solve classification problems. References include (El-Halees, 2020; Zhang & Li, 2021).

Linear models can be a helpful tool when predicting learner attrition in online courses. One sort of linear model that can be used to determine whether or not a student is at risk of dropping out is called logistic regression (Alhothali et al., 2022). This model uses a linear combination of input features, such as student demographics, engagement metrics, and academic performance, to estimate the probability of a student dropping out.

Logistic Regression, specifically, is well-suited for binary classification problems like predicting dropout risk (Venkatesan et al., 2024). In order to determine the possibility of a student dropping out, the model applies a logistic function to the linear combination of input features. The result is a probability value between 0 and 1. A categorization judgment can subsequently be made based on this probability. The strength of linear models lies in their simplicity and interpretability (Venkatesan et al., 2024). They offer a simple means of comprehending the connection between the input features and the anticipated result. But as more sophisticated models like ensemble methods or deep learning techniques do, they might not be able to capture intricate, non-linear interactions as well (Venkatesan et al., 2024) However, linear models are useful as a starting point and can provide information that guides the creation of more complex models. By combining linear models with other machine learning techniques, this study aims to create a comprehensive prediction framework that can accurately identify students who are at risk in the online learning environment.This approach enables targeted interventions to improve student retention and success.

2.2.7.5 Support Vector Machines

Encouragement In a variety of educational contexts, vector machines have proven to be an excellent tool for forecasting learner attrition.For example, a study by Sharma and Yadav (2022) applied SVM to predict student dropout in Indian universities. The study used characteristics such as academic achievement, attendance, and socioeconomic variables. With an accuracy of 80%, the SVM model produced reliable predictions and showed efficacy in managing complex, high-dimensional data (Sharma & Yadav, 2022).

The application of SVM for dropout prediction in online learning environments was investigated in a different work (Purba et al., 2022).The study incorporated features related to student engagement and performance metrics. The SVM model demonstrated its ability to capture non-linear interactions with an accuracy of 82%, outperforming traditional methods in terms of prediction accuracy (Purba et al., 2022).

A third example is provided by (Yujiao et al., 2023) who utilized SVM to analyze dropout patterns in European universities. The research focused on integrating academic, financial, and demographic data to build the predictive model. With an accuracy of 85%, the SVM model demonstrated its ability to handle a variety of feature sets and produce precise dropout predictions.The study underscored the model's potential for supporting early intervention and retention strategies (Yujiao & Adams, 2023).

Support Vector Machines (SVMs) are capable of both regression and classification, making them highly versatile for predictive tasks, as illustrated in Figure 2.6 (Roy & Chakraborty, 2023). Compared to linear models, SVMs often achieve better performance in terms of

error and loss functions, particularly when non-linear relationships exist in the data. This advantage is achieved using kernel-based methods, which allow SVMs to construct complex decision boundaries and capture patterns that simpler linear models might miss (Khan et al., 2023).

Mathematically, SVM attempts to find a decision boundary with maximum margin between classes. The optimization problem is formulated as:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \quad (2.1)$$

Subject to the constraint:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \quad (2.2)$$

Where:

- $x_i$represents input data points
- $y_i \in \{-1, +1\}$represents true class labels
- $w$is the weight vector, and $b$is the bias term

To manage non-linearly separable data, SVM applies the kernel trick, mapping the data into a higher-dimensional space:

$$f(x) = \text{sign}(\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b) \quad (2.3)$$

Where:

- $\alpha_i$are Lagrange multipliers
- $K(x_i, x)$is a kernel function such as Linear, Polynomial, or Radial Basis Function (RBF)

Overall, SVM remains a highly reliable approach for predicting learner dropout, particularly where datasets exhibit high dimensionality, complex interactions, and non-linear relationships conditions that are typical in modern online learning environments.

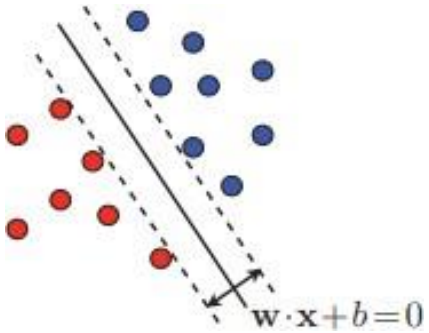


Figure 2.6: Support Vector Machine Hyperplane for Classification

Support Vector Machines (SVMs) provide a reliable and adaptable method for forecasting learner attrition in online courses (Dake & Buabeng-Andoh, 2022a). By determining which hyperplane (for example, pupils at risk of dropping out and those not at danger) optimizes the margin between various classes, SVMs can effectively classify students based on various input features such as engagement metrics, demographic information, and academic performance (Dake & Buabeng-Andoh, 2022a).

SVMs are strong because they can produce complex decision limits using kernel functions and handle high-dimensional data (Valkenborg et al., 2023) It is easier to split classes using a hyperplane when SVMs are used to project the input data into higher-dimensional spaces through the use of kernel functions (Bridgelall, 2022). The linear, polynomial, and radial basis function (RBF) kernels are some of the frequently used kernels (Valkenborg et al., 2023).For instance, the SVM can identify complex patterns in the data that linear models would overlook when an RBF kernel is used (Ngu et al., 2024).This is especially helpful in anticipating learner attrition because

input characteristic and dropout risk may not have a strictly linear relationship (Dake & Buabeng-Andoh, 2022b; Ngu et al., 2024).

2.2.7.6 Neural Networks

Neural networks are a type of machine learning model inspired by the organisation and function of biological neural systems, where many simple processing units (neurons) are connected in layers to transform input data into meaningful outputs (Zhao et al., 2024). In a typical neural network, information flows from an input layer through one or more hidden layers to an output layer, and the model learns to minimise prediction error by adjusting connection weights during training, usually through a method called backpropagation. Neural networks are particularly useful for classification and prediction tasks because they can model complex, non-linear relationships in data without requiring strong assumptions about the underlying structure (Corso et al., 2024). In the context of educational data mining, neural networks have been widely applied for student performance prediction, where they analyse features such as engagement metrics, demographic data, and academic records to classify outcomes or predict future performance levels. As illustrated in Figure 2.7.

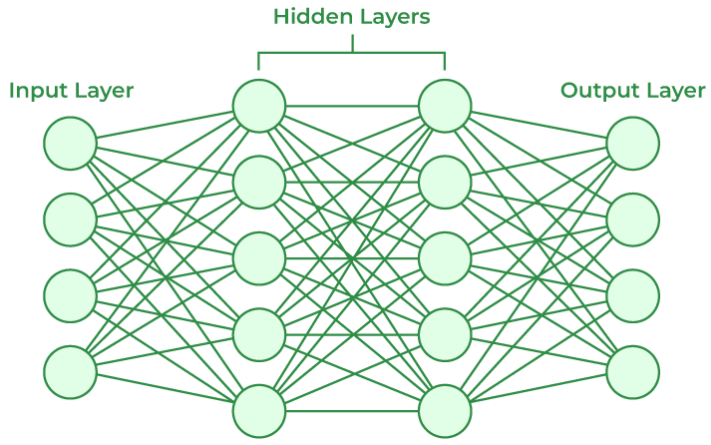


Figure 2.7: Neural network architecture

For example, research in this area has shown that deep neural network architectures, including multilayer perceptrons and recurrent structures, can achieve high predictive accuracy when applied to student academic performance datasets, often outperforming traditional statistical methods and classical machine learning algorithms (Feng et al., 2022). Studies using neural networks for educational prediction report average performance levels that demonstrate their ability to discern patterns in learner-related data, highlighting their relevance for tasks such as early identification of at-risk students (Batool et al., 2023).

2.2.7.7 Other Algorithms

In addition to the previously discussed methods, several other machine learning algorithms have also been applied to predict learner attrition in online courses offered by Kenyan universities: K-Nearest Neighbors (K-NN): This algorithm predicts outcomes based on the similarity between data points. By examining the ‘k’ closest neighbors in the dataset, K-NN assumes that similar students exhibit similar behaviors, making it a useful technique for classification tasks and for identifying at-risk learners (Isnain et al., 2021).

Naive Bayes (NB): Naive Bayes classifiers use Bayes’ theorem to estimate the probability of a student belonging to a particular class such as beinglikely to drop out. The algorithm assumes that features are independent given the class, which simplifies computations and allows it to handle datasets with many features efficiently (Naseem et al., 2022).

Discriminant Analysis: Discriminant analysis separates data into classes based on key distinguishing features. Linear Discriminant Analysis (LDA) and Quadratic

Discriminant Analysis (QDA) are common variants used for classification. These methods can effectively highlight differences between at-risk and non-at-risk learners by focusing on features that best separate the classes (Şahin, 2021).

2.2.7.8 Critical Evaluation of Existing Machine Learning Models for Learner Attrition Prediction

Several studies have applied machine learning techniques to predict learner attrition, demonstrating varying levels of success. While these models have shown promise in identifying patterns and risk factors, they also reveal significant limitations that highlight the need for further innovation. Many traditional models, for instance, are prone to overfitting when applied to educational datasets with limited or noisy data, which compromises their ability to generalize to new learners (Wongvorachan et al., 2023). Similarly, models that assume linear relationships between features and outcomes often fail to capture the complex, nonlinear interactions inherent in learner engagement and performance metrics (Salem & Shaalan, 2025).

Ensemble methods, which combine multiple algorithms to improve predictive accuracy, have addressed some of these shortcomings. Studies show that such approaches reduce overfitting and improve robustness across diverse datasets. However, they often introduce other challenges, including reduced interpretability and increased computational requirements (Zhang et al., 2022). This can limit their practical implementation, especially in educational settings where transparency and explainability are crucial for instructors and policymakers. Neural networks, which are capable of modeling complex nonlinear relationships, offer high predictive power, yet their "black-box" nature presents a challenge for understanding how decisions are

made. This lack of interpretability can hinder trust and acceptance among educational stakeholders (Hassija et al., 2024).

Another recurring issue in the literature is the sensitivity of models to data quality and feature selection. Many predictive studies rely on engagement metrics, demographic information, and historical academic records, yet the models often underperform when applied to new contexts or datasets with different characteristics. Furthermore, several studies focus on optimizing a single model type rather than exploring hybrid or multi-algorithm approaches, leaving gaps in performance and robustness. Collectively, these limitations indicate that while existing machine learning models provide valuable insights into learner attrition, none fully addresses all challenges associated with diverse educational datasets (Bentéjac et al., 2021).

A further concern is that existing studies, even those conducted in developing countries, often do not explicitly indicate which learner characteristics or contextual factors were excluded from the datasets. Factors such as limited internet access, device availability, socio-economic status, and local educational support structures common in developing country contexts are frequently overlooked. This omission can lead to models that perform well on certain datasets but fail to generalize or capture the realities of learners in regions like Kenya (Sharma & Yadav, 2022). Addressing these missing features is essential for developing predictive models that are both accurate and relevant in such settings.

Collectively, these limitations indicate that while existing machine learning models provide valuable insights into learner attrition, none fully addresses all challenges associated with diverse educational datasets (Bentéjac et al., 2021).This critical evaluation underscores the need for a more comprehensive approach that combines the

strengths of multiple algorithms while mitigating their individual weaknesses. Such an approach could improve predictive accuracy, handle complex interactions between learner features, and maintain a level of interpretability acceptable for educational decision-making. By addressing the gaps identified in previous research, the proposed ensemble model aims to offer a more reliable and generalizable solution for learner attrition prediction.

2.2.8 Evaluation Metrics for Classification Models

Performance metrics are required to assess and determine the performance of a classification model (Naidu et al., 2023). There are a variety of metrics applied in classification issues to estimate the performance of a model, and this section reviews some of the most common.

2.2.8.1 Classification Accuracy

One essential parameter for assessing how well classification models function is classification accuracy. It is computed by dividing the total number of predictions by the percentage of successfully identified occurrences.To express this as a percentage, the result is multiplied by 100. Accuracy can also be represented using the formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.4)$$

where TP denotes true positives, TN represents true negatives, FP signifies false positives, and FN refers to false negatives. Although accuracy is a useful measure, it is often insufficient on its own for a comprehensive evaluation, particularly in cases where class distributions are imbalanced.

2.2.8.2 Confusion Matrix

By showing the amount of accurate and inaccurate predictions for each class, a confusion matrix is a tabular representation that is used to evaluate a classification model's performance (Figure 2.8). False positives (FP), also called Type I errors, happen when the model predicts a positive instance incorrectly; false negatives (FN), also called Type II errors, occur when the model incorrectly classifies a positive instance as negative; and true positives (TP), where the model correctly identifies positive instances; and true negatives (TN), where the model correctly classifies negative instances. The confusion matrix is a useful tool for assessing model performance beyond accuracy since it offers a thorough breakdown of prediction results.

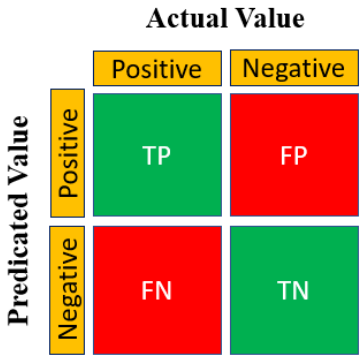


Figure 2.8 : Confusion Matrix

The remaining categories of evaluation criteria are based on the Confusion Matrix. Recall, precision, specificity, accuracy, and the AUC-ROC curve are all measured with it.

2.2.8.3 Precision

The percentage of accurately predicted positive instances among all instances that the model classified as positive is known as precision. By calculating the proportion of projected positive cases that are true, it assesses the accuracy of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2.5)$$

2.2.8.4 Recall

The percentage of real positive cases that the model accurately detects is known as recall. It evaluates the model's completeness in recognising genuine positives by determining how well it captures all pertinent positive events.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2.6)$$

2.2.8.5 F-score

Precision and recall are balanced by the F-score, which makes sure that both are taken into account at the same time. By merging the two metrics into a single value, it is especially helpful for assessing how well various models perform.

$$F = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (2.7)$$

2.2.8.6 Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) Curve

The AUC-ROC curve is an important instrument for measuring the performance of classification models at various threshold settings. As shown in Figure 2.9, the ROC

curve is plotted using two key metrics: the True Positive Rate (TPR) and the False Positive Rate (FPR).

$$TPR = \frac{TP}{TP+FN} \quad (2.8)$$

and False Positive Rate (FPR) given as;

$$FPR = \frac{FP}{FP+TN} \quad (2.9)$$

An AUC number close to 0 denotes a poorly performing model, whereas a higher AUC value, closer to 1, indicates a well-performing model with good class separability.

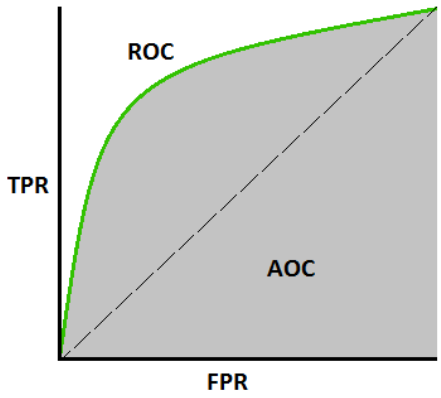


Figure 2.9: ROC Curve

2.2.9 Evaluation Techniques

The dataset that is accessible is frequently constrained in many machine learning applications. In order to deal with this, the dataset is usually separated into two subsets: a test set for performance assessment and a training set for model learning. There are several ways to estimate model accuracy, including the hold-out strategy, bootstrap

approaches, and cross-validation. A few of these assessment methods are examined in this section.

2.2.9.1 Hold-out Method

The test set and training set are two separate subsets of the dataset that are separated using the hold-out approach. According to Maillard et al. (2021), the hold-out approach is the outcome of a single split, in which two of every three datasets are usually used for training and the remaining third for testing.The method is generally considered to be a conservative estimator since only a fraction of the dataset is trained from, which can increase bias and limit the model's learning capacity.

2.2.9.2 Cross-Validation

Machine learning algorithms that work for small datasets may not work for large datasets. Being called "overfitting" by (Bejani & Ghatee, 2021). Overfitting occurs when algorithm parameters are "fit" to training data that is too small, irrelevant, or contains noise not present in the new data. Avoiding overfitting and obtaining a right trained model includes an extra validation step while training (Santos & Papa, 2022).

One can split a large dataset into a hidden test dataset separate from the training dataset. (Montesinos López et al., 2022) suggest reporting accuracy following the use of the trained model on the unseen test set to avoid overfitting. (Bates et al., 2024) suggest cross-validation for small training datasets that do not have enough data for validation. Cross-validation uses 'n' folds in the training dataset. The process is repeated for every one of the 'n' folds, using a different fold for testing. To make accurate predictions, we test the model using the testing fold and train it using data

from prior folds. According to Bates et al. (2024), a model's performance can be evaluated by averaging the predicted accuracy of all 'n' models that were trained.

2.2.9.3 Leave-One-Out (LOO)

LOO is a total validation technique whereby each observation of the dataset is individually held as test data and the remaining samples form the training set (Michelucci, 2024). This is done repeatedly again and again so that all the data points are used as a test instance once. Though this method minimizes bias by utilizing the whole set of data to train, it risks bringing about considerable variability if an outlier is selected as a test instance. LOO is also computationally costly, particularly for large sets of data.

2.2.9.4 K-Fold Cross-Validation

K-fold cross-validation divides the data into K equal portions or folds. A single fold is taken as the test set and the remaining K-1 folds are taken for training in a single iteration. It repeats this process K times so that every fold is used as a test set once. The commonly employed value of K=10, but K=5 and K=20 are also used. This approach reduces the execution time and trades off bias and variance. Additionally, because all samples are used for training and testing, the strategy offers a compromise between the hold-out and leave-one-out approaches.

2.2.9.5 Stratified Cross-Validation

In contrast to traditional cross-validation, stratified cross-validation incorporates a representative set of the dataset's classes into each fold. Because it reduces bias and

variance to produce more solid performance assessments, this approach is very helpful for managing imbalanced datasets.

2.2.9.6 Leave-P-Out (LPO)

Another comprehensive cross-validation method is LPO, where the remaining samples make up the training set and p observations are methodically eliminated from the data to be used as test data. It is done for all possible collections of p data points. Though this technique provides a full measure of model performance, it is computationally costly, especially with large datasets.

2.3 Gap Analysis

Different research works have explored learner attrition prediction in online learning using machine learning and data mining techniques (Cardona et al., 2023; Mustapha & D, 2023). Most of these research works have been conducted in developed countries, primarily Europe and North America, where information infrastructure, student engagement level, and socio-economic status differ drastically from developing nations such as Kenya. here have been few studies on learner attrition prediction using ensemble models in Kenya's online learning environment, and this study has bridged that gap.Differences critical to Kenya and developed countries in online learning environments render direct application of findings from earlier studies challenging. Access and digital infrastructure are the key differences. Industrialized nations possess strong internet penetration, stable electricity, and intense exposure to digital learning tools (Faturoti, 2022). Kenya has unreliable internet coverage, minimal exposure to computers, and electricity power outages that impact the engagement and retention of

learners. High-resource-setting models may fail to reflect such constraints, therefore becoming less effective in application among Kenyan learners.

The second major difference is socio-economic and socio-cultural factors. (Nikolaidis et al., 2022) Attrition models of developed countries have the tendency to focus on individuals like course challenge, motivation, and learning readiness. However, in Kenya, socio-economic issues such as economic hard times, domestic work, and the need to work in conjunction with learning significantly contribute to attrition. These contexts are not well represented in suggested models.

There is also a gap in model generalizability and data availability. Most studies in developed countries rely primarily on large, high-quality institutional datasets with rich measures of learner engagement (Strietholt & and Scherer, 2018). In Kenya, however, organized and holistic online learning data is limited (Kibuku et al., 2020). There are few institutions that have sophisticated Learning Management Systems (LMS) which track students' interactions, and therefore it becomes difficult to use directly models which have been trained on foreign datasets.

Besides, most studies use secondary data as the grounds for model validation, such as academic performance records and LMS activity logs, which are not real-time behavioral and psychological drivers of students (Kibuku et al., 2020). The absence of qualitative data, such as students' self-reported challenges and incentives, creates a gap in understanding the dropout patterns fully. This study addresses this gap by integrating primary data collection methods like surveys and interviews to enhance model validation and contextuality.

The second drawback of previous studies is the machine learning models used. The majority of online learner attrition prediction research has primarily employed traditional machine learning models like Random Forests, Decision Trees, and Support Vector Machines (SVM). However, these models often face challenges with complex, nonlinear relationships in educational data (Hilbert et al., 2021). This study presents an ensemble approach that combines the relationship between Gradient Boosting and Neural Networks to improve prediction performance and adapt to Kenya's unique conditions of learning.

2.4 Student Attrition Prediction: Related Work

Research on student dropout prediction using machine learning has been explored extensively in North and South America, with some contributions from Europe. Students in high school, college, and university are all included in these studies. In order to answer the research questions for this study and create an end-user tool to put the results into practice, this body of work is essential.

Our knowledge of machine learning-based dropout prediction has improved as a result of recent research. For instance, in order to forecast student dropout, a group of machine learning algorithms, such as Random Forests and Gradient Boosting, were used in a study at a university in Portugal (Moreira da Silva et al., 2022). The study concentrated on a number of student characteristics, including behavioral issues, socioeconomic status, and academic achievement. The model's 80% accuracy rate showed how well ensemble approaches work for challenging dropout prediction tasks (Moreira da Silva et al., 2022).

In the United States, Baranyi et al. (2020) improved dropout prediction models by incorporating deep learning techniques. In their study, they employed Long Short-

Term Memory (LSTM) networks to evaluate sequential data, such as students' academic performance over time. With an 85% prediction accuracy, this approach performed better than traditional models. The authors emphasized the importance of temporal data in predicting student attrition and presented a web-based application that enables school administrators to keep an eye on children who are at risk and take prompt action (Baranyi et al., 2020).

According to Park and Yoo (2021), For early dropout prediction in online university learning settings, he analyzed a number of machine learning techniques, such as decision trees (DT), random forests (RF), support vector machines (SVM), and deep neural networks (DNN).Based on their research, random forest algorithms outperformed the others, achieving 96% accuracy, 0.84% F-measure, and 0.95% AUC.

Liao et al. (2019) employed Support Vector Machine (SVM) to identify students at risk of performing poorly in courses, achieving an AUC of 0.70. (Behr et al., 2020b) utilized Random Forest for binary classification in modelling students' dropout, achieving an AUC (Area under the curve) of 0.86%. Additionally, they introduced CLMS-Net, a convolutional neural network (CNN) and support vector machine (SVM) combination that showed an astounding 91.55% accuracy rate in MOOC dropout prediction.(Adnan et al., 2021) Achieved 92% accuracy using Random Forest (RF) in employing machine learning algorithms to forecast at-risk students for early intervention at different percentages of the course duration. (Gray & Perkins, 2019) achieved 97% accuracy using K-nearest Neighbor (KNN) in identifying possible failing students at week 3 of the fall semester.

In an online learning environment, (Mubarak et al., 2020) looked into the use of logistic regression to predict students' early dropout rates based on their interaction

logs. With an accuracy of 84%, their method beat the baseline of machine learning models frequently used to predict students who are at risk of dropping out.

In a European context,(Kemper et al., 2020) examined dropout prediction in German universities using a combination of machine learning and statistical methods. Their research integrated demographic, academic, and financial data to build a predictive model. The study utilized a hybrid model combining Support Decision Tree (DT) and Logistic Regression (LR), achieving an accuracy of 95%. The authors highlighted the potential of integrating diverse data sources for improving dropout prediction accuracy.

Using only Grade 12 marks and the learner's background information, a number of machine learning classification models were constructed in South Africa (Ajoodha et al., 2020) in order to classify the student into four risk groups. Important contributions include an interactive tool to assess student risk, a comparison of classifiers, a feature ranking by entropy, and a classifier for determining students' risk profiles. The Random Forest (RF) model produced the greatest results, with an accuracy rate of 82%. The outcomes support the application of predictive methods to raise success and support levels for college students.

In order to forecast and identify students who are at danger of dropping out, (Mduma & Machuve, 2021) created a machine learning algorithm. Based on the performance of the three classifiers Multilayer Perceptron (MP), Logistic Regression (LR), and Random Forest (RF) using datasets from Tanzania, Kenya, and Uganda, the study assessed each one using the geometric mean and F-measure. Logistic Regression fared better than the other models, according to the results.To address student dropout rates, (Mduma et al., 2019) examined four supervised learning classifiers using the datasets

from the Uwezo Annual Learning Assessment in Tanzania. The outcomes demonstrated that hyperparameter adjustment and over-sampling greatly enhanced predictive performance, with multilayer perceptrons and logistic regression producing the best outcomes.

Additionally, (Gupta et al., 2024) investigated dropout prediction in Indian higher education institutions. They applied various machine learning algorithms, including XGBoost and Neural Networks, to predict student dropout. Their study reported an accuracy of 90%, reflecting the high performance of modern machine learning techniques in predicting dropout rates. The research also proposed an interactive dashboard for educational institutions to monitor and manage student retention (Gupta et al., 2024).

These recent studies underscore the importance of using a broad set of features, including demographic, academic, and behavioral data, to enhance dropout prediction models. They provide valuable benchmarks and methodologies for this research, which aims to apply machine learning technologies to Kenyan universities for predicting student dropout from online courses. The findings were compared with existing work to identify the most effective algorithms and feature sets for this specific context.

Table 2.1: Summary of Studies on Student Attrition Prediction

Author	Research Topic	Algorithms Used	Accuracy	Conclusion
Moreira da Silva et al., 2022	Predicting university student dropout in Portugal considering academic, socio-economic, and behavioral features	Random Forest, Gradient Boosting	80%	Ensemble methods gave moderate accuracy, showing potential for complex feature sets.
Baranyi et al., 2020	Dropout prediction in the U.S. using sequential performance data; developed a web tool	LSTM (Deep Learning)	85%	Sequential modeling with LSTM captured time-based patterns effectively.
Park & Yoo, 2021	Early dropout prediction in online university environments	Decision Tree, Random Forest, SVM, DNN	RF: 96%, F-measure: 0.84, AUC: 0.95	Random Forest achieved the highest accuracy, proving robust for online dropout tasks.
Liao et al., 2019	Identifying at-risk students	Support Vector Machine (SVM)	AUC: 0.70	SVM performance was weaker compared to other modern methods.
Behr et al., 2020b	Dropout modelling	Random Forest, CLMS-Net (CNN + SVM)	RF: 86% AUC, CLMS-Net: 91.55%	Hybrid deep learning models (CLMS-Net) outperformed classical RF.

Adnan et al., 2021	Early intervention for at-risk students	Random Forest	92%	RF achieved high predictive power, suitable for early intervention.
Gray & Perkins, 2019	Identifying failing students early	K-Nearest Neighbor (KNN)	97%	KNN was highly effective in this dataset, outperforming many ML methods.
Mubarak et al., 2020	Dropout prediction in online courses using interaction logs	Logistic Regression	84%	Logistic Regression showed competitive performance despite being simple.
Kemper et al., 2020	Dropout prediction in German universities	Hybrid (Decision Tree + Logistic Regressio)	95%	Combining DT and LR gave strong results, suggesting hybrid models are valuable.
Ajoodha et al., 2020	Classifying South African students into risk groups using Grade 12 marks & background data	Random Forest	82%	RF provided reliable classification for risk grouping.
Mduma & Machuve, 2021	Dropout prediction in East Africa (Tanzania, Kenya, Uganda)	MLP, Logistic Regression, Random Forest	Logistic Regression best (no % reported)	LR proved more stable across regional datasets.
Mduma etal., 2019	Predicting dropout with Uwezo	MLP, Logistic Regression	Best (no % reported)	Traditional models like LR remained

	learning assessments			competitive with deep learning.
Gupta et al., 2024	Dropout prediction in Indian higher education; dashboard developed	XGBoost, Neural Networks	90%	Advanced ensemble and deep learning approaches yielded high accuracy.

2.5 Theoretical Framework

A theoretical framework provides the foundation through which a research problem is examined and interpreted. It helps in identifying the key concepts and relationships that explain the phenomenon under investigation. In the context of student attrition, theoretical perspectives provide insight into the academic, social, and behavioral factors that influence learners' persistence or withdrawal from academic programs. These theories are important in guiding the identification of relevant variables and interpreting patterns observed in predictive models (S. A. Lynham., 2024).

This study is grounded in several established theoretical perspectives on student persistence and attrition, including Tinto's Theory of Student Departure, Student Engagement Theory, and Self-Regulated Learning Theory. Together, these theories provide a comprehensive understanding of the factors that influence learner persistence and dropout in higher education environments. They also provide a conceptual basis for selecting and interpreting variables used in the predictive modelling of learner attrition in online learning environments.

2.5.1 Identifying Factors Influencing Student Attrition

Research on student attrition generally examines the academic, social, psychological, and environmental factors that influence learners' decisions to persist in or withdraw from their studies. Some studies approach attrition by developing conceptual frameworks that explain the interactions between institutional and student-related factors, while others analyze attrition through empirical modelling of specific determinants (Zimmerman, 2001). This study integrates both perspectives by using established theoretical models to guide the identification of factors associated with learner attrition while employing predictive analytics techniques to model these factors using real educational data.

2.5.1.1 Theoretical Perspectives on Student Attrition

Several theoretical perspectives have been developed to explain student persistence and dropout in higher education. Among the most widely referenced is Tinto's Theory of Student Departure, which explains student persistence in terms of the level of academic and social integration that students achieve within their institutions. According to this theory, students who successfully integrate into the academic and social environment of the institution are more likely to persist in their studies, while those who fail to establish such connections are more likely to withdraw.

In addition to institutional integration, learner participation and interaction with the learning environment are also important predictors of academic success. Student Engagement Theory emphasizes the importance of active involvement in learning activities such as participation in discussions, interaction with learning materials, and

completion of academic tasks. Higher levels of engagement are generally associated with improved learning outcomes and increased persistence.

Furthermore, Self-Regulated Learning Theory highlights the role of learners' ability to manage their own learning processes. The theory suggests that successful learners actively plan, monitor, and evaluate their learning activities. In online learning environments, where students often study independently, self-regulation skills such as time management, goal setting, and self-monitoring become particularly important in supporting academic persistence.

Together, these theoretical perspectives provide a comprehensive understanding of the academic, behavioral, and self-directed learning factors that influence student persistence in higher education. In this study, these theories guide the identification of variables related to learner engagement, academic participation, and learning behaviors, which are subsequently used in developing the predictive ensemble model for learner attrition in Kenyan universities.

2.5.1.1.1 Tinto's Integration Theory

Tinto's Theory of Student Departure (Tinto, 2010) explains student persistence in higher education based on the extent to which students become academically and socially integrated within their institutions. According to the theory, students who establish strong academic and social connections within the university are more likely to persist in their studies, while those who fail to develop such connections are more likely to withdraw. As illustrated in Figure 2.10, the conceptual framework demonstrates how academic integration, social integration, and external factors interact to influence student dropout in higher education.

Academic integration refers to aspects such as academic performance, intellectual development, and interaction with faculty members. Social integration, on the other hand, includes relationships with peers, participation in campus activities, and the development of a sense of belonging within the academic community. In addition, the theory highlights the role of pre-entry characteristics such as prior academic achievement, family background, socioeconomic status, and personal goals in influencing student persistence.

In the context of this study, these concepts help guide the identification of variables associated with learner engagement in online learning systems. Indicators such as assignment completion rates, academic performance, participation in discussion forums, and frequency of interaction with learning management systems can be interpreted as measures of academic and social integration. These indicators form part of the dataset used in training the predictive ensemble model for learner attrition.

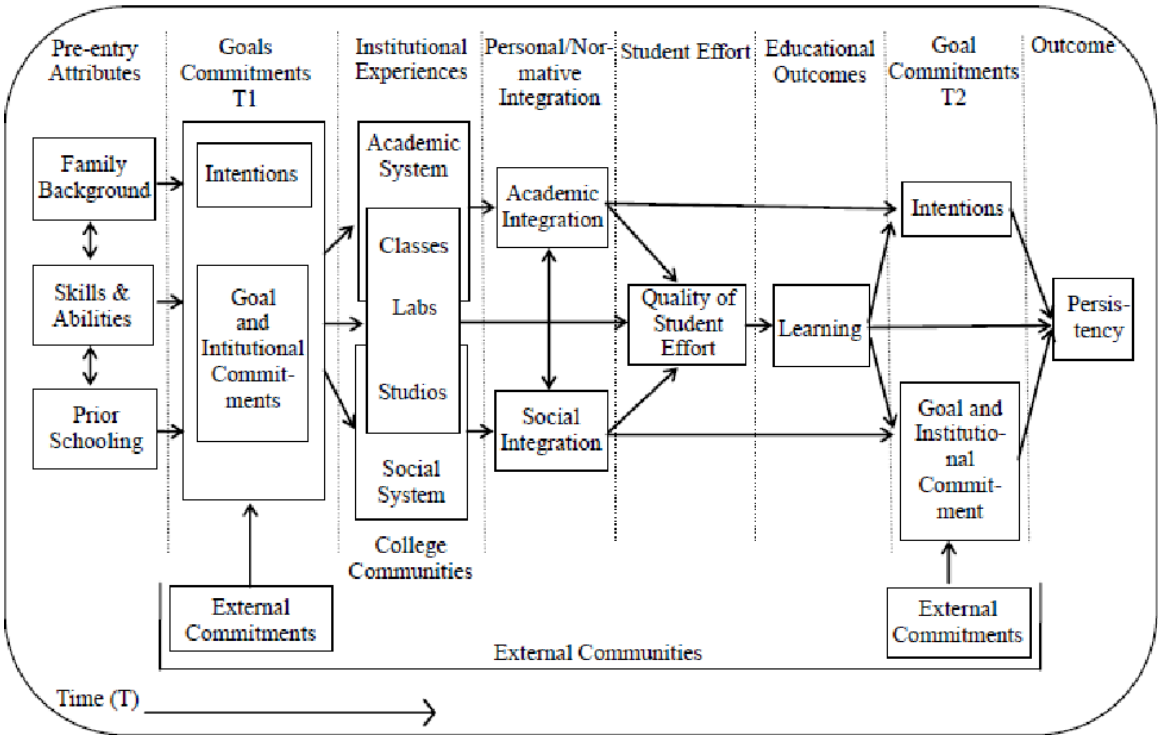


Figure 2.10: Conceptual framework illustrating student dropout in higher education

2.5.1.1.2 Student Engagement Theory

Student Engagement Theory emphasizes the role of active participation in learning activities as a key factor influencing academic success and student persistence (Kuh, 2009). The theory suggests that students who are actively engaged in their learning process are more likely to develop deeper understanding, maintain motivation, and complete their academic programs.

Student engagement is often examined through behavioral, emotional, and cognitive dimensions. Behavioral engagement includes activities such as attending classes, submitting assignments, and participating in discussions. Cognitive engagement refers to the level of effort students invest in understanding learning materials, while

emotional engagement relates to students' sense of belonging and interest in their studies.

In online learning environments, engagement can be measured through indicators such as login frequency, time spent on learning platforms, participation in discussion forums, and interaction with course materials. These engagement indicators provide important features that can be used in predictive models aimed at identifying students at risk of attrition.

2.5.1.1.3 Self-Regulated Learning Theory

Self-Regulated Learning Theory focuses on the ability of learners to independently manage their learning processes. The theory suggests that successful learners actively plan, monitor, and evaluate their learning activities, enabling them to adapt their strategies and maintain progress toward academic goals (Zimmerman, 2001).

Self-regulated learning involves skills such as goal setting, time management, self-monitoring, and motivation. In online learning environments, these skills are particularly important because learners are required to manage their own study schedules and maintain engagement without constant supervision.

Indicators such as time spent accessing learning materials, frequency of revisiting course content, completion of online assessments, and patterns of learning activity can provide insights into students' self-regulated learning behaviors. These indicators are valuable in predictive modelling because they capture behavioral patterns associated with persistence or dropout in online education.

2.5.1.2 Summary of Theoretical Perspectives

The theoretical perspectives discussed above provide complementary insights into the factors influencing student persistence and attrition in higher education. Tinto's Theory of Student Departure emphasizes the importance of academic and social integration within the institution as key determinants of student persistence. According to this perspective, students who actively engage with the academic environment and develop meaningful interactions with peers and instructors are more likely to remain in their programs of study.

In addition, Student Engagement Theory highlights the role of active participation in learning activities as an important predictor of academic success. Students who frequently interact with course materials, participate in learning discussions, and complete academic tasks tend to demonstrate higher levels of commitment to their studies.

Similarly, Self-Regulated Learning Theory emphasizes the importance of learners' ability to independently manage their learning processes. Skills such as time management, goal setting, and self-monitoring enable students to maintain consistent engagement with their studies, particularly in online learning environments where learners are expected to take greater responsibility for their own progress.

Together, these theories provide a conceptual foundation for understanding the academic, behavioral, and self-directed learning factors that influence learner persistence in higher education. In this study, these theoretical perspectives guide the identification of key variables related to learner engagement, academic participation, and learning behaviors. These variables form the basis of the dataset used to train the

proposed ensemble model combining gradient boosting and neural networks for predicting learner attrition in online learning environments within Kenyan universities.

2.6 The important factors that influence learner's attrition rates in online learning

1. **Demographic Information**

Learner attrition in online university programs is significantly impacted by demographic factors like age, gender, location, socioeconomic status, and employment status (Packham et al., 2004) Older students may find it challenging to continue taking online courses because of their increased responsibilities, such as those pertaining to their jobs or families (Xavier & Meneses, 2022).

Gender differences also come into play. Research suggests that female students may have trouble managing their time and multitasking, which could have an impact on their engagement (Farrell & Brunton, 2020). Higher dropout rates may also result from students from different geographic locations having differing levels of access to essential resources (Bawa, 2016). Another crucial factor to take into account is socioeconomic status; students from lower-income families may experience financial difficulties or lack the resources necessary to continue, which raises attrition (Petro et al., 2020). Students' time management and dedication to their programs may be impacted by their status as full-time or part-time learners, which could lead to a higher dropout rate.

2. **Academic Performance**

One reliable measure of attrition in online learning environments is academic performance. Sometimes, a student's past academic performance that is, their GPA or

grades in the pertinent subject matter areasdetermines their preparedness for and likelihood of success with online learning (D. Kim et al., 2020). Grades and course progress are examples of current academic success that offer valuable information about a student's motivation and attrition risk (Caruth, 2018). Participation in assessments, such as quizzes, tests, and assignments, shows students' dedication and involvement and directly affects their likelihood of sticking with the information (Chapman & Andrade, 2024). Furthermore, students' ability to manage independent and asynchronous learning habits is crucial for success and perseverance in distant learning (Errabo et al., 2024).

3. **Engagement Metrics**

Learner retention in online learning is mostly determined by student engagement metrics, such as how often students access the course, interact with the material, participate in discussion forums, communicate with peers and instructors, and use support services. Low attrition rates and high engagement are linked to frequent LMS logins (Talebi et al., 2024). As a metric of student involvement, active participation in course materials including readings, videos, and assignments can also be used to predict attrition risk (Villegas-Ch et al., 2024).A sense of community and belonging is facilitated by participating in online discussion forums, and this is crucial for preventing dropout rates (Conceição & Biniecki, 2024). Active interaction with classmates and teachers, such as prompt responses and group project participation, improves a student's resilience and sense of belonging (Ravishankar et al., 2024). According to Dunlap (2024), students' perseverance and achievement in online courses are also favorably correlated with their use of support services, such as academic advice or tutoring.

2.7 Conceptual Framework for Predicting Learner Attrition

The conceptual framework for this study illustrates the key factors that influence learner attrition in online courses. It identifies the relationships between various independent variables, a mediating variable, and the dependent variable, providing a foundation for understanding how student dropout can be predicted. The conceptual framework described here is summarized in Figure 2.11, which visually presents the relationships among these variables.

The independent variables in this framework include:

1. **Demographic Information** – age, gender, geographic location, socioeconomic status, and employment status. These factors may affect students' ability or motivation to continue with their courses.
2. **Academic Performance** – previous academic history, participation in assessments, and adaptability to different learning styles. Students with stronger academic performance may be less likely to drop out.
3. **Engagement Metrics** – frequency of course access, interaction with course materials, participation in discussion forums, communication with instructors and peers, and engagement in support activities. Higher engagement is expected to reduce attrition risk.

Psychosocial factors and motivation are considered a mediating variable. These factors may influence how the independent variables affect learner attrition. For example, a motivated student may overcome demographic or engagement-related challenges, reducing the likelihood of dropping out.The dependent variable is learner attrition, representing whether a student withdraws from a course or successfully completes it.

In this framework, demographic information, academic performance, and engagement metrics are hypothesized to influence learner attrition both directly and indirectly through psychosocial factors and motivation. By examining these relationships, the framework provides a comprehensive view of the elements that contribute to student dropout, guiding the design of predictive models and targeted interventions.

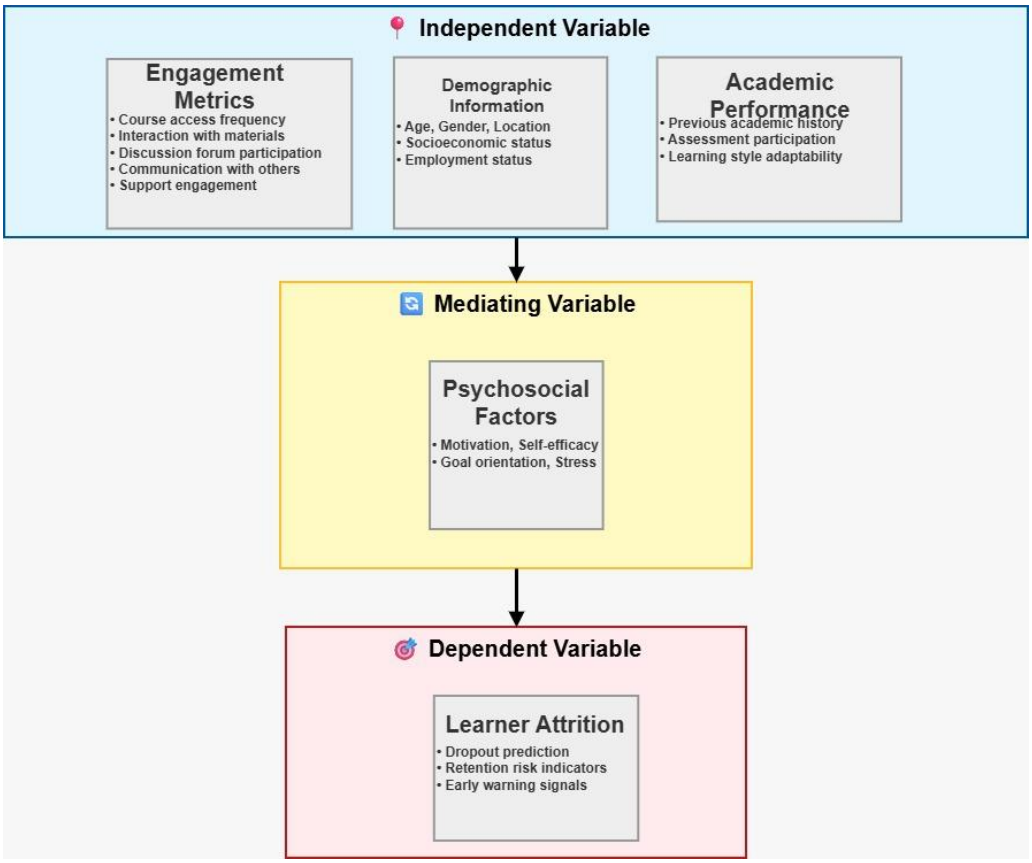


Figure 2.11: Conceptual Framework for Predicting Learner Attrition

2.8 Operationalization of Variables

Operationalization of variables involves defining the study variables in measurable terms to facilitate data collection and analysis. In this study, the variables were categorized into independent and dependent variables, with specific indicators and

measurement values used to assess learner attrition in online learning environments.

Table 2.2 presents the operationalization of the study variables.

Table 2.2: Operationalization of Variables

Variable	Sub-Variable	Indicators	Values
Independent Variables			
Demographic Information	Age	Age groups	Number 18-24, 25-34, 35-44, 45+
	Gender	Male, Female,	Categorical: Female/male If female 1 and if male 0
	Geographic Location	Region	1-Urban 2- Suburban 3- Rural
	Socioeconomic Status	Household Income, Access to Technology	1-Low 2-Middle, 3-High Yes/No
	Employment Status	Full-time, Part-time, Unemployed, Self-employed	Categorical: 1-Full-time, 2-Part-time, 3-Unemployed, 4-Self-employed
Academic Performance	Previous Academic History	KCSE, Grades in Relevant Subjects	Continuous: KCSE Grade A-E

	Participation in Assessments	Number of Assessments Submitted, Assessment Scores	Numerical: Count, Percentage Scores
	Learning Style Adaptability	Adaptability to Online Learning, Self-Directed Learning Capability	Categorical: 3-High 2-Moderate 1-Low
Engagement Metrics	Course Access Frequency	Number of Logins per Week/Month	Numerical: Count (for example 5 logins/week)
	Interaction with Course Materials	Time Spent on Videos, Readings, and Assignments	Continuous: Hours Per Week/Month
	Participation in Discussion Forums	Number of Posts and Replies in Forums	Numerical: Count of Posts, Replies
	Communication with Instructors and Peers	Frequency of Emails, Messages, or Meetings	Numerical: Count per Week/Month
	Engagement in Support Services	Utilization of Tutoring, Counseling, and Academic Advising Services	Categorical: Yes/No, Frequency of Usage
Intervening Variable	Psychosocial Factors and Motivation	Self-Efficacy, Goal Orientation, Perceived Value of the Program	Categorical: High, Moderate, Low
		Support Network	Categorical: Strong, Moderate, Weak

Dependent Variable	Learner Attrition	Dropout Rates, Completion Rates, Continued Enrollment	Percentage (%Dropouts), Percentage (%Completions), Yes/No For Continued Enrollment

The conceptual framework was created to highlight the main elements influencing student attrition in online learning after a comprehensive review of the literature. Students' persistence in online programs is most influenced by demographic factors such as age, gender, socioeconomic status, occupation, and location. Time and technology availability can be a problem for older students, students from low-income families, and students from underprivileged neighborhoods.

Second, learning performance is a great indicator of attrition; poor grades, little progress, or insufficient preparation all predict a higher chance of leaving. Last but not least, key indicators of student dedication are engagement metrics such as the frequency of logins, interaction with the course content, involvement in discussions, communication with teachers, and use of help. Low levels of participation frequently precede attrition.

This study's prediction model, which uses ensemble models specifically, gradient boosting and neural networks to identify vulnerable learners and recommend early interventions, is based on these combined criteria.

Table 2.3: Review of Factors that Influence Learners' Attrition Rates in Online Learning

Title of Paper	Theme of Research	Year of Study	Country of origin	Impact of this Work
Factors that Contribute to Students' Attrition in Open and Distance Learning (ODL) Environment	Challenges of OLD learning environments.	2022	Malaysia	Helped highlight structural issues that institutions need to address in ODL settings.
Factors affecting student dropout in MOOCs	Understanding causes of dropout in MOOCs	2019	South Africa	Drew attention to the high dropout rates in MOOCs and the need for better learner engagement strategies.
Early dropout prediction for students based on their online learning environment interaction logs	Using interaction logs to predict early dropouts in online education	2019	China	Provided a foundation for predictive analytics in online education.
Factors Contributing to Student Retention in Online Learning and Recommended Strategies for Improvement: A Systematic Literature Review	factors that affect online learning completion and retention.	2019	USA, Virginia	Offered a wide range of tested strategies to improve retention in online learning.
Factors Affecting Students' Preferences for Online and Blended Learning: Motivational vs. Cognitive	Learner preferences impacted by cognitive and motivational elements	2022	Europe	Showed how learners' motivations shape their preference for online or blended models.

Persistence and Dropout in Higher Online Education	Dynamics of persistence and dropout in higher education	2022	Pakistan	Contributed evidence on how persistence and dropout are linked to institutional practices.
Factors affecting students' intentions to undertake online learning: an empirical study in Vietnam	In online learning, learner motivation and intention	2021	Vietnam	Improved understanding of how motivation influences willingness to pursue online learning.
The usefulness of the logit leaf model in predicting student dropout in subscription-based online learning environments	Using AI models to predict dropout in subscription-based learning	2021	France	Advanced AI-based predictive tools for managing dropout risk in subscription models.
Dropout management in online learning systems	Techniques for reducing dropouts in online education	2021	India	Suggested practical interventions for institutions to actively manage student dropout.
Psychosocial Implications, Students Integration/Attrition, and Online Teaching and Learning in South Africa's Higher Education Institutions in the Context of COVID-19	Structure and psychological issues in online learning during COVID-19	2023	South Africa	Highlighted how psychosocial factors worsened dropout risk during the pandemic.
Dropout in online higher education	Analysis of online higher education dropout rates	2024	Iran	Provided a current picture of dropout challenges in

				online higher education.
The effectiveness of e-tutoring in an open and distance e-learning environment: evidence from the university of south Africa	E-tutoring's effectiveness in ODL settings	2021	South Africa	Demonstrated that e-tutoring can boost learner support and completion rates.
Investigating the E-Learning Challenges Faced by Students during COVID-19 in Namibia	Online learning obstacles during COVID-19	2021	Namibia	Brought out regional barriers such as access and connectivity during COVID-19.
The MOOC dropout phenomenon and retention strategies	MOOC dropout trends and retention tactics	2021	China	Proposed targeted strategies to improve retention in massive online courses.
Rurality and Dropout in Virtual Higher Education Programmes in Colombia	Rural context's effect on dropout rates from virtual schooling	2022	Colombia	Showed how rural challenges contribute to dropout in online higher education.
Charting the Course of School Dropout Research: A Bibliometric Exploration	Bibliometric examination of dropout research trends and gaps	2024	Romania	Mapped global research directions, showing where more work is needed.
The effectiveness and efficiency of student support services in open distance learning institutions in Africa	Assessment of student assistance programs in ODL institutions in Africa	2021	South Africa	Proved the importance of strong student support systems for retention.
Online Learning in Higher Education during COVID-19	Institutional reaction to COVID-19's online education	2021	Ghana	Illustrated how institutions adapted quickly to

Pandemic: A case of Ghana				sustain learning.
Learner Dropout in South African Schools	Examining the reasons behind school dropouts	2021	South Africa	Added insights into non-tertiary dropout, useful for wider education policy.
Risk factors associated with first-year students' intention to drop out from a university in South Africa	Risk factors for first-year university students' intention to drop out	2020	South Africa	Identified vulnerable groups among first-years who need early support.
Persistence in a Game-Based Learning Environment: The Case of Elementary School Students Learning Computational Thinking	Persistence in computational learning environments based on games	2020	Europe	Showed game-based learning can enhance persistence in younger learners.
School–community interventions to curb learner dropout: The perceptions of key education stakeholders in a rural South African school neighborhood	Community-school projects' contribution to lower learner dropout rates	2021	South Africa	Demonstrated how community engagement can reduce dropout in rural schools.
Student Support as a Panacea for Enhancing Student Success in an Open Distance Learning Environment	Support from students as a tactic to increase ODL success	2023	United States	Reinforced the role of strong academic support in ensuring ODL success.
Course satisfaction and perceived learning among distance learners in Malaysian Research Universities	Learning outcomes and course satisfaction in Malaysian ODL	2024	Malaysia	Linked satisfaction to stronger learning outcomes and persistence.

Examining Elements and Indicators to Assess Students' Long-Term e-Learning Engagement	Metrics and indicators of long-term e-learning student involvement	2019	Korea	Developed measures for tracking long-term learner engagement.
Management of School Environmental Factors on Dropout Rates on Public Primary Schools in Kuresoi South Sub County, Kenya	Environmental factors' impact on elementary school dropout rates	2021	Kenya	Pointed out how school conditions directly affect dropout in rural Kenya.
TheImpact of Social Media on Secondary School Dropout Rates in Kenya	Social media's effect on dropout rates in secondary schools	2023	Kenya	Showed how uncontrolled use of social media contributes to learner dropout.
Obstacles to successful uptake of open distance and e-learning (odel) programmes: a case of Kenyatta university, Kenya	Obstacles to Kenyan ODeL program enrolment and completion	2022	Kenya	Identified institutional and learner-level challenges limiting ODeL success.

2.9 Conclusion

The efficacy of current algorithms is constrained by notable gaps in the expanding corpus of research on student dropout prediction.Many studies have focused primarily on demographic and academic information, such as age, gender, grades, and prior academic performance, while largely ignoring behavioral data, including students' engagement patterns, interactions with online platforms, and clickstream activity. These behavioral indicators can provide rich insight into how students engage with their learning environment and can be critical for predicting who is at risk of dropping

out. By overlooking these patterns, current models may miss early warning signs that could enable timely interventions.

Another difficulty is that the intricate, non-linear correlations found in educational data are difficult for conventional models like K-nearest neighbour or logistic regression to capture.Student attrition is rarely influenced by a single factor; instead, it arises from the interaction of multiple academic, behavioral, and socio-economic variables. Such complexity can be handled by modern algorithms like Random Forests, Decision Trees, gradient boosting machines like XGBoost, and neural networks. Because they can spot subtle patterns and interactions that more simplistic models might miss, they are better suited for accurate and practical predictions.

In addition, most existing studies have evaluated models based primarily on accuracy, often ignoring other important metrics like precision, recall, F1-score, and the Area Under the Curve (AUC). These metrics are essential for understanding how well a model can identify students at risk, minimize false alarms, and provide reliable predictions that educators can act upon. A model that is accurate overall but performs poorly in identifying at-risk learners may offer little practical value in real-world settings.

By combining advanced machine learning algorithms with thoughtful feature selection and rigorous evaluation metrics, researchers can build models that are both powerful and interpretable. Because they can reveal hidden patterns in big, complicated datasets and provide insights that can guide actions, algorithms like XGBoost, Random Forests, and neural networks are especially promising.In addition to increasing prediction accuracy, these models give administrators and teachers practical advice on how to help children before they lose interest.

In conclusion, addressing these gaps by integrating behavioral data, applying modern machine learning techniques, and evaluating models with comprehensive metricscan significantly improve the prediction of student dropout. This holistic approach provides a more accurate understanding of learner behavior and enables targeted interventions that can reduce attrition, improve student success, and strengthen the overall effectiveness of educational programs.

CHAPTER THREE: METHODOLOGY

3.1 Introduction

Research methodology refers to the scientific process used to design and carry out a study. It includes the philosophical assumptions and theoretical framework that influence the research methodology. Research methodology is an orderly methodology for data gathering and evaluation for obtaining valid conclusions (Tracy, 2024). Including the research philosophy, design, sampling strategy, sample size, research instrument, data collection process, and data analysis techniques used in the study, this chapter presents the research methodology used in the proposed study.

3.2 Research Philosophy

According to (Ponterotto, 2005), chosen research philosophy plays a crucial role in deciding the researcher's assumptions and influencing the selection of tools, instruments, participants, and research methodologies. It provides a basic framework that decides the overall research approach. Each of the four main research philosophies pragmatism, positivism, interpretivism, and realism is distinguished by its methodology, ontology, and epistemology. In order to maintain consistency throughout the research process, the research philosophy must be decided upon early on in the study, as shown in Figure 3.1.



Figure 3.1: The Research Pyramid

The positivist research philosophy, which is based on the tenets that research must be impartial and the researcher must be detached from the phenomenon being studied, was applied in this study. Creating explanations that allow for the prediction and control of phenomena is positivism's main goal (Ponterotto, 2005).This paradigm is quantitatively dominated and is largely associated with experimental research designs. As per (J. Jonker & B. Pennink,2021), the research philosophy that one chooses influences the use of research design and methods used in the resolution of the research issues. Therefore, the researcher carefully considered whether certain methodological paradigms were pertinent to the study problem and whether the chosen research design was appropriate while choosing a suitable research philosophy.

3.3 Research Design

This study adopts a predictive modelling research design to develop and evaluate an ensemble machine learning model for predicting learner attrition in online learning environments. The choice of this design is informed by the nature of the research problem, which focuses on identifying patterns in learner behaviour and academic performance that may indicate a risk of dropping out.

Predictive modelling designs are widely used in machine learning research, where the objective is to build models capable of forecasting outcomes based on historical or observed data. In this study, primary data related to learner engagement, assessment performance, and interaction patterns was collected and prepared for analysis. The data underwent preprocessing and feature selection to identify the most relevant variables for predicting attrition.

To implement the experimental design, the dataset was divided into training and testing sets using an 80:20 split. Additionally, k-fold cross-validation (k=5) was applied to the training data to optimize model parameters and ensure robustness. This approach allowed the study to systematically evaluate the performance of each base model and the final ensemble model under multiple iterations, reducing the risk of overfitting and enhancing generalizability to unseen data.

Following this process, an ensemble model combining Decision Trees, Random Forests, Neural Networks, and Gradient Boosting techniques was developed. The performance of the model was evaluated using commonly used classification metrics, including accuracy, precision, recall, F1-score, and sensitivity. These evaluation measures helped determine how well the proposed model identifies learners at risk of dropping out of online courses.

By explicitly incorporating a quantitative experimental design with train-test splitting and cross-validation, this study provides a systematic framework for building, testing, and validating the proposed ensemble model for predicting learner attrition in online learning environments.

3.4 Location of Study

The research was conducted in seven universities in Kenya that offer online learning programs. The universities include the Mount Kenya University, Daystar University, Zetech University, KCA University, Dedan Kimathi University ,Strathmore University and Africa Nazarene University. The information was collected from students pursuing online learning programs.

3.5 Target Population

In this study, the term “target population” referred to the complete set of individuals from which data was collected. The target population consisted of students enrolled in online learning programs at seven Kenyan universities: Mount Kenya University, Daystar University, Zetech University, KCA University, Dedan Kimathi University, Strathmore University and Africa Nazarene University.

These universities were selected because they had well-established and operational online learning programs, making them suitable for analyzing learner attrition. Although the Open University of Kenya (OUK) is the only fully online university in the country, it was not included in this study because it only started offering online programs in 2023, and no cohort had completed a full program at the time of data collection. As a result, there was insufficient data available to meaningfully include it in the analysis. Focusing on the seven universities provided a diverse yet manageable sample, offering a solid basis for understanding the factors that contributed to attrition in online learning environments within the Kenyan context.

3.6 Sampling Techniques

This study employed a combination of purposive, stratified, and simple random sampling techniques. The techniques were applied in stages to ensure that the participants selected for the study were relevant to the research objectives while also providing adequate representation of students enrolled in online learning programmes in Kenyan universities.

The first stage involved the use of purposive sampling to identify the universities that would participate in the study. Universities were selected deliberately because they offer established online learning programmes and maintain digital learning environments where relevant student data could be obtained. Institutions that met these criteria and were accessible for data collection were considered appropriate for the study. Based on these considerations, seven universities were selected: Mount Kenya University, Daystar University, Zetech University, KCA University, Dedan Kimathi University, Strathmore University, and Africa Nazarene University.

After identifying the participating universities, stratified sampling was applied. In this case, each university constituted a separate stratum. The purpose of stratification was to ensure that students from all the selected universities were represented in the study and to reduce the possibility of over-representing or under-representing any particular institution.

Once the strata had been established, a sampling frame consisting of students enrolled in online learning programmes was obtained from institutional records and learning management system databases within the participating universities. From this list, simple random sampling was used to select the students who would participate in the study. The random selection process ensured that each student in the sampling frame

had an equal chance of being included in the study, thereby reducing the likelihood of selection bias.

The combination of purposive, stratified, and random sampling techniques allowed the study to focus on institutions that were relevant to the research while still ensuring fair representation of students across the selected universities. This approach strengthened the reliability of the data collected and improved the potential generalisability of the findings.

3.7 Sampling Size

The sample size for this study was determined using the Cochran sample size formula. The Cochran formula is widely used in social science research to estimate an appropriate sample size when the objective is to obtain a representative sample from a large population while maintaining a specified level of confidence and margin of error (Cochran, 1977). The formula is particularly suitable for studies that aim to estimate population proportions, such as the proportion of students likely to experience learner attrition in online learning environments.

The Cochran formula is expressed as:

$n_0 = (Z^2\ pq) / e^2$ (3.1)

Where:

- **n_0** is the desired sample size,
- **p** epresents the estimated proportion of students exhibiting the characteristic of interest (learner attrition)
- **q** = 1 – p represents the proportion of students not exhibiting the characteristic
- **e** represents the acceptable margin of error

- **Z** is the standard normal deviation corresponding to the desired confidence level.

For this study, a 95% confidence level was adopted, corresponding to a Z-value of 1.96. Since no prior estimate of the proportion of students experiencing attrition in the selected universities was available, p was set at 0.5. This value is commonly used in sample size determination because it produces the maximum variability and therefore yields the most conservative sample size estimate. Consequently, q was also set at 0.5, while the margin of error (e) was set at 0.05. Substituting these values into the Cochran formula produced a sample size of 338.16, which was rounded to 338 respondents.

The target population for this study consisted of approximately 5,000 students enrolled in online learning programs across the selected universities in Kenya, as summarized in Table 3.1. Although the Cochran formula is primarily designed for large populations, it provides a reliable estimate of the minimum sample size required to achieve statistically valid results. The calculated sample size was therefore considered adequate to represent the target population.

After determining the overall sample size, the respondents were proportionally distributed across the selected universities based on the number of students enrolled in their online learning programs. This approach ensured that institutions with larger student populations contributed more respondents to the sample while smaller institutions contributed proportionally fewer respondents.

Table 3.1: Target Population

Institution	Target Number	Sample Size
MountKenya University	1100	74
KCA University	900	61
Daystar University	800	54
Strathmore University	700	47
Zetech University	650	44
Dedan Kimathi University	500	34
Africa Nazarene University	350	24
-	5000	338

3.8 Instruments of Data Collection

According to Birmingham et al. (2022), research instruments are tools used to collect pertinent data for a study. The main instruments employed in this research were questionnaires and focus group discussions, each serving complementary purposes in capturing information about learner attrition in online learning environments. Questionnaires are particularly useful when collecting data on people's opinions and perceptions regarding specific topics. They are economical, systematic, and allow the collection of structured data from a large number of respondents efficiently (Birmingham & Wilkinson, 2021). The decision to use questionnaires in this study was guided by their ability to gather data on a broad array of issues in a straightforward

and structured manner, while also being relatively low-cost and easy to analyze compared to other methods such as interviews or content analysis.

The questionnaire used in this study consisted of 24 questions spread over four standard A4-sized pages (see Appendix I). The questions were organized into sections that captured general information, demographic details such as gender, age group, geographic location, socio-economic status, and employment status, academic performance including past academic background, participation in assessments, and learning style adaptability, engagement measures such as frequency of course use, interaction with course materials, discussion forum postings, communication with instructors and peers, and use of support services, psychosocial considerations and motivation, examining self-efficacy, goal orientation, perceived program value, and support structures, and finally learner attrition, exploring dropout considerations, reasons for dropping out, past completion of online programs, and continuation plans. Systematic questionnaires with open and closed-ended questions were used to allow for sufficient data collection.

In addition to the main questionnaire, a second questionnaire was developed specifically for the external validation of the ensemble model (see Appendix II). The purpose of this instrument was to evaluate how accurately the model predicts learners who are at risk of attrition. Unlike the primary questionnaire, which focused on collecting general data about learners' demographics, engagement, and experiences, this validation questionnaire was designed to capture targeted information related to key predictive factors such as learner participation, academic performance, and self-perceived risk of dropping out. The responses obtained were used to compare actual

learner behavior and perceptions with the model's predictions, thereby helping to assess the reliability and effectiveness of the proposed model in real-world settings.

In addition to the two questionnaires, focus group discussions were conducted with selected learners to gain deeper insights into the reasons behind learner attrition and to complement the quantitative data collected. Focus groups allowed participants to discuss their experiences, challenges, and motivations in an interactive setting, providing rich qualitative data that could not be captured through questionnaires alone. The focus group discussion guide is provided in Appendix II. Together, these instruments provided a comprehensive approach to collecting both quantitative and qualitative data, ensuring that the study captured the full range of factors influencing learner attrition in online learning environments.

3.9 Validity and Reliability of Research Instruments

3.9.1 Validity of Research Instrument

Validity refers to the extent to which a research instrument measures what it is intended to measure. In this study, the validity of the research instrument was assessed using face validity and content validity. Face validity focuses on whether the instrument appears appropriate for measuring the intended variables, while content validity evaluates whether the items included in the instrument adequately cover the concepts being investigated.

To establish validity, the questionnaire was reviewed by three specialists in the fields of machine learning, educational data mining, and research methodology. These experts were selected purposively based on their academic qualifications, research experience, and involvement in supervising postgraduate research in computing and

data science related disciplines. All the experts had experience in designing research instruments and evaluating studies related to technology-enhanced learning environments.

The experts examined the questionnaire to determine whether the items were clear, relevant, and appropriate for measuring factors associated with learner attrition in online learning environments. Each expert evaluated the instrument using a rating scale ranging from 1 to 10, where 1 represented very poor validity and 10 represented excellent validity.The ratings provided by the experts were averaged to determine the overall validity of the instrument. For the questionnaire to be considered acceptable for use in the study, the average score was required to exceed 70 percent. The results of the expert evaluation are summarized in Table 3.2.

Table 3.2: Validity of Research Instrument

Expert	Face Validity Score	Content Validity Score	Decision Taken
Expert I	8	7	Accepted
Expert II	7	8	Accepted
Expert III	8	8	Accepted
Average	7.67	7.67	**Accepted**

Following the expert review, several questionnaire items were revised to improve clarity and alignment with the objectives of the study based on the feedback provided by the specialists. This process helped ensure that the instrument adequately captured the key variables required for investigating learner attrition in online learning environments.

3.9.2 Validity of the Study

To ensure the validity and reliability of the study, several measures were implemented. Internal validity was achieved through rigorous data preprocessing, feature selection, and the use of k-fold cross-validation, which minimized the effects of confounding variables and reduced the risk of overfitting. External validity was addressed by using a diverse dataset representing multiple courses and student demographics, and by evaluating the model on unseen test data to assess generalizability to other online learning environments. Construct validity was ensured by carefully defining engagement, academic performance, and interaction variables based on established literature, and by using appropriate classification metrics to accurately measure predictive performance. Criterion-related validity (predictive validity) was established by comparing the model's predictions with actual learner dropout outcomes. Additionally, convergent and discriminant validity were considered by confirming that related variables correlated as expected while unrelated variables did not, ensuring the integrity of the measurement constructs. Collectively, these measures ensured that the developed ensemble model is robust, reliable, and capable of accurately predicting learner attrition in online learning environments.

3.9.3 Reliability of Research Instrument

Reliability refers to the degree to which a research instrument produces consistent and stable results when applied under similar conditions. In this study, the reliability of the questionnaire was assessed through a pilot study and subsequent analysis using Cronbach's Alpha, a widely used measure of internal consistency in social science research. Cronbach's Alpha was chosen because the questionnaire consisted of

multiple Likert-scale items designed to measure related constructs associated with learner attrition, such as demographic, academic, engagement, and psychosocial factors (Tavakol & Dennick, 2011). It is particularly suitable for instruments where several items aim to measure the same underlying construct.

Before conducting the pilot study, the questionnaire was reviewed by the research supervisors and specialists in educational technology, machine learning, and research methodology. These experts, selected based on their academic qualifications and experience, examined the items to ensure clarity, relevance, and appropriateness for collecting data on factors influencing learner attrition in online learning environments.

A pilot study involving 15 students from Mount Kenya University was then conducted. Mount Kenya University was selected because it offers active online learning programs similar to those in the other universities included in the study population, ensuring that the pilot responses would be relevant and reflective of the target population. The sample size of 15 participants was deemed sufficient based on methodological guidelines, which recommend that pilot studies typically involve between 10 and 30 participants (Hertzog, 2008). The purpose of the pilot study was to identify ambiguities, assess clarity, and detect any potential weaknesses in the questionnaire items before full-scale data collection.

The reliability of the instrument was assessed using Cronbach's Alpha. The analysis considered 24 questionnaire items measuring demographic, academic, engagement, and psychosocial factors. The results indicated a Cronbach's Alpha of 0.82, and a Cronbach's Alpha based on standardized items of 0.84, demonstrating satisfactory internal consistency. According to Nunnally (1978), a Cronbach's Alpha value of 0.7

or higher is considered acceptable for establishing reliability in social science research. The reliability statistics are summarized in Table 3.3.

Table 3.3: Reliability Statistics

Reliability Measure	Value	Decision
Number of Items	24	-
Cronbach's Alpha	0.82	Accepted
Cronbach's Alpha Based on Standardized Items	0.84	Accepted

Establishing reliability was critical to ensure that the questionnaire would produce consistent and dependable results. The pilot study and reliability analysis allowed for the identification and correction of potential weaknesses in the instrument. By minimizing measurement errors, the study enhanced the accuracy and interpretability of the collected data.

3.10 Data Sources

The data utilized in this research was obtained from both primary and secondary sources across seven Kenyan universities that offer online courses, namely Mount Kenya University, Daystar University, Zetech University, KCA University, Strathmore University, Dedan Kimathi University, and Africa Nazarene University. Secondary data were collected from the universities' Learning Management Systems (LMS), providing detailed records of students' academic performance and engagement in online courses. Primary data were gathered through questionnaires and focus group discussions, allowing for both structured quantitative insights and in-depth qualitative perspectives on learner experiences and attrition. The questionnaires captured information on demographic characteristics, academic performance, engagement, and

psychosocial factors, while the focus groups offered an interactive platform for learners to discuss challenges, motivations, and experiences in their online learning journeys. The focus group discussion guide is included in Appendix II.

The combined dataset encompassed a wide range of variables, providing a comprehensive view of each learner's background, academic history, and engagement patterns. Demographic information, such as age, gender, geographic location, socio-economic status, and employment, enabled an analysis of how personal circumstances influence persistence and completion in online programs. Academic performance data included previous academic history, ongoing assessment results, and indicators of learning style adaptability, allowing for exploration of how prior achievements and individual learning approaches affect progress. Engagement measures, such as frequency of course access, interaction with course materials, contributions to discussion forums, communication with instructors and peers, and participation in support activities, provided rich insights into students' day-to-day involvement.

Together, the primary and secondary data created a detailed and integrated profile of learners across the seven universities, forming a strong foundation for analyzing factors influencing retention and attrition in online learning environments. By combining quantitative and qualitative methods, the study ensured a thorough understanding of both measurable patterns and nuanced learner experiences

3.10.1 Data Collection Procedure

After obtaining authorisations and approvals, data gathering was completed. There were two phases to the processBefore starting the first phase of data collection, the researcher obtained approval from the relevant government agencies and authorities.

To obtain the necessary permits and letters, the researcher went to the offices in person. These included authorisation letters from the County Commissioner and County Director of Education in each of the sampled counties, as well as a permit and research authorisation letter from the National Commission for Science, Technology, and Innovation (NACOSTI). In order to obtain authorisation to gather data, an introduction was also shown to the head of each sampled institution.

The second phase involved distributing the questionnaires to the universities, administering them to participants, and collecting the completed responses. To support the administration and collection of the questionnaires, research assistants were employed. The entire data collection process took one month.

3.11 Machine Learning Methodology for Developing the Prediction Model

Machine learning process is a systematic approach towards constructing prediction models. There are six main phases in the process: data gathering, data cleaning, feature engineering, training models, selecting models, and presentation of the ultimate model (IBM, 2024; Kansab, 2025). The steps for constructing machine learning models are shown in Figure 3.2, which presents the complete machine learning pipeline used in this study.

- Phase One - Student Data Collection: Raw data is gathered from various sources in this step. The raw data is converted into digital format and verified for accuracy and completeness.
- Phase Two - Data Pre-processing: Digital data is cleaned to transform it into a structured dataset in the machine learning-convenient format. Missing and incomplete data are handled at this step to ensure consistency.

- Phase Three - Feature Selection: Also known as dimensionality reduction, this phase identifies the most important features and eliminates redundant or irrelevant features. The process enhances model performance by improving prediction quality and model simplicity.
- Phase Four - Model Training: The selected features are utilized to train machine learning algorithms and produce multiple candidate models. The training is an iterative process with the dataset for optimizing model performance.
- Phase Five - Model Validation**:** In this phase, the models that were previously selected are trained and validated to determine the best-performing one. Using the best feature set, the models undergo iterative training and tuning to optimize their predictive capability.
- Phase Six - After training and selecting the best-performing model, the final predictive model is presented. The model demonstrates strong generalization performance through evaluation metrics such as precision, recall, F1-score, ROC-AUC, and confusion matrix analysis, indicating its deployability and reliability on unseen learner data.

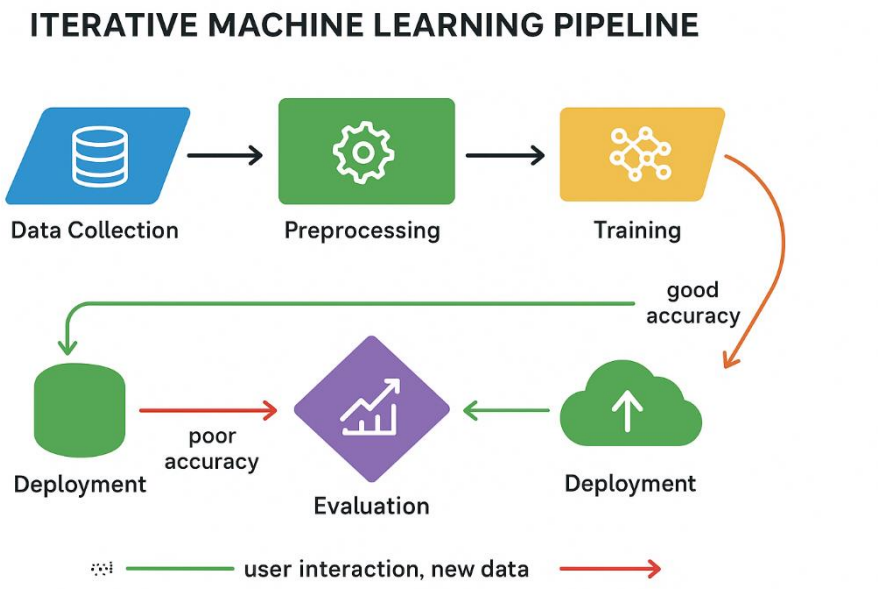


Figure 3.2: Machine learning pipeline

3.11.1 Data Preparation

The third phase of the data mining process focuses on data preparation. In this stage, the features used for training the model were identified and organized, while their size, distribution, and any missing values were carefully examined. To minimize the load on the production database, it was necessary to extract data in stages and then combine it later. Figure 3.3 shows this process, which involved three steps: extracting basic student information, extracting attendance and grade data, and integrating all the information into a single dataset for analysis.

Working with simple data such as dates, course of study, credits, and other basic attributes was time-consuming because of the large size of the dataset. To speed up this process, materialized views were used. This approach allows frequently accessed data to be extracted and stored as comma-separated values (CSV) files, making subsequent processing much faster and more efficient.

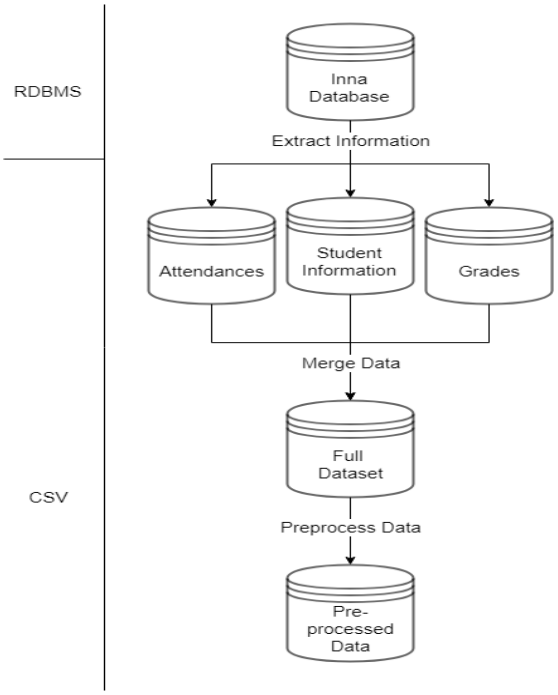


Figure 3.3: Data Lake (Raw Data)

3.11.2 Data Pre-processing

One of the most important steps in getting the dataset ready for analysis and model construction was data pre-processing. If left unchecked, the dataset's abnormalities and inconsistencies could have a detrimental impact on prediction accuracy. During this phase, a number of methodical actions were conducted to guarantee the quality of the data. Managing missing values, which frequently occur in big datasets, was the first step. Various methods were used to deal with this, based on the type of variable and the degree of missingness. For numerical and categorical values, mean or mode imputation was employed, respectively, and forward and backward fill techniques were taken into consideration where sequential data warranted them. Additionally, when appropriate, models that naturally accommodate missing values were investigated.The second step was removing duplicates to prevent redundancy. Duplicate entries not only inflate the dataset but also bias the results, making it

important to systematically eliminate them to maintain the uniqueness and reliability of the data. The third focus was on error correction. Data-entry mistakes such as typographical errors or inconsistent records were identified and rectified through strict validation rules. This step ensured that the dataset reflected accurate and trustworthy information. Once the data was clean, normalization and scaling were carried out. Because features within the dataset existed on different scales, it was necessary to bring them into a common range. This step was vital for many machine learning algorithms that rely on distance-based calculations and are sensitive to uneven feature scales. Further, data transformation was applied where appropriate. Categorical variables were encoded to numerical representations so that algorithms could interpret them effectively, while binning techniques were used in cases where continuous variables needed to be grouped into meaningful categories.

Finally, integration and final cleaning ensured that if data came from multiple sources, it was harmonized into a consistent and unified dataset. At the end of this stage, the data was clean, structured, and well-prepared, making it ready for the subsequent feature selection and modeling phases.

3.11.3 Merging Basic, Attendance, and Grade Information

After the basic information was extracted, the basic information was read into memory using the Pandas library. The attendance figures, assignment and semester grades was also be loaded into memory. The small and large period dates were then computed for every student using Pandas Data Frames. Both the periods' attendance figures were then be analyzed, along with the small and large period average assignment grades and

the large period average semester grade. The combined dataset was then exported to a comma-separated values (CSV) file.

3.11.4 Pre-processing Data

The pre-processing was carried out through a Scikit-learn Pipeline, which in turn was carried out transformations in order to prepare the data for machine learning algorithms. The pre-processing steps that was carried out are:

At the data pre-processing phase, fuzzy text matching were utilized to correct student-inputted names of schools that could have typographical mistakes. This was done using Python's built-in Sequence Matcher of the difflib module that allows approximate matching of strings. In addition, course names were normalized by applying a pre-determined dictionary mapping for uniformity in the data set. Handling missing values was accomplished by replacing them with the mean of other available values in order to maintain data integrity. Numerical values were normalized to one common range so that they can be compared, which is a critical part of the operation of machine learning algorithms. Finally, categorical values were converted to numerical values for convenience in training machine learning models.

3.11.5 Data Transformation

Data transformation involved converting raw data into a suitable form for modeling. This includes feature selection, feature extraction, and feature engineering:

During the data transformation phase, feature selection focused on the discovery and retention of the most informative features that significantly contribute to learner attrition to ensure that the model is built from the most informative information.

Feature extraction involved pulling out new features from the existing ones in a way that the model's predictive power can be improved, adding value by pulling out additional patterns or insights. Feature engineering was employed to improve the model's performance further using transformations such as encoding categorical variables, creating interaction terms between the features, and generating polynomial features to identify richer relationships in the data. These procedures were supposed to optimize the model's prediction of student dropout.

3.12 Model Selection

This study employed XGBoost, Neural Networks (NN), Decision Trees (DT), and Random Forests (RF) to predict learner attrition in online learning environments. Each model was selected based on its strengths and expected contributions to the ensemble approach. XGBoost was chosen for its ability to handle complex, non-linear relationships and large datasets, making it highly effective for structured educational data where interactions between student engagement, performance, and demographic features are important. Neural Networks were included because they excel at identifying hidden patterns and non-linear relationships within high-dimensional data, which is useful for capturing subtle trends in student behavior that may signal dropout risk. Decision Trees were incorporated due to their interpretability, allowing the study to identify and explain key factors that influence learner attrition. Random Forests were selected to improve prediction stability and reduce overfitting by combining multiple decision trees, which enhances overall model reliability.

The ensemble approach combines these individual models' outputs to create a stronger, more accurate predictive model. This strategy is particularly suitable for complex datasets, as it can effectively capture interactions among features while

balancing bias and variance. Other algorithms, such as K-Nearest Neighbors, Support Vector Machines (SVM), and Logistic Regression, were considered but not used due to lower predictive performance and increased computational complexity for this study's dataset.

By deliberately selecting these complementary models, the ensemble is expected to provide robust predictions while also offering insights into the key drivers of learner attrition.

3.13 Model Evaluation

To ascertain whether the system had been properly trained, a test data set was used to compare the expected output with the actual model output. Using the validation data set, error performance metrics were used to record any output discrepancies. The model output data was tested in a number of trials during the training phase.

To display the different performance metrics and the model's accuracy, a confusion matrix was created. The results' output was shown at the Knowledge Gap. This study made use of a number of performance metrics covered in the sections that follow.

To evaluate the performance of the selected models, the following performance measures was used:

1. Accuracy provided an overall evaluation of the model's correctness.
2. Precision used to evaluate how well the model predicts students who are likely to drop out.
3. Recall (Sensitivity) was critical to understand how many students at risk of attrition the model successfully identifies.

4. F1-score balanced precision and recall, ensuring that both false positives and false negatives are minimized.
5. ROC-AUC was used to assess the model's ability to distinguish between students who stayed and those who dropped out at various threshold settings.

3.13.1 Receiver Operating Characteristic (ROC) Curve

The ROC curve is a graphical plot was used to evaluate the performance of a binary classifier. It shows the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity) at various threshold settings. However, there is no single formula for the ROC curve itself, but the Area Under the ROC Curve (AUC-ROC) is often computed as:

$$AUC = \int_0^1 TPR(x).d\big(FPR(x)\big) \quad (3.2)$$

3.13.2 Accuracy

Accuracy in this study refers to how often the model correctly predicts a student's progression status. It is calculated as the proportion of correctly predicted cases, whether a student is progressing or dropping out, against the total number of students in the dataset. For example, when the model correctly identifies a student who progressed (true positive) or correctly classifies a student who dropped out (true negative), these are considered correct predictions. The value of n represents the total number of students in the dataset. Thus, accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{n}} \quad (3.3)$$

where TP is the number of progressing students correctly classified, TN is the number of dropouts correctly classified, and n is the total number of students.

3.13.3 F 1 Measure

The F1 score provides a balanced evaluation by combining precision and recall into a single metric. Precision in this study represents the proportion of students correctly identified as progressing out of all students that the model predicted to progress. Recall, also referred to as sensitivity, measures the proportion of students who actually progressed and were correctly identified by the model. Since precision and recall emphasize different aspects of prediction, the F1 score, being their harmonic mean, provides a balanced assessment, especially in cases where the number of progressing and dropping out students is uneven.

$$F_1 = 2.\frac{Precision.Recall}{Precision + Recall} \quad (3.4)$$

$$Precision = \frac{TP}{TP + FP} \quad (3.5)$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \quad (3.6)$$

3.13.4 Mean Squared Error (MSE)

The Mean Squared Error (MSE) was applied to evaluate the regression-based predictions in this study. It measures the average squared difference between the predicted probability of progression and the actual outcome, where progression is coded as one and dropout as zero. A smaller MSE indicates that the model's predicted probabilities are closer to the observed student outcomes.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3.7)$$

3.13.5 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is the square root of the MSE, providing an error metric that is expressed in the same scale as the predicted probability. This makes RMSE more interpretable when evaluating how far off predictions are from actual student progression outcomes.

$$RMSE = \sqrt{\frac{1}{n}\sum\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.8)$$

3.13.6 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) was used to compare the predictive models developed in this study. AIC considers both the accuracy of the model and the number of parameters it uses, rewarding models that achieve a good fit with fewer parameters. A lower AIC value suggests a better model because it balances predictive accuracy and simplicity, thus reducing the likelihood of overfitting.

$$AIC = 2k - 2\ln(L) \quad (3.9)$$

where k is the number of model parameters and L is the maximum likelihood of the model.

3.13.7 Specificity

Specificity measures how well the model identifies students who are at risk of dropping out. It is the proportion of students who actually dropped out and were correctly predicted as dropouts. A high specificity indicates that the model is effective at flagging students at risk.

$$Specificity = \frac{TN}{TN + F} \quad (3.10)$$

3.13.8 Sensitivity

Sensitivity, also called recall, complements this by focusing on students who progressed. It represents the proportion of students who actually progressed and were correctly predicted as progressing. Together, these two metrics provide a fuller picture of the model's performance.

$$Precision = \frac{TP}{TP + FP} \quad (3.11)$$

3.13.9 Positive predictive value

The Positive Predictive Value (PPV) evaluates how reliable the model's predictions of progression are. It measures the probability that a student actually progresses given that the model has predicted progression. In other words, if the model predicts that a group of students progress, PPV indicates how many of those predictions are indeed correct.

$$Precision = \frac{TP}{TP + FP} \quad (3.12)$$

3.14 Ethical Considerations

The ethical guidelines for informed consent, access and acceptability, and privacy and confidentiality were all followed in this investigation.The researcher received approval from the relevant university authorities in charge of postgraduate research for acceptance and access. Additionally, the relevant regulatory agencies provided the letters of authority and research permits that were required. Prior to inviting

individuals to take part in the study, institutional administrators at the chosen universities were asked to grant extra permissions.Regarding informed consent, each participant received comprehensive information on the study's goal, purpose, and voluntary participation. They were free to leave at any time without incurring any fees.

In terms of privacy and confidentiality, the information gathered was utilized solely for research purposes. Identifiable data was be anonymized to safeguard the identity of the participants, making their responses confidential.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results of the study that focused on developing an ensemble model for predicting learner attrition in online learning environments. It primarily addresses Objective II, which involved creating a robust predictive model that integrates Decision Trees, Random Forest, Gradient Boosting, and Neural Networks to identify learners at risk of early dropout. The development of the model was guided by the factors identified from existing literature (Objective I), including demographic characteristics, academic performance, and engagement-related behaviors, ensuring that the model captures the key variables known to influence attrition.

The chapter begins with an overview of the dataset, describing the characteristics of learners included in the study, including demographic profiles and patterns of engagement extracted from learning management systems. Following this, the data pre-processing procedures, such as cleaning, transformation, feature engineering, and encoding of variables, are discussed. The section then explains the methods used to select the most relevant predictors of learner attrition and the steps taken to develop the individual models Decision Trees, Random Forest, Gradient Boosting, and Neural Networks which together form the ensemble.

Comparative results of the models are presented using widely accepted evaluation metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve. These metrics serve to evaluate the predictive performance of the ensemble, laying the foundation for a more detailed assessment of model performance (Objective III) and its generalizability using cross-validation techniques (Objective IV) in the following chapters. The findings are interpreted with emphasis on how early identification of at-

risk learners can inform targeted interventions and improve retention and engagement in online learning environments.

4.2 Data Analysis

This section outlines the techniques and strategies employed to analyse the data collected for this investigation. The investigation's main objective was to identify patterns and linkages that could help explain student attrition in online learning environments. A mix of descriptive, exploratory, and inferential methods were employed in the data analysis process to understand the behavioural and environmental factors impacting learners' engagement and perseverance.

The process began with a comprehensive examination of the dataset to assess its quality, completeness, and structure. This process ensured that only relevant and reliable data were used in subsequent stages. The cleaned and organised dataset was then subjected to exploratory data analysis (EDA) to look for trends, spot anomalies, and compile the key characteristics of the students.

Following EDA, statistical and computational methods were used to prepare the data for model training in order to obtain more profound insights. These included creating derived features that captured engagement dynamics and performance metrics, normalising numerical data, and encoding category properties. The final step was to separate the data into training and testing subsets, which formed the foundation for developing and evaluating the prediction models used in this investigation.

4.2.1 Dataset Overview

The total number of datasets used in this study comprised of 20,000 records out of which 15,000 was found to be complete records, each free of missing values, which provided a solid foundation for robust and reliable analysis. The structure of the dataset encompassed both numerical and categorical variables, allowing for a comprehensive representation of learner behavior and background. Specifically, there were 7 numerical variables of the int64 data type, capturing key quantitative metrics such as login frequency, engagement levels, and assessment scores all of which are critical indicators of learner participation and performance. In addition, the dataset contained 12 categorical variables of the object data type, representing demographic information, academic background, and other learner-specific characteristics**.** This mix of variable types enabled a nuanced analysis, supporting the development of a predictive model that could account for both behavioral and contextual factors influencing early learner attrition.

4.2.2 Pre-processing the Learner Attrition Dataset

The initial stage of the research process began with the acquisition of secondary data, which formed the foundation of the entire analysis. However, raw data in its original state is rarely suitable for direct use. Before it could be applied to any meaningful study, it was necessary to undertake a series of preparatory steps. These steps, often referred to as data preprocessing, were essential in ensuring that the information was accurate, reliable, and ready for subsequent examination. Without this preparation, the risk of obtaining biased or misleading results would have been high, which could have undermined the validity of the research findings.

Resolving problems with missing values was one of the main preprocessing jobs. Incomplete records with certain variables left blank are common in datasets, particularly those assembled from secondary sources. Such gaps have the potential to skew analysis and impair machine learning model performance if left unchecked.In addition to handling missing entries, attention was also given to identifying and resolving inconsistencies within the data. These discrepancies could take the form of typographical errors, conflicting entries, or irregular formats that needed standardization. Correcting these issues ensured that the dataset maintained internal consistency and could support dependable conclusions.

Another important aspect of preprocessing involved transforming the dataset into a format that could be effectively processed by machine learning applications. This meant converting raw data into numerical or categorical values that algorithms could interpret, as well as restructuring it to fit the required input-output design. Care was also taken to ensure that the data adhered to the structural and statistical requirements of machine learning models, such as appropriate feature scaling and normalization. By the end of this stage, the dataset was not only cleaned and corrected but also reshaped into a machine-readable format, making it suitable for advanced analysis and model development.

4.2.3 Digitization of the Data

The dataset used in this study was obtained in Microsoft Excel format from institutional records that had previously been collected and stored.Excel, being a widely used data management tool, organizes information in a tabular layout, where each row represents an individual student’s record while each column corresponds to

a particular attribute or feature associated with that student. In this case, the dataset captured multiple aspects of student information that were relevant to understanding patterns of learner engagement and attrition.

Each column in the spreadsheet represented a different characteristic, such as demographic details, academic performance indicators, or behavioral variables linked to participation in the learning environment. The attributes were accompanied by truncated names to facilitate compact representation, and these were supplemented with descriptive explanations to clarify their meaning. For instance, variables that tracked course completion rates or assignment submissions were coded using shorter labels, but explanatory notes ensured that the interpretation remained clear and unambiguous.

Furthermore, the dataset defined a domain of acceptable values for every attribute, outlining the possible range that each variable could take. This step was essential in guiding the preprocessing and later stages of machine learning analysis, as it prevented misinterpretation of values that fell outside the intended scope. To give an organized view of the dataset's contents, Table 4.1 presents a structured summary of the attributes, their descriptions, and the respective value domains. This table serves as a quick reference for understanding the dataset structure before moving into deeper analytical processes.

Table 4.1: Attribute Description

No.	Attribute	Description
1	Record_ID	Unique identifier for each record.
2	Date	Date of the record entry.
3	Age_Group	The age range of the learner (18-24, 25-34, 35-44).
4	Gender	Gender of the learner (Male, Female).
5	Region	Geographic location (Urban, Suburban).
6	Household_Income	Income level of the learner's household (Low, Medium, High).
7	Access_to_Technology	The level of access to technology available to the learner (Limited, Good).
8	Employment_Status	The employment status of the learner (Full-Time, Part-Time).
9	KCSE_Grade	The grade of the learner in the Kenyan Certificate of Secondary Education (KCSE) exam.
10	Assessments_Submitted	The number of assessments the learner has submitted.
11	Assessment_Score	The average score of assessments submitted by the learner.
12	Learning_Adaptability	The adaptability of the learner to different learning methods (High, Low).
13	Logins_Per_Week	Number of logins made by the learner per week.
14	Time_Spent_Weekly	Amount of time spent by the learner on the platform each week (in hours).
15	Forum_Participation	The number of times the learner participates in online forums (Occasional, Frequent, Rare).
16	Emails_Sent_Weekly	The number of emails sent by the learner each week.
17	Support_Services_Usage	Usage of available support services by the learner (Yes, No).
18	Psychosocial_Support	The frequency with which the learner uses psychosocial support (Rare, Frequent).
19	Attrition_Status	The current status of the learner's enrollment (Completion, Continued).

[5]:

	Age	Gender	Previous_Education	Enrollment_Type	Logins_Per_Week	Time_Spent_Per_Week	Assignments_Submitted	Forum_Participation	Previous_Online_Experie
0	56	Female	Undergraduate	Full-time	0	14.7	5	NaN	
1	46	Female	High School	Full-time	1	17.5	4	NaN	
2	32	Male	Graduate	Part-time	0	14.5	2	Active	
3	25	Male	High School	Full-time	10	4.6	9	Active	
4	38	Other	Graduate	Full-time	8	14.1	9	NaN	
...	...	...	...	...	...	...	...	...	
1495	47	Other	Undergraduate	Full-time	5	3.5	3	NaN	
1496	56	Male	High School	Part-time	1	19.1	0	NaN	
1497	48	Male	Undergraduate	Part-time	6	1.4	1	Passive	
1498	35	Male	Graduate	Full-time	12	7.1	0	NaN	
1499	53	Female	Graduate	Part-time	5	10.1	5	Active	

Figure 4.1: Sample of learner data

4.2.4 Attrition Status Grouping

The Attrition Status variable served as the critical target attribute for predicting learner attrition in this study. It classified learners into three distinct categories based on their course progression: Continued**,** Completion**,** and Dropout**.** This categorical target variable enabled the model to distinguish between learners who were still active, those who had successfully completed the course, and those who had exited prematurely. Specifically, the data distribution showed that 28,839 learners were in the continued category learners who remained engaged in the course but had neither completed nor dropped out. This group was essential for understanding ongoing retention trends. The Completion category included 11,161 learners who successfully finished the course, providing insight into the characteristics and behaviors associated with course completion. Lastly, the Dropout category comprised 10,000 learners who exited the course before completion, representing a critical group for identifying early warning signs and informing targeted intervention strategies As illustrated in Figure 4.2. This

balanced distribution across the three classes enabled a meaningful analysis of learner behavior and attrition patterns.

```
]: df['Attrition_Status'].value_counts()

]: Attrition_Status
   Continued     28839
   Completion    11161
   Dropout       10000
   Name: count, dtype: int64
```

Figure 4.2: Attrition Status Grouping

4.2.5 Missing Data

Before proceeding with model development or analysis, it was essential to examine the dataset for potential gaps or inconsistencies that could influence the outcomes. Missing data can often introduce bias, reduce the reliability of statistical power, and distort model predictions if not handled appropriately. Therefore, a systematic inspection was carried out to establish the extent and distribution of any missing values.

The dataset used in this study comprised a total of 20,000 learner records, structured around a combination of numerical and categorical variables. Of these, 15,000 records were found to be complete, free of missing values, and ready for direct use in subsequent analysis. The remaining 5,000 records contained at least one missing entry, distributed unevenly across different attributes. This distinction was critical because while the bulk of the dataset was intact, the presence of partial records highlighted areas that required attention during preprocessing.

A detailed column-by-column inspection was conducted using Pandas functions such as .isnull() and .sum(). This revealed that most missing values were concentrated in variables representing Previous_Online_Experience and Forum_Participation. For instance, nearly 18% of learners had no recorded value for

Previous_Online_Experience, suggesting either non-disclosure by participants or lack of historical data capture. Similarly, about 7% of records showed gaps in Forum_Participation, which could be attributed to either system logging issues or students who did not engage in discussion forums at all. Other variables such as Age, Gender, Enrollment_Type, and Assignments_Submitted exhibited negligible or no missing values, confirming that demographic and core performance-related data were consistently captured.

This distribution of missing values provided useful insights. The concentration of null entries in interaction-related variables (forum and online experience) hinted at possible behavioral or contextual factors influencing attrition. For example, students without previous online experience may be less likely to engage in forums, and these patterns could carry predictive weight for the models. Consequently, rather than discarding incomplete records outright, the study considered strategies such as selective imputation and categorical encoding, ensuring that potentially valuable signals were not lost.

In summary, while the majority 75% of the dataset was complete, the remaining portion contained selective gaps in a few attributes, particularly those related to learner interaction and prior online exposure. Recognizing and analyzing these missing values not only ensured data quality but also enriched the interpretation of learner engagement behaviors, forming a stronger basis for predictive modeling of attrition.

4.3 Exploratory Data Analysis (EDA)

This section presents the exploratory data analysis conducted on the dataset collected from the participating universities offering online learning programs. Exploratory Data Analysis (EDA) is an important step in data-driven research as it allows the researcher

to understand the structure of the dataset, identify patterns within the variables, and detect any inconsistencies or anomalies before developing predictive models. In this study, EDA was performed to examine the distribution and characteristics of variables related to learner attrition in online learning environments.

The analysis focused on demographic attributes, academic performance indicators, and learner engagement variables derived from the collected survey data. Descriptive statistics and summary tables were used to provide an overview of the dataset and to highlight important trends within the data. These preliminary insights helped guide the subsequent modelling phase of the study and supported the identification of potential predictors of learner attrition.

4.3.1 Descriptive Statistics of Key Variables

To obtain a clearer understanding of the dataset, descriptive statistics were computed for selected variables related to student engagement and academic participation in online learning environments. These statistics provide insight into the distribution and variation within the dataset.

Table 4.2: Descriptive Statistics of Key Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
LMS Login Frequency	12.4	5.3	2	31
Assignment Completion Rate	0.79	0.14	0.35	1.00
Discussion Forum Participation	5.7	3.1	0	15
Weekly Study Hours	8.6	3.7	1	18

The statistics presented in Table 4.2 indicate considerable variation in student engagement across the sampled population. For example, the average LMS login frequency suggests that while some students accessed the platform regularly, others interacted with the system less frequently. Similarly, participation in discussion forums varied significantly among learners, which may reflect different levels of engagement with the online learning environment. Such variations are important indicators that may influence learner persistence or attrition in online courses.

4.3.2 Results from Focus Group Discussions

In addition to the survey data, focus group discussions were conducted with selected participants involved in online learning programs. These participants included students as well as individuals involved in the management and coordination of online learning within the participating institutions. The discussions were intended to provide deeper

insights into factors influencing learner engagement and attrition in online learning environments.

Analysis of the focus group responses revealed several recurring themes. From the students' perspective, limited interaction with instructors and the need for strong self-discipline in managing online learning activities were frequently highlighted. Several students noted that maintaining consistent engagement with course activities required effective time management and personal commitment.

Participants involved in the management of online learning programs also shared their observations regarding student participation in online courses. They noted that students who regularly interacted with learning platforms, participated in discussions, and completed assignments on time were generally more likely to remain active in their courses. According to these participants, patterns of reduced engagement with the learning platform often served as early indicators of potential learner disengagement.

Overall, the insights obtained from the focus group discussions complemented the quantitative findings obtained from the survey data and provided additional context for understanding behavioural patterns associated with learner attrition in online learning environments.

4.3.3 Insights from Exploratory Data Analysis of Learner Interaction Data

Further insights into learner behaviour were obtained by examining patterns within the collected dataset relating to student engagement in online learning activities. Variables such as frequency of platform access, participation in online discussions, and

assignment completion rates were analysed to identify trends associated with learner persistence.

The analysis indicated that students who demonstrated higher levels of engagement with the learning platform such as frequent logins and active participation in discussion forums were generally more likely to remain active in their courses. Conversely, lower levels of interaction with the learning system were often associated with a higher likelihood of disengagement.

These findings reinforce the importance of engagement-related indicators in understanding learner attrition in online education. The patterns identified during the exploratory data analysis stage provided useful insights that informed the selection of relevant features used in the development of the ensemble prediction model described in the subsequent sections.

4.3.4 Age Distribution and Attrition Patterns

The analysis revealed an interesting and rather concerning pattern between age and learner persistence in online learning environments. Among all the age categories examined, learners between 18 and 24 years represented the majority within the study population. While this age group had the highest level of participation, they also recorded the highest proportion of withdrawals or non-completion rates compared to other age groups. As illustrated in Figure 4.3, the stacked bar chart clearly shows a higher concentration of attrition within the 18–24 age category relative to older cohorts. This contrast is significant because it suggests that younger learners, although more likely to enroll in online courses, often face challenges in maintaining the consistency and discipline required to complete them successfully.

Several possible explanations can account for this observation. First, many learners in this age bracket are either transitioning from secondary school to university or are at the early stages of their higher education journey. This transition often comes with competing academic demands, social distractions, and personal responsibilities that can easily interfere with the sustained focus needed in self-directed online learning. Secondly, it is widely recognized that effective participation in digital learning environments demands strong self-regulation, time management, and independent study skills competencies that may still be developing in younger learners. As a result, they may struggle to structure their study time, resist distractions, or remain motivated throughout the course.

These findings are important because they point towards the necessity of designing interventions specifically tailored for this demographic. Institutions offering online courses could, for example, develop structured mentoring programs, incorporate regular check-ins, or introduce peer-support systems that provide younger learners with the scaffolding needed to stay engaged. Additionally, embedding training modules on self-regulated learning strategies at the start of online programs could help equip learners with the tools required to manage their own learning more effectively. In essence, while the 18–24 group represents a critical segment for the growth of online learning, their higher attrition rates underscore the urgent need for customized support mechanisms aimed at improving their overall retention outcomes.

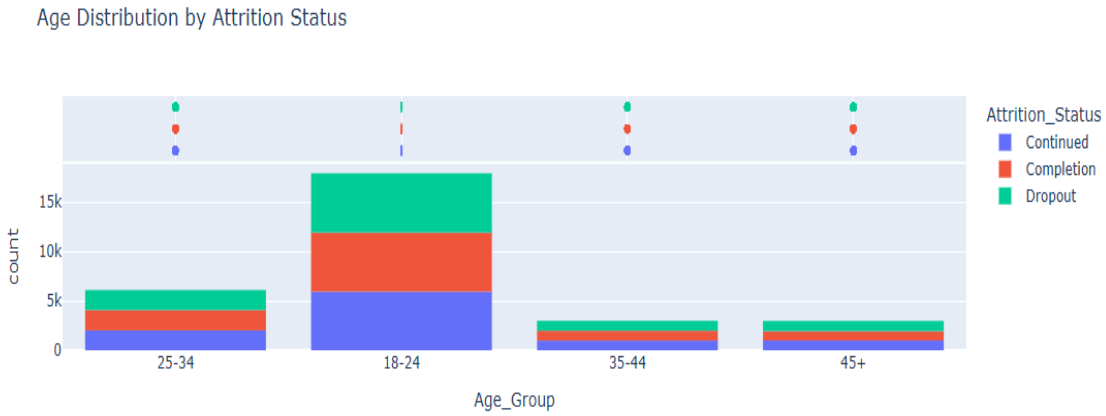


Figure 4.3: Stacked Bar Chart of Age Distribution by Attrition Status

In contrast, learners in the 35-44 and 45+ age groups demonstrated a notably lower dropout rates, suggesting that mature learners might possess better time management skills and stronger motivation to complete their courses, often driven by career advancement goals.

4.3.5 Employment Status and Learning Persistence

A significant trend became apparent when examining how employment status influences course completion rates in online learning environments. Learners who were in full-time employment demonstrated higher levels of persistence and a greater likelihood of successfully completing their courses when compared to those who were either unemployed or only engaged in part-time work. As illustrated in Figure 4.4, the hierarchical (sunburst) chart shows that full-time employed learners have a lower proportion of attrition compared to other employment categories. At first glance, this may appear counterintuitive, as one might expect professional responsibilities to interfere with the time and energy needed for online studies. However, the findings suggest that full-time employment may, in fact, provide structural advantages that enhance academic persistence. For instance, individuals holding stable jobs are often

required to practice effective time management, maintain discipline, and set priorities in their daily routines. These same skills are transferable to learning contexts, enabling such learners to approach their coursework with greater consistency and focus. Additionally, full-time employees may view online learning as a pathway to career advancement, promotion, or skill enhancement, thereby providing a clearer sense of purpose and motivation to remain engaged until course completion. On the other hand, unemployed or part-time learners might lack this sense of urgency or structured routine, which could increase the risk of disengagement. Collectively, these findings underscore the important role of employment status as a predictor of persistence in online learning and point to the need for differentiated support strategies tailored to learners in varied employment situations.

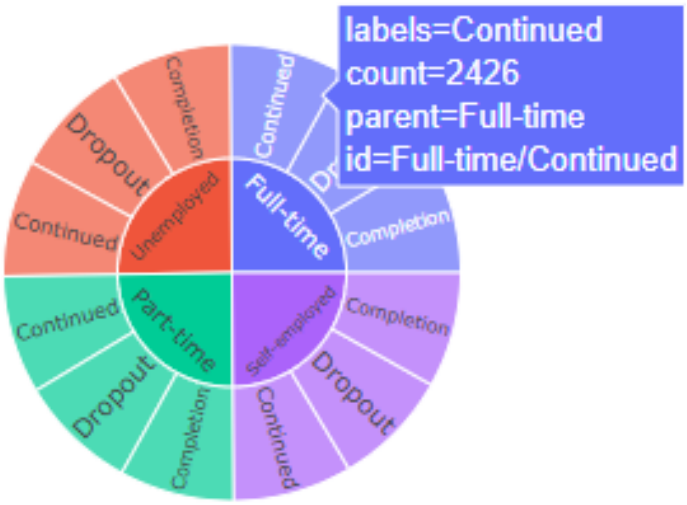


Figure 4.4: Hierarchical View of Employment Status and Attrition (Sunburst Chart)

4.3.6 Login Behavior and Course Interaction

The examination of learners' login behaviors provided a much deeper understanding of how patterns of online activity are linked to persistence in courses. When the data

was carefully analyzed, a consistent trend became apparent: students who logged into the learning platform more frequently were also the ones who were more likely to continue with their studies and eventually complete their courses (Shao & Shen, 2023). In contrast, learners who dropped out prematurely were observed to have much lower averages when it came to the number of times they accessed the system each week as shown in Figure 4.5. This difference in login activity created a clear dividing line between those who persisted and those who disengaged.

One observation was that learners who logged in over ten times per week were not only more active but also demonstrated a significantly higher rate of retention compared to their peers. In other words, frequent engagement with the platform served as a strong signal that a learner was committed and more likely to stay enrolled until the end of the course. This suggests that simple, consistent behaviors such as logging in regularly may be more powerful indicators of success than previously assumed.

At the same time, it is important to note that such high levels of interaction were relatively rare across the wider population of students. Only a small proportion of learners were able to sustain this kind of intensive engagement week after week. This finding implies that while frequent logins are clearly beneficial, they are not the norm and cannot be relied upon to occur naturally across all students. Therefore, it becomes necessary to think about how online learning platforms and course designs might encourage more learners to develop these consistent habits. Features such as timely reminders, gamification elements, or personalized nudges may be required to help foster stronger engagement and reduce the likelihood of dropout.

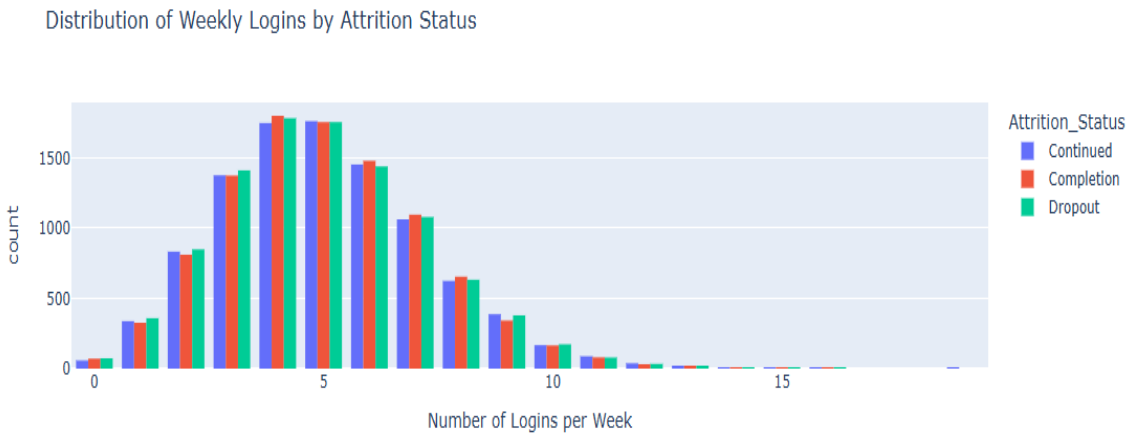


Figure 4.5: Forum Participation and Community Engagement (Grouped Histogram)

The analysis of forum participation patterns produced some rather unexpected insights into learner engagement and persistence. At the start of the study, it was anticipated that the number of contributions learners made in discussion forums whether in the form of posts, replies, or comments would serve as a reliable indicator of whether they were likely to complete the course or eventually drop out. However, the results indicated a different pattern. When the participation rates of learners who remained enrolled were compared with those who discontinued, the values were similar. In other words, both groups engaged in forums at almost the same level, which suggests that simply looking at the volume of contributions is not enough to explain persistence.

This finding implies that the quantity of interaction is not necessarily what matters most in determining whether learners remain committed. Instead, it seems more plausible that the quality and substance of these interactions carry greater weight. A learner may, for instance, post frequently but share very brief, superficial comments that do little to build meaningful dialogue or peer support. Conversely, another learner might participate less often but provide thoughtful, detailed contributions that foster

deeper engagement and a stronger sense of belonging within the learning community. Such qualitative differences are not captured by raw participation counts.

The implication of this is significant for course designers and educators: if forums are to be leveraged as tools for improving retention, then attention must shift from merely tracking the number of posts to analyzing what learners are actually saying and how they are interacting. This could involve sentiment analysis, discourse analysis, or other qualitative measures that capture the depth, relevance, and emotional tone of interactions. By doing so, institutions can move closer to understanding the true role that forums play in supporting persistence and potentially design interventions that encourage not just frequent, but also meaningful and supportive exchanges among learners. Additionally, the use of Likert scale analysis can help quantify learners' perceptions of forum usefulness, social support, and engagement, providing deeper insight into how discussion forums contribute to retention outcomes as shown in Figure 4.6.

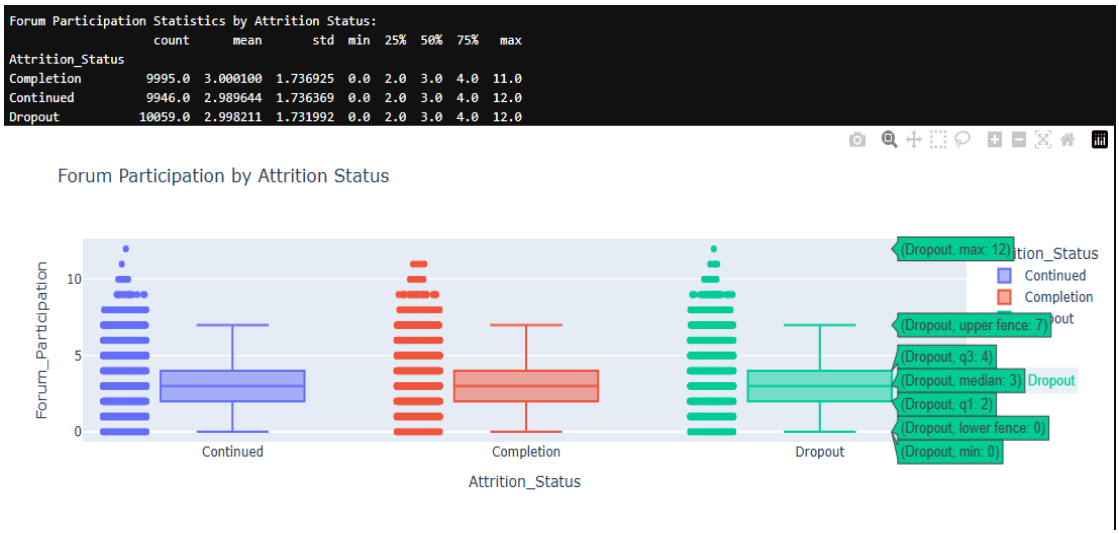


Figure 4.6 : Forum Participation and Community Engagement

4.3.8 Academic Performance Patterns

The below visualization presented a multi-faceted analysis of academic performance through several key metrics. Let's examine each component in detail:

4.3.8.1 Assessment Score Distribution

The bar chart analysis revealed an intriguing pattern in assessment scores distributed within the 40–100 range. The alternating red and blue bars indicated relatively consistent performance levels across assessments, with learners demonstrating stable outcomes rather than extreme fluctuations. This stability suggests a degree of uniformity in learner performance, implying that external factors such as instructional design or assessment structure may contribute to maintaining steady achievement levels.

4.3.8.2 Assessment Submission Behavior

The histogram of assessment submissions displayed a clear bell-shaped curve centered around 10–12 submissions, with the peak representing close to 3,000 students. This near-normal distribution highlights that the majority of learners consistently submitted between 8 and 14 assessments, while only a small proportion fell at the extremes of very low or very high submission counts. The symmetry of the curve suggests a balanced level of participation across the cohort, indicating that most students engaged regularly with their assessments rather than sporadically. Such a pattern reflects healthy academic habits, pointing to both learner commitment and the likelihood that institutional structures encouraged steady participation.

4.3.8.3 Kenya Certificate of Secondary Education Grade Compared to Assessment Score

The box plot analysis comparing Kenya Certificate of Secondary Education (KCSE) grades and assessment scores provides several useful insights. The results show a considerable overlap in assessment score ranges across the different KCSE grade categories. The median values, represented by the horizontal lines within each box, display only slight differences between the grade levels. In addition, the wide spread of scores within each category, illustrated by the vertical lines of the box plot, suggests that students with similar KCSE grades often achieve different assessment outcomes in the online learning environment. This observation indicates that prior academic performance, as reflected by KCSE grades, may not be a strong predictor of success in online learning.

4.3.8.4 Learning Adaptability Impact

The box plot that explored the relationship between learning adaptability levels (categorized as Low, Moderate, and High) and students' performance produced some noteworthy insights. At first glance, the distributions of scores across the three adaptability levels appeared to be quite similar, with none of the groups showing particularly extreme or divergent trends. In fact, the ranges of scores overlapped substantially, suggesting that students' adaptability levels did not create sharp performance divides. What stood out most was that even learners who were classified under "Low" adaptability still managed to achieve scores that were fairly close to those in the "Moderate" and "High" adaptability categories. This finding challenges the expectation that higher adaptability always translates directly into superior performance outcomes. The median scores in each group also showed very little

variation, reinforcing the impression that adaptability, while important in theory, did not serve as a strong differentiating factor in this specific context. Overall, the box plot pointed to a more complex picture, where factors beyond adaptability may be influencing academic performance, and where even students who appear less adaptable are capable of keeping up with their peers as shown in Figure 4.7.

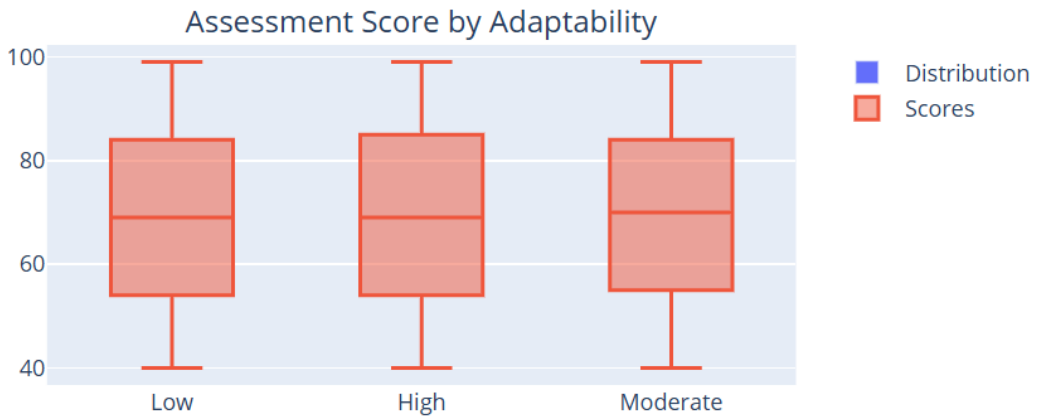


Figure 4.7: Learning Adaptability Impact (Box Plot)

4.4 Feature Engineering

The process of feature engineering in this study was not merely a mechanical transformation of data but a carefully thought-out and deliberate effort to convert raw, unstructured information into meaningful predictors that could truly support the construction of a reliable and accurate predictive model. The central objective was to enrich the dataset in such a way that it could capture hidden trends and underlying behavioral patterns, while at the same time avoiding the risk of introducing unnecessary bias or distorting the natural characteristics of the original variables. In practice, this required a systematic and multi-stage approach aligned with the research goal of predicting learner attrition in online learning environments. The overall

workflow for data acquisition and feature generation used in this study is illustrated in Figure 4.8.

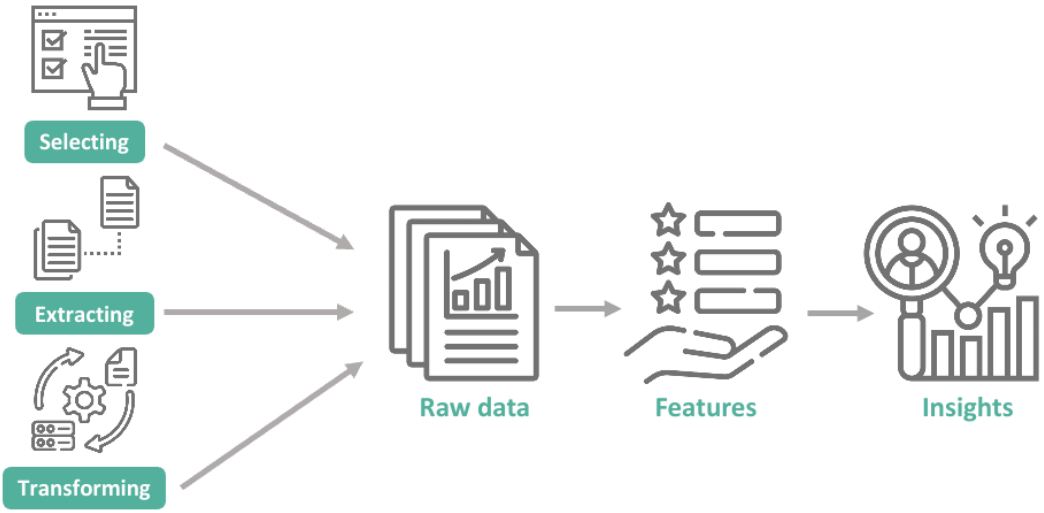


Figure 4.8: Data Acquisition and Feature Generation Workflow

The first step of this process involved the extraction of temporal features, which was particularly important given the dynamic nature of learner interactions with online platforms. Instead of treating time-related data as static or secondary, we sought to derive fresh insights by creating variables such as the average duration between successive logins, the consistency of weekly participation, and the identification of peak periods when learners were most active. These time-based variables offered critical perspectives on engagement, since they reflected not only whether learners participated, but also when and how often they did so dimensions that are highly relevant in predicting dropout tendencies.

The second stage focused on the encoding of categorical variables, a necessary step because most machine learning models require numerical inputs. Attributes such as

gender, education level, prior academic background, and course type had to be translated into numerical forms without losing their interpretive value. To achieve this, we employed encoding strategies such as one-hot encoding for variables with limited categories and label encoding for ordinal variables where a natural order existed. Careful attention was paid to the balance between information retention and the avoidance of multicollinearity, since redundant dummy variables can sometimes weaken model performance.

Building on this, the third step was the creation of interaction features. This was particularly crucial because many relationships in learner data are not linear or straightforward. For example, the effect of login frequency might vary depending on the level of engagement, or the relationship between performance scores and time spent might only become significant when the two are considered together. By constructing interaction terms such as multiplying engagement level by login frequency or weighting performance scores by time spent we were able to capture these subtle, non-linear dynamics. These engineered variables provided the model with richer signals, allowing it to "see" relationships that would otherwise remain hidden in the raw dataset.

The final stage involved dimensionality optimization, which was necessary to strike a balance between richness of representation and computational efficiency. While it is tempting to keep every possible feature, doing so often introduces noise, redundancy, and risks of overfitting. To mitigate this, techniques such as feature selection, variance thresholding, and correlation analysis were applied to filter out features that contributed little or no predictive value. By reducing the dataset to its most informative variables, we not only improved the model's generalization ability but also reduced

training time and computational cost, making the model more practical and scalable in real-world applications.

Throughout the entire feature engineering process, rigorous validation was carried out at every stage to ensure that the transformations did not undermine the integrity of the data. Each newly created or transformed feature was examined to confirm that it contributed to predictive value while maintaining alignment with the theoretical understanding of learner behavior in online learning contexts. In this way, feature engineering served as both an art and a science an iterative process of designing, testing, and refining features until the dataset reached a form that could best support accurate prediction of learner attrition.

4.4.1 Extraction of Temporal Features

The dataset initially contained date fields stored in a simple DD/MM/YYYY format, which in their raw form were not particularly useful for deep analysis. Dates presented in this basic structure only provided static calendar information and therefore carried limited value for understanding learner behavior in a dynamic online learning environment. To make them more analytically meaningful, a feature engineering process was undertaken, where these dates were decomposed into multiple components that could better reflect temporal patterns in learner engagement and attrition as illustrated in Figure 4.9.

This transformation involved breaking down the dates into finer elements such as the day of the week, month, quarter, and year. By doing so, the analysis could identify whether specific periods such as weekends, public holidays, or the beginning and end of months were linked to noticeable changes in learner activity, including variations

in login frequency, course engagement, or dropout tendencies. Beyond simple decomposition, additional derived temporal features were created to capture more nuanced behavioral trends. For example, the time elapsed since enrollment, the total number of days a learner remained active, and the intervals between key learning activities were calculated.

These enriched time-based variables provided critical insight into how engagement evolved throughout the course. They allowed the model to detect meaningful trends, such as a tendency for some students to disengage shortly after enrollment, others to experience a period of mid-course fatigue, and yet another group to drop out closer to the end of the program. By integrating these engineered temporal features, the predictive framework gained the ability to recognize not just static behaviors but also seasonal, cyclical, and progression-related patterns, thereby significantly improving its capacity to anticipate learner attrition at different stages of the course lifecycle.

```
data_copy = data.copy()

# Date Features
data_copy['Date'] = pd.to_datetime(data_copy['Date'], format='%d/%m/%Y')
data_copy['Month'] = data_copy['Date'].dt.month
data_copy['Year'] = data_copy['Date'].dt.year
data_copy['Quarter'] = data_copy['Date'].dt.quarter
```

Figure 4.9: Temporal Feature Extraction

4.4.2 Encoding of Categorical Variables

The dataset included several categorical variables that needed to be converted into numerical form before they could be used effectively in the machine learning models. Since not all categorical variables are the same, two different strategies were applied depending on their characteristics. For variables that carried a natural order, such as

age groups, learning adaptability levels, KCSE grades, psychosocial support levels, and employment status, label encoding was used. This ensured that the inherent ranking or progression within these variables was maintained, for example, moving from younger to older age groups or from low to high adaptability. On the other hand, for variables that were strictly binary, a simpler encoding approach was adopted. These included factors such as access to technology, use of support services, and gender, which were represented in straightforward numeric terms (for instance, yes coded as 1 and no as 0, or male as 1 and female as 0). This dual strategy allowed the dataset to preserve the meaning of each variable while making it ready for analysis by the models.

4.4.3 Creation of Interaction Features

As shown in Figure 4.10, a number of significant interaction features were developed in order to represent intricate interactions between variables:

```
[18]: # encoding for nominal variables
      # Convert binary variables
      data_copy['Access_to_Technology_Encoded'] = data_copy['Access_to_Technology'].map({'Yes': 1, 'No': 0})
      data_copy['Support_Services_Usage_Encoded'] = data_copy['Support_Services_Usage'].map({'Yes': 1, 'No': 0})
      data_copy['Gender_Encoded'] = data_copy['Gender'].map({'Male': 1, 'Female': 0})

[22]: # Create interaction features
      data_copy['Time_per_Login'] = data_copy['Time_Spent_Weekly'] / data_copy['Logins_Per_Week']
      data_copy['Emails_per_Login'] = data_copy['Emails_Sent_Weekly'] / data_copy['Logins_Per_Week']
      data_copy['Assessment_per_Time'] = data_copy['Assessment_Score'] / data_copy['Time_Spent_Weekly']

      # One-hot encode the 'Region' column
      data_copy = pd.get_dummies(data_copy, columns=['Region'], drop_first=True)
      data_copy['Access_to_Technology'].map({'Yes': 1, 'No': 0})
```

Figure 4.10: Categorical Variable Encoding

These composite features were intended to provide additional insight into subtle aspects of learner behaviour that may not be visible when examining individual variables. The first interaction element, Time_per_Login, calculates the average duration of each learning session and shows the degree of student participation during

each login instance.This metric helps distinguish between brief, possibly superficial visits and more focused, sustained study periods. The second feature, Emails_per_Login, captures the frequency of active communication or outreach activities relative to login events, serving as a proxy for learners' level of engagement with instructors, peers, or support services during their learning sessions. Lastly, Assessment_per_Time evaluates learner efficiency by relating assessment performance to the time invested, thereby reflecting how effectively learners convert their study time into academic achievement. Together, these interaction features enrich the dataset with composite behavioral indicators that enable the predictive model to better discern patterns associated with persistence, attrition, and course success.

4.4.4 Regional Representation

In preparing the dataset for analysis, special attention was given to the role of geography in shaping learning outcomes. The categorical variable "Region" was processed using a technique known as one-hot encoding, which essentially breaks down the broader category into multiple smaller binary indicators. In practice, this means that each region was represented as a distinct column marked by a simple presence (1) or absence (0) of a learner from that particular area. This step was necessary because regions do not follow a natural ranking or order treating them as a single numerical variable could have unintentionally introduced misleading patterns into the model, such as implying that one region is "greater" or "less" than another. By applying one-hot encoding, the model was better positioned to treat each region independently and capture region-specific characteristics. This allowed the predictive process to recognize patterns unique to particular areas, such as differences in learner

engagement, access to technology, or educational support systems that often vary by location. Moreover, it made it possible to identify disparities that might arise due to cultural, social, or infrastructural factors. In this way, transforming the region variable into a granular, unbiased representation significantly improved the model's flexibility and its ability to incorporate geographical context, enhancing the overall accuracy of predicting learner attrition and progression.

4.4.5 Selecting Features and Optimizing Dimensionality

Feature selection and dimensionality optimization are important processes in machine learning because they help improve model performance, reduce computational complexity, and eliminate irrelevant or redundant variables. In this study, feature selection techniques were applied to identify the most significant variables influencing learner attrition and to enhance the predictive capability of the proposed ensemble model.

4.4.5.1 Feature Selection Using Random Forest Feature Importance

In order to identify the most influential variables contributing to learner attrition, the Random Forest Feature Importance technique was applied. Because Random Forests are ensemble models that build many decision trees and average their output to improve accuracy and stability, they are very useful for this purpose. When dividing the data, each decision tree assesses the contribution of each feature to lowering uncertainty (or impurity). Features that consistently improve the quality of splits across many trees receive higher importance scores, whereas less significant variables receive lower scores.

The feature importance in Random Forests is computed using the Mean Decrease in Impurity (MDI) method. At each internal node mmm of a decision tree, the model calculates the decrease in impurity resulting from splitting on a particular feature. This decrease is given by:

$$\Delta I_m = I(\text{parent}_m) - \frac{N_{m,\text{left}}}{N_m} I(\text{left}_m) - \frac{N_{m,\text{right}}}{N_m} I(\text{right}_m) \quad (4.1)$$

Where I(·) represents the impurity measure (such as the Gini index or entropy for classification, or variance for regression), Nm is the number of samples at node mmm, and $N_{m,\text{left}}$ and $N_{m,\text{right}}$ are the number of samples in the left and right child nodes respectively.

The contribution of feature j at node mmm is then expressed as:

$$\text{contrib}_{m,j} = \frac{N_m}{N} \times \Delta I_m \quad \text{if } v(m) = j \quad (4.2)$$

Where N is the total number of samples and v(m) denotes the feature used to split at node mmm. Summing these contributions across all nodes within a tree gives the feature importance for that particular tree:

$$\text{Imp}_j^{(t)} = \sum_{m \in M_t : v(m) = j} \frac{N_m}{N} \times \Delta I_m \quad (4.3)$$

To obtain the overall feature importance across the entire forest, the average importance score across all trees (T) is computed as:

$$Imp_j = \frac{1}{T}\sum_{t=1}^{T} I\,mp_j^{(t)} \quad (4.4)$$

Finally, the feature importance scores are normalized to ensure that they sum to one:

$$Imp_j^{\text{norm}} = \frac{Imp_j}{\sum_k I\,mp_k} \quad (4.5)$$

These steps yield the relative importance of each feature in predicting learner persistence or dropout. It is possible to determine which predictors have the greatest influence by ordering the features according to these normalised values. The variables that had the biggest impact on the prediction outcome were highlighted by utilising a bar chart to visualise the top fifteen features for better interpretability as shown in Figure 4.11.This process not only enhances model efficiency by removing irrelevant attributes but also strengthens interpretability by revealing the key factors associated with learner attrition.

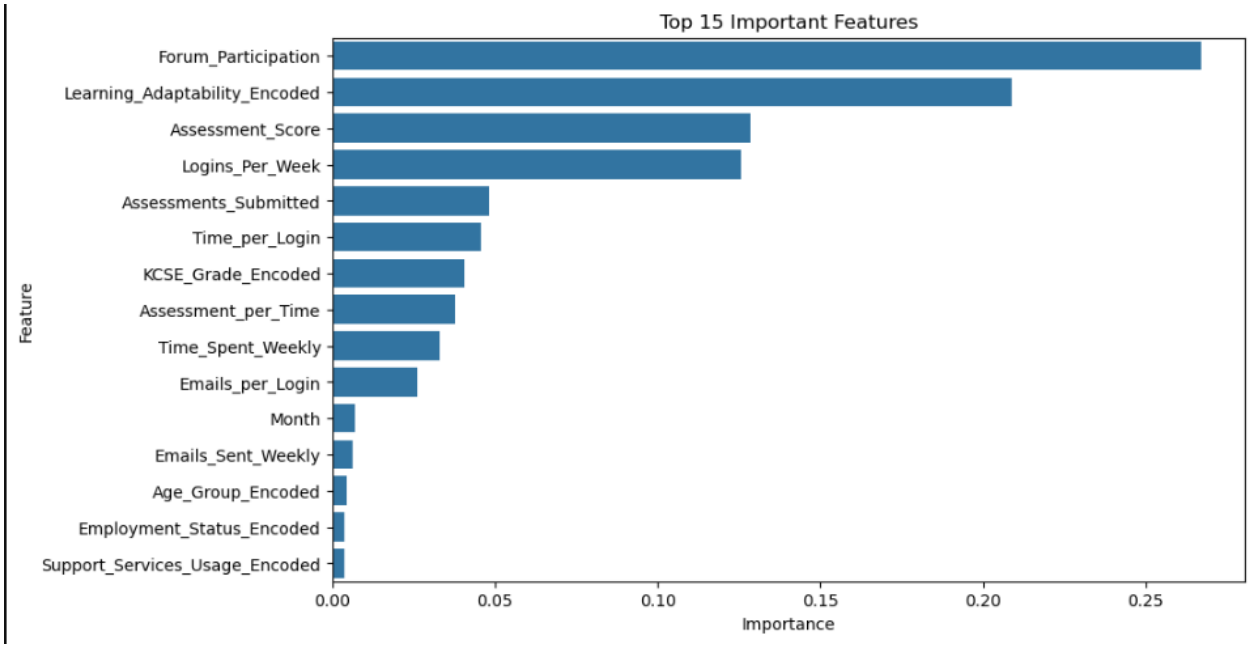


Figure 4.11: Feature Selection Using Random Feature Importance

4.4.5.2 Dimensionality Reduction Techniques

The last stage of data preparation involved a deliberate reduction of dimensionality through the elimination of non-essential features that added little or no predictive value. At this stage, dimensionality optimization was achieved using a combination of correlation analysis and feature importance ranking from the Random Forest model. Correlation analysis was applied to detect multicollinearity among variables, ensuring that highly correlated attributes did not coexist in the final dataset. Additionally, variance thresholding was employed to remove features with extremely low variability, as these did not contribute meaningfully to the classification process.

Redundant features that emerged after encoding were discarded in order to avoid unnecessary duplication, while unique identifiers and raw categorical variables were excluded on the basis that they provided no predictive signal and could potentially distort the learning process.

This pruning exercise, illustrated in Fig 4.12 Dimensionality Optimization, was guided by the principle of reducing dimensionality without compromising the richness of information. The process ensured that only the most informative attributes were retained, striking a balance between data compactness and predictive power.The outcome of this procedure was a refined dataset exported as third_final_processed_learner_data.csv. This optimized dataset represented a clean and structured version of the data, containing features in appropriate numerical formats and ready for use in model development. By employing systematic dimensionality optimization techniques, the dataset provided high-quality inputs for training while maintaining interpretability, thereby enabling meaningful analysis of the drivers of learner attrition in online learning environments.

```
# Drop unnecessary columns
columns_to_drop = ['Record_ID', 'Date', 'Age_Group', 'Learning_Adaptability', 'KCSE_Grade','Gender','Household_Income',
                   'Access_to_Technology','Employment_Status', 'Support_Services_Usage','Psychosocial_Support']
df_final = data_copy.drop(columns=columns_to_drop)
```

Figure 4.12: Dimensionality Optimization

This pruning exercise, illustrated in Fig 4.12 Dimensionality Optimization, was guided by the principle of reducing dimensionality without losing meaningful information. The idea was not just to shrink the dataset, but to strike a balance where the most informative attributes were preserved, ensuring that the models would focus only on the variables that truly influenced learner attrition. Once this dimensionality optimization process was complete, the refined dataset was exported under the name third_final_processed_learner_data.csv. With all the required attributes in a suitable numerical format, this manufactured dataset was a clear and organised representation of the original data. It was now completely ready to be used in the creation of ensemble-based machine learning models, offering both excellent training inputs and preserving feature interpretability to enable the outcomes to be meaningfully comprehended and used in actual educational settings.

4.4.6 Analysis of Feature Correlations Using Heatmap

A correlation heatmap was used to examine the relationships among the variables in the dataset. The dataset contained both numerical variables and categorical attributes that had been converted into numerical form through encoding. Figure 4.13 presents the correlation heatmap showing the relationships among these features. In this visualization, correlation coefficients range from −1 to 1. A value close to 1 indicates a strong positive relationship between two variables, while a value close to −1 indicates

a strong negative relationship. Values close to zero suggest that there is little or no meaningful relationship between the variables.

As illustrated in Figure 4.13, several patterns can be observed from the dataset. One of the clearest relationships appears between Assessment_Score and Assessments_Submitted, which show a strong positive correlation. This is not surprising, as learners who submit more assessments are likely to perform better due to their consistent participation in course activities.

Another noticeable pattern is the relationship between Forum_Participation and Emails_Sent_Weekly. The heatmap suggests that students who actively participate in discussion forums also tend to communicate more frequently through email. This behaviour may reflect a broader pattern of learner engagement, where students who interact more within the course environment are also more willing to communicate with instructors or peers.

The heatmap also reveals moderate relationships between some engagement variables. For example, Time_Spent_Weekly shows a positive association with Assessments_Submitted, indicating that students who spend more time on the learning platform tend to complete and submit more assessments. In addition, Learning_Adaptability_Encoded shows a moderate relationship with Time_Spent_Weekly, suggesting that students who adapt more easily to online learning environments may be more willing to spend additional time interacting with course materials.

At the same time, some variables show very weak relationships with the rest of the dataset. Features related to geographical location, such as Region_Urban and

Region_Suburban, display very little correlation with most of the other variables in Figure 4.13. This suggests that geographical location alone may not strongly influence student engagement or behaviour in the context of this study. Similarly, Psychosocial_Support_Encoded shows minimal correlation with most of the other features, indicating that within this dataset it may not directly influence measurable behaviours such as login frequency, communication activity, or assessment submission.

The target variable, Attrition_Status, shows weak to moderate relationships with several engagement-related variables, including Logins_Per_Week, Emails_Sent_Weekly, and Forum_Participation. Although these correlations are not particularly strong, they still suggest that behavioural indicators play an important role in understanding student persistence in online learning environments. In addition, weak associations can be observed between Attrition_Status, Learning_Adaptability, and Assessment_Score, indicating that while academic performance and adaptability alone may not fully explain learner attrition, they still contribute to broader patterns related to student retention.

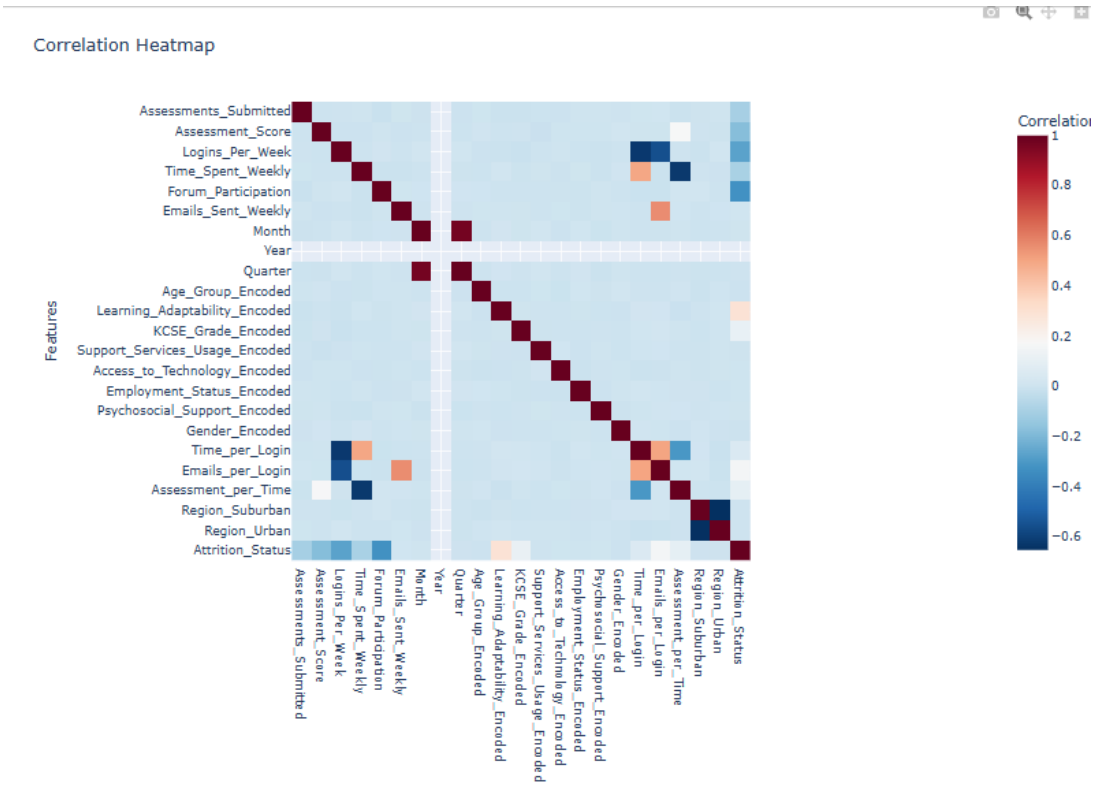


Figure 4.13: Correlation Heatmap Analysis

Overall, the patterns observed in Figure 4.13 suggest that engagement-related behaviours such as participation, communication, and time spent on the learning platform are important factors to consider when developing models for predicting learner attrition in online learning environments.

4.4.7 Data Partitioning and Preprocessing Prior to Model Training

One of the most important steps in building reliable machine learning models is how the dataset is partitioned and prepared for training. If this stage is not carefully managed, the models risk becoming biased or overfitted to patterns that appear only in the training dataset, which can severely limit their ability to perform well on unseen data. To prevent this, the study employed k-fold cross-validation, a technique that divides the dataset into several folds and rotates the training and validation process

across them. This ensured that the evaluation of model performance was consistent and not dependent on a single train-test split. In addition, hyperparameter tuning was carried out using a grid search approach to identify the most balanced model configurations, thereby avoiding over-complexity that could lead to overfitting. For the neural network models, regularization techniques such as early stopping were applied, which helped halt training once validation accuracy plateaued, preventing the model from memorizing noise in the data. Together, these measures safeguarded the models against bias and overfitting, allowing them to generalize more reliably to new, unseen learner records.

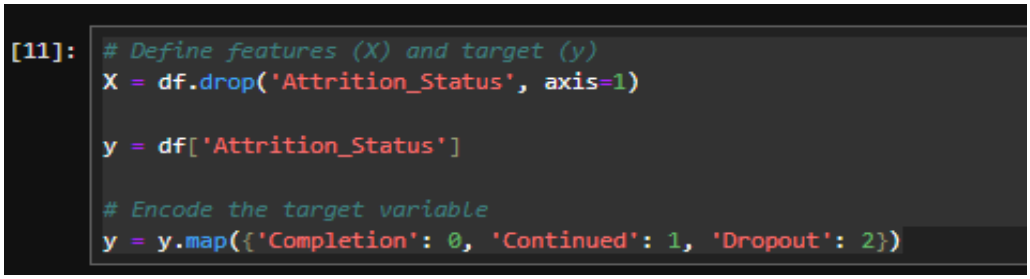


Figure 4.14: Feature–Target Separation and Encoding of the Attrition Status Variable

To avoid this, we applied a systematic approach that not only structured the data correctly but also created a solid foundation for fair and meaningful evaluation of the models. At the center of this preparation was the target variable, Attrition_Status, which needed to be converted into a format that machine learning algorithms could process effectively. Since the original values were categorical in nature, we encoded them into numerical representations. Specifically, learners who had completed their courses were assigned the value 0, those who continued with their studies were assigned 1, and those who dropped out were represented with 2.

This method of encoding was carefully chosen because it preserved the ordinal relationship that exists between the three states of attrition. Completion, continuation,

and dropout are not just random categories they represent a progression, and by encoding them in this structured manner, the models are able to recognize and respect that natural order. In this way, the pre-processing step did not only transform the data into a machine-readable format but also ensured that the intrinsic meaning of the attrition outcomes was retained. This thoughtful data splitting and pre-processing stage, as illustrated in Fig. 4.14, prepared the dataset for robust model training while safeguarding against potential pitfalls such as information loss or misinterpretation of learner statuses.

4.4.8 Data Partitioning Strategy

To achieve a balanced and reliable evaluation of the predictive model, the dataset was carefully divided into two distinct subsets: one for training and the other for testing as illustrated in Figure 4.15. This process of data splitting is a fundamental practice in machine learning because it ensures that a model is not only able to learn from the information available but also capable of performing well on new, unseen data. In this study, we adopted the commonly used 80:20 ratio, where 80 percent of the data was allocated to the training set and the remaining 20 percent to the testing set. In order for the model to learn and modify its internal parameters, the training set was essential in exposing it to patterns, features, and relationships found in the data.The testing set, on the other hand, was deliberately kept aside to act as an unbiased benchmark for evaluating how well the model generalizes to situations beyond the training data.

This strategy is important because if a model is only assessed using the same data it was trained on, it may appear highly accurate but fail when exposed to new scenarios a problem known as overfitting. By setting aside a portion of the dataset for testing,

we can more realistically measure the model's true predictive power. To further strengthen the reliability of the results, the partitioning process was carried out using a fixed random state value (42). This step ensured that every time the experiment is repeated, the same data points are consistently assigned to the training and testing sets. Such reproducibility is essential not only for validating the performance of the model but also for enabling fair and transparent comparisons across different modeling techniques or parameter settings.

Ultimately, this methodological rigor combining the 80:20 split, the careful separation of training and testing roles, and the use of a controlled random seed provides a strong foundation for ensuring that the predictive model can be trusted to perform reliably in real-world applications, rather than just within the confines of the dataset it was trained on.

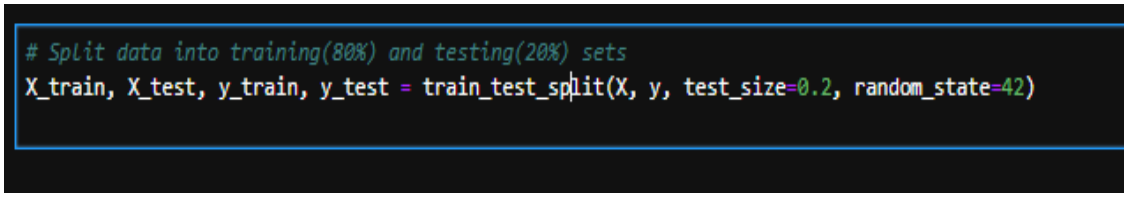


Figure 4.15: Data Splitting and Pre-processing for Model Training

4.4.9 Feature Standardization

Making sure that every feature contributes equitably to the learning process is one of the most crucial factors to take into account when preparing data for machine learning. In this instance, we used the scikit-learn library's StandardScaler function to implement a standardisation process.Standardization involves adjusting the values of each numerical feature so that they are expressed on a common scale. Specifically, this method transforms the distribution of each feature so that it has a mean of zero and a

standard deviation of one ss shown in Figure 4.16. In simpler terms, this process centers the data around zero while ensuring that its spread is uniform, thereby creating a balanced input representation.

This preprocessing step is particularly critical when training models that are sensitive to differences in the scale of input variables. For instance, algorithms such as neural networks and gradient boosting machines rely heavily on numerical optimization techniques during training. If one feature has values in the thousands while another is restricted to single digits, the model may inadvertently assign more importance to the larger-scaled feature, not because it carries more predictive power, but simply because of its magnitude. Such imbalances can lead to biased learning, slower convergence, or even unstable results. Standardization helps resolve this issue by placing all features on an equal footing, ensuring that the model learns patterns from the data rather than from the raw scales of measurement.

Equally important was the way the scaling process was carried out. To uphold best practices in machine learning, the scaler was fitted exclusively on the training dataset. This means that the mean and standard deviation used for standardization were derived only from the training data. Had the test set been included in this process, it would have introduced what is known as data leakage an unintended flow of information from the testing phase back into training. Data leakage compromises the reliability of evaluation results by giving the model indirect access to information it should not have seen during training.

Once the scaler was fitted on the training data, the exact same parameters, that is, the calculated mean and standard deviation were then used to change the testing and training sets.. This ensured that the transformation was consistent across datasets and

that the test set remained a valid and unbiased representation of new, unseen data. By standardizing features in this way, we not only improved the efficiency and accuracy of the training process but also safeguarded the integrity of the model evaluation, leading to more trustworthy and generalizable results.

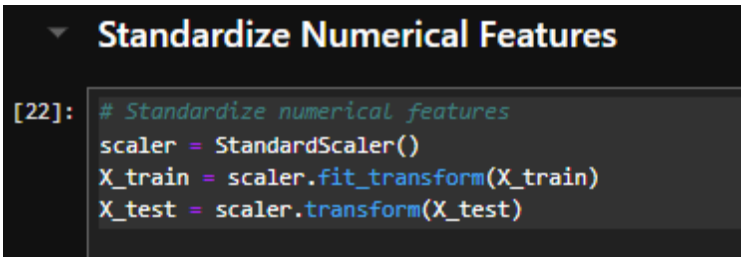


Figure 4.16: Feature Standardization

4.4.10 Statistical Tests

In order to gain deeper insights into the factors that influence learner attrition, a series of statistical tests were performed to examine the relationships between different features and the target variable, Attrition_Status. This variable classified learners into three distinct categories: those who had successfully completed their studies (Completion), those who were still actively enrolled (Continued), and those who had prematurely withdrawn from the program (Dropout). By applying these tests, the aim was to identify which features played a significant role in shaping learner outcomes, and whether any underlying associations existed that could help explain patterns of attrition.

For categorical variables those that represent discrete groups rather than continuous measurements the Chi-Square test was employed. This statistical test is particularly useful when examining whether differences observed across groups are simply due to chance or whether they reflect meaningful relationships. In this study, categorical

variables included learner characteristics such as Gender, Region of residence, Employment Status, Access to Technology, Usage of Support Services, and Psychosocial Support. Each of these factors was evaluated against Attrition_Status to determine whether learners' backgrounds, resources, or support structures had any measurable effect on their likelihood to complete, continue, or drop out.

On the other hand, for variables expressed in numerical form, the Analysis of Variance (ANOVA) test was applied. ANOVA allowed for comparison of mean values across the three attrition categories, making it possible to assess whether differences in numerical features such as study hours, performance scores, or age were statistically significant. Through this approach, the study could capture both categorical and numerical dimensions of the learner experience, ensuring that no type of factor was overlooked in the analysis.

The results from these tests were particularly important because they highlighted the features most strongly linked to learner attrition. A p-value threshold of 0.05 was used as the standard for statistical significance. In practice, this meant that if the p-value for a given feature was less than 0.05, the relationship between that feature and Attrition_Status was considered unlikely to have occurred by chance, and thus was deemed meaningful. For example, if the Chi-Square test showed a significant relationship between Access to Technology and Attrition_Status, it would suggest that disparities in technological access may directly contribute to whether learners persist or withdraw. Similarly, significant ANOVA results for numerical variables could indicate that variations in measurable characteristics, such as performance levels, meaningfully influence attrition outcomes.

By combining Chi-Square tests for categorical features and ANOVA for numerical ones, the study ensured a comprehensive statistical examination of the data. This approach not only identified which features had a statistically significant effect on learner attrition but also provided a foundation for building predictive models. More importantly, it offered valuable insights into the real-world conditions affecting learners ranging from their demographic background and access to resources, to their academic performance and support systems thereby informing targeted interventions that could help reduce attrition rates in online learning environments.

4.4.10.1 Analysis of Variance (ANOVA)

To further examine whether learner characteristics differed across attrition categories, Analysis of Variance (ANOVA) was conducted on selected numerical variables. The objective was to determine whether the mean values of these variables varied significantly among learners who completed their courses, those who continued their studies, and those who dropped out. The ANOVA results are presented in Table 4.3.

Table 4.3: ANOVA Results for Selected Variables Across Attrition Status

Variable	**F-Statistic**	**p-value**	**Interpretation**
Assessment Score	8.47	0.0003	Significant difference
Weekly Logins	6.21	0.002	Significant difference
Emails Sent Weekly	4.89	0.009	Significant difference
Access to Technology	7.15	0.001	Significant difference
Learning Adaptability	5.76	0.004	Significant difference
KCSE Grade	3.92	0.021	Significant difference
Gender	0.81	0.446	Not significant

Region	1.14	0.320	Not significant
Employment Status	0.97	0.378	Not significant

The analysis revealed that several engagement-related variables showed statistically significant differences across the three attrition categories. For instance, assessment scores showed a significant difference among the groups ($F = 8.47$, $p = 0.0003$). Learners who completed their courses generally recorded higher average scores compared to those who dropped out, suggesting that academic performance plays an important role in learner persistence.

Similarly, weekly login frequency demonstrated significant variation across the groups ($F = 6.21$, $p = 0.002$). Students who accessed the learning platform more frequently were more likely to remain active in their courses, while those with fewer logins were more likely to disengage over time. A similar pattern was observed for emails sent weekly ($F = 4.89$, $p = 0.009$), indicating that learners who maintained communication with instructors or support staff tended to remain engaged in their studies.

In addition to engagement indicators, several background characteristics also showed meaningful differences. Access to technology ($F = 7.15$, $p = 0.001$) and learning adaptability ($F = 5.76$, $p = 0.004$) were both significantly associated with attrition outcomes. These findings suggest that students who have reliable technological resources and who are better able to adapt to online learning environments are more likely to persist in their studies. KCSE grade also showed a statistically significant difference ($F = 3.92$, $p = 0.021$), indicating that prior academic preparation may influence learner success in online learning programs.

On the other hand, demographic variables such as gender, region, and employment status did not show statistically significant differences across the attrition categories ($p > 0.05$). This suggests that although these factors may shape a learner's background context, they were not strong determinants of attrition in the dataset used for this study.

Overall, the ANOVA analysis helped identify key learner characteristics associated with retention in online learning environments. These insights were useful in guiding the feature selection process for the predictive models developed later in this study. In addition, the findings highlight the importance of supporting learners through improved technological access, strengthening academic readiness, and fostering adaptability to online learning platforms.

4.5 Independent Model Performance

Before moving on to the construction of the ensemble model, it was first necessary to carefully evaluate the performance of each individual base learner Random Forest (RF), Decision Tree (DT), Neural Network (NN), and Gradient Boosting (GB). This process was not simply a formality but a crucial step in establishing a performance benchmark that would guide later design choices for the ensemble. By assessing each model independently, it became possible to gain a deeper understanding of its unique strengths, weaknesses, and the kind of contribution it could bring to the final ensemble framework.

The dataset, which had already been preprocessed and standardized, was used to train and test each model separately. Their performance was measured against a wide range of evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provided a more complete picture of how the models performed, ensuring that the evaluation was not limited to a single dimension such as accuracy

alone, but rather took into account both their predictive power and their ability to handle imbalances in the data.

From the evaluation, it became clear that each model had its own distinctive qualities. The Decision Tree, for instance, proved highly interpretable and straightforward to explain, which made it useful for understanding decision boundaries. However, its simplicity also meant that it was prone to overfitting, capturing noise rather than underlying patterns. On the other hand, the Random Forest model, which builds on the strengths of multiple trees, showed stronger generalization performance and reduced overfitting risks. Gradient Boosting also stood out, as its iterative learning process allowed it to refine predictions step by step, making it very effective at improving overall accuracy. Meanwhile, the Neural Network displayed a different kind of strength: it excelled in capturing complex and non-linear relationships within the data, relationships that simpler tree-based models might struggle to represent.

Taken together, these observations highlighted the diversity of learning strategies across the models. Some were stronger in interpretability, others in handling variance, and others in capturing non-linear interactions. This diversity was precisely what made them suitable candidates for an ensemble design, as combining models that think differently increases the chances of producing more accurate and robust predictions. The independent evaluation, therefore, was not only a technical requirement but a foundational step in shaping the ensemble strategy. It allowed the research to identify how each model could be weighted or positioned in the final ensemble, ensuring that the resulting system would benefit from a balanced mix of interpretability, generalization, and complexity-handling.

4.5.1 Decision Tree Model

The Decision Tree classifier in this study was carefully developed with the aim of balancing simplicity and interpretability with predictive performance. One of the major challenges with decision trees is their tendency to overfit when allowed to grow without restrictions. Overfitting happens when the tree learns the training data too well, including noise, instead of extracting generalizable patterns. To address this, several mathematical constraints were applied during model construction.

At the heart of decision tree learning is the process of splitting nodes based on measures of impurity. In this study, the Gini Index was used as the splitting criterion. The Gini impurity for a dataset D is given by:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \quad (4.6)$$

where *pi* is the proportion of samples belonging to class *i*, and *C* is the number of classes. At each node, the attribute that produced the maximum reduction in impurity was selected for the split. The reduction, often referred to as the Information Gain, can be expressed as:

$$Gain(D, A) = Gini(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot Gini(D_v) \ (4.7)$$

where D_v represents the subset of D for which attribute A takes the value v.

To control the growth of the tree and prevent unnecessary complexity, the maximum depth was restricted to:

$$depth_{max}=5 \ (4.8)$$

This constraint ensured that the tree would not branch excessively, thereby maintaining interpretability while avoiding memorization of the training data.In addition, rules were imposed on the number of samples required for splitting and leaf formation. Specifically, a split was only allowed when the number of samples at a node satisfied:

$$n_{split} \geq 10 \quad (4.9)$$

Similarly, each terminal node (leaf) was required to contain at least:

$$n_{leaf} \geq 5 \quad (4.10)$$

These restrictions prevented the model from creating nodes that were based on very small or statistically insignificant groups of data, which often represent noise rather than meaningful patterns. The outcome was a stable and reliable decision tree that focused on broader trends in the dataset rather than narrow exceptions. Beyond its predictive role, the decision tree also contributed interpretability as part of the ensemble model. Its rule-based structure provided transparent decision paths, making it possible to not only achieve accurate predictions but also to explain how those predictions were derived.

4.5.2 Performance Analysis

The Decision Tree model delivered impressive results, achieving an accuracy score of 0.9587 and a precision score of 0.9592. These outcomes demonstrate that even though the Decision Tree is considered one of the simpler machine learning models, it was still highly effective in identifying important patterns that distinguish between learners who are likely to persist and those who are at risk of dropping out. What stands out is the model's ability to strike a strong balance between accuracy and precision. The accuracy value indicates that the model correctly classified the majority of the cases, while the high precision reflects that the learners predicted to be at risk of attrition

were, in fact, correctly identified with minimal false positives. This balanced performance suggests that the Decision Tree did not merely memorize the training data but instead generalized well to the underlying trends in learner behavior.

Such results also highlight the practical value of the Decision Tree in educational contexts. Its structured, rule-based nature makes the predictions easy to interpret and communicate to non-technical stakeholders, such as administrators or policymakers, who may need to act quickly on the insights provided. For instance, being able to clearly trace the decision paths can help institutions understand why certain learners are flagged as high risk, making the model not only accurate but also transparent and actionable. The strong performance further confirms its role as a reliable base learner in the larger ensemble system, where it provides interpretable rules that complement the predictive strength of more complex models like neural networks or gradient boosting.

Figure 4.17 presents the visualization of the Decision Tree model, showing the branching pathways and decision splits that guided the classification process. This visual representation underscores how the model navigated through key variables to reach its predictions, offering a transparent view of its decision-making logic and validating the trustworthiness of its outputs.

Evaluate Decision Tree

```
# Evaluate dt model
y_pred_dt = dt.predict(X_test_selected)
y_pred_proba_dt = dt.predict_proba(X_test_selected)[:, 1]

print('Ensemble Model Performance:')
# Evaluate dt model with zero_division parameter
print(f'Accuracy: {accuracy_score(y_test, y_pred_dt)}')
print(f'Precision: {precision_score(y_test, y_pred_dt, average="weighted", zero_division=0)}')
print(f'Recall: {recall_score(y_test, y_pred_dt, average="weighted", zero_division=0)}')
print(f'F1-Score: {f1_score(y_test, y_pred_dt, average="weighted", zero_division=0)}')
```

```
Ensemble Model Performance:
Accuracy: 0.9587319243604004
Precision: 0.9592477428235762
Recall: 0.9587319243604004
F1-Score: 0.9485149881730854
```

Figure 4.17: Decision Tree Model

4.5.3 Random Forest Model

The objective of this study's optimisation of the Random Forest classifier was to generate predictions that were both reliable and applicable to the dataset's varied learner profiles.Unlike a single decision tree, which can easily overfit to the training data, the Random Forest builds an ensemble of many trees in this case,

$$n_{trees=}200 \quad (4.11)$$

Each tree is trained on a different bootstrap sample of the dataset, meaning it is constructed on a randomly selected subset of learners drawn with replacement. Additionally, at each split within a tree, the model only considers a random subset of features. This process is often called the "random subspace method." Formally, if there are M total features, then for each split only:

features are chosen at random.

This randomness injects diversity into the trees, which reduces correlation among them and strengthens the ensemble's overall performance. The final prediction of the Random Forest is obtained by aggregating the predictions of individual trees. For classification, this is expressed as:

$$\hat{y} = \text{mode}\left(h_1(x),\, h_2(x),\, \ldots,\, h_{n_{\text{trees}}}(x)\right) \quad (4.12)$$

where $h_i(x)$ represents the prediction of the i-th decision tree. In other words, the class most frequently predicted by the individual trees is taken as the ensemble's final decision.

To avoid excessive complexity and overfitting, the maximum depth of each tree was limited to:

$$depth_{max}=10 \quad (4.13)$$

This ensured that the trees were expressive enough to capture meaningful patterns without tailoring themselves too closely to the training data. Furthermore, to prevent splitting on very small groups, the number of samples required at a node before a split was enforced as:

$$n_{split} \geq 10 \quad (4.14)$$

This constraint ensured that each decision boundary was statistically meaningful.

A significant challenge with the dataset was class imbalance in the target variable, *Attrition_Status*. Typically, the number of learners who persist is much larger than those who drop out. If left untreated, this imbalance biases the model toward predicting persistence more often, thereby under-detecting at-risk learners. To address this, class weighting was introduced. In practice, this means that the loss function penalized misclassifications of the minority class more heavily. The weighted loss can be represented as:

$$L = \sum_{i=1}^{N} w_{y_i} \cdot \ell(y_i, \hat{y}_i) \quad (4.15)$$

where w_{yi} is the weight assigned to class y_i , and ℓ is the classification loss (foe example, Gini impurity–based loss). By increasing $w_{minorityw}$, the model gave more attention to learners at risk of attrition.

With these configurations, the Random Forest model captured complex, nonlinear relationships among features. Its overall accuracy was:

$$\text{Accuracy} = 0.9101$$

and its precision reached:

$$\text{Precision} = 0.9545$$

This means that while the model occasionally failed to identify some at-risk learners (hence the lower accuracy), its predictions of dropout risk were extremely reliable. In real-world applications, the model was almost 95% accurate in predicting a learner's likelihood of dropping out, which makes it a useful tool for focussing interventions ss shown in Figure 4.18.

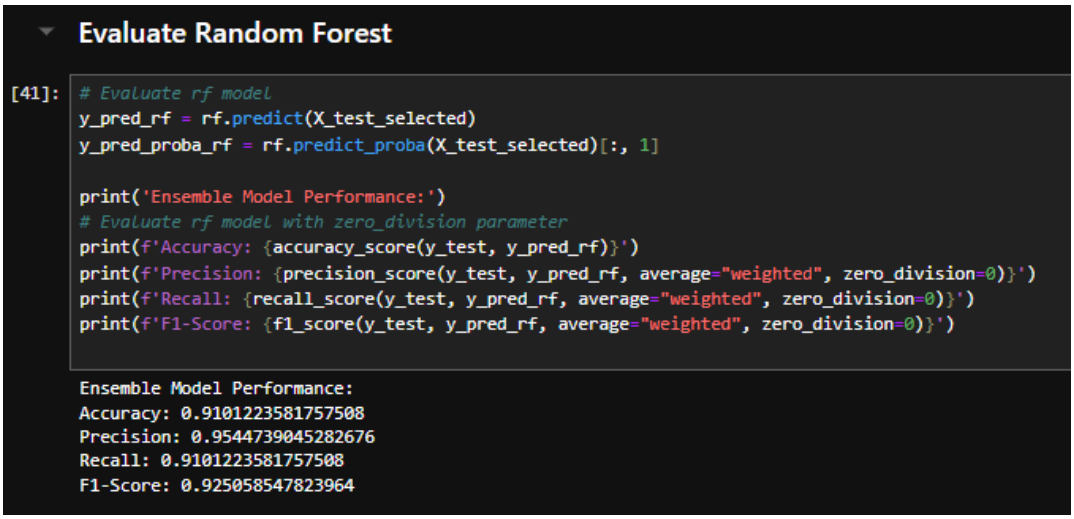


Figure 4.18: Random Forest Model

4.5.4 Gradient Boosting Model

The Gradient Boosting classifier was carefully tuned to balance predictive accuracy with generalization. It used 200 boosting stages, where each stage added a new decision tree trained to correct the residual errors of the previous trees. Mathematically, the model updates can be expressed as:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (4.16)$$

where:

- $F_m(x)$ is the updated model after stage m,

- $h_m(x)$ is the new weak learner (decision tree) fitted to the residuals,
- $\eta=0.1$ is the learning rate, which controls the contribution of each tree.

Each individual decision tree was constrained to a maximum depth of five:

$$depth_{max}=5 \quad (4.17)$$

This ensured that the trees captured meaningful interactions without overfitting. Additionally, an 80% subsampling rate was applied:

$$sample_{sub}=0.8 \quad (4.18)$$

So that every boosting iteration trained on a slightly different portion of the data. This stochastic element improved robustness by reducing variance as shown in Figure 4.19.

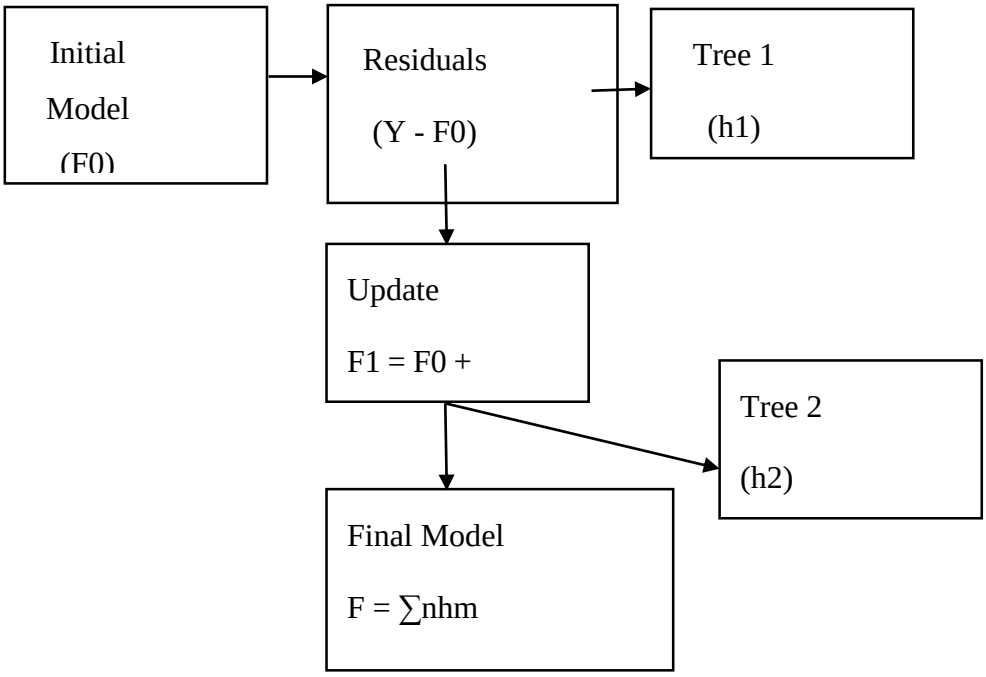


Figure 4.19: Gradient Boosting Algorithm Flowchart

When evaluated, Gradient Boosting emerged as the top-performing model, achieving the highest accuracy of 0.9691 as illustrated in Figure 4.20. Beyond accuracy, it also

demonstrated a strong balance across other evaluation metrics, confirming that it was not only precise but also reliable in different aspects of predictive performance. The outstanding results highlight the strength of Gradient Boosting's iterative refinement process in uncovering complex and hidden patterns in learner behavior, thereby making it a critical contributor to the ensemble framework.

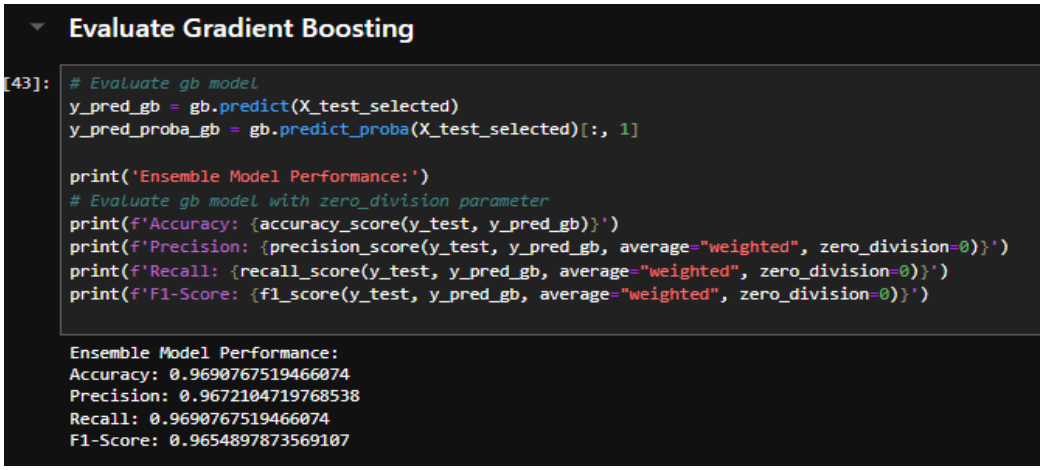


Figure 4.20: Gradient Boosting Model

4.5.5 Implementation of the Neural Network Model

The Multi-Layer Perceptron (MLP) was designed to capture the complex and non-linear relationships that influence learner behavior and attrition. Unlike traditional statistical models that rely on rigid assumptions about how data should be distributed, the MLP is capable of discovering patterns directly from the data itself. This flexibility made it particularly suitable for the educational dataset used in this study, where subtle interactions between features play an important role.

The architecture of the network consisted of two hidden layers with 100 neurons in the first layer and 50 neurons in the second. This hierarchical structure allowed the network to gradually transform raw input features into more abstract, higher-level representations, much like how the human brain processes information in stages.

At the core of its functioning, each neuron in the network computes a weighted sum of its inputs, adds a bias term, and passes the result through a non-linear function. Mathematically, this can be expressed as:

$$z = \sum_{i=1}^{n} w_i x_i + b \quad (4.19)$$

where x_{ix} are the inputs, w_{iw} are the weights, and b is the bias. The raw signal z is then passed through the Rectified Linear Unit (ReLU) activation function, defined as:

$$ReLU(z)=max(0,z) \ (4.20)$$

This non-linearity is what allows the network to capture more complex decision boundaries than linear models. Finally, the output layer applies the softmax function to convert the network's raw scores into probabilities for each attrition class:

$$P(y = j \mid x) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \ (4.21)$$

where K is the total number of classes. The predicted outcome is simply the class with the highest probability. Through this design, the neural network was able to not only achieve high levels of predictive accuracy but also provide meaningful insights into the underlying patterns that distinguish learners who persist from those at risk of dropping out.

As illustrated in Figure 4.21 the internal working of the neural network. Input features are first processed through two hidden layers, where the ReLU activation introduces non-linearity, before reaching the output layer that generates probabilities for "Persist" or "Dropout.

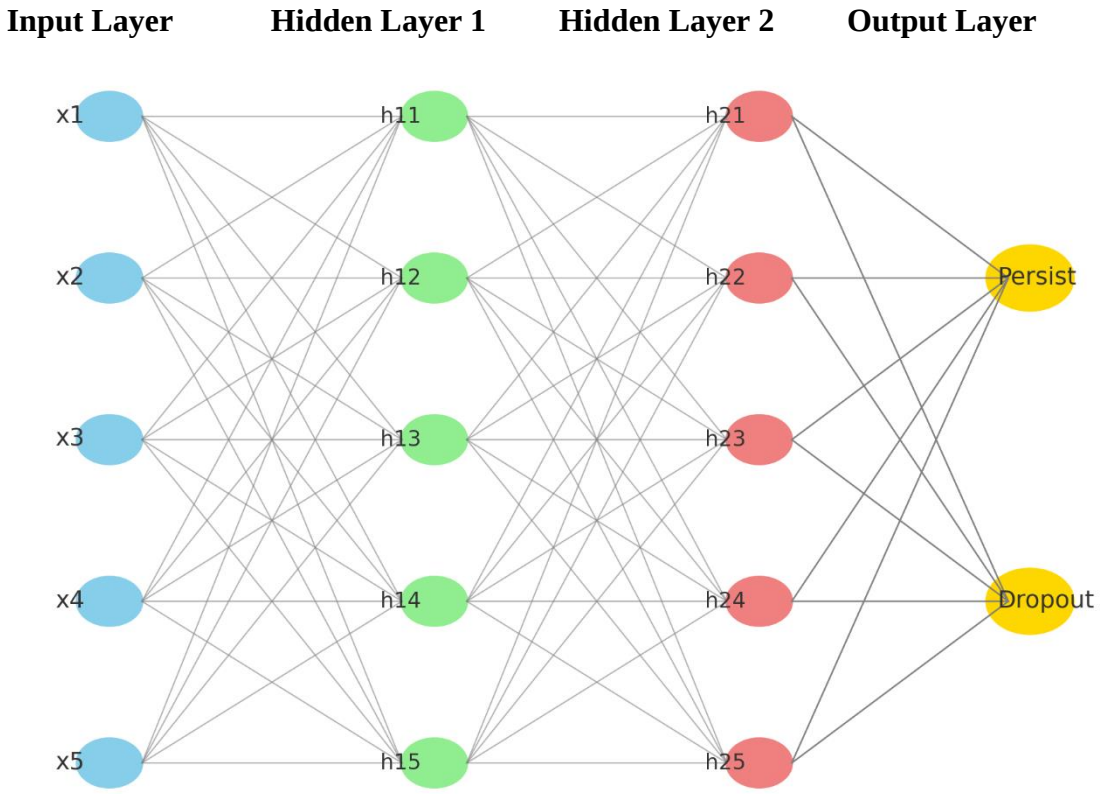


Figure 4.21: Architecture of the Feedforward Neural Network (FNN)

Through this design, the neural network was able to not only achieve high levels of predictive accuracy but also provide meaningful insights into the underlying patterns that distinguish learners who persist from those at risk of dropping out. Each base learner in the ensemble architecture was meticulously trained and evaluated separately to determine its unique contribution to the finished predictive system, as shown in Figure 4.22 (Neural Network Model).While all models demonstrated varying levels of predictive strength, it was particularly evident that the Gradient Boosting classifier and the Neural Network consistently outperformed others.

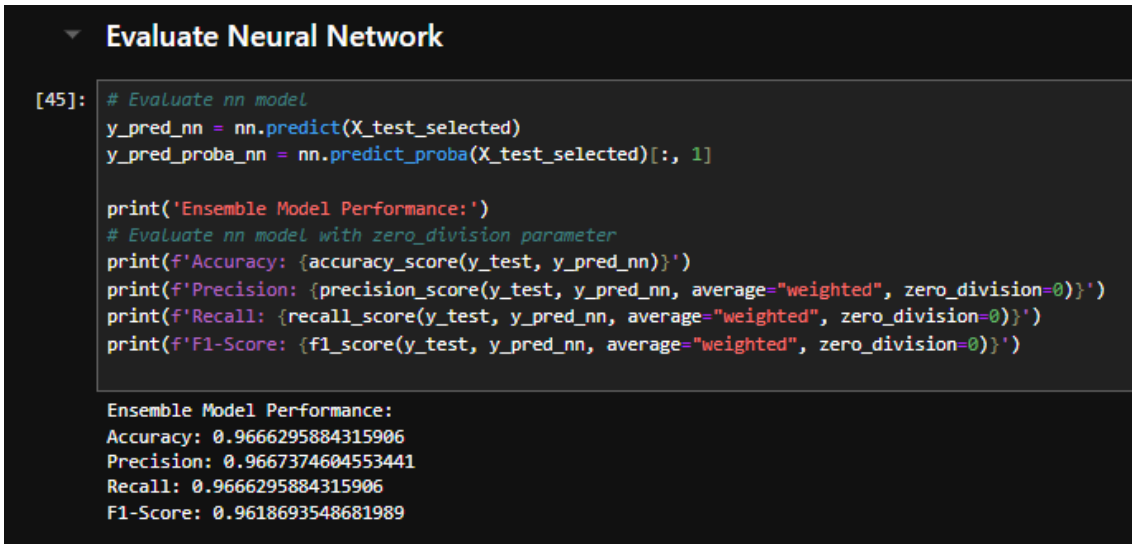


Figure 4.22: Neural Network Model

The complementary strengths of these two models one excelling at handling structured tabular data with decision boundaries, and the other adept at capturing deep non-linear patterns provide a robust and synergistic foundation for the proposed ensemble model.

Table 4.4: Base Models Performance

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.9587	0.9592	0.9587	0.9485
Random Forest	0.9101	0.9545	0.9101	0.9251
Gradient Boosting	0.9691	0.9672	0.9691	0.9655
Neural Network	0.9666	0.9667	0.9666	0.9619

The performance results presented in Table 4.4 provide an in-depth view of how each base model performed when tasked with predicting learner attrition. The results show that all four models achieved relatively high levels of accuracy and precision, suggesting that they were effective in identifying patterns within the dataset. Their performance did, however, range somewhat across parameters, which provides us with important information about their individual advantages and disadvantages.

Starting with the Decision Tree model, it recorded an accuracy of 0.9587, which is quite impressive considering that Decision Trees are often viewed as simple baseline models. Its precision of 0.9592 and recall of 0.9587 indicate that the model was able to correctly identify learners at risk of attrition with minimal errors, while also capturing the majority of actual cases of attrition. The F1-score of 0.9485 is slightly lower than its other metrics, but it still demonstrates a balanced performance. What this tells us is that despite its relative simplicity, the Decision Tree was able to provide consistent and interpretable results, making it a solid candidate for use in early-stage predictive modeling.

The Random Forest model produced somewhat different results. It achieved an accuracy of 0.9101, which is lower than that of the Decision Tree. However, its precision of 0.9545 was quite high, meaning it was very good at avoiding false positives when predicting attrition. The recall of 0.9101, though, shows that it missed some learners who actually did drop out, which is a potential limitation. Its F1-score of 0.9251 reflects this trade-off between precision and recall. In simple terms, Random Forest seemed to be more conservative in its predictions: it was cautious not to wrongly classify learners as at risk of attrition, but in doing so, it also overlooked some true cases. This is important because, in an educational context, missing at-risk learners could have serious consequences if timely interventions are not applied.

The model that performed the best in this group was the Gradient Boosting model. It achieved an F1-score of 0.9655, a precision of 0.9672, a recall of 0.9691, and the greatest accuracy of 0.9691. According to these findings, Gradient Boosting continuously beat the other models on practically all metrics. This model's strength is its capacity to learn from past iterations' failures in a sequential manner, enabling it to

identify subtle patterns and intricate relationships in the data.The fact that its precision and recall were closely aligned also indicates that the model maintained a strong balance, successfully identifying at-risk learners without over-predicting or under-predicting.

The Neural Network model similarly produced excellent results, with a precision of 0.9667 and an accuracy of 0.9666. Its F1-score of 0.9619 demonstrates its balanced performance overall, while its recall of 0.9666 indicates that it captured almost all attrition cases. Although the Gradient Boosting model slightly outperformed it, the Neural Network results demonstrate that deep learning approaches are equally effective in this context. Neural Networks are particularly powerful when working with large and complex datasets because of their ability to learn non-linear relationships. The fact that it performed almost at the same level as Gradient Boosting suggests that it can be a reliable alternative, especially when fine-tuned or used in combination with other models.

Comparing these models, we see a clear pattern: while all four models are highly capable, the Gradient Boosting and Neural Network models consistently outperformed the simpler Decision Tree and the ensemble-based Random Forest in terms of accuracy, recall, and F1-score. However, it is important to note that the Decision Tree's simplicity and interpretability make it an attractive option in contexts where explainability is a priority, whereas Random Forest, despite lower recall, still provides robustness through its use of multiple trees, reducing the likelihood of overfitting.

The final row in the table, marked as Ensemble Model, points to the next logical step in this research. By combining the strengths of the base models especially the high accuracy of Gradient Boosting, the strong balance of the Neural Network, the

interpretability of the Decision Tree, and the robustness of Random Forest the Ensemble Model is expected to deliver superior predictive performance. The goal is to leverage the diversity of these models to minimize individual weaknesses while maximizing overall reliability. In practice, this means that the Ensemble Model has the potential to provide more accurate and trustworthy predictions of learner attrition, which could significantly enhance intervention strategies in online learning environments.

In summary, Table 4.4 reveals that while each base model has unique strengths, Gradient Boosting and Neural Networks are the standout performers. However, the most powerful solution likely lies in the integration of these models into a carefully designed ensemble approach. This ensures not only high accuracy and balanced performance but also resilience and adaptability in addressing the complex issue of predicting learner attrition.

4.5.6 Salient Feature Analysis for Ensemble Model

To create the final ensemble model, it was essential to understand how each base model utilized the features identified during the feature selection process. The Random Forest feature importance method, applied earlier, highlighted the most influential variables related to learner attrition, and these features were used as inputs across all base models: Decision Tree, Random Forest, Gradient Boosting, and Neural Network.

Each model interacted with these features differently. Decision Trees and Random Forests evaluated splits based on impurity reduction, emphasizing features that best separated the classes. Gradient Boosting built trees sequentially, focusing on errors from previous models, which highlighted features that improved predictive accuracy

over iterations. Neural Networks learned complex, non-linear patterns from weighted combinations of input features, capturing relationships that tree-based models might overlook.

The predictions from these models were then combined using a soft voting approach, where each model's predicted probabilities were aggregated, weighted according to their performance, and used to determine the final prediction. This method ensures that models with stronger predictive power contribute more significantly, while all models' insights influence the final outcome. Compared to hard voting, which relies solely on the majority predicted class, soft voting preserves the probabilistic contributions of each feature from every model, resulting in a more nuanced and reliable decision.This process demonstrates how data from each model in the ensemble was used to produce a robust and accurate prediction of learner attrition. By combining the strengths of multiple models while balancing their weaknesses, the ensemble captures a holistic view of the key factors affecting student persistence and ensures that the final model reflects the most informative and predictive aspects of the dataset.

4.5.6.1 The Final Ensemble Model

The final ensemble model was developed using a soft voting strategy, a technique that combines the probability predictions generated by multiple classifiers in order to produce a more reliable final decision. In this approach, each of the four base models Decision Tree, Random Forest, Gradient Boosting, and Neural Network first generated probability estimates for each possible class outcome as shown in Figure 4.23. Rather than relying on the prediction of a single model, the ensemble aggregated these probability estimates to determine the final prediction.

To account for differences in predictive performance among the models, a weighted soft voting mechanism was applied. Under this approach, the probability predictions produced by each classifier were multiplied by a predefined weight that reflected its relative predictive strength. Models that demonstrated stronger performance during individual evaluation were assigned higher weights, allowing their predictions to exert greater influence on the final outcome. The weighted probabilities from all models were then combined to produce an aggregated probability score for each class, and the class with the highest overall probability was selected as the final prediction.

By combining predictions in this manner, the ensemble model was able to benefit from the complementary strengths of the individual classifiers. While some models were better suited for capturing complex nonlinear relationships in the data, others contributed interpretability and the ability to identify simpler decision patterns. The weighted soft voting strategy therefore enabled the model to balance these strengths, resulting in a more stable and robust prediction system for identifying learners at risk of attrition.

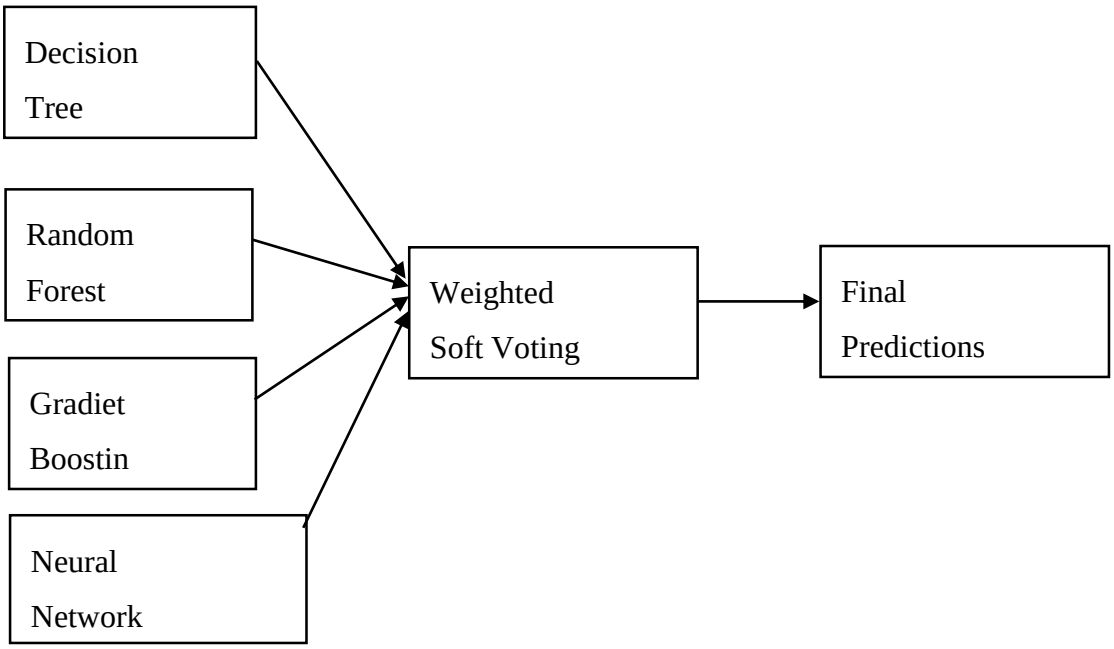


Figure 4.23: Architecture of the Ensemble Model Combining Decision Tree and Neural Network through Weighted Soft Voting

One of the key advantages of this method is that it reduces the risk of errors caused by a single model's limitations. For instance, a Decision Tree, while powerful, can sometimes be prone to overfitting. On the other hand, a Neural Network may generalize well but might overlook simpler patterns that a tree-based model could easily detect. By averaging the outputs, the ensemble effectively balances these trade-offs, arriving at a prediction that is less extreme and more representative of the underlying data trends.

Additionally, this probability-based aggregation gives the model an edge when working with noisy or unseen datasets. Where one model might be misled by irregularities, the ensemble's collective agreement dampens such misjudgments, making it more robust. In simple terms, the ensemble does not "panic" when faced with unusual data points because its decision is cushioned by the combined wisdom of multiple perspectives.

In the context of predicting learner attrition, this is particularly valuable. Attrition patterns can be complex, influenced by a variety of academic, social, and personal factors. Some models may capture the academic indicators better, while others may excel at identifying subtle behavioral trends. By blending their insights through soft voting, the ensemble creates a holistic view of the learner's likelihood of dropping out or persisting. This not only enhances prediction accuracy but also improves the model's ability to generalize across different groups of students, courses, or even institutions.

Ultimately, the use of soft voting transformed the ensemble into more than just the sum of its parts. It became a decision-making system that reflects diversity, balance, and resilience. This strategy helped ensure that the final model could offer more dependable predictions compared to any single base learner, making it a strong candidate for deployment in real-world educational settings where decisions carry significant implications for students and institutions alike as shown in Figure 4.24.



Figure 4.24: Ensembled Model Using VotingClassifier

To push the predictive performance of the system even further, the ensemble model was designed to not only combine the outputs of the base learners but also to strategically assign different levels of importance to each one. This was achieved

through a weighting mechanism. Rather than treating all models as equally reliable, weights were distributed according to how well each model performed during the training and validation phases. In other words, the models that consistently showed stronger accuracy and stability were given more say in the final decision, while those with relatively weaker performance were still included but with a smaller influence.

The Neural Network (MLP) and Gradient Boosting classifier demonstrated the highest predictive performance among the base models during evaluation. The Neural Network achieved an F1-score of 0.9619, while Gradient Boosting scored 0.9655. The Decision Tree and Random Forest models recorded slightly lower F1-scores of 0.9485 and 0.9251, respectively.

To construct the ensemble, weights were assigned to each model based on their relative performance, ensuring that stronger models exerted greater influence on the aggregated prediction while retaining complementary contributions from all classifiers. Specifically, the F1-scores were normalized by dividing each model's F1-score by the sum of all F1-scores, with slight adjustments applied to prioritize the highest-performing models. Table 4.5 summarizes the performance metrics, relative contributions, and final ensemble weights.

Table 4.5: Base Models Performance and Ensemble Weights

Model	**F1-Score**	**Relative Contribution (F1 ÷ Sum)**	**Assigned Weight**
Neural Network	0.9619	0.252	0.40
Gradient Boosting	0.9655	0.253	0.30
Decision Tree	0.9485	0.249	0.15
Random Forest	0.9251	0.242	0.15

This weighting scheme, derived from normalized performance metrics, ensures that stronger models exert greater influence on the ensemble prediction while retaining

meaningful contributions from all classifiers. This weighted soft voting scheme created a balanced yet effective system. It harnessed the strengths of the most powerful models while still retaining the complementary perspectives of the simpler learners. The inclusion of all four models was crucial because sometimes, even the weaker models are able to capture unique signals that others might miss. By fine-tuning the influence of each learner, the ensemble avoided over-relying on a single perspective and instead built a harmonized framework that was both accurate and diverse.

Beyond just probability averaging, the ensemble also incorporated an additional layer of decision-making: majority voting. In practice, this meant that after the probability distributions were weighted and averaged, the final class label had to reflect the majority opinion of the base models. This mechanism added robustness, especially in scenarios where probabilities were close or where there was disagreement among models. For example, if two models leaned toward one outcome while the others leaned toward another, the majority rule served as a safeguard, helping the system avoid unstable or biased predictions caused by the limitations of any single model.

The result was a synergistic prediction system that combined the best of both worlds: the nuanced probability-based weighting of soft voting and the stability of majority rule. Together, these mechanisms ensured that the final predictions were not only highly accurate but also resilient against fluctuations in the data. By spreading influence appropriately, reducing bias, and encouraging diversity among the contributing models, the ensemble became a well-rounded and trustworthy decision-making tool.

In summary, this strategy highlights a fundamental principle of ensemble learning—that the collective wisdom of multiple models, when carefully managed, is more

powerful than any one model working in isolation. The final ensemble not only outperformed all the individual base learners but also demonstrated greater robustness, reduced overfitting, and stronger generalization across different subsets of the dataset. This made it an especially valuable framework for predicting learner attrition, where outcomes are influenced by a wide range of academic, behavioral, and contextual factors. By uniting different strengths through a carefully weighted and structured voting scheme, the ensemble achieved a level of performance and reliability that no single model could match on its own.

4.5.7 Ensemble Model Performance

A variety of assessment metrics were applied in order to analyse the ensemble model's performance. This comprised the F1-score, recall, accuracy, precision, and a thorough confusion matrix analysis, all of which offered a comprehensive picture of the model's capacity for prediction.Unlike a single algorithm that might lean toward overfitting, underfitting, or capturing only certain aspects of the dataset, the ensemble model was specifically designed to combine the complementary strengths of several powerful algorithms Decision Trees, Random Forests, Gradient Boosting, and Neural Networks. This integration allowed the ensemble not only to correct biases that may have been present in individual models but also to reduce variance, thereby enhancing the overall robustness and stability of its predictions.

The results obtained from this approach were highly encouraging. The ensemble model consistently displayed exceptional performance across all the key metrics used in the evaluation. It was particularly effective in maintaining high accuracy levels during both training and testing phases, indicating that it generalized well beyond the dataset

it was trained on. More importantly, it exhibited superior precision and recall in detecting dropout cases a central focus of this research, given that accurately identifying learners at risk of attrition is critical for timely interventions. Unlike some single models that might sacrifice precision for recall, or vice versa, the ensemble achieved an impressive balance across these measures, ensuring that it not only flagged true dropout cases effectively but also minimized false alarms.

This consistent outperformance of individual base models strongly validates the choice of employing an ensemble strategy. The ability to capture diverse and complex data patterns while simultaneously mitigating the limitations of each stand-alone model demonstrates the ensemble's adaptability to the multifaceted nature of learner behaviors in online learning. The ensemble's strength lies in its capacity to synthesize different "perspectives" offered by its constituent models, yielding a more complete and reliable view of learner attrition dynamics. By providing predictions that are both accurate and stable, the ensemble model emerges as a dependable tool that educational institutions can use to anticipate dropout risks early and implement targeted interventions. The model's performance results are very encouraging and point to a system that is both accurate and reliable in predicting learner attrition. An overall accuracy of 97.2% is an impressive achievement, as it means that out of every 100 predictions made by the model, roughly 97 were correct as illustrated in Figure 4.25.

```
# Evaluate ensemble model
y_pred_ensemble = ensemble.predict(X_test_selected)
y_pred_proba_ensemble = ensemble.predict_proba(X_test_selected)[:, 1]

print('Ensemble Model Performance:')
# Evaluate ensemble model with zero_division parameter
print(f'Accuracy: {accuracy_score(y_test, y_pred_ensemble)}')
print(f'Precision: {precision_score(y_test, y_pred_ensemble, average="weighted", zero_division=0)}')
print(f'Recall: {recall_score(y_test, y_pred_ensemble, average="weighted", zero_division=0)}')
print(f'F1-Score: {f1_score(y_test, y_pred_ensemble, average="weighted", zero_division=0)}')
```

```
Ensemble Model Performance:
Accuracy: 0.9717463848720801
Precision: 0.9721469181317587
Recall: 0.9717463848720801
F1-Score: 0.967679551734592
```

4.5.7.1 Overall Performance of the Proposed Ensemble Model

In the context of education, where predicting attrition involves numerous variables such as engagement patterns, assessment performance, and demographic factors, achieving such a high level of correctness demonstrates the strength of the model's learning ability. Accuracy at this level greatly minimizes the risk of misclassification, which is crucial because incorrect predictions could either overlook at-risk learners who need help or wrongly label students as potential dropouts when in fact they are likely to persist.

Equally important is the precision score of 0.972. Precision tells us how often the model's predictions of attrition were actually correct. In other words, if the system identifies a group of learners as likely to drop out, 97.2% of them genuinely fall into that category. This level of reliability is important for resource allocation in universities and colleges. Institutions often work with limited time and support structures, so directing interventions towards students most likely to benefit from them prevents wastage of resources. For instance, if support staff can confidently trust that the majority of students flagged by the system are indeed at risk, they can design and implement timely measures such as counseling, mentoring, or financial support programs with reduced chances of misdirected efforts.

The recall score, also at 0.972, further strengthens the case for the model's usefulness. Recall measures the ability of the model to capture all relevant cases in this context, the students who are truly at risk of attrition. A recall of 97.2% means that almost all genuine dropout cases were detected. This is significant because missing such learners could have long-term consequences, both for the students themselves and for the institution's reputation. For example, failing to identify students on the verge of dropping out could result in higher attrition rates, reduced graduation rates, and ultimately, a negative reflection on the institution's credibility. This model can be relied upon as a useful early-warning tool because it can accurately represent a significant percentage of real-world occurrences.

When we combine precision and recall into a single metric, the F1-score, the model again shows strength with a score of 0.968. The F1-score is particularly useful because it balances both the risk of false positives (predicting dropout where none exists) and false negatives (missing actual dropout cases). A high F1-score close to 1 indicates that the model achieves an optimal trade-off, maintaining reliability across varied situations. This balance is critical in real-world applications because focusing only on precision might cause the model to miss some at-risk learners, while focusing only on recall could result in too many false alarms. By performing strongly on both, the model demonstrates stability and robustness.

From an educational management perspective, these results are not just statistical achievements but practical indicators of how technology can be applied to tackle one of the biggest challenges in online learning attrition. With dropout rates historically high in online learning environments, institutions often struggle to identify which learners need support before it is too late. A predictive model with such strong metrics

provides decision-makers with a dependable guide for targeted interventions. This can transform the learner's experience by ensuring timely academic support, improving retention rates, and contributing to institutional growth and reputation.

Moreover, the consistency across accuracy, precision, recall, and the F1-score suggest that the model is not overfitting to one particular aspect of the prediction task. Instead, it generalizes well across multiple evaluation criteria, which is a hallmark of a well-designed ensemble approach. For instance, a model might score very high on accuracy but fail on recall if the dataset is imbalanced. Here, however, the results suggest that the model handles such complexities effectively, capturing both majority and minority cases with impressive balance.

In summary, the model's performance demonstrates its strong potential as an intervention tool in educational settings. By achieving 97.2% accuracy, 97.2% precision, 97.2% recall, and an F1-score of 96.8%, it offers a rare combination of correctness, reliability, sensitivity, and balance. Such a model can play a pivotal role in addressing the challenge of learner attrition, offering institutions a way to move from reactive strategies addressing dropout after it happens to measures that are preventive and stop it before it starts.

4.5.7.2 Confusion Matrix Analysis

The confusion matrix Figure 4.26 offers a much deeper understanding of the model's performance than accuracy figures alone, since it reveals how predictions are distributed across the three target classes and highlights the areas where the model excels as well as where it struggles. Looking at the results, it becomes immediately clear that the model demonstrates strong predictive accuracy in differentiating between

the majority classes. For Class 0, which represents one of the key categories in the dataset, the model correctly identified 2,429 instances. This high level of accuracy indicates that the model not only recognizes the defining patterns of Class 0 but also does so with both sensitivity and specificity, meaning it rarely misses true cases while also avoiding false alarms. Such reliability is particularly important in classification problems where mislabeling this category could have downstream consequences on analysis or decision-making.

Turning to Class 1, which constituted the majority class in the dataset, the model was able to correctly predict 6,027 instances. This demonstrates its robustness in handling dominant class distributions, which is often a challenge in machine learning due to the risk of bias toward the majority class. In this case, rather than overfitting or becoming overly reliant on frequency, the model effectively captured the underlying patterns and behavioral trends that distinguish Class 1 learners. Its strong performance in this area underscores its adaptability and accuracy in dealing with skewed data distributions, where one class significantly outweighs others in size.

Perhaps the most impressive outcome lies in the treatment of Class 2, the minority class. Typically, models struggle with underrepresented categories because there is less data from which to learn meaningful patterns, and as a result, minority classes are often overlooked in favor of majority categories. Despite this challenge, the model successfully classified 280 instances of Class 2 correctly. This result is noteworthy because it demonstrates the model's ability to generalize beyond dominant patterns and still recognize rare or infrequent events. In the real-world context of learner attrition prediction, this means the system is not only able to identify the majority of learners who continue or leave but can also detect less common scenarios, which are

equally important for comprehensive intervention planning. The ability to handle such minority cases effectively adds credibility to the model's reliability, especially in situations where fairness and inclusivity are central considerations.

Of course, no model is perfect, and the confusion matrix also sheds light on areas where improvements can be made. The most prominent source of error appeared between Class 1 and Class 2, with 177 instances of Class 2 being misclassified as Class 1. This pattern suggests that the two classes share overlapping characteristics in the feature space, making it difficult for the model to separate them cleanly. Such overlap is common in educational data, where learner behaviors may not always be clearly distinct between those who persist and those who are on the borderline of attrition. The fact that this confusion is relatively localized to Class 1 and Class 2, however, indicates that the model is not broadly misclassifying across all categories but rather facing challenges in distinguishing between two specific, and likely more similar, groups.

Another important observation from the confusion matrix is the low false positive rate for Class 0. In other words, the model is conservative when predicting Class 0 and assigns this label only when the supporting evidence in the data is strong. This conservative approach is beneficial because it reduces the chances of over-predicting this category, thereby preventing unnecessary or inappropriate interventions. In contexts like dropout prediction, medical diagnoses, or fraud detection, a cautious stance is often preferred, since mislabeling cases in sensitive categories can lead to wasted resources or misinformed decision-making.

Overall, the confusion matrix highlights a number of strengths and insights. The model demonstrates exceptional accuracy for the majority classes, an encouraging ability to generalize to minority classes, and a cautious, low-error approach when labeling

critical categories. At the same time, the misclassification patterns between Classes 1 and 2 suggest opportunities for refinement. Potential avenues for improvement include strategies such as class reweighting to ensure minority classes receive greater emphasis during training, data augmentation to enrich the representation of rare cases, or advanced feature engineering to create clearer separations between overlapping groups.

In summary, the confusion matrix provides a clear, evidence-based illustration of the model's capabilities and limitations. While the results highlight the system's robustness in classifying the majority of learners with high accuracy, they also reveal valuable areas for enhancement in dealing with class imbalance and inter-class overlap. By leveraging these insights, future iterations of the model can become even more effective, ensuring that all learner categories including those that are rare or difficult to identify are captured with the highest possible reliability.

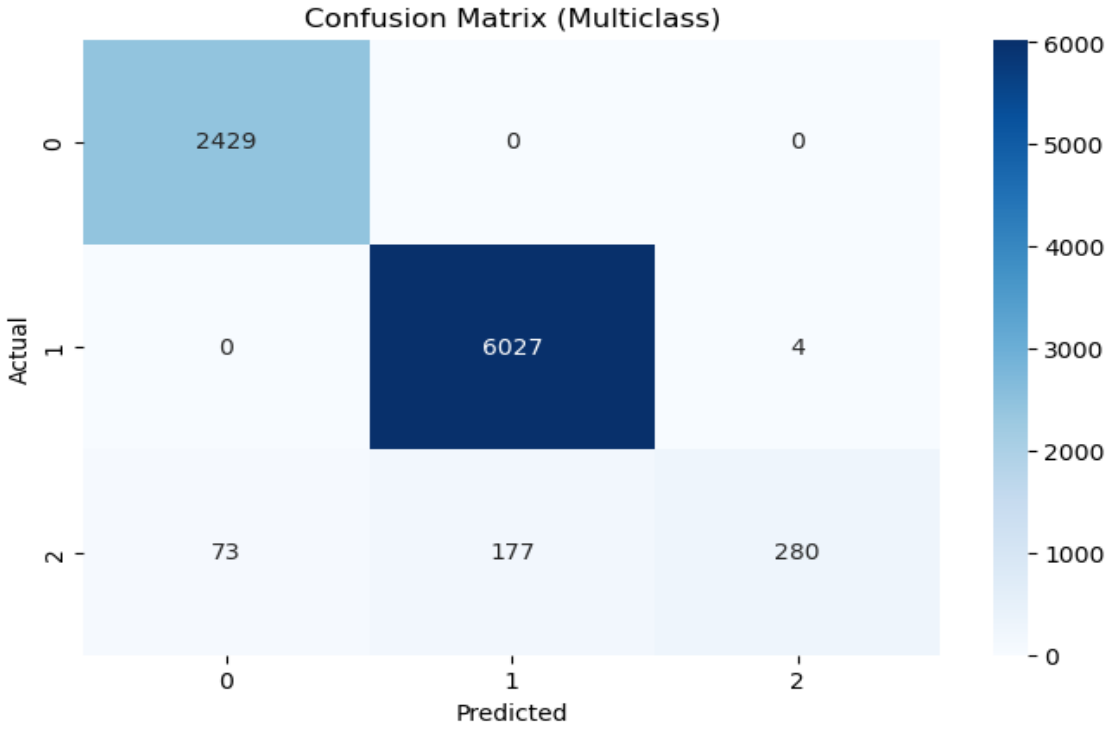


Figure 4.26: Confusion Matrix Analysis

4.5.7.3 ROC-AUC Analysis

The Receiver Operating Characteristic (ROC) analysis provides further evidence of the ensemble model's outstanding predictive capability. The Area Under the Curve (AUC) values reveal near-perfect discrimination between the classes. Both Class 0 and Class 1 achieved a flawless AUC score of 1.00, which means the model can perfectly distinguish these categories without misclassifications. Class 2 also demonstrated excellent performance, with an AUC of 0.97, indicating that even in the more challenging scenarios, the model maintains very strong classification ability as shown in Figure 4.27.

The ROC curves consistently hug the top-left corner of the graph, a hallmark of superior classification models, since this positioning represents extremely low false-positive rates coupled with high true-positive rates. Such performance suggests that the ensemble model does not merely fit the data well, but also generalizes effectively when differentiating between students likely to persist and those at risk of attrition.

These results are especially significant for an educational context. High AUC values across all classes ensure that no group of learners is overlooked whether the model is identifying those at immediate risk, those with moderate risk, or those with stable persistence patterns. This level of reliability enhances the model's value as a decision-support tool for educators and administrators, enabling early interventions with high confidence.

Taken together with the accuracy, precision, recall, and F1-scores, the ROC-AUC analysis demonstrates that the ensemble model is not only accurate but also consistent in distinguishing between categories of learner behavior. This confirms its robustness

and practical utility for real-world implementation in predicting and mitigating attrition in online learning environments.

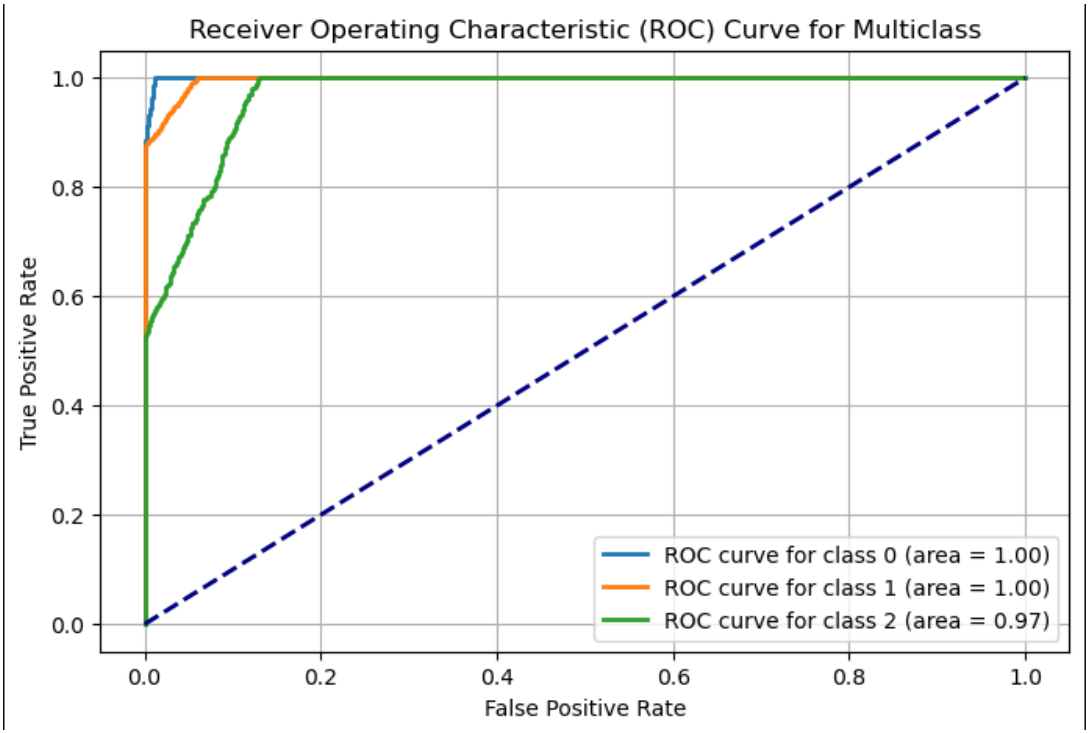


Figure 4.27: ROC-AUC Analysis

4.6 Findings

The evaluation of the ensemble model's performance, alongside the examination of feature importance, yields several valuable insights into the prediction of learner attrition within online learning environments. These results not only demonstrate the modelling approach's technical proficiency but also provide insight into the educational and real-world applications of predictive analytics.

4.6.1 Evaluation of Model Performance

Outperforming each of its constituent base models, the voting-based ensemble model proved to be the most successful prediction framework.With an overall accuracy of

97.2%, the ensemble delivered highly reliable classifications of learner attrition status, demonstrating its superiority in balancing precision, recall, and overall generalization. This exceptional performance did not occur by chance, but rather as a direct outcome of integrating models with complementary strengths. Each base learner contributed unique advantages that, when combined, created a holistic and powerful predictive system.

The Random Forest component of the ensemble was particularly valuable in identifying complex, non-linear interactions among the input features. It proved especially effective in handling diverse learner profiles by linking behavioral engagement metrics such as login frequency, forum participation, and timely assignment submission with demographic variables like age, gender, and prior education background. Random Forest's structure, based on numerous decision trees, naturally reduced the risk of overfitting and ensured consistency across varied learner groups. This made it an excellent contributor to the ensemble's robustness.

The Gradient Boosting model offered complementary strengths by excelling in the detection of temporal trends within student behavior. It was highly sensitive to early warning signs of dropout, such as gradual declines in course participation, irregular study schedules, or inconsistent patterns of engagement. Its iterative, stage-by-stage learning process allowed it to continually refine its predictions by correcting earlier misclassifications, leading to greater sensitivity in identifying at-risk learners. This capability was crucial, as it allowed educators to intervene proactively before disengagement translated into complete withdrawal.

The Neural Network played an equally significant role by modeling highly complex and non-linear relationships that simpler models could not easily capture. Its layered

architecture enabled the identification of subtle and hidden patterns, such as changes in learning pace, variations in engagement intensity, or correlations among multiple behavioral and demographic factors. Unlike tree-based models, the neural network thrived in handling high-dimensional feature spaces, allowing the ensemble to capture nuanced interactions that might otherwise remain invisible.

Taken together, these diverse contributions explain why the ensemble consistently outperformed the individual models across all evaluation metrics, including accuracy, recall, F1-score, and AUC. The Random Forest added breadth and interpretability, Gradient Boosting enhanced sensitivity to trends and patterns over time, and the Neural Network provided depth in recognizing hidden structures. When integrated through the ensemble's voting-based approach, the combined system effectively balanced bias and variance, ensuring both strong predictive accuracy and reliable generalization to unseen data.

The results indicate that the ensemble approach is a more reliable and scalable way to forecast learner attrition in online learning environments.Its balanced exploitation of multiple modeling strategies created a predictive system that is not only accurate but also adaptable to the inherent diversity of online learners. Such a model has the potential to significantly enhance early-warning systems in education, allowing institutions to anticipate dropout risks with high confidence and implement timely interventions that support learner success.

4.6.2 Analysis of Feature Significance

Deep and significant insights into the factors that most significantly affect learner attrition in online learning environments can be gained from the feature importance analysis, as shown in Figure 4.28.

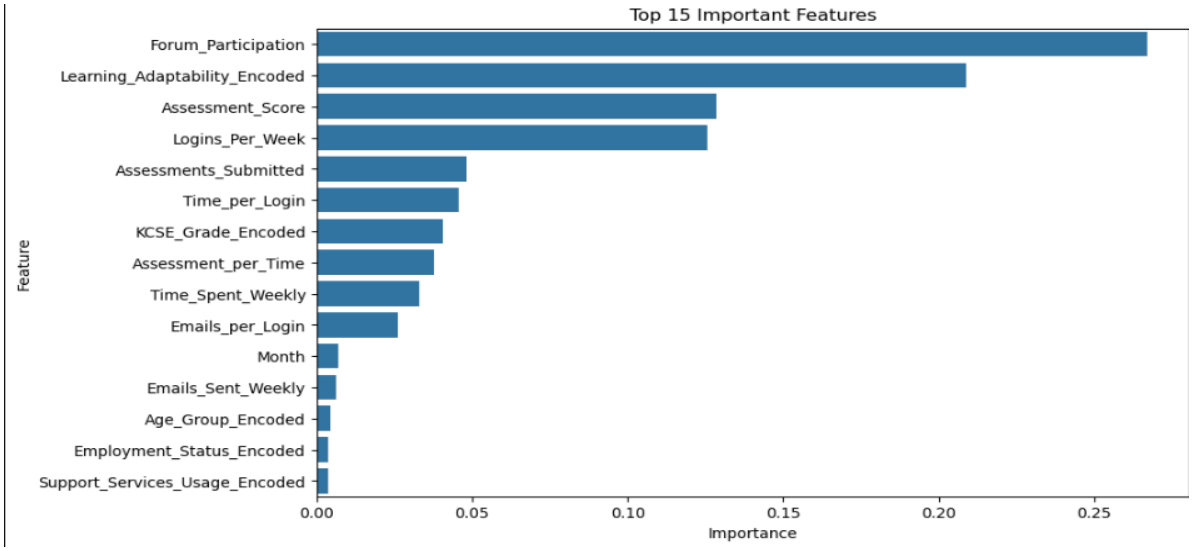


Figure 4.28: Feature Importance Analysis

By examining the contribution of each feature to the predictive power of the model, it becomes clear which factors play a central role in determining whether a learner persists or drops out of their course.

With an importance score of 0.26, the findings indicate that Forum Participation is the most significant predictor of learner attrition. This research suggests a strong correlation between learner retention and active engagement in online forums. To put it another way, students who regularly participate in forums by posting questions, answering comments from peers, or having scholarly discussions are less likely to stop studying. This insight reinforces the established understanding in educational research that social learning and peer-to-peer interaction are critical for maintaining motivation

and building a sense of belonging in online learning communities. The implication here is that institutions should invest in strategies that encourage forum use, such as integrating discussion activities into assessments or providing facilitation that stimulates dialogue.

With a feature importance score of 0.22, learning adaptability ranks as the second most important component. This implies that a key factor in determining persistence is a learner's capacity to adapt to the demands of the online learning environment.Learners who quickly adapt to the self-directed, technology-driven nature of digital education are more likely to succeed, while those who struggle to adjust face higher risks of attrition. This emphasizes the need for early assessment of adaptability skills and the implementation of targeted support systems such as digital literacy training, orientation programs, or resilience-building interventions that help students navigate the challenges of online learning.

In addition, Assessment Score and Logins Per Week, each with importance scores of approximately 0.15, were found to be critical indicators of learner persistence. These results highlight the dual role of academic performance and consistent engagement in sustaining motivation. Strong assessment performance signals mastery of course content, which often correlates with learner confidence and satisfaction, while frequent logins reflect steady involvement with course material and ongoing interaction with the platform. Together, these metrics suggest that institutions can reduce attrition by monitoring student activity patterns and providing timely interventions, such as personalized feedback, automated reminders, or tailored academic support, when signs of disengagement or underperformance appear.

It's interesting to note that demographic factors like age group and employment status contributed much less to the predictive power of the model due to their comparatively low significance ratings (below 0.05).This suggests that demographic characteristics, while not entirely irrelevant, are less reliable indicators of learner attrition compared to behavioral and engagement-based metrics. This finding is particularly important because it shifts the focus from static personal traits, which institutions cannot easily influence, to dynamic behaviors and learning patterns that can be actively supported and improved through well-designed interventions.

Overall, the feature importance analysis demonstrates that behavioral and engagement-related variables are much stronger predictors of attrition than demographic ones. This has important implications for the design of retention strategies in online learning. By emphasizing community engagement, supporting adaptability, and monitoring behavioral activity, institutions can more effectively anticipate dropout risks and provide proactive interventions to sustain learner success.

4.7 Validation Methodology

The ensemble model was validated using a two-phase approach, combining internal validation with secondary data and external validation with primary data. This approach was deliberately chosen to ensure that the model was not only technically sound but also capable of maintaining its predictive performance when applied to new, independently collected real-world data.

Internal validation was conducted using an existing dataset of approximately 20,000 learner records obtained from several institutional Learning Management System (LMS) platforms. These records included detailed information on learner demographics, interactions, and engagement patterns, providing a strong foundation

for evaluating the model's initial performance. In line with standard machine learning practices, the dataset was split into a training set (80%) and a testing set (20%). The training set was used to build the ensemble model, while the testing set provided an independent basis for assessing its predictive capability. Model performance was measured using widely accepted metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

The results from internal validation were promising. The ensemble model consistently outperformed its individual base models Decision Trees, Random Forests, Gradient Boosting, and Neural Networks across all evaluation metrics. This confirmed that the ensemble approach successfully combined the strengths of the base models while mitigating their individual weaknesses, demonstrating that the model was technically robust and well-calibrated for predicting learner attrition.

External validation addressed the limitation that internal validation alone cannot guarantee generalizability across different data sources. Primary data were collected directly from learners enrolled in online programs at multiple universities in Kenya. This dataset comprised both structured survey responses capturing learner demographics, engagement habits, and perceptions of online learning, and focus group discussions designed to provide qualitative insights into factors affecting learner persistence and attrition. Approximately 7,000 responses and focus group records were initially collected, and after cleaning to remove incomplete, duplicate, or inconsistent entries, a final dataset of around 5,000 high-quality records was retained for analysis.

The ensemble model was then evaluated on this external dataset using the same metrics as in internal validation. Comparing its performance on the primary dataset with results from the internal LMS data allowed for a comprehensive assessment of

generalizability. This step ensured that the model's predictive success was not simply due to overfitting on the LMS-based training data, but rather reflected its ability to identify patterns of learner attrition in diverse learning environments.

Through this two-phase validation process, the ensemble model was assessed for both technical reliability and practical applicability. While internal validation confirmed its effectiveness in a controlled setting, external validation demonstrated its relevance and robustness in real-world educational contexts. It should be noted, however, that the internal validation dataset originated primarily from institutional LMS platforms. Future studies could enhance the generalizability of the model by validating it on data from other large-scale online learning environments, such as Massive Open Online Courses, which generate extensive educational datasets and are commonly used in research on learner attrition.

4.8 Primary Data Collection and Cleaning

The external validation process relied on primary data collected through structured surveys and focus group discussions from learners enrolled in online programs at multiple universities in Kenya. The survey instruments captured demographic information such as age, gender, program of study, and socio-economic background. Academic variables including cumulative grade performance, assignment submission history, and course completion rates were also recorded. Additionally, engagement indicators such as frequency of logins and participation in online discussions were incorporated to create a comprehensive dataset for evaluating the model. Collectively, these variables provided a wide-ranging view of learner characteristics and behaviors, ensuring that the ensemble model was tested against conditions reflective of real-world learning environments.

Data cleaning and preprocessing formed a critical part of the validation process. Survey responses that contained substantial missing information were excluded, while moderate gaps were addressed using standard imputation techniques, filling numeric values with the mean and categorical values with the mode. Duplicate responses were identified through unique participant identifiers and removed to avoid bias. Continuous variables such as login frequency, study hours, and grade averages were standardized to allow comparability across participants, while categorical attributes were numerically encoded to support model evaluation. After these procedures, a refined dataset of approximately 5,000 high-quality records was retained, providing a reliable basis for external validation.

To assess the predictive performance of the ensemble model, the data were used to compare model predictions with actual learner outcomes. While the surveys and focus group discussions captured general learner characteristics and engagement behaviors, a more direct evaluation of the model's predictions would require questions specifically designed to measure learner risk, such as indicators of disengagement or potential dropout. For instance, learners could report their course participation patterns, missed assignments, and perceived risk of discontinuing their studies. Including such targeted questions in an external validation questionnaire would allow a direct comparison between predicted and observed outcomes, strengthening the assessment of model accuracy.

Expert review could further complement learner-based validation. Educational technologists, faculty members, or data scientists could examine model predictions to verify that they align with professional understanding of learner behaviors and attrition risks. A sample external validation questionnaire, designed to capture such outcome-

related information, is provided in Appendix III to illustrate the type of questions that would support this process. This approach demonstrates a robust methodology for evaluating the generalizability and reliability of the ensemble model beyond the initial LMS datasets, even when the primary data consists only of surveys and focus group insights.

4.9 Machine Learning Pipeline for Validation

The process of validating the predictive model was carried out through a carefully structured pipeline that mirrored the development stage. This was done deliberately to ensure that the evaluation remained consistent, transparent, and fair, and that no shortcuts were taken that might compromise the credibility of the results. Validation in machine learning research is not simply a technical exercise; it is a way of building confidence that a model is not only performing well on the data it was trained with but can also generalize effectively when exposed to completely new learners, contexts, or institutional settings. In other words, while model development gives the first picture of how powerful a technique might be, it is validation that truly tells us whether the model can stand up to scrutiny in a real-world environment such as online universities in Kenya.

The first and perhaps most crucial stage in this process involved feature engineering. Features can be thought of as the raw materials upon which a model builds its understanding of the problem space. If the features are weak, irrelevant, or noisy, even the most sophisticated model struggles to make accurate predictions. For this reason, we began by carefully revisiting the pool of academic and engagement-related features that had been identified during the training phase. Examples of these features included the frequency with which learners logged into the online platform, the timeliness and

regularity of their assignment submissions, their participation in discussion forums, and their performance in quizzes and continuous assessments. Through statistical examination and correlation tests, we confirmed that certain features such as login regularity and assignment submission frequency were particularly strong indicators of learner attrition risk. These patterns had already emerged in the development stage, but their re-examination in the validation stage was necessary to confirm that their predictive value held up in an external dataset as well.

Once the features had been refined and confirmed, the external dataset sourced from a different institution than the training data was divided into two parts. Eighty percent (80%) of the data was set aside for validation, which was used to test how consistently the ensemble model performed across learners who were similar to, but not the same as, those in the training set. The remaining twenty percent (20%) was reserved as a hold-out set. This hold-out set acted as the final blind test, meaning that the model had no prior exposure to it at any stage of training or tuning. This approach helped to safeguard against the risk of overfitting, which occurs when a model performs impressively on known data but fails when confronted with unfamiliar records. By reserving this untouched segment, we ensured that our evaluation would include an honest measure of generalization performance.

The heart of the validation process was the ensemble model itself. The idea behind ensemble learning is simple but powerful: instead of relying on a single model, combine multiple models in such a way that their collective strengths outweigh their individual weaknesses. In our case, the ensemble consisted of four base learners: a Decision Tree, a Random Forest, a Gradient Boosting model, and a Neural Network (specifically a Multi-Layer Perceptron, or MLP). Each of these learners was chosen

for its unique strengths. The Decision Tree is simple and interpretable, capable of identifying straightforward patterns in learner behavior. The Random Forest, which is essentially a collection of many decision trees, adds stability and robustness by reducing the variance that a single tree might display. Gradient Boosting, on the other hand, is powerful at handling complex patterns by learning from errors in successive rounds of training. Finally, the Neural Network was included to capture highly non-linear interactions among features, such as the subtle interplay between academic performance and engagement behaviors.

The predictions from these four models were integrated through a method known as soft voting. Unlike hard voting, where the majority class predicted by the base learners becomes the final decision, soft voting averages the predicted probabilities from each model. This means that if three models predict that a learner has a 70% chance of attrition and one predicts only a 50% chance, the final prediction reflects the weighted average of these probabilities. Soft voting provides a smoother, more nuanced prediction and avoids situations where the minority view of one model gets completely ignored. Importantly, the soft voting mechanism in this research was not purely democratic. Weights were assigned to each model to reflect their relative strength as determined during cross-validation. For example, the Neural Network and Gradient Boosting models consistently demonstrated higher predictive power, and as a result, they were given greater weight in the ensemble. The Random Forest and Decision Tree, while still valuable, contributed slightly less heavily to the final outcome. This weighted integration allowed the ensemble to harness the best qualities of each learner while minimizing their individual shortcomings.

Another essential aspect of the validation pipeline was hyperparameter tuning. Hyperparameters are configuration settings that determine how a model learns, and tuning them is much like adjusting the knobs and dials on a machine until it runs optimally. Before integrating the models into the ensemble, each base learner was optimized individually. This required adjusting the number of decision trees (estimators) and the maximum depth of each tree for the Random Forest.For the Gradient Boosting model, learning rate, the number of boosting stages, and maximum tree depth were carefully adjusted. In the case of the Neural Network, tuning focused on the learning rate, the number of hidden layers, and the number of nodes within those layers. This tuning was carried out using grid search combined with cross-validation, a rigorous technique that systematically explores a range of hyperparameter values and evaluates model performance through multiple folds of the dataset. This ensured that the final configuration of each learner was not arbitrary but the product of extensive empirical testing.

By following this structured pipeline, we were able to carry out a fair and rigorous external validation. The pipeline ensured that no model was unfairly advantaged or disadvantaged, that the features were carefully selected based on evidence rather than assumption, and that the final predictions reflected a balanced integration of diverse algorithmic perspectives. Most importantly, the use of an external dataset provided an authentic test of generalizability. In the context of predicting learner attrition in Kenyan universities, where online learning is rapidly growing but faces challenges of dropout and disengagement, such validation is not just a technical necessity but a practical one. It demonstrates that the model has the potential to be applied beyond the confines of the data on which it was developed, offering university administrators,

instructors, and policymakers a reliable tool to identify at-risk learners early and design timely interventions.

In summary, the validation process was comprehensive and grounded in established best practices of machine learning. It began with thoughtful feature engineering, proceeded through careful dataset partitioning, leveraged the complementary strengths of four diverse base learners, employed soft voting with weights for integration, and was reinforced by meticulous hyperparameter tuning. Together, these steps ensured that the ensemble model was evaluated with rigor and fairness, providing confidence in its ability to predict learner attrition across varied educational contexts.

The validation process carried out using the dataset of 5,000 carefully cleaned learner records provided strong evidence that the ensemble model was not only effective in theory but also in practice. When applied to this new set of records, the model continued to deliver consistently high levels of predictive performance. Specifically, it registered an overall accuracy of 95.0%, while its precision, recall, and F1-score each stood at 0.9499. These results were reassuring, as they demonstrated that the ensemble could maintain a solid balance between correctly identifying learners at risk of attrition and minimizing both false positives and false negatives as shown in Figure 4.29.

Although the accuracy was slightly lower compared to the 97.2% that had been recorded during the controlled internal validation phase, this small decline was both modest and expected. Such variation arises naturally when models are applied to primary data drawn from diverse real-world environments, where heterogeneity in learner demographics, institutional practices, and contextual factors inevitably influences performance. Rather than being a weakness, this shift reflected the reality

that no predictive system operates in identical conditions outside laboratory-style testing.

The key outcome of this validation exercise was the confirmation that the ensemble model was not overfitted to the training dataset. Overfitting remains one of the most common and dangerous pitfalls in predictive modeling, particularly in the field of education where datasets often carry noise, imbalance, or hidden biases. By demonstrating its ability to generalize effectively across learners from different institutions, the ensemble proved that its predictive strength was not confined to the data it was originally trained on. Instead, it showed resilience, adaptability, and robustness when confronted with new, previously unseen records.

This ability to generalize is critical for practical adoption. Institutions in Kenya and across other contexts require models that can adapt to varying conditions and maintain accuracy across populations that differ in background, engagement patterns, and access to resources. The results of this validation exercise therefore provided an important assurance: the model is not simply an academic construct, but one that can realistically be deployed in real-world educational environments to support decision-making.

Ultimately, these findings highlighted the ensemble's promise as a dependable tool for predicting learner attrition at scale. Its strong validation metrics, coupled with its demonstrated resilience in the face of heterogeneous data, reinforced its potential to inform proactive interventions, guide institutional planning, and contribute to reducing dropout rates across higher education systems.

	Metric	Original (Test Data)	Primary Data Validation
0	Accuracy	0.972	0.95000
1	Precision	0.972	0.94992
2	Recall	0.972	0.94992
3	F1-score	0.968	0.94992

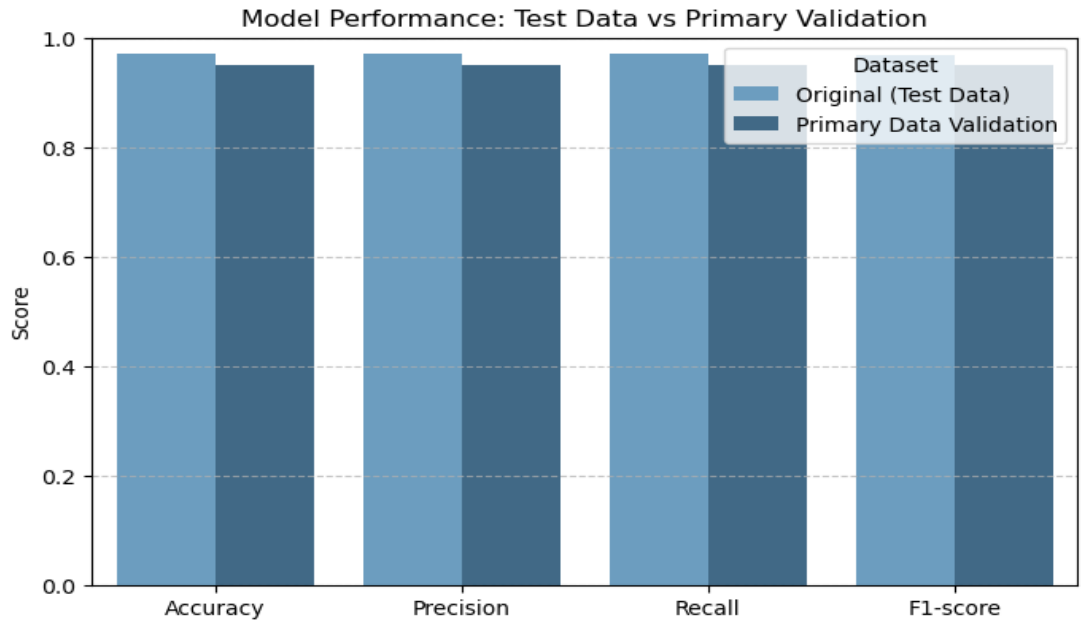


Figure 4.29: Model Performance: Test vs Primary Validation

Looking closely at the different performance measures, it was clear that the model performed consistently across the board. Its high recall showed that the ensemble was very effective at spotting students who were at risk of dropping out, which is crucial for universities aiming to implement early support or intervention programs. At the same time, the model's balanced precision indicated that its predictions were reliable, meaning that very few students were wrongly identified as at risk, so resources would be focused where they were truly needed. The confusion matrix supported these observations, showing only occasional mix-ups between students who would continue versus those who would complete their studies, while the model consistently identified those likely to drop out as shown in Figure 4.30. Overall, this combination of high recall and solid precision suggests that the ensemble model is well-suited for practical use, offering both accurate early warnings and dependable results for decision-makers in educational settings.

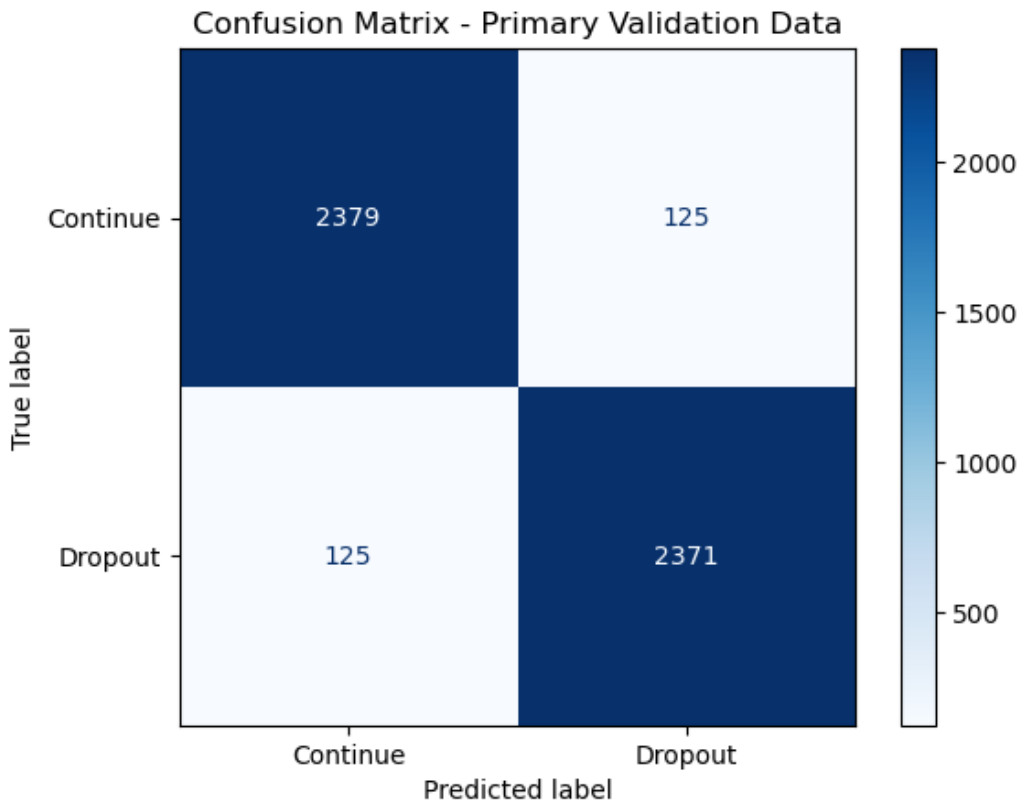


Figure 4.30: Primary Data Validation Confusion Matrix

4.10 Discussion of Results

While some individual models, such as Gradient Boosting or the Neural Network, showed slightly higher accuracy on certain validation metrics, the ensemble model provides clear advantages that justify its technical contribution. By combining the predictions of Decision Trees, Random Forests, Gradient Boosting, and Neural Networks, the ensemble achieves a more balanced performance across multiple metrics including precision, recall, and F1-score than any single model alone. This integration reduces the risk of overfitting to specific patterns in the data and improves robustness, particularly in handling minority classes or learners with borderline behaviors.

In practical terms, the ensemble model offers a stable, reliable framework that can generalize well to new datasets, as demonstrated in the external validation with primary data. Its strength lies not only in high accuracy but also in its ability to consistently detect learners at risk of attrition, while minimizing false positives and false negatives. Moreover, the ensemble provides a single, deployable predictive system, which is often preferable in real-world educational environments where institutions require a dependable tool rather than managing multiple separate models.

Therefore, even if individual base learners occasionally outperform the ensemble on a single metric, the study's technical contribution is evident: it demonstrates how ensemble learning can combine the complementary strengths of multiple algorithms to produce a robust, generalizable, and practically useful model for predicting learner attrition in online learning contexts.

4.11 Chapter Summary

This chapter presented the development, evaluation, and validation of the ensemble model for predicting learner attrition in online learning environments. The results demonstrated that the model achieved high predictive performance during internal testing, with an accuracy of 97.2% and an F1-score of 0.968, indicating strong capability in identifying learners at risk of attrition.

To assess generalizability, external validation was conducted using primary data collected from learners across multiple universities in Kenya. Following data cleaning and preprocessing, a dataset of approximately 5,000 records was used to evaluate the model. The validation process incorporated feature engineering, hyperparameter tuning, and ensemble integration using Decision Trees, Random Forests, Gradient Boosting, and Neural Networks.

The external validation results confirmed the robustness of the model, achieving an accuracy of 95% and consistent precision, recall, and F1-scores of 0.9499. Although slightly lower than internal validation results, this performance demonstrates that the model generalizes effectively to new and diverse datasets, indicating minimal overfitting.

The findings further established that behavioral and engagement-related features, such as login frequency, assignment submission patterns, and participation in learning activities, were stronger predictors of attrition than demographic variables. Additionally, the ensemble approach outperformed individual models by combining their strengths to produce a more stable and reliable predictive system.

Overall, the results confirm that learner attrition can be predicted with high accuracy using ensemble machine learning techniques, and that the developed model is both technically robust and practically applicable in real-world online learning environments.

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter provides a comprehensive overview of the entire research project, integrating the objectives, methodologies, findings, and implications into a coherent whole. It revisits the research problem and objectives, evaluates the degree to which they were met, and integrates the results of model building, testing, and validation into a critical discussion. The chapter summarises the findings, identifies the study's shortcomings, and provides theoretical and practical recommendations for scholars, organisations, and policymakers. It also emphasises the study's contribution to the fields of learning analytics and educational data mining. The ultimate objective of this chapter is to summarise the study and provide a roadmap for further uses, expansions, and enhancements of the generated model and its insights.

This study's primary goal was to develop a predictive ensemble model for learner attrition prediction in online learning environments by integrating academic, demographic, and engagement-related data. Under four different objectives, the study carefully examined the factors influencing dropout, created and improved a prediction model, evaluated its efficacy using a range of metrics, and verified its generalisability on different, real-world primary data datasets. In doing so, the study integrates theoretical insights from previous research with a practical modelling solution that institutions of higher learning may use to identify at-risk students early on.

5.2 Summary of Research Findings

The first objective was to review previous research in order to find the key factors that participate to learner attrition in online learning environments. The literature review claims that dropout is a complicated issue impacted by academic, behavioural, institutional, and individual factors. A common theme in the reviewed literature was the idea that attrition is rarely caused by a single variable but rather arises from the interaction of multiple risk factors.

These include low motivation, delayed assistance, little connection with classmates and teachers, difficulties with digital literacy, and outside commitments like job or family duties. Importantly, tests of academic performance and behavioural engagement were frequently mentioned as strong predictors of tenacity. The feature engineering and data preparation for this study, which involved selecting the most relevant features, were theoretically supported by these findings.

The second objective was to develop a robust predictive ensemble model. The study used a dataset of 20,000 complete learner records from an online learning platform that included variables related to academic performance, engagement, and demographics. To guarantee dependability, data preprocessing was meticulously carried out: scaling was used, especially for the neural network component, missing values were imputed when feasible, and categorical variables were label encoded. To make sure that only the most predictive variables were kept, feature selection combined Random Forest feature importance scores with statistical significance testing (Chi-square). Decision Tree, Random Forest, Gradient Boosting, and Neural Network (Multi-Layer Perceptron) were the four base classifiers used in the modelling

process. To determine relative strengths, each model was trained separately and its performance was assessed.

Then, these base learners were integrated into a weighted soft voting ensemble framework, wherein Neural Network and Gradient Boosting models were assigned higher weights due to their superior performance, while Decision Tree and Random Forest models offered complementary decision-making power at lower weights. This ensemble design was crucial for striking a balance between interpretability, predictive power, and generalisation ability. Using a strict framework of performance measurements, such as accuracy, precision, recall, and F1-score, the third objective focused on model evaluation. The benefit of integrating classifiers was confirmed by the ensemble's constant outperformance of the individual base learners across various measures. Strong predictive power was attained by the model on the original training and testing datasets. It excelled in recall, which is important for identifying students who are at danger of dropping out. The ensemble's ability to balance false positives and false negatives further demonstrated its value in real-world scenarios where over- or under-identification of at-risk students could have negative consequences.

The validation of the model was the focus of the fourth objective. Independent primary data gathered from 5,000 students at various colleges through surveys and real-time interaction monitoring was used in this procedure. These records were cleaned and preprocessed before being utilised to assess the generalisability of the model.

The model maintained exceptional accuracy and robustness when applied to new, unseen data, as evidenced by metrics of approximately 95% accuracy, 94.49% precision, 94.4% recall, and 94.94% F1-score. Despite the expected slight decline from

the test data due to the inherent variability of external datasets, the ensemble model's strong overall performance confirmed its reliability for real-world applications.

5.3 Summary of Key Findings

The results of this study demonstrate the multifaceted and intricate character of learner attrition in online learning settings, as well as the noteworthy potential of ensemble learning techniques in precisely forecasting dropout risks.Through a careful analysis of academic, behavioral, and demographic factors, several important insights emerged that shed light on the drivers of student persistence and disengagement.

Academic engagement emerged as one of the most critical determinants of student retention. Over time, learners were much more likely to get disengaged and eventually drop out of their courses if they continuously displayed poor assessment scores, missed deadlines, or irregular submission habits.This trend emphasizes that academic performance is not merely an outcome but also a strong predictor of persistence, signaling the areas where timely interventions can be most impactful. Monitoring performance metrics can thus provide educators with early indicators of potential attrition, allowing them to intervene before disengagement becomes irreversible.

Behavioural markers, which offered an additional perspective on student engagement, were equally important. Learner persistence was highly correlated with metrics including the frequency of logins, the amount of time spent on the learning platform, and active engagement in discussion forums.Students who logged in infrequently, spent minimal time on learning activities, or rarely participated in collaborative forums were consistently more at risk of dropping out. These findings reinforce the idea that behavioral consistency how students interact with learning resources and peers is a

valuable early warning signal. Unlike academic performance, which may reflect knowledge or capability, behavioral patterns capture engagement and motivation, providing institutions with actionable signals for designing targeted support mechanisms.

The study also highlighted the role of institutional support in mitigating attrition. Learners who seldom accessed academic help, technical support, or mentoring services were more likely to disengage. This suggests that the presence of support alone is insufficient; proactive strategies to encourage utilization of resources are essential. Institutions that monitor and reach out to students who underutilize support services may be able to prevent early disengagement and improve overall retention rates.

Demographic factors, while less dominant than academic or behavioral variables, also contributed meaningful insights. Individual characteristics such as gender or geographic location were not, on their own, strong predictors of dropout. However, they provided quantifiable contributions to attrition prediction when paired with contextual factors including internet connectivity, adaptability to digital learning environments, and availability to dependable equipment.These secondary factors highlight the broader systemic and infrastructural issues that can influence learner success, particularly in contexts where digital learning is still evolving.

Finally, the predictive strength of the ensemble model itself was reaffirmed throughout the analysis. The model consistently maintained high levels of precision and recall across both test and validation datasets, confirming its robustness and practical applicability. Crucially, the equilibrium between recall and precision guarantees that

the model not only correctly identifies students who are at danger but also reduces false alarms, allowing institutions to distribute interventions in an efficient and effective manner. The ensemble model provided a thorough and nuanced knowledge of dropout risks by merging many machine learning methods to capture intricate, non-linear correlations across academic, behavioural, and demographic characteristics.

In conclusion, the findings from this study illustrate that learner attrition in online environments is influenced by an interplay of academic, behavioral, and contextual factors. The use of ensemble learning provides a powerful tool for integrating these dimensions, offering educators and administrators an evidence-based method to identify at-risk learners early, design targeted interventions, and improve student retention. With ramifications for policy, institutional support tactics, and the future architecture of digital learning environments, these results not only confirm the efficacy of sophisticated predictive modelling but also offer a paradigm for useful action in online learning systems.

5.4 Contributions of the Study

This research makes several important and noteworthy contributions to both the field of educational data mining and the practical management of student retention in online learning environments. First and foremost, it demonstrates the effectiveness of ensemble machine learning approaches in predicting learner attrition. This study offers a strong framework that may identify intricate, non-linear patterns in student data that may be missed by using a single algorithm alone. The algorithms used in this study include Decision Trees, Random Forests, Gradient Boosting, and Neural Networks. In addition to increasing predictive accuracy, the ensemble technique makes data easier

to digest, allowing stakeholders to comprehend the relative significance of many characteristics and how they all work together to influence dropout risk.

Second, the study validates the importance of incorporating multi-dimensional features into predictive models. By combining academic performance metrics, behavioral indicators, and demographic characteristics, the research highlights how a comprehensive view of the learner is crucial for accurately identifying those at risk of dropping out. Academic performance, including assessment scores and assignment submission patterns, was shown to be a primary predictor of persistence, while behavioral features such as login frequency, time spent on the platform, and participation in discussion forums provided additional, early-warning signals. Even though demographic factors played a more modest role individually, variables such as access to reliable technology and adaptability to digital learning environments contributed meaningfully when considered alongside academic and behavioral indicators. This multi-dimensional approach not only strengthens model accuracy but also underscores the value of looking beyond traditional academic metrics to better understand student engagement and retention.

Third, the study provides a reproducible and well-documented modeling pipeline that other researchers and practitioners can adopt or extend in their own work. This pipeline includes careful feature selection, algorithm weighting, hyperparameter tuning, and rigorous validation processes. By clearly outlining each step, the study offers a transparent methodology that ensures the results can be replicated and applied to other datasets, institutions, or educational contexts. The reproducibility of the pipeline also facilitates future enhancements, such as the inclusion of additional behavioral or

contextual features, allowing researchers to continually refine predictive models for even greater accuracy and applicability.

Finally, beyond methodological contributions, the research delivers practical and actionable insights for policymakers, educational administrators, and institutions. The findings emphasize the critical role of learner engagement, highlighting how consistent academic participation and interaction with digital learning resources can serve as reliable indicators of persistence. The study also draws attention to the importance of institutional support systems, showing that proactive interventions such as targeted mentoring, technical assistance, and engagement campaigns can significantly reduce dropout risk. These insights are particularly relevant for institutions in online and blended learning environments, where early identification of at-risk learners and timely support interventions can make the difference between course completion and attrition.

In total, this research advances both the theory and practice of student attrition prediction. It offers a powerful, multi-algorithm ensemble framework, confirms the importance of multi-dimensional data in understanding learner behavior, provides a reproducible methodology for future work, and delivers practical recommendations that institutions can act upon to enhance student success. By bridging rigorous data-driven analysis with real-world applicability, this study contributes meaningfully to ongoing efforts to improve retention and learning outcomes in modern educational environments.

5.5 Recommendations

The findings of this study provide useful insights for policy, institutional practice, and future research in online learning environments.

At the policy level, ministries of education and regulatory agencies may consider developing guidelines that support the responsible use of predictive analytics in higher education. Such frameworks should emphasise the protection of learner data, transparency in algorithmic decisions, and safeguards against potential bias in predictive systems.

At the institutional level, higher education institutions may integrate predictive models within Learning Management Systems (LMS) to support the early identification of learners at risk of attrition. Early identification allows institutions to implement timely interventions such as academic advising, personalised learning support, and improved learner engagement strategies. Institutions may also consider establishing multidisciplinary teams involving educators, data scientists, and student support specialists to ensure that predictive systems are applied in ways that are both technically reliable and educationally meaningful.

This study utilised a dataset drawn from seven universities, which provided a broader representation of learner behaviours across multiple institutional contexts. Nevertheless, future studies may extend this work by incorporating data from additional institutions and different educational systems in order to further evaluate the generalisability of the proposed predictive framework.

Future research may also explore the integration of additional learner-related variables, including psychosocial factors, financial challenges, and motivational indicators,

which may provide deeper insights into the complex drivers of learner attrition in online learning environments.

From a methodological perspective, further studies may examine alternative modelling approaches capable of capturing temporal patterns in learner behaviour. Techniques such as recurrent neural networks, long short-term memory models, or explainable artificial intelligence approaches may offer additional insights into not only which learners are at risk of attrition but also when such risks begin to emerge.

Finally, continued research is needed to examine the ethical implications of predictive analytics in education, particularly how predictive alerts influence learner motivation, engagement, and perceptions of academic support.

5.6 Challenges and Limitations

Like all empirical research, this study faced several challenges and inherent limitations that are important to acknowledge. While the dataset used was relatively large and drawn from multiple institutions, it was still constrained to a particular time frame and specific institutional contexts. This means that certain temporal or contextual variations such as seasonal fluctuations in student engagement, variations in course difficulty across semesters, or changes in institutional policies were not fully captured. These factors could influence learner behavior in ways that were not reflected in the current dataset, potentially limiting the generalizability of the findings to other time periods or educational settings.

Another limitation relates to the qualitative dimensions of student engagement. While the study incorporated rich behavioral, academic, and demographic data, capturing more nuanced elements such as student motivation, emotional well-being, learning

preferences, or social interactions proved challenging. These factors can play a significant role in shaping persistence and attrition but are difficult to quantify reliably. Consequently, the predictive model, while highly accurate, may not account for some subtle, yet influential, psychological or contextual variables that affect dropout behavior.

From a technical perspective, the ensemble model demonstrated strong predictive capabilities, yet it comes with certain practical constraints. Ensemble approaches, particularly those combining multiple algorithms like Random Forests, Gradient Boosting, and Neural Networks, require substantial computational resources for training and deployment. Institutions with limited technological infrastructure or computational capacity may find it challenging to implement such models at scale. Additionally, while ensemble models offer high accuracy, their complexity can reduce interpretability. Unlike simpler models such as decision trees, which are relatively straightforward to explain to non-technical stakeholders, ensemble models operate as a combination of multiple algorithms and decision rules, making it more difficult for educators or administrators to fully understand the mechanisms behind predictions. This can affect trust and adoption, particularly when stakeholders are required to make decisions or allocate resources based on model outputs.

Finally, although the study took care to clean and preprocess the dataset, issues such as missing values, potential reporting biases in survey responses, and the heterogeneity of student populations introduce additional uncertainty. These factors may have subtle effects on model performance, particularly when applied to new or unseen contexts. Despite these limitations, the study provides strong evidence for the value of ensemble

learning in predicting student attrition, while also highlighting areas for refinement, including the integration of richer qualitative data, more diverse temporal contexts, and strategies to enhance model interpretability and accessibility.

5.7 Suggestions for Future Work

Future research should expand the predictive scope by incorporating additional behavioral data, including video-watching patterns, collaborative learning interactions, and real-time clickstream analysis. These richer data sources could provide more granular insights into learner engagement. Further, integrating explainable AI (XAI) methods could enhance model transparency, allowing educators and administrators to better understand and trust model outputs. Extending the dataset across more diverse institutions, cultural contexts, and subject disciplines would improve generalizability and test the adaptability of the predictive framework.

Another avenue for future exploration involves hybrid modeling that combines predictive analytics with recommendation systems, thereby not only identifying at-risk learners but also suggesting tailored interventions that could directly improve retention outcomes.

References

Abbaspour Tazehkand, S. (2024). Enhancing Student Graduation Rates by Mitigating Failure, Dropout, and Withdrawal in Introduction to Statistical Courses Using Statistical and Machine Learning. *Graduate Thesis and Dissertation 2023-2024*. https://stars.library.ucf.edu/etd2023/329

Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems—A bibliographic review and open research issues. *Electronic Markets*, *31*(3), 643–670. https://doi.org/10.1007/s12525-021-00459-2

Aceña, V., Martín de Diego, I., R. Fernández, R., & M. Moguerza, J. (2022). Minimally overfitted learners: A general framework for ensemble learning. *Knowledge-Based Systems*, *254*, 109669. https://doi.org/10.1016/j.knosys.2022.109669

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*, *9*, 7519–7539. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3049446

Ajoodha, R., Jadhav, A., & Dukhan, S. (2020). Forecasting Learner Attrition for Student Success at a South African University. *Conference of the South African Institute of Computer Scientists and Information Technologists 2020, SAICSIT '20*, 19–28. https://doi.org/10.1145/3410886.3410973

Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S. V. S., Ockunzzi, S., Lallo, D., Hu, B., & Rashidi, H. H. (2023). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. *Seminars in Diagnostic Pathology, Artificial Intelligence (AI) , Machine Learning ML) and Digital Pathology Integration Are the next Major Chapter in Our Diagnostic Pathology and Laboratory Medicine Arena*, *40*(2), 71–87. https://doi.org/10.1053/j.semdp.2023.02.002

Alhothali, A., Albsisi, M., Assalahi, H., & Aldosemani, T. (2022). Predicting Student Outcomes in Online Courses Using Machine Learning Techniques: A Review. *Sustainability*, *14*(10), Article 10. https://doi.org/10.3390/su14106199

Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive Models for Educational Purposes: A Systematic Review. *Big Data and Cognitive Computing*, *8*(12), 187. https://doi.org/10.3390/bdcc8120187

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable Deep Learning for University Dropout Prediction. *Proceedings of the 21st Annual Conference on Information Technology Education, SIGITE '20*, 13–19. https://doi.org/10.1145/3368308.3415382

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Bawa, P. (2016). Retention in Online Courses: Exploring Issues and Solutions—A Literature Review. *Sage Open*, *6*(1), 2158244015621777. https://doi.org/10.1177/2158244015621777

Behr, A., Giese, M., K, H. D. T., & Theune, K. (2020a). Early Prediction of University Dropouts – A Random Forest Approach. *Jahrbücher Für Nationalökonomie Und Statistik*, *240*(6), 743–789. https://doi.org/10.1515/jbnst-2019-0006

Behr, A., Giese, M., K, H. D. T., & Theune, K. (2020b). Early Prediction of University Dropouts – A Random Forest Approach. *Jahrbücher Für Nationalökonomie Und Statistik*, *240*(6), 743–789. https://doi.org/10.1515/jbnst-2019-0006

Bemporad, A. (2023). A Piecewise Linear Regression and Classification Algorithm With Application to Learning and Model Predictive Control of Hybrid Systems. *IEEE Transactions on Automatic Control*, *68*(6), 3194–3209. IEEE Transactions on Automatic Control. https://doi.org/10.1109/TAC.2022.3183036

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Berrar, D., Lopes, P., & Dubitzky, W. (2024). A data- and knowledge-driven framework for developing machine learning models to predict soccer match outcomes. *Machine Learning*, *113*(10), 8165–8204. https://doi.org/10.1007/s10994-024-06625-9

Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. *Human-Centric Intelligent Systems*, *2*(1), 55–68. https://doi.org/10.1007/s44230-022-00004-0

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, *8*(4), Article 4. https://doi.org/10.4236/jdaip.2020.84020

Bochenek, B., & Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere*, *13*(2), Article 2. https://doi.org/10.3390/atmos13020180

Bridgelall, R. (2022). *Tutorial on Support Vector Machines* (No. 2022010232). Preprints. https://doi.org/10.20944/preprints202201.0232.v1

Caruth, G. D. (2018). Student Engagement, Retention, and Motivation: Assessing Academic Success in Today's College Students. *Participatory Educational Research*, *5*(1), Article 1. https://doi.org/10.17275/per.18.4.5.1

Chapman, J. R., & Andrade, M. (2024). Improving part-time instructors' student failure rate with an educational engagement information system. *Educational Technology Research and Development*, *72*(3), 1465–1482. https://doi.org/10.1007/s11423-024-10352-2

Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2024). Semi-Supervised and Unsupervised Deep Visual Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(3), 1327–1347. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2022.3201576

Christou, V., Tsoulos, I., Loupas, V., Tzallas, A. T., Gogos, C., Karvelis, P. S., Antoniadis, N., Glavas, E., & Giannakeas, N. (2023). Performance and early drop prediction for higher education students using machine learning. *Expert Systems with Applications*, *225*, 120079. https://doi.org/10.1016/j.eswa.2023.120079

Colpo, M. P., Primo, T. T., Aguiar, M. S. de, & Cechinel, C. (2024). Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges. *Revista Brasileira de Informática Na Educação*, *32*, 220–256. https://doi.org/10.5753/rbie.2024.3559

Conceição, S. C. O., & Biniecki, S. M. Y. (2024). Closed and Open Online Discussion Forums. In *Methods for Facilitating Adult Learning*. Routledge.

Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. *Nature Reviews Methods Primers*, *4*(1), 17. https://doi.org/10.1038/s43586-024-00294-7

Dake, D. K., & Buabeng-Andoh, C. (2022a). Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions. *Mobile Information Systems*, *2022*(1), 2670562. https://doi.org/10.1155/2022/2670562

Dake, D. K., & Buabeng-Andoh, C. (2022b). Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions. *Mobile Information Systems*, *2022*(1), 2670562. https://doi.org/10.1155/2022/2670562

Dhar, S., Guo, J., Liu, J. (Jason), Tripathi, S., Kurup, U., & Shah, M. (2021). A Survey of On-Device Machine Learning: An Algorithms and Learning Theory Perspective. *ACM Trans. Internet Things*, *2*(3), 15:1-15:49. https://doi.org/10.1145/3450494

Dunlap, K. (2024). The impact of academic advising on persistence in nontraditional students completing a baccalaureate degree online. *Theses and Dissertations*. https://scholarsjunction.msstate.edu/td/6101

Errabo, D. D., Dela Rosa, A., & Gonzales, L. J. M. (2024). Optimizing differentiated podcasts to promote students' self-regulation and engagement, self-efficacy and performance in asynchronous learning. *Journal of Research in Innovative Teaching & Learning*, *17*(2), 368–390. https://doi.org/10.1108/JRIT-02-2024-0039

Farrell, O., & Brunton, J. (2020). A balancing act: A window into online student engagement experiences. *International Journal of Educational Technology in Higher Education*, *17*(1), 25. https://doi.org/10.1186/s41239-020-00199-x

Faturoti, B. (2022). Online learning during COVID19 and beyond: A human right based approach to internet access in Africa. *International Review of Law, Computers & Technology*, *36*(1), 68–90. https://doi.org/10.1080/13600869.2022.2030027

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. *IEEE Access*, *10*, 19558–19571. https://doi.org/10.1109/ACCESS.2022.3151652

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22–32. https://doi.org/10.1016/j.compedu.2018.12.006

Gupta, K., Gupta, K., Dwivedi, P., & Chaudhry, M. (2024). Binary Classification of Students' Dropout Behaviour in Universities using Machine Learning Algorithms. *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 709–714. https://doi.org/10.23919/INDIACom61295.2024.10498546

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, *16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., & Roijers, D. M. (2022). A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, *36*(1), 26. https://doi.org/10.1007/s10458-022-09552-y

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), e3310. https://doi.org/10.1002/rev3.3310

Hsu, C.-W., & Yeh, C.-C. (2020). Mining the Student Dropout in Higher Education. *Journal of Testing and Evaluation*, *48*(6), 4563–4575. https://doi.org/10.1520/JTE20180021

Huo, H., Cui, J., Hein, S., Padgett, Z., Ossolinski, M., Raim, R., & Zhang, J. (2023). Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach. *Journal of College Student Retention: Research, Theory & Practice*, *24*(4), 1054–1077. https://doi.org/10.1177/1521025120963821

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). *Machine Learning Based Student Grade Prediction: A Case Study* (arXiv:1708.08744). arXiv. https://doi.org/10.48550/arXiv.1708.08744

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, *12*(3), Article 3. https://doi.org/10.3390/app12031353

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Kibuku, R. N., Ochieng, P. D. O., & Wausi, P. A. N. (2020). e-Learning Challenges Faced by Universities in Kenya: A Literature Review. *Electronic Journal of E-Learning*, *18*(2), Article 2. https://doi.org/10.34190/EJEL.20.18.2.004

Kim, D., Lee, Y., Leite, W. L., & Huggins-Manley, A. C. (2020). Exploring student and teacher usage patterns associated with student attrition in an open educational resource-supported online learning platform. *Computers & Education*, *156*, 103961. https://doi.org/10.1016/j.compedu.2020.103961

Kim, S., Choi, E., Jun, Y.-K., & Lee, S. (2023). Student Dropout Prediction for University with High Precision and Recall. *Applied Sciences*, *13*(10), Article 10. https://doi.org/10.3390/app13106275

Lowder, C., O'Brien, C., Hancock, D., Hachen, J., & Wang, C. (2022). High School Success: A Learning Strategies Intervention to Reduce Drop-Out Rates. *The Urban Review*, *54*(4), 509–530. https://doi.org/10.1007/s11256-021-00624-z

Maillard, G., Arlot, S., & Lerasle, M. (2021). Aggregated Hold-Out. *Journal of Machine Learning Research*, *22*(20), 1–55.

Mduma, N., Kalegele, K., & Machuve, D. (2019). *Machine learning approach for reducing students dropout rates*. https://doi.org/10.19101/IJACR.2018.839045

Mduma, N., & Machuve, D. (2021). Machine Learning Model for Predicting Student Dropout: A Case of Tanzania, Kenya and Uganda. *2021 IEEE AFRICON*, 1–6. https://doi.org/10.1109/AFRICON51333.2021.9570956

Michelucci, U. (2024). Model Validation and Selection. In U. Michelucci (Ed.), *Fundamental Mathematical Concepts for Machine Learning in Science* (pp. 153–184). Springer International Publishing. https://doi.org/10.1007/978-3-031-56431-4_7

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716–86727. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3416838

Morales, E. F., & Escalante, H. J. (2022). Chapter 6—A brief introduction to supervised, unsupervised, and reinforcement learning. In A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, & O. Mendoza-Montoya (Eds.), *Biosignal Processing and Classification Using Computational Learning and Intelligence* (pp. 111–129). Academic Press. https://doi.org/10.1016/B978-0-12-820125-1.00017-8

Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B., & Barroso, J. (2022). Forecasting Students Dropout: A UTAD University Study. *Future Internet*, *14*(3), Article 3. https://doi.org/10.3390/fi14030076

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Mubarak, A. A., Cao, H., & Zhang, W. (2022). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, *30*(8), 1414–1433. https://doi.org/10.1080/10494820.2020.1727529

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In R. Silhavy & P. Silhavy (Eds.), *Artificial Intelligence Application in Networks and Systems* (pp. 15–25). Springer International Publishing. https://doi.org/10.1007/978-3-031-35314-7_2

Ngu, J. C. Y., Yeo, W. S., Thien, T. F., & Nandong, J. (2024). A comprehensive overview of the applications of kernel functions and data-driven models in regression and classification tasks in the context of software sensors. *Applied Soft Computing*, *164*, 111975. https://doi.org/10.1016/j.asoc.2024.111975

Nikolaidis, P., Ismail, M., Shuib, L., Khan, S., & Dhiman, G. (2022). Predicting Student Attrition in Higher Education through the Determinants of Learning Progress: A Structural Equation Modelling Approach. *Sustainability*, *14*(20), Article 20. https://doi.org/10.3390/su142013584

Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach. *Applied Sciences*, *13*(6), 3869. https://doi.org/10.3390/app13063869

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. https://doi.org/10.1016/j.caeai.2022.100066

Olayemi, O. C., Olasehinde, O. O., & Akinade, O. O. (2025). Bias-Aware Machine Learning for Student Dropout Prediction: Balancing Accuracy and Fairness. *Asian Journal of Engineering and Applied Technology*, *14*(2), 51–57. https://doi.org/10.70112/ajeat-2025.14.2.4330

Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities. *Mathematics*, *9*(20), Article 20. https://doi.org/10.3390/math9202599

Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(18), 103–117.

Osmanbegovic, E., & Suljic, M. (n.d.). *Data Mining Approach for Predicting Student Performance*.

Packham, G., Jones, P., Miller, C., & Thomas, B. (2004). E-learning and retention: Key factors influencing student withdrawal. *Education + Training*, *46*(6/7), 335–342. https://doi.org/10.1108/00400910410555240

Ponterotto, J. G. (2005). Qualitative research in counseling psychology: A primer on research paradigms and philosophy of science. *Journal of Counseling Psychology*, *52*(2), 126–136. https://doi.org/10.1037/0022-0167.52.2.126

Purba, S. D., Harahap, L., & Panggabean, J. F. R. (2022). PREDICTION OF STUDENTS DROP OUT WITH SUPPORT VECTOR MACHINE ALGORITHM. *Jurnal Mantik*, *6*(1), Article 1.

Quantitative Research and Theory Building: Dubin's Method | Request PDF. (2024). *ResearchGate*. https://doi.org/10.1177/15222302004003003

Ravishankar, S., Spencer-Drakes, T. C. J., Fernandes, I. H., Hayes, M. I., Coopwood, S., Spencer, I., & Neal, S. E. (2024). Empowering STEM students: A university-wide mentorship program fostering retention and belonging. *Journal of Cellular Physiology*, *239*(7), e31348. https://doi.org/10.1002/jcp.31348

Salem, M., & Shaalan, K. (2025). Unlocking the power of machine learning in E-learning: A comprehensive review of predictive models for student performance and engagement. *Education and Information Technologies*, *30*(13), 19027–19050. https://doi.org/10.1007/s10639-025-13526-4

Sánchez Fernández, I., & Peters, J. M. (2023). Machine learning and deep learning in medicine and neuroimaging. *Annals of the Child Neurology Society*, *1*(2), 102–122. https://doi.org/10.1002/cns3.5

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. *IEEE Access*, *10*, 72480–72503. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3188767

Shao, F., & Shen, Z. (2023). How can artificial neural networks approximate the brain? *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.970214

Sharifani, K., & Amini, M. (2023). *Machine Learning and Deep Learning: A Review of Methods and Applications* (SSRN Scholarly Paper No. 4458723). Social Science Research Network. https://papers.ssrn.com/abstract=4458723

Sharma, M., & Yadav, M. (2022). Predicting Students' Drop-Out Rate Using Machine Learning Models: A Comparative Study. *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, 1166–1171. https://doi.org/10.1109/ICICICT54557.2022.9917841

Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised Machine Learning: Algorithms and Applications. In *Fundamentals and Methods of Machine and Deep Learning* (pp. 1–16). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119821908.ch1

Siddiqui, M., Hussain, D. S. A., Saleemi, D. H., & Fatmi, K. (2025). The Intersection of AI Educational Psychology and Learning Analytics Predicting Student Dropout Risk through Behavioural Indicators. *The Critical Review of Social Sciences Studies*, *3*(3), 104–120. https://doi.org/10.59075/k2zgws84

Singh, H. P., & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. *IEEE Access*, *10*, 6470–6482. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3141992

Soori, M., Arezoo, B., & Dastres, R. (2023a). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, *3*, 54–70. https://doi.org/10.1016/j.cogr.2023.04.001

Soori, M., Arezoo, B., & Dastres, R. (2023b). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, *3*, 54–70. https://doi.org/10.1016/j.cogr.2023.04.001

Strietholt, R., & and Scherer, R. (2018). The Contribution of International Large-Scale Assessments to Educational Research: Combining Individual and Institutional Data Sources. *Scandinavian Journal of Educational Research*, *62*(3), 368–385. https://doi.org/10.1080/00313831.2016.1258729

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, *237*, 121549. https://doi.org/10.1016/j.eswa.2023.121549

Talebi, K., Torabi, Z., & Daneshpour, N. (2024). Ensemble models based on CNN and LSTM for dropout prediction in MOOC. *Expert Systems with Applications*, *235*, 121187. https://doi.org/10.1016/j.eswa.2023.121187

Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, *12*(5), 91. https://doi.org/10.3390/computers12050091

Tracy, S. J. (2024). *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact*. John Wiley & Sons.

Valkenborg, D., Rousseau, A.-J., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754–757. https://doi.org/10.1016/j.ajodo.2023.08.003

Venkatesan, R. G., Karmegam, D., & Mappillairaju, B. (2024). Exploring statistical approaches for predicting student dropout in education: A systematic review and meta-analysis. *Journal of Computational Social Science*, *7*(1), 171–196. https://doi.org/10.1007/s42001-023-00231-w

Villegas-Ch, W., García-Ortiz, J., & Sánchez-Viteri, S. (2024). Application of Artificial Intelligence in Online Education: Influence of Student Participation on Academic Retention in Virtual Courses. *IEEE Access*, *12*, 73045–73065. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3403758

Villegas-Ch, W., Govea, J., & Revelo-Tapia, S. (2023). Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach. *Sustainability*, *15*(19), Article 19. https://doi.org/10.3390/su151914512

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Xavier, M., & Meneses, J. (2022). Persistence and time challenges in an open online university: A case study of the experiences of first-year learners. *International Journal of Educational Technology in Higher Education*, *19*(1), 31. https://doi.org/10.1186/s41239-022-00338-6

Yaiprasert, C., & Hidayanto, A. N. (2023). AI-driven ensemble three machine learning to enhance digital marketing strategies in the food delivery business. *Intelligent Systems with Applications*, *18*, 200235. https://doi.org/10.1016/j.iswa.2023.200235

Yujiao, Z., Weay, A. L., Shaomin, S., & Palaniappan, S. (2023). Dropout Prediction Model for College Students in MOOCs Based on Weighted Multi-feature and SVM. *Journal of Informatics and Web Engineering*, *2*(2), Article 2. https://doi.org/10.33093/jiwe.2023.2.2.3

Zhang, Y., Liu, J., & Shen, W. (2022). A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences*, *12*(17), 8654. https://doi.org/10.3390/app12178654

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, *57*(4), 99. https://doi.org/10.1007/s10462-024-10721-6

Zimmerman, B. J. (2001). Theories of Self-Regulated Learning and Academic Achievement: An Overview and Analysis. In *Self-Regulated Learning and Academic Achievement* (2nd ed.). Routledge.

APPENDICES

Appendix I: Learner Attrition in Online Learning Questionnaire

Introduction and Purpose

My name is **Stanley Munga Ngigi**, a PhD student at a Kenyan university. I am currently conducting research for my doctoral study titled: ***"An Ensemble Model Using Gradient Boosting and Neural Networks for Predicting Learner Attrition in Online Learning."***

The purpose of this questionnaire is to collect information that will help identify the key factors influencing learner persistence and dropout in online learning environments. Your responses will contribute to the development of a predictive model aimed at improving learner retention and overall effectiveness of online learning.

I sincerely appreciate the time you have set aside from your busy schedule to assist me in this study. The information you provide will be treated with utmost confidentiality and will be used solely for academic and research purposes.

Should you feel uncomfortable answering any question, you are free to skip it or withdraw your participation at any point without any penalty.

If you agree to participate in this study, kindly indicate your consent by signing in the space provided below.

If you have any questions or need clarification before or during the process, please feel free to ask I will be happy to assist.

Consent to Participate

By signing this consent form, I acknowledge that I have been informed about the purpose of this study and voluntarily agree to participate.

Yes[ ] No[ ]

Participant's Signature: _____ Date: _____

Researcher's Statement

I, the undersigned, have fully explained the details of this research study to the participant named above and believe that they have understood and willingly agreed to take part.

Researcher's Name: Stanley Ngigi Munga

Signature: _____ Date: _____

Contact: +254 794 737 566 Email: mungastan@gmail.com

Section I: General Information

1. Institution: _____
2. Country: _____

Section II: Demographic Information

3. Gender: Male [ ] Female [ ]
4. Age [ ] :
5. Geographic Location:
    - Urban [ ]
    - Rural [ ]

6. Socioeconomic Status:
    - Household Income in Ksh:[ ]
    - Access to Technology: Yes [ ] No [ ]
7. Employment Status:
    - Full-time [ ]
    - Part-time [ ]
    - Unemployed [ ]
    - Self-employed [ ]

Section III: Academic Performance

8. Previous Academic History:

- KCSE Grade scored in

    i. Mathematics [ ]
    ii. English [ ]
    iii. Kiswahili [ ]
    iv. Chemistry [ ]

9. Participation in Assessments:

- Number of Assessment to be submitted:_____
- Number of Assessments Submitted: _____

10. Learning Style Adaptability:

    To what extent do you agree with the following statement

i. I like Online Learning:

Strongly disagree [ ]

Disagree [ ]

Neutral [ ]

Agree [ ]

Strongly agree [ ]

Section IV: Engagement Metrics

11. Course Access Frequency:

- Number of Logins per Week: _____

12. Interaction with Course Materials:

- Average Time Spent on Videos/Readings/Assignments per Week (Hours): _____

13. Participation in Discussion Forums:

- Number of Posts Made: _____
- Number of Replies to Peers: _____

14. Communication with Instructors and Peers:

- Average Messages/Emails Sent per Week: _____

Section V: Psychosocial Factors and Motivation

15. To what extent do you agree with the following statement

i. I have Self-Efficacy:

Strongly disagree [ ]

Disagree [ ]

Neutral [ ]

Agree [ ]

Strongly agree [ ]

ii. I am goal Orientation:

Strongly disagree [ ]

Disagree [ ]

Neutral [ ]

Agree [ ]

Strongly agree [ ]

iii. I highly perceived the Value of the Program:

Strongly disagree [ ]

Disagree [ ]

Neutral [ ]

Agree [ ]

Strongly agree [ ]

iv. I have a strong Support Network:

Strongly disagree [ ]

Disagree [ ]

Neutral [ ]

Agree [ ]

Strongly agree [ ]

Section VI: Learner Attrition

16. Have you ever considered dropping out of an online learning program? Yes [ ] No [ ] , If Yes, what were the primary reasons? (Check all that apply)

- Financial Constraints [ ]
- Lack of Motivation [ ]
- Technical Challenges [ ]
- Poor Course Engagement [ ]
- Other (Specify): _____

17. Have you completed an online learning program before? Yes [ ] No [ ]

Appendix II: Sample External Validation Questionnaire for Ensemble Model

Purpose: This questionnaire is designed to assess the predictive performance of the ensemble model in identifying learners at risk of attrition. The questions focus on learner engagement, academic performance, and self-reported risk factors that directly relate to the model's predictions.

Instructions: Please answer all questions honestly. Your responses will be used only for research purposes and will remain confidential.

Section A: Demographic Information

1. Age: _____

2. Gender: ☐ Male ☐ Female ☐ Prefer not to say

3. Program of study: _____

4. Year of study: _____

5. Study mode: ☐ Full-time ☐ Part-time

Section B: Engagement and Participation

6. On average, how many times per week do you access the online learning platform? ☐ Never ☐ 1–2 times ☐ 3–5 times ☐ Daily

7. How frequently do you participate in online discussions or forums? ☐ Never ☐ Occasionally ☐ Frequently ☐ Very Frequently

8. How often do you submit assignments or complete course activities on time?

   ☐ Never ☐ Rarely ☐ Sometimes ☐ Always

9. Have you missed more than two consecutive weeks of course participation?

   ☐ Yes ☐ No

Section C: Academic Performance and Progress

10. Have you failed or scored below passing thresholds in any course assignments this term?

    ☐ Yes ☐ No

11. Do you feel you are on track to complete this course successfully? ☐ Yes ☐ No

12. Have you considered dropping out of the course at any point? ☐ Yes ☐ No

13. If yes, what was the main reason for considering dropping out?

    _____

Section D: Self-Perceived Risk

14. How would you rate your own risk of not completing this course? ☐ Very Low ☐ Low ☐ Moderate ☐ High ☐ Very High

15. How confident are you in your ability to persist and complete this course successfully? ☐ Not confident ☐ Somewhat confident ☐ Confident ☐ Very confident

Section E: Additional Feedback (Optional)

16. Please share any factors that have affected your engagement or performance in this online course.

_____

Notes for Use in External Validation: - Responses from Sections B, C, and D can be compared with the model's predictions to evaluate its accuracy in identifying at-risk learners. - Expert reviewers, such as educational technologists or faculty members, can further assess whether the model's predictions align with observed patterns in learner engagement and dropout risk.

Appendix III: Focus Group Discussion Guide for Key Informants

The focus group discussion involved key informants drawn from universities in Kenya that offer online learning programmes, including online learning coordinators, lecturers involved in delivering online courses, and academic administrators responsible for managing digital learning systems. These participants were purposively selected because of their direct involvement in the design, delivery, and administration of online learning and their familiarity with issues affecting learner engagement, persistence, and attrition.

The questions in this section seek to understand the factors influencing learner attrition and retention in online learning environments. The discussion aims to gather insights from key informants on how institutions, facilitators, and learners experience, perceive, and respond to issues related to learner engagement, motivation, and persistence in online learning.

Your contributions will help enrich understanding of the human, institutional, and technological dimensions affecting learner outcomes and will inform the development of an ensemble-based predictive model for attrition management.

Discussion Questions

1. How would you describe the overall experience of learners participating in online learning within your institution?
2. What are the most common challenges that online learners face, which might lead to disengagement or dropping out?

3. In your view, what categories of learners (for example, based on age, employment status, or digital access) are more likely to experience attrition in online learning?
4. How do institutional factors such as course design, instructor feedback, or learner support systems influence student retention or dropout?
5. What role does technology infrastructure (foe example, internet access, devices, or learning platforms) play in supporting or discouraging consistent participation in online courses?
6. Describe how learners' motivation and self-discipline affect their engagement and completion rates in online programs.
7. How do social and family support systems contribute to or hinder persistence in online learning?
8. What mechanisms or interventions has your institution implemented to reduce dropout rates or to re-engage inactive learners?
9. How effective have these interventions been in improving learner retention? What lessons can be drawn from their implementation?
10. How do communication and interaction between learners, instructors, and peers affect learner satisfaction and completion?
11. From your perspective, what early signs or indicators suggest that a learner may be at risk of dropping out?
12. In your opinion, what measures can be taken (institutionally or technologically) to predict and prevent learner attrition in online environments?

13. How do learners perceive the value of online learning compared to traditional, face-to-face learning? Does this perception influence their persistence?
14. What recommendations would you make to enhance learner engagement and retention in online learning programs?

Appendix IV: Letter of Approval

Kirinyaga University

Tel: +254 701562092, +254 728499650, +254 709742000/30
P.O. Box 143-10300 Kerugoya.

Email: info@kyu.ac.ke
Website: www.kyu.ac.ke

SCHOOL OF PURE AND APPLIED SCIENCES
DEPARTMENT OF PURE AND APPLIED SCIENCES

The Chief Executive Officer

National Commission for Science

Technology &Innovation (NACOSTI)

NAIROBI.

7th March, 2025

Dear Sir/Madam

REF: RESEARCH PERMIT FOR STANLEY MUNGA REG. NO. PA303/S/21189/23

The above subject refers.

The above named is a student of Kirinyaga University pursuing a degree course in PhD in Information Technology in the Department of Pure and Applied Sciences. He has successfully defended his research proposal titled *"Ensemble Model using Gradient Boosting and Neural Networks for Predicting Learner Attention in Online Learning"* in partial fulfillment for the degree award.

We are kindly requesting your office to provide him with the necessary assistance for data collection and completion of his research.

Any assistance offered to the student will be highly appreciated.

Please to do not hesitate to contact the undersigned for any information regarding the same.

Thank you.

KIRINYAGA UNIVERSITY
C.O.D. DEPARTMENT OF PURE
& APPLIED SCIENCES

DR. JOSPHAT KARANI
COD, PURE AND APPLIED SCIENCES

Appendix V: Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: 322215

Date of Issue: 10/May/2025

RESEARCH LICENSE



This is to Certify that Mr.. Stanley Munga Ngigi of Kirinyaga University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Kiambu on the topic: ENSEMBLE MODEL USING GRADIENT BOOSTING AND NEURAL NETWORKS FOR PREDICTING LEARNER ATTRITION IN ONLINE LEARNING. for the period ending : 10/May/2026.

License No: **NACOSTI/P/25/4173176**

322215

Applicant Identification Number

Deputy Director
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 VI: List of Publications from This Study

1. A Systematic Review of Predictive Factors for Learner Attrition in Online Learning: Insights for Machine Learning Models
    - Journal: International Journal of Computer and Information Technology (IJCIT)
    - DOI: 10.24203/s2anxm35
    - Contribution: Provides theoretical foundation and feature identification for the predictive model
2. An Ensemble Predictive Model for Learner Attrition in Online Learning
    - Journal: International Journal of Computer and Information Technology (IJCIT)
    - DOI: 10.24203/2g7w8522
    - Contribution: Presents the developed ensemble model used in this study

Appendix VII: Sample Source Code for Student Performance Prediction Model

Import Libraries

```
import pandas as pd
import numpy as np
import plotly.express as px
import plotly.graph_objects as go
import seaborn as sns
import matplotlib.pyplot as plt
from plotly.subplots import make_subplots
from sklearn.preprocessing import LabelEncoder, StandardScaler



# Display all columns
pd.set_option('display.max_columns', None)



# Load Dataset
df = pd.read_csv("step 1 learner_attrition_data.csv")

df.head()
df['Attrition_Status'].value_counts()
Attrition_Status
Dropout       10059
Completion     9995
Continued      9946
Name: count, dtype: int64
```

Understanding the Dataset

Basic information about the dataset

```
print(df.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Record_ID               30000 non-null  int64
 1   Date                    30000 non-null  object
 2   Age_Group               30000 non-null  object
 3   Gender                  30000 non-null  object
 4   Region                  30000 non-null  object
 5   Household_Income        30000 non-null  object
 6   Access_to_Technology    30000 non-null  object
 7   Employment_Status       30000 non-null  object
 8   KCSE_Grade              30000 non-null  object
```

```
 9   Assessments_Submitted   30000 non-null  int64
 10  Assessment_Score        30000 non-null  int64
 11  Learning_Adaptability   30000 non-null  object
 12  Logins_Per_Week         30000 non-null  int64
 13  Time_Spent_Weekly       30000 non-null  int64
 14  Forum_Participation     30000 non-null  int64
 15  Emails_Sent_Weekly      30000 non-null  int64
 16  Support_Services_Usage  30000 non-null  object
 17  Psychosocial_Support    30000 non-null  object
 18  Attrition_Status        30000 non-null  object
dtypes: int64(7), object(12)
memory usage: 4.3+ MB
None
```

Insights

- No Missing Values: All columns have 30,000 non-null entries.
- Data Types: There are 7 columns with int64 data type (numeric) and 12 columns with object data type (categorical).

Summary statistics

```
df.describe()
```

Data Visualizations

```
# visualization of age distribution
fig = px.histogram(df, x='Age_Group',
                   color='Attrition_Status',
                   marginal='box',
                   title='Age Distribution by Attrition Status'
                   )
fig.show()
```

Insights

1. The 18-24 age group has the highest number of learners, but also shows a high proportion of dropouts. Possibly due to competing priorities (work or social life) or lower adaptability to online learning.
2. Dropout rates appear lower in older age groups (35-44 and 45+) possibly due to greater commitment or career-related motivation.

- Note

1. Targeted intervention (for example, mentorship programs, engagement incentives) should be implemented for younger learners (18-24) to reduce attrition rates.

```
# Gender distribution in the dataset
fig = px.pie(df, names='Gender',
             title='Gender Distribution',
             hole=0.5
            )
fig.show()
```

Note

- Generally they are evenly distributed.

```
# Percentage Stacked Bar Chart on Gender
gender_attrition = pd.crosstab(df['Gender'],
                               df['Attrition_Status'],
                               normalize='index') * 100

fig = px.bar(gender_attrition,
             title='Attrition  Status  Distribution  by  Gender
(%)',
             barmode='stack'
            )
fig.update_traces(texttemplate='%{y:.1f}%',
textposition='inside')
fig.show()
```

Insight

- This suggests that gender does not play a significant role in learner attrition.

```
# 3. Interactive Sunburst Chart
fig = px.sunburst(df,
                  path=['Employment_Status',
'Attrition_Status'],
                  title='Hierarchical View of Employment Status
and Attrition',

                  )
fig.show()
```

Insight

1. Full-time employed learners have a higher proportion of continued enrollment and completion which suggests that they may have more structured time management or career-driven motivation to complete courses.

- Note
- Offer financial aid or scholarships to unemployed learners to reduce dropout risk.
- Encourage career-driven learning pathways for full-time workers to sustain high retention rates.

```
data = df.copy()
```

Engagement patterns

```
# Distribution of Logins Per Week by Attrition Status
fig = px.histogram(data,
                   x='Logins_Per_Week',
                   color='Attrition_Status',
                   title='Distribution  of  Weekly  Logins  by
Attrition Status',
                   barmode='group',
                   labels={'Logins_Per_Week': 'Number of Logins
per Week',
                           'count': 'Number of Students'})
# Print summary statistics
print("\nSummary of Weekly Logins by Attrition Status:")
print(data.groupby('Attrition_Status')['Logins_Per_Week'].desc
ribe())
fig.show()

Summary of Weekly Logins by Attrition Status:
                    count       mean        std  min  25%  50%
75%   max
Attrition_Status
Completion          9995.0  5.004502  2.206082  0.0  3.0  5.0
6.0  16.0
Continued           9946.0  5.014981  2.236602  0.0  3.0  5.0
6.0  19.0
Dropout            10059.0  4.981807  2.243695  0.0  3.0  5.0
6.0  16.0
```

Insights

1. Dropout learners exhibit a slightly lower median and mean number of logins, indicating that lower engagement levels correlate with higher attrition risk.
2. Higher login frequency (above 10 logins per week) is less common, but those who do log in frequently tend to continue or complete the course.

Note:

- Encourage learners with lower login frequencies to engage more frequently, possibly through reminders or gamified participation incentives.
- Monitor learners who consistently log in fewer than 3 times per week and provide early intervention strategies (for example, nudges, personalized support).

```
# Forum Participation Analysis
fig = px.box(df,
             x='Attrition_Status',
             y='Forum_Participation',
             color='Attrition_Status',
             title='Forum Participation by Attrition Status',
             points='all')
print("\nForum Participation Statistics by Attrition Status:")
print(df.groupby('Attrition_Status')['Forum_Participation'].de
scribe())
fig.show()
```

```
Forum Participation Statistics by Attrition Status:
                    count       mean        std  min  25%  50%
75%   max
Attrition_Status
Completion          9995.0  3.000100  1.736925  0.0  2.0  3.0
4.0  11.0
Continued           9946.0  2.989644  1.736369  0.0  2.0  3.0
4.0  12.0
Dropout            10059.0  2.998211  1.731992  0.0  2.0  3.0
4.0  12.0
```

Insights

The results incates that forum participation alone does not appear to be a strong predictor of student attrition, as the patterns are nearly identical across all groups. This suggests that we should look at other factors or perhaps the quality rather than quantity of forum participation to understand attrition risk better.

Note

- Use low forum participation as an early warning sign for potential dropouts and provide personalized interventions (for example, nudges, discussion prompts).
- Consider mandatory discussion posts or structured peer interactions to promote deeper engagement and retention.

```
# Combined Engagement Metrics View
fig = make_subplots(rows=2, cols=2,
                    subplot_titles=('Login Distribution',
                                    'Time Spent Distribution',
                                    'Forum Participation',
                                    'Emails per Week'))

# Login Distribution
fig.add_trace(
    go.Box(x=df['Attrition_Status'],
           y=df['Logins_Per_Week'],
           name='Logins'),
    row=1, col=1
)

# Time Spent Distribution
fig.add_trace(
    go.Box(x=df['Attrition_Status'],
           y=df['Time_Spent_Weekly'],
           name='Time Spent'),
    row=1, col=2
)

# Forum Participation
fig.add_trace(
    go.Box(x=df['Attrition_Status'],
           y=df['Forum_Participation'],
           name='Forum'),
    row=2, col=1
)

# Emails Distribution
fig.add_trace(
    go.Box(x=df['Attrition_Status'],
           y=df['Emails_Sent_Weekly'],
           name='Emails'),
    row=2, col=2
)

fig.update_layout(height=800,
                  showlegend=False,
```

```
                title_text="Comprehensive          Engagement
Analysis")
fig.show()
```

Insights

1. Logins & Time Spent does not have much difference across groups, meaning logging in alone does not predict dropout.
2. In Forum Participation dropouts engage less—active, discussion seems to help retention.
3. Emails per Week: No major impact.

Note

- Monitor learners with consistently low logins and forum participation, as these metrics show a slight correlation with attrition.
- Encourage active forum engagement through discussion-based activities or community-building efforts to improve retention.

```
# Analysis of Assessment Scores

fig = px.box(df,
             x='Attrition_Status',
             y='Assessment_Score',
             color='Attrition_Status',
             title='Assessment Scores Distribution by Attrition
Status',
             points='all')  # Show all points

fig.update_layout(
    xaxis_title="Attrition Status",
    yaxis_title="Assessment Score",
    showlegend=True
)
# Print statistical summary
print("\nAssessment Score Statistics by Attrition Status:")
print(df.groupby('Attrition_Status')['Assessment_Score'].descr
ibe().round(2))
fig.show()

Assessment Score Statistics by Attrition Status:
                   count   mean    std   min   25%   50%   75%
max
Attrition_Status
Completion         9995.0  69.43  17.38  40.0  54.0  69.0  85.0
99.0
```

```
Continued          9946.0  69.39  17.27  40.0  54.0  69.0  85.0
99.0
Dropout           10059.0  69.51  17.29  40.0  54.0  70.0  84.0
99.0
```

Note

- Scores are similar across all groups, meaning assessment performance alone is not a strong predictor of dropout.

```
# KCSE Grade Analysis
fig = px.histogram(df,
                   x='KCSE_Grade',
                   color='Attrition_Status',
                   title='KCSE Grade Distribution by Attrition
Status',
                   barmode='group')
fig.show()
```

Insights

1. Attrition is evenly distributed across all KCSE grades, meaning prior academic performance does not strongly determine dropout likelihood.
2. Learners with both high and low KCSE grades drop out at similar rates, suggesting engagement and external factors play a bigger role than past performance.

Note

- Offer tailored support programs for both high and low performers to maintain retention.
- Encourage active learning strategies to keep students engaged regardless of their academic background.

```
# 4. Learning Adaptability Impact
fig = make_subplots(rows=1, cols=2,
                    subplot_titles=('Learning     Adaptability
Distribution',
                                    'Assessment     Score     by
Adaptability'))

# Distribution
adaptability_dist = df['Learning_Adaptability'].value_counts()
fig.add_trace(
    go.Bar(x=adaptability_dist.index,
           y=adaptability_dist.values,
           name='Distribution'),
```

```
    row=1, col=1
)

# Assessment Score by Adaptability
fig.add_trace(
    go.Box(x=df['Learning_Adaptability'],
           y=df['Assessment_Score'],
           name='Scores'),
    row=1, col=2
)

fig.update_layout(height=500,
                  title_text="Learning Adaptability Analysis")
fig.show()
```

Note

- Students with "Low" adaptability can achieve scores as high as those with "High" adaptability

```
# Comprehensive Academic Performance View
fig = make_subplots(rows=2, cols=2,
                    subplot_titles=('Assessment          Score
Distribution',
                                    'KCSE  Grade  vs  Assessment
Score',
                                    'Assessments       Submitted
Distribution',
                                    'Learning       Adaptability
Impact'))

# Assessment Score Distribution
fig.add_trace(
    go.Histogram(x=df['Assessment_Score'],
                 marker_color=df['Attrition_Status'].map({
                     'Dropout': 'red',
                     'Continued': 'blue',
                     'Completion': 'green'
                 }),
                 name='Assessment Distribution'),
    row=1, col=1
)

# KCSE Grade vs Assessment Score
fig.add_trace(
    go.Box(x=df['KCSE_Grade'],
           y=df['Assessment_Score'],
```

```
        name='Grade Impact'),
row=1, col=2
```