PREDICTING STUDENT ACADEMIC OUTCOMES USING MACHINE LEARNING

BY

IZUNWANNE NELSON CHISOM
GOU/U22/CSC/784

DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS, FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY, GODFREY OKOYE UNIVERSITY, ENUGU STATE.

IN PARTIAL FULFILLMENT OF THE AWARD OF BACHELOR OF SCIENCE (B.Sc), DEGREE IN COMPUTER SCIENCE.

SUPERVISOR: DR. IFEANYI JAMES ODEGWO

JUNE, 2026.

APPROVAL

This research, titled "**PREDICTING STUDENT ACADEMIC OUTCOMES USING MACHINE LEARNING**," has been assessed and approved by the Department of Computer Science and Mathematics Committees in the Faculty of Computing and Information Technology, Godfrey Okoye University, Enugu, Nigeria.

Dr. Ifeanyi James Odegwo Supervisor	**.....** Signature	**.....** Date
External Examiner External Examiner	**.....** Signature	**.....** Date
Dr. Frank Okebanama HOD, Department of Computer Science and Mathematics	**.....** Signature	**.....** Date
Prof. F. I. Ajah Dean, Faculty of Computing And Information Technology.	**.....** Signature	**.....** Date

CERTIFICATION

I certify that I completed the research included therein and that it was primarily based on my observations and investigation. Consequently, I will fully accept the University's decision if I am found to have cheated on the assignment.

_____

IZUNWANNE NELSON CHISOM

GOU/U22/CSC/784

DEDICATION

This work is dedicated to the Almighty God, the source of all wisdom and knowledge for His love, grace and guidance throughout the duration of this study.

ACKNOWLEDGMENTS

I first of all credit and praise the Most- High God for His infinite mercy, wisdom, strength and guidance in my academic life and especially during the process of this research. None of this would have been possible if it were not for His grace.

I wish to thank my Project supervisor Dr. Ifeanyi James Odegwo for his judicious guidance and constructive criticisms, patience and continuous support throughout the period of this research. He guided and gave very useful advice in directing the work and in its quality. I also thank the Head of Department, Dr. Frank Okebanama and all the lecturers of the Department of Computer Science and Mathematics, Godfrey Okoye University for the knowledge they have provided me with during my time here. It has been an inspiration to them to be so committed to teaching and to being excellent at it.

I want to express my gratitude to my dear parents, Mr. Bartholomew Eluwa and Mrs. Helen Eluwa for their unconditional love, financial support, encouragement and prayers throughout my academic journey. They have been my supporters and have believed in me, which has created my success. Also, I would like to thank my brother for his help and favor and my friends who have been by my side in the difficulties of this academic period. I have never been able to articulate how much your encouragement, motivation and moral support have meant to me. To all those who have helped in any aspect to make this project a success, God bless you and may God reward you.

ABSTRACT

A dropout prediction system is a data-driven tool that uses machine learning algorithms to analyze student data and classify whether a student is likely to complete their academic programme or leave before graduation. In the context of Nigerian higher education, where manual monitoring lacks predictive capability, this study developed a predictive analytics system for final-year Computer Science students at Godfrey Okoye University. The study employed a CRISP-DM approach to examine 3,630 student records with 37 features covering academic, demographic and lifestyle details. The Decision Tree Classifier performed best in terms of accuracy (94.21%), recall (100.00%) and F1-Score (92.95%) among the five classification algorithms tested. The trained model was deployed as a batch and interactive Streamlit web application, which provides visual analytics and compares model performance. Finally, this system gives administrators an efficient and data-based resource to proactively identify students at risk and take effective retention measures.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The world higher education is experiencing remarkable changes as it adopts digital approaches and evidence-based methods to boost academic success. Nigerian universities are expected to achieve the twin mission of not only retaining students to completion but also providing quality academic instruction in all courses. Dropout in Nigerian higher education is a pervasive problem, and to add to this problem institutional response to dropout is often ad hoc and affected by lack of resources. The massive dropout does lead to problems of productivity, wastage of financial resources and impacts negatively the fulfillment of developmental aspirations of Nigerian higher education (World Bank, 2021). This is especially alarming for Computer Science courses which are sequential and a failure in one stage may lead to an accumulation into an abyss of failure.

Student dropout and poor performance are not merely waste of opportunity to us, but also a colossal waste of the opportunity to invest our educational resources. Students who drop out or do not complete their programmes deny the universities tuition fees, government subsidies are not well spent and the economy at large loses potential skilled graduates. Dropout represents a loss of money, career and psychological trauma for the students and their families. A country sector analysis carried out by the World Bank (2021) found that Nigeria suffers from a variety of systemic inefficiencies, resource misallocation, and student dropouts at all levels of the education system, which results in significant economic losses and impedes human capital development in the country.

Traditional approaches to monitoring student progress rely primarily on after-the-fact measures such as reviewing end-of-semester results or responding to individual student complaints. By the

time these conventional methods reveal that a student is underperforming, the window for meaningful academic support has often narrowed considerably. The student may have taken several courses unable to finish or may have received carryovers and no longer be interested in attending. What is needed is a classification system that can reliably assign each student a likelihood of completing or not completing their degree based on observable data.

In fact, over the years, the Nigerian university sector has built up a large database of student academic performance data such as admission data, semester results, GPA, course enrolment, socioeconomic parameters and other student information. When analyzed well, this wealth of historical data can show patterns that can help to differentiate between successful graduates and dropouts. Institutions can use this information to take advantage of predictive analytics and identify students with the characteristics of a dropout and offer targeted support.

Predictive analytics appears to be one way to resolve the challenge: through the application of computational models to large quantities of past student data, universities can produce evidence-based forecasts of individual student futures. In higher education, this approach can incorporate various student data inputs to predict students' probabilities for different academic events, such as whether or not a student will finish a program. Academic studies of prediction in higher education have shown that properly calibrated classification models, when provided with sufficiently complete student data, can predict events with over 85% accuracy (Romero & Ventura, 2020).

Two closely related fields have emerged over the past ten years, both applying computational models to the massive quantities of student data generated by universities. EDM and LA have both seen increases in their popularity with universities across America, Europe, and Asia, where adoption has shown results like increased program completion (Kurzweil & Wu, 2015), and Georgia State University is a famous example. The system they implemented monitored students

daily for warning indicators and has had a significant effect on both increasing graduation rates overall and decreasing achievement gaps between student groups. Systems such as these, however, have not been adopted at Nigerian universities. While recent studies like the Osemwegie & Amadin (2023) that used six different machine learning algorithms to predict whether Computer Science students will drop out from the University of Benin have shown that it is possible to predict outcomes, there is little system adoption and this study seeks to fill that void. Several families of classification algorithms have demonstrated strong performance on educational prediction tasks, among them tree-based methods, ensemble approaches, linear classifiers, and margin-based models. These computational methods can uncover multivariable relationships and nonlinear patterns that would be impractical to identify through manual inspection alone. In addition, they are constantly adaptable since the data is constantly changing, and as such, they are adaptable to the dynamic educational environment.

The population under study is the most suitable population to study this kind of analysis and the population to which an accurate outcome prediction needs to be made is the population of Computer Science students in their final year. By that point, students have a history of their academic performance that allows for reasonable inferences, and making predictions about their success or failure to graduate or drop out at this point can help guide important decisions about interventions. Courses in Computer Science typically feature heavy workloads, dynamic course content and tasks that emphasize practical skills and need sustained attention.

Building on this premise, the present study develops a classification system capable of determining whether individual final-year Computer Science students are likely to complete their degree or leave the programme without graduating. The resulting tool is delivered as a browser-accessible application built on the Streamlit framework, providing academic staff with an intuitive interface

through which they can input student details and receive outcome predictions. The implementation of such a system will provide universities with a data-driven tool that can help to guide their decisions on student support, enabling a more targeted distribution of institutional support to strengthen completion rates and academic achievement.

1.2 Statement of the Problem

The problem of student drop-outs and poor academic performances is faced in Nigerian Universities particularly in challenging courses like Computer Science at Godfrey Okoye University. Notwithstanding the resources invested in teaching infrastructure and learning materials, a considerable share of enrolled students continue to underperform academically, or to make up the necessary gaps to secure a good level of attainment at the end of their programme, or drop out of the school altogether. This is not only a personal issue, but also an institutional issue as regards reputation of the institution, accreditation and human capital development for the country. Most Nigerian universities are using the conventional method of manual review in monitoring students' progress. Academic advisers and course coordinators generally judge students' performance once the final marks have been made public at the end of each semester, with no structured mechanism for distinguishing between students on track to graduate and those whose data profiles suggest elevated dropout probability. The lack of predictive capability leaves the university with a deficit of data-informed knowledge of what factors are associated with student dropouts.

Furthermore, the current means of student monitoring are extensively manual, and the judgement is subjective. It is possible that certain students may be missing or not doing well in continuous assessment, but there is no structured way of integrating the observations from various courses, and there is no means of identifying those students who are at high risk because of their combination of

warning factors. This disjointed strategy implies that a great number of at-risk students end up passing through the cracks and are not assisted until the problem has become acute. For years, Nigerian universities have collected students' academic information in their database and records, however, this information is not exploited for predictions. These records are stored in most institutions for administrative reasons, including the production of transcripts and clearing a student for graduation, without making use of the predictive power of these past records. This appears to be a resource which has not yet been utilized, as cumulative academic performance has been proved consistently with course completion. Also, there is a lot of unavailability of locally-designed and tested predictive models appropriate for the Nigerian Higher education context. Various studies have built and tested such models in the western education institutions, but there is much doubt about the applicability of such models in the Nigerian context because of the different educational culture, infrastructure, and student demographics. It is therefore essential to develop predictive systems that are based on local data and are tested locally. This research will address the problem of lack of a predictive system, which could identify a final year Computer Science student who is at risk of not completing their degree at the university due to his/her academic records and demographic characteristics. Without such a system, universities cannot effectively distribute the limited resources of their support staff to the students who need them the most, resulting in unnecessary academic failures and dropouts.

1.3 Aim and Objectives of the Study

The aim of this study is to develop a predictive analytics system to predict whether final-year Computer Science students will graduate or dropout based on their academic records and demographic characteristics. The study pursues the following specific objectives:

i. Extract and prepare academic records, and convert raw data to features suitable for machine learning analysis.

ii. Identify and analyze important factors that influence student dropout risk and academic outcomes.

iii. Develop and train 5 machine learning algorithms which can classify students into their expected outcome in the academic field.

iv. Evaluate the trained model using standard classification metrics which are accuracy, precision, recall, F1-score and deploy the strongest performer.

1.4 Significance of the Study

The creation of a predictive analytics system to predict students' outcomes holds significance for different parties both within and beyond the university context. This study could have the following meaning:

i. This research has direct impact for students because of its capacity to flag students whose data profiles indicate a higher probability of non-completion. Those who are flagged as being at risk may be referred to a tutoring scheme, counselling or financial support as appropriate.

ii. This research helps Academic Staff members and gives them a tool they can use to identify students that need attention, so that they can put their limited time and energy into the students that need it the most. Academic staff may have to take the initiative and contact at-risk students proactively to provide support rather than to wait for students to approach them with problems.

iii. The findings of this research will assist the administrators in boosting the performance of institutions and at the same time serve the educational mission of the university and

increase retention rates which will in turn affect institutional ranking, accreditation, and reputation.

iv. This study has relevance to national objectives that will be useful in the reduction of dropouts and improvement of academic performance. In addition, the study shows that predictive analytics can be implemented effectively in the Nigerian educational environment, and could serve as a motivation to other schools to follow suit.

v. This research adds to the EDM and LAM research that is gaining traction within the disciplines of Educational Data Mining and Learning Analytics, especially in Africa where this area of research is relatively under-represented. The methodology, findings and lessons learnt can serve as a reference to other research within Nigerian schools and similar environments. This research is also a platform for further research like use of adaptive learning system and tailor-made solutions for learning.

1.5 Scope of the Study

This study aims to build and design a predictive analytical system to predict student outcomes within a particular range that specifies the boundaries of this study. It's crucial to know these boundaries if you're going to interpret the results and make it relevant to other contexts.

i. Subject Scope: Final year Computer science students. This group was chosen because by the end of the senior year students have had enough years of schooling to make meaningful predictions about their outcomes upon graduation. The prediction task is a binary classification problem; whether a student is going to graduate or drop out by the end of the last year. Patterns found may or may not be transferable to other programming and other stages of the students' academic experiences.

ii. Data Scope: Academic records and demographic data adapted to the Nigerian university context – Godfrey Okoye University's grading system. All the following data is included: demographic information, for example: gender, marital status, out of state status. Family background such as parent's education, occupation. Financial aspects like; monthly allowance, status of school fees, school scholarship status. Academic entry qualifications such as: JAMB score, O level credits. Semester performance data like courses registered, courses passed and failed, GPA, assessment scores. Engagement data namely, attendance, character assessment, hours spent studying and Lifestyle data like hours spent playing games, clubbing frequency, substance use status. Overall academic standing is an important metric to consider: the cumulative GPA (CGPA).

iii. Technical Scope : The classification methods used in this work are supervised machine learning models which are Decision Tree, Random Forest and Logistic Regression. Work is performed in Python utilizing Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn and Pickle libraries for data handling, training of models, plotting visualization graphs and serializing the trained models. Top ranked model is then hosted using a Streamlit web interface to support single student-prediction, batch prediction, data visualization dashboards and model performance reports.

iv. Geographical Boundary: The technique may have potential for use elsewhere, however, the boundary of this research work is confined to the Nigerian environment of higher education industry, considering the features and the problems facing the universities in Nigeria. Care should, however be taken while applying conclusions to the generalization of the result to other education systems with other structures and other sets of students.

1.6 Limitations of the Study

While this contribution is valuable to the body of research in educational data mining in Nigeria, the following limitations are noted:

i. Dataset limitations: The dataset is not actually the records from the university, but adapted from an already available dataset for the Nigerian university system. This way data is consistently available but models developed based on such data would need to be tested on a dataset drawn from the actual institution before they are implemented fully.

ii. Generalizability Concerns: Models that are trained on information from one institution or programme might not perform well on the different one. Predictors for dropouts and/or graduation for Computer Science students at one university may not be the same as at another university with different admission standards, teaching, or demographics.

iii. Temporal Constraints: Since education is a dynamic concept, what is useful and appropriate today might not be as appropriate in the future due to changing curriculum, methods of instruction, or student groups. The models created as a part of this study may require future updates in order to keep the predictive model up-to-date.

iv. Ethical/Privacy issues: Using student records for algorithm predictions also raises ethical and privacy issues related to consent and data privacy as well as concerns about potential algorithmic bias. The ethical protocols have been respected for this study, but these should also be considered during the implementation of such systems to avoid potential adverse effects.

v. Implementation Gap: This study does not consider full-scale implementation and evaluation of intervention strategies, but rather focuses on developing and validating predictive models to be implemented as a web application. The ultimate effect of such a

system is dependent on the accuracy of the prediction, as well as the effectiveness with which predictions are converted to supportive action.

1.7 Operational Definition of Terms

i. Predictive Analytics: A data-driven approach that leverages computational models, pattern recognition, and historical records to generate forecasts about what is likely to happen next. The use of predictive analytics in this study will be determining if a student will graduate or drop out of his/her final year.

ii. Student Dropout: The departure of a student from their academic programme without earning the intended degree or certificate. Drop out in this study is the voluntary withdrawal of student from the Computer Science programme or the academic dismissal.

iii. At-Risk Student: A student whose combination of academic indicators and personal characteristics suggests an elevated probability of not completing their programme. There are predictive models, using academic records and demographic data, that flag at-risk students in this research.

iv. Machine Learning: A computational discipline in which algorithms iteratively refine their internal parameters by processing datasets, enabling them to recognize complex relationships and improve their predictive accuracy over time. Performance typically improves as additional training examples become available.

v. Feature: a quantifiable property or observable attribute extracted from raw data and supplied as an input column to a predictive model. This study includes the following features: GPA, attendance scores, JAMB scores, and other demographic and lifestyle factors.

vi. Cumulative Grade Point Average (CGPA): A numerical representation of overall academic performance of a student across all courses he/she has been studying, based on the weighted average of the grade points secured in the courses. The CGPA of students in Nigerian Universities is normally computed on a scale of 0.0 to 5.0.

vii. A category of supervised learning in which the model assigns each input record to one of a finite set of predefined class labels. In this study, the author uses binary classification and categorizes students into two different classes; Graduate or Dropout.

viii. Decision Tree: An algorithm that partitions the input space through a hierarchy of if-then rules, splitting records at each internal node based on the feature and threshold that yield the greatest class separation. Its output is readable as a flowchart, and it handles both numeric and categorical inputs natively.

ix. Random Forest: An ensemble strategy that trains many independent tree classifiers on bootstrapped data samples and aggregates their individual outputs through majority voting. By averaging over a diverse set of trees, this approach reduces variance and typically yields stronger generalization than any single tree.

x. Logistic Regression: a classifier that maps a weighted sum of input features through a sigmoid curve, producing a probability estimate between zero and one for a given class. Despite its name suggesting a regression method, it is principally employed for categorical outcome prediction.

xi. Gradient Boosting: A sequential ensemble technique in which each successive weak learner focuses on the residual errors left by its predecessors, progressively reducing the overall prediction error. Many prediction tasks have been solved with state-of-the-art

performance using algorithms like the gradient boosting algorithms XGBoost and LightGBM.

xii. Support Vector Machine (SVM): A classifier that searches for the decision surface maximizing the gap between the closest data points of each class. In high-dimensional spaces, SVMs perform well when class boundaries are distinct, and through kernel functions can also handle cases where the separation is nonlinear.

xiii. Streamlit: A Python-based toolkit that converts data scripts into shareable web interfaces with minimal front-end coding. Its design philosophy prioritizes rapid prototyping, enabling researchers and analysts to publish interactive dashboards and input forms without needing HTML, CSS, or JavaScript expertise.

xiv. CRISP-DM: A popular and commonly used data mining and analytics project methodology is CRISP-DM (Cross-Industry Standard Process for Data Mining). It has six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

xv. Educational Data Mining (EDM): a growing discipline that applies computational techniques to the rich data streams generated within schools and universities, with the goal of uncovering insights about learner behavior, instructional effectiveness, and institutional processes.

xvi. F1-Score: A single metric that balances the trade-off between how many of the predicted positives are correct and how many of the actual positives are found, computed as their harmonic average. The F1 score is especially important for assessing classifiers on imbalanced data sets when there's a large difference in the number of positive and negative examples.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The current chapter offers a critical review of the literature available on how well student academic outcomes can be predicted using machine learning techniques. The review is embedded in the context of the broader research areas of educational data mining and learning analytics, which aim to leverage institutional data to support evidence-based decision making in higher education in a rapidly developing manner. This chapter provides a theoretical underpinning, technical approaches and empirical evidence from previous research to provide a foundation for the current research and to pinpoint gaps in knowledge to be addressed in this study.

The review is broken down into a number of sections. The theoretical background section includes ideas, theories, and technical frameworks as the basis of this research, including educational data mining, learning analytics, and machine learning algorithms. This literature review critically examines the previous research studies that have tried to predict student performance or dropouts based on various methodological procedures. A tabular comparison of reviewed studies is provided in the summary section and a gap section describes the specific contributions that this study will make.

The chapter demonstrates through a critical and scholarly review of academic papers, research articles, conference papers and institutional reports that efforts in predictive accuracy have been made in predicting the outcomes of students using data-driven approaches, but this is not the case when applied to the local context of the Nigerian higher education institutions. This review will help to confirm that this study will add to the field and build upon previous studies.

2.2 Theoretical Background

This section describes the key theories, concepts and technical supports to be used in the research. Having an adequate knowledge about these theories are vital in order to explore how predictive analytics can be used in the Higher Education to predict student performances and at-risk students.

2.2.1 Theory of Student Retention and Departure

In order to develop the knowledge needed to explain the reasons for student dropout and/or student continuation, a theoretical framework that explains the complex interaction of the many variables that influence a student's choice to stay or drop out of school is needed. The two theoretical accounts that have historically been influential in student departure literature are the Tinto Model of Student Departure and the theory of Student Involvement (Bean and Metzner, 1985).

Initially formulated in 1975 and revised in subsequent years, Tinto's (1993) framework is considered the most commonly cited theoretical explanation for why students leave college (Tinto, 1993). The model states that each student enters university with individual characteristics and family background circumstances which will shape his or her initial goals, aspirations, commitment to college, etc. Once a student enters university, they either become academically integrated and/or socially integrated into the institution. Academic integration can be a more or less formal and personal interaction and is defined by engagement with the academic system. Social integration is informal and social in nature and comes from membership in college activities, organizations, and from friendships. Both the extent of the student's academic and social integration with the university institution will be reflected in their future engagement with the institution and in their success or failure at the institution (Tinto, 1993). Students with little academic and social integration will most likely voluntarily drop out of school or academically withdraw from the institution. Astin's (1984) theory of Student Involvement provides a

complementary approach which is an elaboration of Tinto's ideas that emphasizes what students do at university (Astin, 1984). Involvement, in Astin's work, is "the amount of physical and psychological energy that a student devotes to the academic and social life of the college" (Astin, 1984). It includes hours spent studying, as well as hours participating in events and activities, talking to professors and peers. Higher amounts of student involvement result in higher amounts of student learning and development; this is highly relevant to this study because amount of involvement is quantifiable and directly related to student success as academic outcome predictors (attendance, study time, and course involvement). As these two theories predict, student withdrawal is a process, and clues or signals that predict dropout can be observed along the way. This is precisely the goal of predictive analytics-that by looking at the chain of signals and cues which signal the extent of academic and social integration (or the lack thereof) schools will be able to predict student disengagement and dropout. The line of thinking developed in Tinto's model and Astin's theory is extended to non-traditional learners in Bean and Metzner's (1985) work, stating that external, non-academic variables such as money, work, and family issues are often more relevant to the decision to drop out or stay enrolled than social integration; this may be particularly true for Nigeria where students are influenced by an array of external socioeconomic factors.

2.2.2 Educational Data Mining

Educational Data Mining (EDM) is a hybrid discipline borrowing methods from the fields of statistical inference, pattern recognition and algorithmic modelling to provide sense for the enormous quantities of data that schools and universities collect (Romero & Ventura, 2020). The field was established in the early 2000s as schools and universities began gathering massive amounts of student data via student information systems, learning management systems and computer-based assessments. The over-arching goal of EDM is to develop analytical pipelines

that convert these unorthodox data streams into knowledge of student progression, where they may be failing and how institutional processes impact student outcomes. The main aims of EDM according to Baker and Inventado (2014) are to predict student performance and behavior; model student learning; examine the effect of pedagogical support; further the science of learning and aid educational and administrative decision-making. The work of this research is prediction, specifically the ability to develop models that can predict future behavior from student demographic and historical information. There are a number of different kinds of techniques commonly used in EDM. These include classification models, which assign individual students to discrete outcome groups; prediction models, which are a form of classification, predict numeric outcome values such as a student's expected semester GPA; clustering techniques which attempt to find naturally existing groups of students; Association Rule Mining which searches for patterns in student behavior; and Sequence Mining which focuses on ordering student behavior patterns. EDM is a discipline as it has become widely adopted with the formation of forums for research in it since 2008; the International Conference on Educational Data Mining, and since 2009; the Journal of Educational Data Mining.

2.2.3 Learning Analytics

Learning Analytics (LA) shares much of its methodological toolkit with EDM but differs in orientation: its primary concern is translating analytical outputs into actions that improve teaching practice and the student experience. The Society for Learning Analytics Research describes the field as the systematic gathering, processing, and interpretation of data about learners and the environments in which they study, to support understanding and improvement of learning and learning environments (Siemens & Gasevic, 2012). EDM is more about creating and testing analytical methods, whereas LA is more about using the results to benefit stakeholders. Shafiq *et*

al. (2022) performed a wide-ranging systematic literature review of the literature related to the use of educational data mining and predictive analytics to improve student retention. They surveyed 100 studies on student retention prediction covering the period from 2017 to 2021, and noted that machine learning and deep learning algorithms were widely adopted for this task. Ensemble approaches and gradient boosting algorithms were among the top-performing methods that were identified in the review, with a lack of use of unsupervised learning methods in identifying homogeneous student groups noted. A typical analytics workflow proceeds through several stages: assembling records from disparate data stores, pre-processing and integrating the data, analyzing the data using statistical or machine learning methods, visualization and interpretation of results and application of interventions based on the results. Importantly, for learning analytics to work it needs to be able to connect the dots between the analysis of the data and taking action that is beneficial for the student. This can occur as automated notifications to advisors, personalized suggestions to students, and strategic resourcing by administrators. Ferguson (2012) outlined a number of challenges within the field of learning analytics, some of which are still pertinent today. This includes technical challenges of data integration across multiple systems, ethical questions about student privacy and surveillance, data ownership and consent, the risk that analytics can further perpetuate inequities, and the need to ensure that analytics results in positive rather than punitive responses. These are taken into consideration and form the ethical standpoint used in the current research.

2.2.4 Machine Learning Algorithms for Prediction

Machine learning is a family of algorithms that enable computational systems to detect structure in data and progressively sharpen their predictions through exposure to examples rather than through hand-coded rules (Mitchell, 1997). When it comes to student performance prediction, the machine

learning algorithms use the previous information of students who have finished their various programs (students with known outcomes) to produce a model that can predict the performance of current students.

The majority of student outcome studies employ supervised classification, in which models are fitted to datasets where the correct label for each record is already known. These algorithms can be divided into classification for categorical outcomes such as graduate/dropout and regression for continuous outcomes such as final GPA. There are several common algorithms implemented in the field of educational prediction research:

Logistic regression converts a linear combination of input variables into a probability score via the sigmoid function, making it a natural fit for two-class prediction tasks. Logistic regression is simple, yet can be very competitive with more sophisticated algorithms if the relationships between predictors and outcomes are more or less linear. The primary benefit is that it is interpretable: each fitted coefficient shows how a one-unit change in the corresponding predictor shifts the log-odds of the outcome (Hosmer *et al*.,2013).

A Decision Tree structures its classification logic as a series of nested conditions: each internal node tests a feature against a threshold, each branch represents one possible outcome, and each terminal node assigns a class. Their visual, rule-based structure makes them among the most interpretable classifiers, and they naturally capture non-linear effects and feature interactions. They tend to overfit, particularly as trees become deep and complex. On the downside, an unrestricted tree may memorize training noise, a tendency that worsens as depth increases (Quinlan, 1986).

Random Forest brings together a large number of independently grown trees, each constructed on a random bootstrap sample, and determines the final label by majority vote across the ensemble. By pooling diverse trees trained on different subsets of rows and columns, the method suppresses

variance and typically outperforms any single constituent tree. An additional benefit is the built-in ranking of predictor relevance, which indicates which input variables contribute most to the ensemble's decisions (Breiman, 2001).

Here Gradient Boosting models like XGBoost, CatBoost and LightGBM are created one by one with every model trying to correct the mistakes of the previous model. These algorithms have been used to tackle a large variety of prediction problems, and perform particularly well when the data is structured into tables. Villar and de Andrade (2024) reported that using boosting algorithms, specifically LightGBM and CatBoost, performed better in predicting student dropout in comparison with the classical classification algorithms.

Standard benchmarks in educational prediction include accuracy which are the share of all cases classified correctly, precision which is the share of predicted dropouts who really did drop out, recall which is the share of actual dropouts the model detected, F1-score which balances precision and recall into a single figure, and Area under the ROC Curve (AUC-ROC). Practitioners also use cross-validation, most commonly k-fold partitioning, to gauge how stable the model's performance is when exposed to different subsets of the data.

2.2.5 Key Variables in Student Performance Prediction

Many factors have been found which differentiate successful students from those who struggle or drop out of school. These variables can be classified according to the nature and source of the variable. Demographic attributes like age, gender, household economic standing, and whether the student is the first in their family to attend university, form one cluster of predictors. These variables could be predictive, but using them involves moral issues of maintaining prejudices or facilitating discrimination. But many researchers believe that focus should be on the malleable factors instead of the immutable demographic factors (Kizilcec & Lee, 2020). Among the strongest

indicators of how a student will fare academically are measures of their previous scholastic record which include secondary school grades, standardized test scores e.g., JAMB in Nigeria and University grades of previous school. The predictive power of past performance is that education is a cumulative process: students who have struggled in the past are more likely to continue to fail unless there are interventions. Engagement and Behavioral Variables record student participation. In the case of Finnish higher education, Vaarma and Li (2024) identified accumulated credits, number of failed courses, and engagement indicators as the most influential features for predicting the dropout from the degree program. Their study showed that when using transcript and demographic data, their behavioral data contributes significantly to the amount of data that can be used to predict. Attendance and Participation have been shown to be predictors of school success for a long time. Regular attendance and active participation are good predictors of student achievement. Relevant predictors in blended learning situation can be physical attendance and online participation. Student mastery is immediately reflected in Assessment Performance for early and mid-course assignments, quiz grades, and midterm exams. Assessment performance becomes especially important for its prediction value as it offers a set of indicators of what a student has mastered which can be measured and can be correlated with final academic results.

2.3 Review of Related Works

This section appraises published investigations that have applied computational classification and pattern-discovery techniques to forecast student outcomes such as graduation and programme withdrawal. The reviewed studies cover a wide range of geographical backgrounds, educational contexts and methodologies, and focus on the most recent studies of the last three years from 2021 to 2024 to provide a current picture of the field.

In IEEE Access, Shafiq *et al*. (2022) undertook a systematic review of the literature on student retention using educational data mining and predictive analytics. The review analyzed 100 studies from 2017 to 2021 that looked at factors and methodological components in developing strong student retention predictive models. The authors proposed a taxonomy of machine learning approaches, and they explored success factors in traditional, blended, and online learning environments. The main findings indicated that supervised learning and deep learning techniques are most commonly used, and the favored algorithms included Random Forest, Decision Trees and Neural Networks. The review showed that ensemble approaches overall gave the best results, and it was observed that few of the studies considered cognitive and personal factors, most of which used the traditional and static features.

The researchers proposed a novel two-layer stacked generalization method for the prediction of student dropout in university programmes in order to solve the problem of student dropout in universities (Niyogisubizo *et al*. 2022). This study, published in Computers and Education: Artificial Intelligence, addressed the problem of predicting a learner's drop-out with a very small data set by constructing a stacking ensemble of Random Forest, XGBoost, Gradient Boosting and Feed-forward Neural Networks. Drawing on data from Constantine the Philosopher University in Slovakia collected over 5 academic years from 2016 to 2020, that their ensemble method was more accurate and attained the highest value of the area under the curve (AUC) compared to individual base models. This was due to the ensemble of these stacking algorithms, for which the improved results were attributed to harnessing the complementary strengths of various algorithms.

In Electronics, Tamada *et al*. (2022) investigated predicting at-risk students in technical classes using LMS logs. The researchers selected technical education programmes where the dropout rate is usually high, and analyzed the behavioral characteristics from Moodle log files like login

patterns, resource accesses, assignment submission behavior, and forum participation. They successfully predicted with machine learning classifiers with accuracy greater than 85% such as Decision Trees, Random Forest and Support Vector Machines. Their findings showed that the LMS engagement data is a valid predictor of drop-out in technical programmes and that there were specific patterns of behavior that could be found among at-risk students.

Flores *et al*. (2022) compared predictive models with balanced classes using the SMOTE method to predict student dropout in higher education. Published in the journal Electronics, this study dealt with the important topic of class imbalance, in which the number of dropouts is usually minimal compared to the number of students in the majority class, and models are prone to predicting the majority class. The authors tried several algorithms such as Decision Trees, Random Forest, Naive Bayes and two types of Neural Networks: one without balancing using SMOTE (Synthetic Minority Over-sampling Technique) and another one balanced with SMOTE. Results demonstrated that the SMOTE technique can enhance the sensitivity (recall) of the dropout class without significantly compromising the overall accuracy. The highest accuracy and recall were obtained by Random Forest + SMOTE algorithm.

Rodriguez-Hernandez *et al*. (2021) carefully applied and tested artificial neural networks to predict academic performance. It was published in Computers and Education: Artificial Intelligence and it contains a rigorous methodology for designing and evaluating neural network models for education prediction. The authors systematically evaluated different neural network architectures and sets of predictors. They showed that the neural networks can provide prediction accuracy close to or better than ensemble algorithms with proper design. Interestingly, the study revealed that the input feature selection significantly impacted the prediction accuracy, where the previous academic performance and engagement indicators were identified as the most powerful factors.

Yagci (2022) looked at the use of machine learning algorithms in predicting students' academic performance through educational data mining. This study was conducted on 1,854 university students with the Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), and K-Nearest Neighbors (KNN) algorithms. Published in the journal "Smart Learning Environments". The Random Forest model had the highest classification accuracy when predicting final exam grades using midterm performance and demographic data around 75%. To select features the study used Information Gain, and prior academic performance turned out to be the most significant predictor.

Realinho *et al*. (2021) have created a benchmark dataset that predicts student dropouts and academic success, which can now be found in the UCI Machine Learning Repository. The data originates from a Portuguese institution, with a sample of 4424 undergraduate students enrolled in different degree programs and covers 37 features related to the students' demographic, socioeconomic and academic performance in the first 2 semesters of study. Three-class prediction tasks (dropout, enrolled, graduate) is a class imbalance problem. This dataset is publicly available and is widely considered to be a benchmark for evaluating different prediction methods, with reported accuracy rates ranging from 75% to 92% in various algorithms and preprocessing techniques.

Vaarma and Li (2024) undertook an empirical investigation into the machine learning prediction of student dropout in Finnish higher education, appearing in Technology in Society. The three contributions of this research were: comparing LMS (Moodle) data to transcript data and demographic data, analyzing the importance of the predictors over the course of the month after enrolment, and measuring the predictive power of the predictors over time. The accumulated credits, not passing classes, and student activity in the online learning environment were the most

important predictive variables identified by the researchers. Importantly, the study found that engagement data is highly predictive and should be taken into account in conjunction with traditional academic data.

In the article "Comparative Study of Supervised Machine Learning Models for Prediction of Student Dropout and Academic Performance" published in Discover Artificial Intelligence, Villar and de Andrade (2024) compared supervised machine learning models for predicting student dropouts and academic performance. Their study methodically evaluated the Decision Tree, Support Vector Machine, Random Forest and boosting algorithms (Gradient Boosting, XGBoost, CatBoost, LightGBM) on the UCI student dropout dataset. The researchers used SMOTE for class balancing, Optuna for hyperparameter tuning, and Isolation Forest for outlier detection, and they were able to optimize boosting algorithm whose accuracy was more than 90%, compared to traditional classification techniques.

In Applied Sciences, Lee *et al*. (2023) have published a high precision and high recall Student Dropout Prediction system for universities. The researchers developed a composite model with several classifiers drawing on data from 67,060 students at the Gyeongsang National University in South Korea for the last six years (2016-2022) so as to improve the precision which is accuracy of predicting dropouts and recall which are proportion of dropouts that were identified of all the classifiers used. In addition, PCA was used to reduce the dimensions of the data and K-means clustering was used to determine specific causes for dropout. High precision and recall with interpretable clusters of reasons for dropping out.

In Scientific Reports, Matz *et al*. (2023) looked at how to use machine learning to forecast student retention based on socio-demographic factors and the degree of app engagement. This study combined the two types of institutional and behavioral data from a mobile application. All models

were run at a large university and compared against each other based on socio-demographic features, app engagement features, and a combination of both. Results of the analysis showed that the app-based engagement measures had a greater predictive power compared to the socio-demographic measures alone. The best accuracy was found with the hybrid model, which demonstrated the importance of combining multiple data sources.

Segura *et al*. (2022) analyzed how to predict university student dropout using machine learning focusing on programme preference alignment. This study, published in the journal Mathematics, was based on data from a large European university and the dropout rates at two stages were predicted: at the beginning of the studies, and after the first semester. The researchers looked at five broad programme areas and compared Support Vector Machines, Decision Trees, Artificial Neural Networks and Logistic Regression. Results showed that there was a significant difference in the accuracy of prediction between the programmes, with Arts and Humanities having the highest dropout rate (45.9%) and Science having the lowest (16.6%). These were the most important predictors, with programme preference alignment and first-semester grades.

In their paper published in the journal Expert Systems with Applications, Waheed *et al*. (2022) introduced an early intervention model for at-risk learners in self-paced learning through neural networks. This research directly tackled the problem of predicting the outcomes of self-paced online learning situations in which typical temporal cues such as "half way through the semester" do not apply. The authors have developed a neural network to classify sequential learning activity data to detect the risk of dropout. The model's AUC scores were >0.85 for early prediction (predicting less than 25% of the expected completion time), and these deep learning techniques could capture the complex temporal patterns in behavioral data effectively.

Oqaidi *et al*. (2022) built a student dropout prediction model for higher education institutions using machine learning algorithms, reported in the International Journal of Emerging Technologies in Learning. This study used a number of classification algorithms to foresee the dropout rate of students in a Moroccan university. Four algorithms were tested by the researchers, namely Logistic Regression, Decision Tree, Random Forest and Support Vector Machine, considering academic and socioeconomic features. Random Forest achieved the best accuracy of 87.5%. The study found that the highest predicting factor was the student's performance in the first year of school followed by socio-economic status. This research is especially relevant to the present investigation as it is an African context study.

The study by Adekitan & Salau (2019) was aimed at engineering students in Nigeria and focused on the influence of certain variables on the performance of students at Covenant University. They employed 1841 students in a multiple regression analysis and neural networks analysis, and concluded that university performance was strongly related to the academic performance of the students in their secondary school (GPA and JAMB scores). Students who were admitted on merit showed better performance than other students. By focusing on performance prediction (not dropout), this study demonstrated the viability and value of using educational data mining in the Nigerian university context and highlighted key factors for prediction that are relevant to this study.

Osemwegie and Amadin (2023) used six Machine Learning Algorithms which are Logistic Regression, SVM, Decision Tree, KNN, Naive Bayes and ANN to predict Computer Science student dropout at the University of Benin, Nigeria, published in the FUDMA Journal of Sciences. This study employed course grades as data to classify students based on their grades in the courses and represented one of the few Nigerian studies that directly focused on addressing dropout

prediction instead of performance prediction. Their study was based on students' academic grades only, and did not account for demographic, financial or lifestyle factors, nor feature a deployed web application for administrators to use in practice.

2.4 Summary of Literature Review

The table below summarizes the key studies reviewed, highlighting their methodologies, contributions, and limitations. This tabular format facilitates comparison across studies and helps identify patterns in the research literature.

S/N	Author(s) & Year	Contribution	Methodology	Limitations
1	Shafiq *et al*. (2022)	Systematic review of 100 studies; taxonomy of ML approaches	Systematic Literature Review	Review only; no empirical validation
2	Niyogisubizo *et al*. (2022)	Novel stacking ensemble combining RF, XGBoost, GB, FNN	Two-layer Stacked Generalization	Single institution; limited sample
3	Tamada *et al*. (2022)	Prediction using LMS logs; >85% accuracy in technical courses	DT, RF, SVM on Moodle logs	Technical courses only
4	Flores *et al*. (2022)	SMOTE improves dropout class recall; RF with SMOTE best	Multiple algorithms with SMOTE	Single institution; Spanish context
5	Rodriguez-Hernandez *et al*. (2021)	Systematic ANN implementation; engagement metrics key	Neural Networks systematic evaluation	Computationally intensive
6	Yagci (2022)	RF achieved 75% accuracy; prior performance most important	RF, SVM, LR, NB, KNN comparison	Moderate accuracy
7	Realinho *et al*. (2021)	UCI benchmark dataset; 4,424 students, 37 features	Dataset creation for ML benchmarking	Portuguese context; class imbalance
8	Vaarma & Li (2024)	LMS data significant predictor; temporal analysis	Three ML models on Moodle + transcript	Finnish context
9	Villar & de Andrade (2024)	LightGBM and CatBoost outperform others; >90% accuracy	DT, SVM, RF, GB, XGBoost, CatBoost	Single dataset
10	Lee *et al*. (2023)	High precision and recall; K-means identifies dropout reasons	Hybrid model with PCA and clustering	Korean context
11	Matz *et al*. (2023)	App engagement data improves prediction; combined model best	ML on socio-demographic + app data	Requires app infrastructure
12	Segura *et al*. (2022)	Dropout varies by programme; Arts highest (45.9%)	SVM, DT, ANN, LR across programmes	European context
13	Waheed *et al*. (2022)	Early prediction in self-paced learning; AUC >0.85	Deep Learning on sequential data	Online/self-paced only
14	Oqaidi *et al*. (2022)	African context (Morocco); RF achieved 87.5% accuracy	LR, DT, RF, SVM on academic data	Limited features
15	Adekitan & Salau (2019)	Nigerian context; JAMB and O-level predict performance	Multiple regression and ANN	Performance only; no dropout
16	Osemwegie & Amadin (2023)	Applied 6 ML algorithms for CS student dropout prediction at University of Benin using course grades	Machine learning (LR, SVM, DT, KNN, NB, ANN)	Limited to academic grades only; no web deployment; single institution

2.5 Research Gap

The review of literature reveals that while predictive analytics has been widely applied to student dropout prediction in North America, Europe, and parts of Asia, significant gaps remain in the Nigerian context. Adekitan and Salau (2019) applied data mining in a Nigerian university, but focused on performance prediction rather than dropout risk. Osemwegie and Amadin (2023) predicted Computer Science student dropout at the University of Benin, but relied solely on course grades without incorporating demographic, financial, or lifestyle variables. No existing study incorporates Nigerian student lifestyle factors such as side hustle hours, gaming hours, and clubbing frequency as predictive features. Furthermore, most studies stop at model development without deploying accessible web-based tools for non-technical academic staff.

This study bridges these gaps by:

i. Developing a dropout prediction model incorporating 37 features (demographics, academics, finances, and lifestyle factors) adapted to the Nigerian university context;
ii. Comparing five machine learning algorithms on the same dataset with consistent evaluation metrics;
iii. Deploying the best-performing model as a six-page Streamlit web application accessible to academic administrators; and
iv. Focusing specifically on Computer Science students at Godfrey Okoye University, Enugu.

CHAPTER THREE

SYSTEM ANALYSIS AND DESIGN

3.0 Introduction

This chapter provides a thorough account of the design decisions and architectural choices behind the Student Dropout Prediction System. All methods currently used to determine if students may be at risk of dropping out are reviewed and limitations of those also explained along with goals for the proposed system. Requirements for the system as well as how it was designed are discussed and system architecture is also presented via UML diagrams. The proposed system will actually be a Streamlit-based web application, which allows for an interactive user interface where one can measure student dropout risk, by means of Machine Learning algorithms. The system will allow users to conduct their own predictions for individual students via web forms, batch predictions by uploading CSV files, providing interactive data visualizations via dashboard views, and provide a thorough analysis of how well the model performed.

3.1 System Analysis

System analysis involves examining the current approach to identifying at-risk students, understanding its limitations, and defining clear objectives for the proposed machine learning-based prediction system.

3.1.1 Analysis of the Existing System

The current system of identifying drop out students at Godfrey Okoye University is a reactive and manual approach. Struggling students are usually identified by the academic advisors and lecturers via personal observation and grading at the end of the semester. The present system comprises of the following features:

i. Manual Identification: At the end of each semester, lecturers and academic advisors will manually check students' results to determine poor students.

ii. Reactive Approach: Student interventions are implemented as a response to a student's demonstrated academic struggles or academic requests for withdrawal.

iii. Available Data Sources: Sources of data like historical academic records and demographic data are not systematically analyzed for prediction.

iv. Periodic Reviews: Student learning is measured on a regular basis typically at the end of each semester rather than on a continuous basis.

v. Informal Process: No standard or automated process is used to identify at-risk students.

3.1.2 Limitations of the Existing System

The manual and reactive approach to student dropout identification has several significant limitations:

i. Delayed Recognition: A student’s risk status only becomes apparent once several course failures have already accumulated, which reduces the effectiveness of intervention.

ii. Absence of Forecasting Tools: The existing process offers no mechanism for estimating which students are likely to leave the programme based on their academic and demographic data.

iii. Inconsistent Identification: Monitoring quality depends on individual staff members, resulting in inconsistent identification of at-risk students.

iv. Time-Consuming: Manual review of student records is labor-intensive and imposes a substantial burden on academic staff.

v. Subjective Assessment: Identification relies on personal judgment rather than data-driven analysis.

vi. No Pattern Discovery: Important patterns in academic performance and student characteristics are not identified through systematic analysis.

3.1.3 Objectives of the Proposed System

The proposed Student Dropout Prediction System addresses the shortcomings of the existing approach by using machine learning algorithms and deploying them through an accessible web application. The specific objectives are as follows:

i. Data-Driven Predictions: To use student academic records, demographic information, and performance data to generate accurate predictions with five machine learning algorithms.

ii. Automated Analysis: To automate student data analysis through a web-based interface, thereby reducing the manual workload of academic staff.

iii. Accurate Classification: To classify students as Graduate or Dropout using the best-performing machine learning algorithm.

iv. Web-Based Accessibility: To make the predictive model available as a Streamlit web application which is accessible to academic administrators without requiring technical expertise.

v. Single and Batch Predictions: To enable both individual student predictions through web forms and bulk predictions through CSV file uploads.

vi. Visual Analytics: To provide interactive dashboards and visualizations that help identify the factors contributing to dropout risk.

vii. Model Transparency: To present model performance metrics and comparisons, enabling users to assess the reliability of the predictions.

3.2 System Requirements Specification

The System Requirements Specification defines the functional and non-functional requirements for the proposed Streamlit-based web application.

3.2.1 Functional Requirements

Functional Requirements describe the functionality the system must support:

Home page module:

- System should present key performance measures: accuracy, total dataset size, distribution of data by class.
- System should present percentage of students who have dropped out/ graduated as interactive pie chart.
- System should show the best performing model and explanation for why it was chosen.

Single student prediction module:

- System should provide an interactive web form for 36 input student attributes.
- System should perform validation on the user input. System should include drop-down options where applicable.
- System should provide predictions, along with probability for each possible class (Dropout and Graduate).
- System should present the confidence in prediction as an interactive gauge chart and a probability distribution chart.

Batch prediction module:

- System should allow user to upload a CSV file for multiple students.
- System should take an uploaded file and predict for all the students.
- System should present a summary, i.e., total students, total students that have dropped out, students at-risk.
- System should allow users to download their results.

Analytics Dashboard module:

- System should show interactive visualizations. Plots are preferably generated using Plotly.

- System should allow users to filter the student data based on gender and their current status (graduated/dropout).

- System should display heat-map for correlation between different features and students' risk of dropping out.

- System should display distributions for key attributes such as CGPA, attendance score, study hours etc.

Model performance module:

- System should compare the five trained algorithms: Decision Tree, Random Forest, Logistic Regression, Gradient Boosting and Support Vector Machine.

- System should also show accuracy, precision, recall, and F1-score for each model.

- System should show confusion matrix for each model.

- System should provide radar charts which give an overall comparison between the performance of the 5 algorithms.

3.2.2 Non-Functional Requirements

Non-functional Requirements: These define the qualities that a system must have:

Performance:

* The single prediction should be delivered within 2 seconds from form submission.

* The system should process up to 5000 batch predictions within 30 seconds.

Usability:

* The web interface will be easy to operate and does not require specialized knowledge of computer systems to be employed.

* Visualizations should be clearly defined, labeled, and interactive.

* The system will provide tips for data input and instructions where needed.

Reliability:

*The system should handle errors smoothly with informative error messages.

* The system should consistently give same outputs for same inputs.

Portability:

* The system should be reachable through any web browser (e.g. Chrome, Fire Fox, Edge, Safari).

* The system should work on any computer that has Python 3.x installed.

3.3 Design Methodology

The development of the Student Dropout Prediction System follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which is the most widely used framework for data mining and machine learning projects (Wirth & Hipp, 2000).

3.3.1 Justification for CRISP-DM Methodology

Reasons for selecting the CRISP-DM Methodology:

Industry Standard: CRISP-DM is widely recognized as a leading procedural framework for data mining project, ensuring a more systematic and tested approach (Shearer, 2000).

Iterative Process: It enables the user to iterate back and forth between phases to re-design the model with consideration to performance during evaluation.

Business Understanding: Focus is first given on the business objective/problem rather than just focusing on the technical implementation.

Overall Aspect: it covers the entire process right from understanding the data to finally deploying it as a web application.

CRISP-DM has the following 6 phases:

1. Business Understanding: The student dropout problem in Godfrey Okoye University and objectives for the prediction.

2. Data Understanding: Examining the data-set, understanding the structure and identification of data quality issues.

3. Data Preparation: Cleaning, transforming and encoding categorical features and scaling of numerical features.

4. Modeling: Training of five machine learning algorithms and comparing performance among them.

5. Evaluation: assessing models using accuracy, precision, recall and F1-score.

6. Deployment: Development of a Streamlit web application to deploy the most suitable model.

3.3.2 Method of Data Collection

This project makes use of a dataset originally sourced from the UCI Machine Learning Repository but has been adapted to the Nigerian university context at Godfrey Okoye University.

Original Data Source:

The underlying dataset was drawn from "Predict Students' Dropout and Academic Success" dataset available on the UCI Machine Learning Repository (Realinho *et al*., 2022), which was originally collected from a higher education institution in Portugal and included various pieces of student academic and demographic information.

Data Transformation:

In order to make the data relevant to the Nigerian academic environment, the following transformations were made:

i. Renamed feature labels to match Nigerian academic terminologies e.g., 'Admission grade' became 'JAMB_Score', 'Curricular units' became 'Registered_Courses'

ii. Adjusted categorical variable values to align with the Nigerian standard e.g., types of scholarship, ranges of parents' occupations

iii. Included new features that would be relevant to a Nigerian university student e.g., 'Out_of_State_Student', 'School_Fees_Cleared', 'Side_Hustle_Hours'

iv. Removed the 'Enrolled' category and only used 'Graduate' and 'Dropout' for binary classification

v. Exchanged Portuguese economic variables with aspects of daily life pertinent to Nigerian university students.

Transformed Data Description:

Dataset Name: Godfrey Okoye University Student Academic Performance Dataset

Number of records: 3,630

Number of features: 37

Data type: CSV file

Target feature: Academic_Status (Graduate or Dropout)

Dataset features:

Demographic features: Gender, marital status, student's residency type (local or out-of-state).

Family background: Highest level of parent's education, parents' occupations.

Financial features: Monthly allowance range, status of school fee payment, amount of unpaid school fees, scholarship type, number of hours spent on side hustles.

Academic entry: JAMB scores, O-level subjects with passed credit.

First semester academic performance: Number of registered courses, number of courses taken in the first semester exams, number of first semester courses passed, number of first semester courses

failed, first semester GPA, marks obtained in the first semester exams, mid-semester marks obtained in the first semester exams.

Second semester academic performance: Number of second semester courses registered, number of courses taken in the second semester exams, number of second semester courses passed, number of second semester courses failed, second semester GPA, marks obtained in the second semester exams, mid-semester marks obtained in the second semester exams.

Engagement features: Attendance rate score, assessment of student's character, average study hours per day.

Lifestyle features: Number of hours spent per week gaming, number of hours per week in clubs and social events, number of nights spent out of school, substance use score. **Table 3.1** shows the distribution of the target variable in the dataset:

Table 3.1: Dataset Class Distribution

Academic Status	**Number of Students**
Graduate	2,209 (60.9%)
Dropout	1,421 (39.1%)
Total	**3,630 (100%)**

3.4 System Modeling Using UML

Unified Modeling Language (UML) diagrams are used to model the structure and behavior of the proposed system (Booch *et al.*, 2005). This section presents the use case diagram and activity diagram that describe the system's functionality and workflow.

3.4.1 Use Case Diagram

The use case diagram shows the relationship between the user (Academic Administrator/Researcher) and the functions of the system. There are 6 use cases to access to within the Streamlit web application via the sidebar menu.

Actors:

Academic Administrator/Researcher: The user who uses the web application to do prediction and explore student information.

System: The auto-pilot system to conduct predictions by using the trained machine learning model.

Use Cases:

UC1 - View Home Dashboard: The user can get the system overview, important metrics, and the distribution of dataset.

UC2 - Predict Single Student: The user fills out a form to predict one student and then receives the prediction result with probabilities.

UC3 - Predict Batch Students: The user uploads a CSV file to predict multiple students in batch. Then, download the result in a csv file.

UC4 - Explore Analytics Dashboard: The user interacts with graphs and filters to get deeper insight.

UC5 - View Model Performance: The user can compare the 5 models and review the confusion matrix for each model.

UC6 - View About Information: The user can get more information about the project, the research approach, and the technologies used.

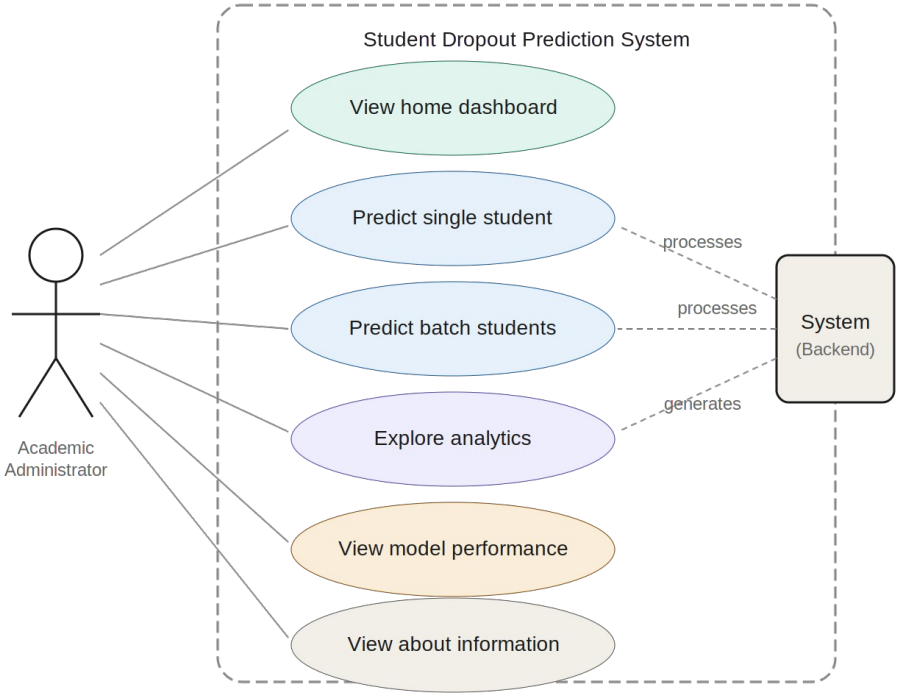

Figure 3.1: Use Case Diagram

The use case diagram above illustrates the two primary actors in the system: the Academic Administrator who interacts with the system by entering student data, uploading CSV files, and viewing predictions, and the System itself which processes inputs, runs the trained model, and generates prediction results and visualizations.

3.4.2 Activity Diagram

The activity diagram provides a high-level overview of the prediction process flow, and explains step-by-step activities from user input until the display of result.

Single Prediction Workflow:

1. User clicks on the 'Single Prediction' page in the website

2. The system displays the single prediction form, including the 36 features which include demographic, academic, financial, and lifestyle attributes

3. User fills in the student data and clicks on the ‘Predict’ button

4. System validates the input data

5. System preprocesses the data by performing the necessary data transformations and encoding. For instance, categorical features are converted to numerical representation, and all features are scaled.

6. The processed data is passed to the trained predictive model, where a prediction is generated. In addition to the predicted class, probabilities for each class are produced.

7. The prediction and the probability values are displayed to the user in a visually attractive format using the gauge chart and the pie chart.

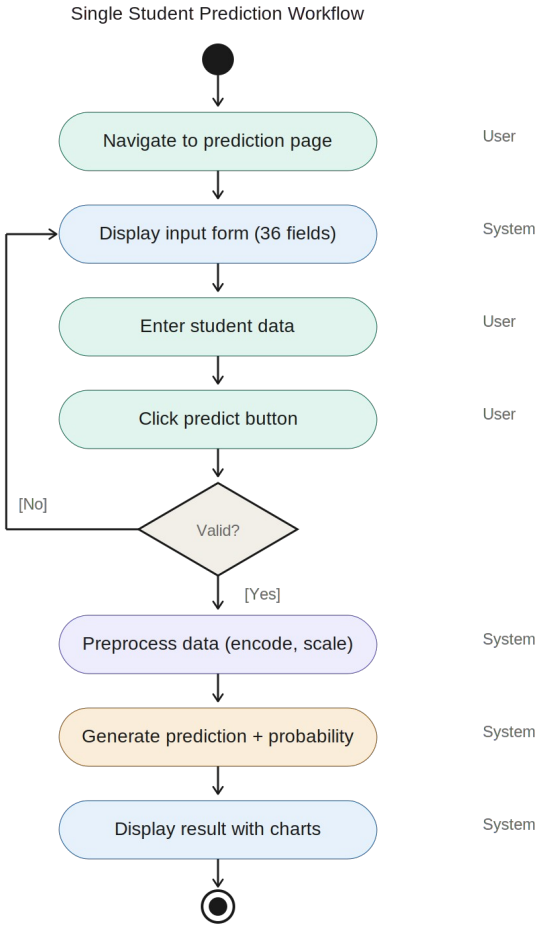

Figure 3.2: Activity Diagram

The activity diagram above depicts the step-by-step workflow of the single student prediction process, from the point where the user navigates to the prediction page, through data entry, input validation, preprocessing, model prediction, and finally the display of results including the predicted outcome, confidence score, and risk gauge.

3.5 System Design

This section presents the design of the system components including input design, output design, data storage, and system architecture.

3.5.1 Input Design

The system takes input through two primary channels: web forms for single predictions and CSV file uploads for batch predictions. Table 3.2 shows the input specifications:

Table 3.2: System Input Specifications

Input Type	Format	Description
Single Student Data	Web Form	36 feature fields including dropdowns for categorical variables and number inputs for numerical variables
Batch Student Data	CSV File	CSV file with 36 columns matching the feature names, one row per student
Dashboard Filters	Dropdowns	Filter selections for gender and academic status

3.5.2 Output Design

The system produces several outputs displayed within the web application interface. Table 3.3 shows the output specifications:

Table 3.3: System Output Specifications

Output Type	Format	Description
Prediction Result	Text + Charts	Graduate/Dropout classification with probability scores displayed via gauge charts and pie charts
Batch Results	Table + CSV	DataFrame showing predictions for all uploaded students, downloadable as CSV
Dashboard Charts	Plotly Charts	Interactive histograms, bar charts, pie charts, and heatmaps
Model Metrics	Tables + Charts	Accuracy, precision, recall, F1-score for all five algorithms with confusion matrices

3.5.3 Data Storage Design

The system uses file-based storage for data persistence. Table 3.4 shows the storage components:

Table 3.4: Data Storage Design

File	Format	Purpose
godfrey_okoye_dataset.csv	CSV	Training dataset with 3,630 student records
trained_model.pkl	Pickle	Serialized trained Decision Tree model
scaler.pkl	Pickle	StandardScaler for feature normalization
feature_columns.pkl	Pickle	List of feature column names after encoding
model_metadata.json	JSON	Performance metrics for all five algorithms
categorical_values.json	JSON	Valid options for categorical dropdowns

3.5.4 System Architecture

The system follows a three-tier architecture namely, the presentation layer (Streamlit UI), application layer (Python processing), and data layer (file storage). The system architecture is displayed in figure 3.3.

Presentation Layer (Streamlit Frontend):

It provides the web-based graphical user interface, with 6 different web pages (Home, Single Prediction, Batch Prediction, Dashboard, Model Performance, About). Accept the user input in terms of forms, file upload, and filters. Display the interactive plots through Plotly charts (Plotly Technologies Inc., 2015).

Application Layer (Python Backend):

Data Preprocessing module: This handles the process of encoding categorical variables, feature scaling using StandardScaler, and validating the data. Prediction engine: Load trained model and scaler, processes input and outputs predictions with probabilities. Visualization engine: Produces interactive plots using Plotly for displaying dashboards and results.

Data Layer (File Storage):

The training dataset (CSV) and trained model and scaler (Pickle files), features metadata (JSON), model performance metrics (JSON) are all stored as files on storage and are loaded during application start for rapid runtime.

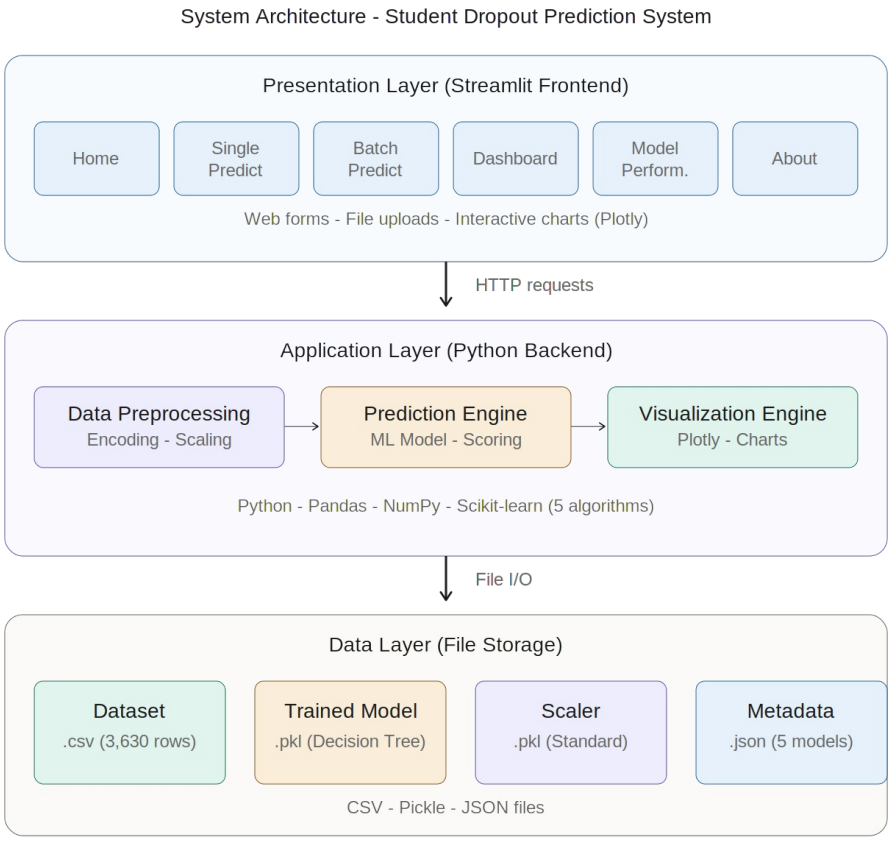

Figure 3.3: System Architecture Diagram

The architecture diagram above shows the three-tier structure of the system: the Presentation Layer (Streamlit UI) which handles user interaction, the Application Layer which contains the data preprocessing module, prediction engine, and visualization engine, and the Data Layer which stores the trained model, scaler, feature metadata, and dataset files.

3.5.5 Technology Stack

Table 3.5 presents the technologies used to implement the system:

Table 3.5: Technology Stack

Component	**Technology**
Programming Language	Python 3.x
Web Framework	Streamlit (Streamlit Inc., 2019)
Data Manipulation	Pandas, NumPy (McKinney, 2010; Harris *et al*., 2020)
Machine Learning	Scikit-learn (Pedregosa *et al*., 2011)
Visualization	Plotly, Matplotlib, Seaborn
Model Serialization	Pickle
Configuration Storage	JSON
Data Storage	CSV Files
Development Environment	VS Code, Jupyter Notebook

3.6 Summary

This chapter presented the system analysis and design of the Student Dropout Prediction System. Student dropout prediction requires a systemic approach in tracking and monitoring students. The weaknesses of the current manual system-late discovery of student issues, lack of prediction ability and unstable identification of at-risk students-are covered. The current system is proposed as a web application built on the Streamlit framework to provide access to the predictions and the use of machine learning for those predictions. The system requirements and specifications are then laid out. The requirements are both functional and non-functional and for a total of 6 modules (Home, Single Prediction, Batch Prediction, Dashboard, Model Performance, and About respectively). The system follows the CRISP-DM methodology, and the usage of the system and the flow of activities in the system using use case and activity diagrams are illustrated. Finally, the system architecture and specifications-including input and output specifications, file storage method and system architecture-as well as the stack of technologies used to build the system (Python, Streamlit, Scikit-learn for Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, Support Vector

Machine and Plotly), data sources, and dataset of 3,630 students and 37 features adapted from the Nigerian university context, are described. The following chapter goes into detail on the system implementation and describes the outcomes from training and testing the machine learning models.

CHAPTER FOUR

SYSTEM IMPLEMENTATION AND RESULTS

4.0 Introduction

This chapter documents the construction of the Student Dropout Prediction System and examines the outcomes obtained from training and benchmarking five classification algorithms. It specifies the environment for the implementation, describes the dataset used, and specifies the data preprocessing steps. The chapter also explains how the models are trained and evaluated, how the web application is created using Streamlit, and the detailed discussion of the experimental results. It was developed as per the CRISP-DM methodology described in Chapter Three of the thesis. The implementation is the conversion of the system design into a working web application that would enable the academic administrator of Godfrey Okoye University to be able to predict students' dropout risk via machine learning classification algorithms (Wirth and Hipp, 2000).

4.1 Implementation Environment

The system was developed and tested using the following hardware and software specifications.

4.1.1 Hardware Specifications

Development took place on a personal computer with the following specifications:

Component	**Specification**
Processor	Intel Core i5 (or equivalent)
RAM	8 GB DDR4
Storage	256 GB SSD
Operating System	Windows 10/11
Display	15.6-inch LCD

4.1.2 Software Requirements

The following software tools, programming languages, libraries, and frameworks were used for system development:

Software	Version	Purpose
Python	3.13	Programming language

Scikit-learn	1.6.x	Machine learning algorithms (Pedregosa et al., 2011)
Streamlit	1.42.x	Web application framework (Streamlit Inc., 2019)
Pandas	2.2.x	Data manipulation and analysis (McKinney, 2010)
NumPy	2.2.x	Numerical computing (Harris et al., 2020)
Plotly	6.0.x	Interactive data visualisation (Plotly Technologies Inc., 2015)
Jupyter Notebook	Latest	Model training environment
Visual Studio Code	Latest	Code editor / IDE
Google Chrome	Latest	Web browser for testing

4.2 Dataset Description

4.2.1 Dataset Source

The source data was drawn from the publicly available dataset titled Predict Students' Dropout and Academic Success, hosted on the UCI Machine Learning Repository and contributed by Realinho et al. (2022). That original collection comprised records from a Portuguese university, covering student background characteristics, semester-level academic indicators, and socio-economic attributes.

The study involved some modifications to suit the educational context of the study especially the educational environment at Godfrey Okoye University, Enugu. The names of the features were modified to fit into the Nigerian academic context, for instance, the name of "Admission grade" was changed to "JAMB_Score". Categorical values were further standardized to reflect Nigerian standards. Also, new features were developed to capture the experiences of Nigerian students: Out_of_State_Student, School_Fees_Cleared, and Side_Hustle_Hours. The "Enrolled" category was dropped to make the problem a binary classification problem, and Portuguese economic indicators were replaced by factors related to Nigerian lifestyles.

4.2.2 Feature Description

The transformed dataset contains 3,630 student records with 37 features (109 features after one-hot encoding). The features are organized into the following categories:

Category	Count	Examples
Demographic	3	Gender, Marital Status, Out_of_State_Student
Financial	4	Monthly_Allowance_Bracket, School_Fees_Cleared, On_Scholarship, Outstanding_Fees
Academic Entry	2	JAMB_Score, O_Level_Credits_Passed
Semester Performance	16	1st/2nd Sem GPA, Courses Passed/Failed, Midsemester Scores, Exam Scores
Engagement	3	Attendance_Score, Character_Assessment, Daily_Study_Hours
Lifestyle	6	Weekly_Gaming_Hours, Weekly_Clubbing_Frequency, Side_Hustle_Hours, Substance_Use_Index
Target	1	Academic_Status (Graduate / Dropout)

4.2.3 Target Variable Distribution

The classification is binary categorical, with two classes (Academic_Status). From the dataset, there are 2209 students classified as Graduate (60.9%) and 1421 students that are classified as Dropout (39.1%). As a result, this yields a moderately imbalanced dataset, which is why the main evaluation metric chosen was the F1-Score which provides a balance between precision and recall, as discussed by Hosmer, Lemeshow and Sturdivant (2013).

4.3 Data Preprocessing

The raw dataset was preprocessed and restructured so it could be used for machine learning model training. This preprocessing step corresponds to the data preparation phase of the CRISP-DM framework (Shearer, 2000).

4.3.1 Target Variable Encoding

The target variable, Academic_Status was converted to binary form. With this combination, Dropout was assigned class label 1 (the positive class) and Graduate was assigned class label 0. In

this way, the machine learning models can treat the task of predicting dropouts as a binary classification problem. The process is illustrated in the code snippet below:

```python
y_encoded = np.where(y == 'Dropout', 1, 0)
```

4.3.2 Feature Encoding

For categorical attributes such as Gender, Marital Status, Highest_Parental_Education, Parent_Occupation, Monthly_Allowance_Bracket and others, one-hot encoding (also called dummy variable encoding) was applied to convert them into numbers. With this approach, each category becomes its own binary column, so the machine learning algorithms can work with data that isn't originally numeric. The parameter drop_first was used to prevent multicollinearity, by dropping the first category of each feature:

X_encoded = pd.get_dummies(X, columns = categorical_cols, drop_first = True)

After the encoding, there were 37 features, and 109 features.

4.3.3 Feature Scaling

Each numeric column was normalized by subtracting its column mean and dividing by its standard deviation, a transformation carried out through scikit-learn's StandardScaler utility. This normalisation step is especially relevant for distance-based and gradient-based classifiers like SVM and Logistic Regression, whose internal computations can be dominated by features with larger numeric ranges (Pedregosa *et al*., 2011):

``python

scaler = StandardScaler()

X_train_scaled = scaler.fittransform(Xtrain)

X_test_scaled = scaler.transform(X_test)

``

4.3.4 Train-Test Split

The dataset was divided into training and test sets using an 80:20 split, with random_state fixed at 42 for reproducibility. This produced 2,904 samples for training and 726 samples for testing:

```python

X_train, X_test, y_train, y_test = train_test_split(

X_encoded, y_encoded, test_size=0.2, random_state=42

)

```

4.4 Model Training and Evaluation

Five machine learning classification algorithms were trained and evaluated on the preprocessed dataset. The algorithms were chosen based on their proven effectiveness in binary classification tasks as discussed in the literature review (Chapter Two).

4.4.1 Algorithms Implemented

Below are five ML algorithms that were trained with the use of scikit-learn.

1. Random Forest Classifier - this an ensemble- based model where many decision trees are trained and the most frequent class assigned to each sample is the class predicted (Breiman, 2001) and used with n_estimators=100.

2. Logistic Regression - a statistical model of a binary outcome, this model calculates the probability of a two-class event occurring using the logistic function (Hosmer, Lemeshow and Sturdivant, 2013) and used with max_iter=1000.

3. Decision Tree Classifier - a non-parametric supervised learning technique, this model learns rules from input features in order to predict the target variable (Quinlan, 1986) and used with default parameter values.

4. Gradient Boosting Classifier - an ensemble-based model which sequentially builds an additive model in each iteration trying to correct the mistakes made by previous models (Friedman, 2001) and used with n_estimators=100.

5. Support Vector Machine - a classification algorithm that finds an optimal hyperplane separating various classes in a high dimensional feature space (Cortes and Vapnik, 1995) and used with RBF kernel and probability=True.

Below is the code snippet for the instantiation of the five ML models:

```python
models = {
"Random Forest": RandomForestClassifier(nestimators=100, random_state=42),
"Logistic Regression": LogisticRegression(maxiter=1000, random_state=42),
"Decision Tree": DecisionTreeClassifier(random_state=42),
 "Gradient Boosting": GradientBoostingClassifier(n_estimators=100, random_state=42),
"Support Vector Machine": SVC(kernel='rbf', probability=True, random_state=42)}
```

4.4.2 Model Comparison Results

Each algorithm was fitted on the scaled training data and evaluated on the test set using the four principal classification benchmarks: Accuracy, Precision, Recall, and F1-Score. Table 4.1 summarises the performance comparison of all five algorithms.

Table 4.1: Performance Comparison of Five Machine Learning Algorithms

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	77.00%	62.39%	100.00%	76.84%
Logistic Regression	89.67%	79.19%	98.92%	87.96%
Decision Tree	94.21%	86.83%	100.00%	92.95%
Gradient Boosting	86.36%	73.80%	99.64%	84.79%
Support Vector Machine	76.45%	61.83%	100.00%	76.41%

4.4.3 Confusion Matrix Analysis

The confusion matrix for each algorithm allows us to visualize the performance and accurately determine how many predictions were right versus wrong. When looking at the highest scoring algorithm, the Decision Tree, and using the 726 test samples, we see that there were 407 correct negatives (graduates correctly classified), 277 correct positives (dropouts correctly classified), 42 incorrect positives (graduates incorrectly classified as dropouts), and 0 incorrect negatives (dropouts were never incorrectly classified). The recall for this algorithm was 100%, meaning it correctly classified every student who was at risk of dropping out, a finding that can be extremely beneficial to the use of this tool in an educational setting.

4.4.4 Best Model Selection

Decision Tree Classifier was picked as the best model as it returned the F1-Score with the maximum value of 92.95%. F1-Score was selected as the primary criterion for deciding the model, because it is the harmonic mean of precision and recall, and gives a fair measure of both parameters

that are necessary considering the moderately imbalanced dataset of the study. In terms of overall performance, of the five algorithms that were tested, Decision Tree came out on top: with the maximum accuracy of 94.21%, the highest precision of 86.83%, recall of 100.00%, and the best F1-Score of 92.95%. The trained Decision Tree model was then pickled using Python's pickle module, and stored in the file trained_model.pkl in order to use it in the Streamlit application.

4.5 System Implementation

The finished system takes the form of a multi-page browser application constructed with the Streamlit library. It consists of six pages, with each page designed for a specific purpose. This section includes screenshots of the system and explains what each page does.

4.5.1 Home Page

The home page is the entry point of the system. The page shows a welcoming statement about the system, four main performance indicators of the system, namely: Total Students Analysed: 3,630, Model Accuracy: 94.2%, Features Used: 109, F1-Score: 93.0%, a brief explanation of the system features, and a dataset overview presenting the distribution of academic outcome and categories of features.

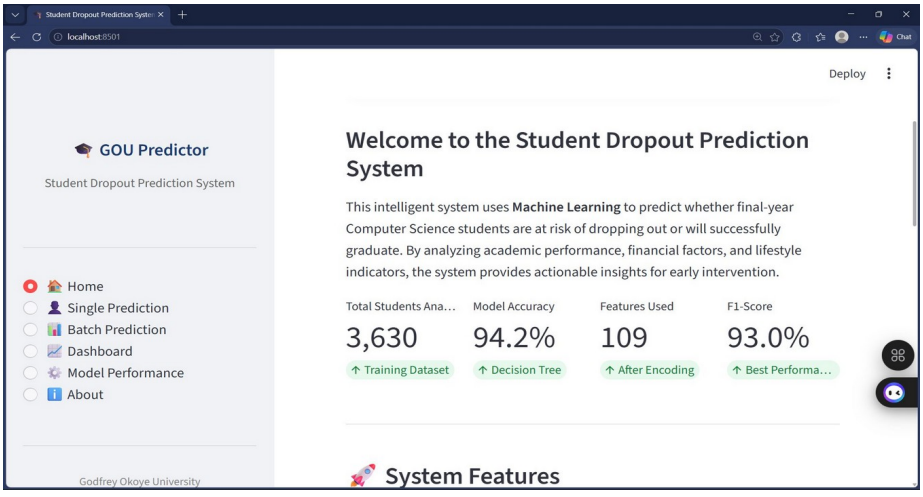

Figure 4.1: Home Page

The landing page displays a welcome message, four key system metrics (3,630 students, 94.2% accuracy, 109 features, 93.0% F1-Score), and an overview of system features.

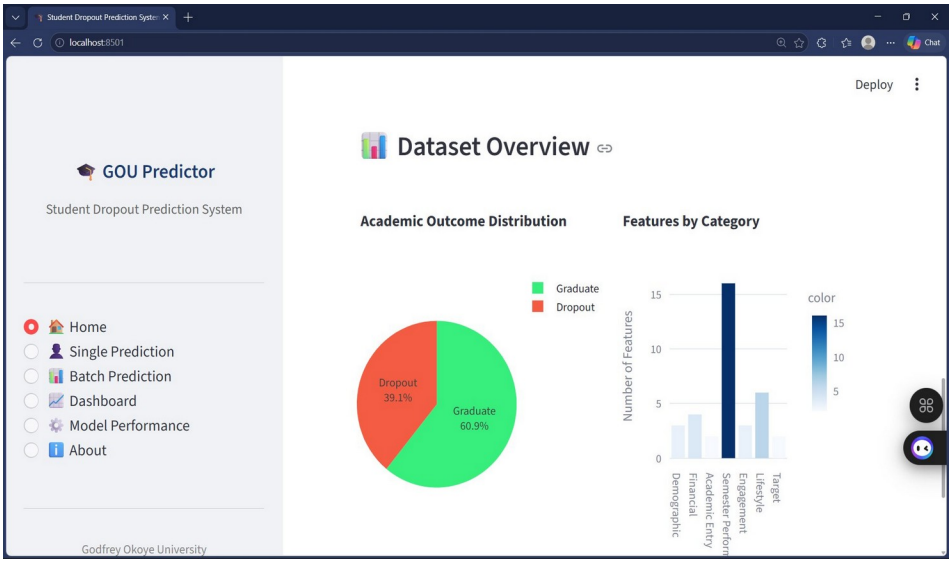

Figure 4.2: Dataset Overview

The academic outcome distribution pie chart (60.9% Graduate, 39.1% Dropout) and a bar chart showing features organisd by category.

4.5.2 Single Prediction Page

The Single Prediction page provides an interface where academic administration can submit individual student data via a web form and receive an immediate dropout risk prediction. The form is comprised of 9 sections, which includes Demographics, Family Back ground, Finance, Academic Entrance qualifications, 1st Semester academic, 2nd Semester academic, Engagement and assessment, Lifestyle and Present academic status. A total of 36 fields are provided in the input form.

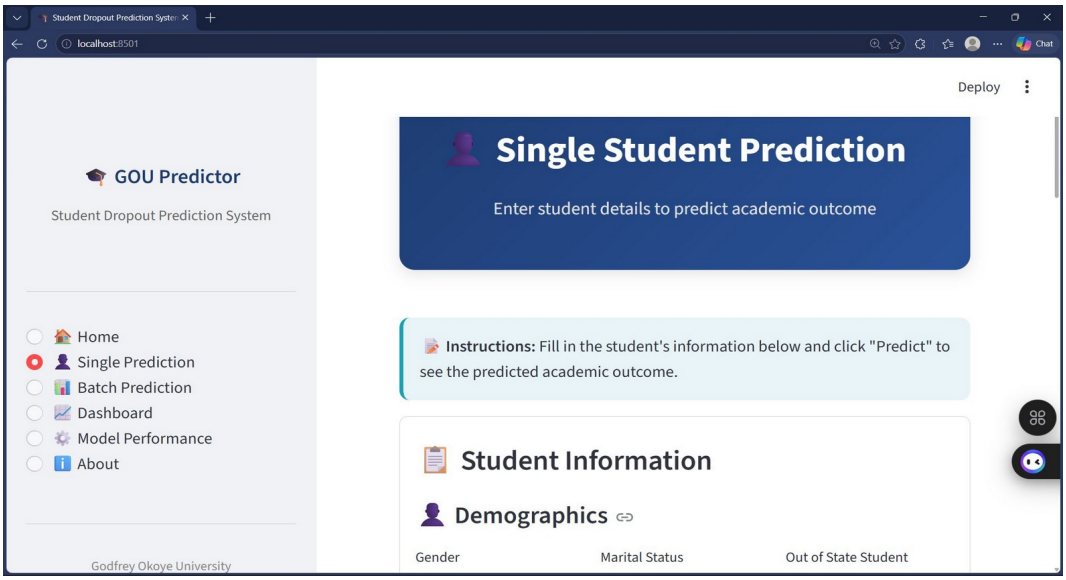

Figure 4.3: Single Prediction Page

The header and input form with the Demographics section, where the administrator enters student details such as Gender, Marital Status, and Out of State status.

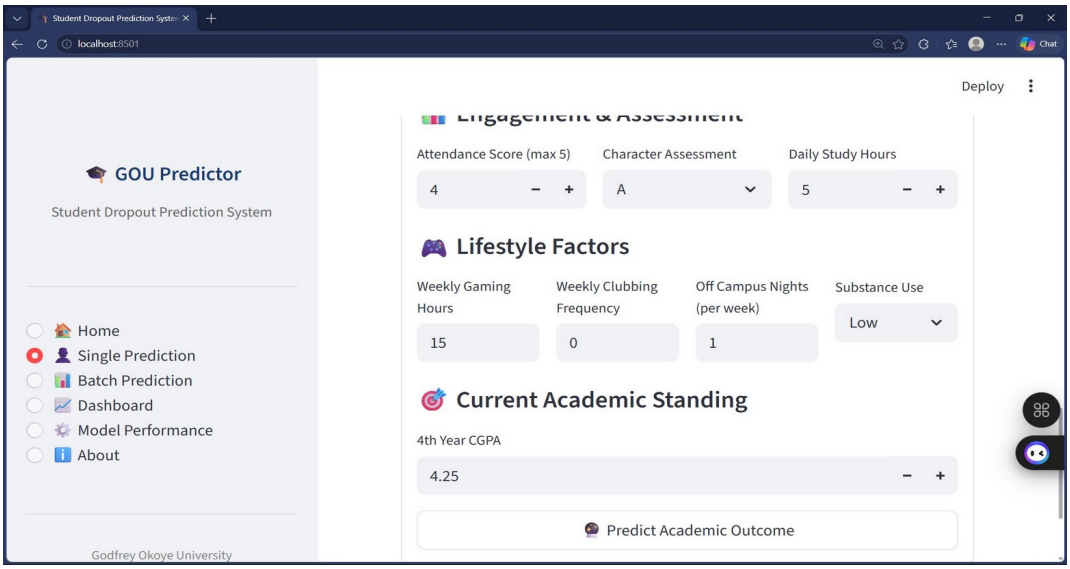

Figure 4.4: Single Prediction Page

This shows the Lifestyle Factors section of the input form with fields for Weekly Gaming Hours, Clubbing Frequency, Off Campus Nights, and Substance Use, along with the Current Academic Standing (4th Year CGPA) and the Predict button.

Upon submission, the system displays the prediction result with a confidence score, a recommendation message, a dropout risk gauge, and an outcome probability pie chart.

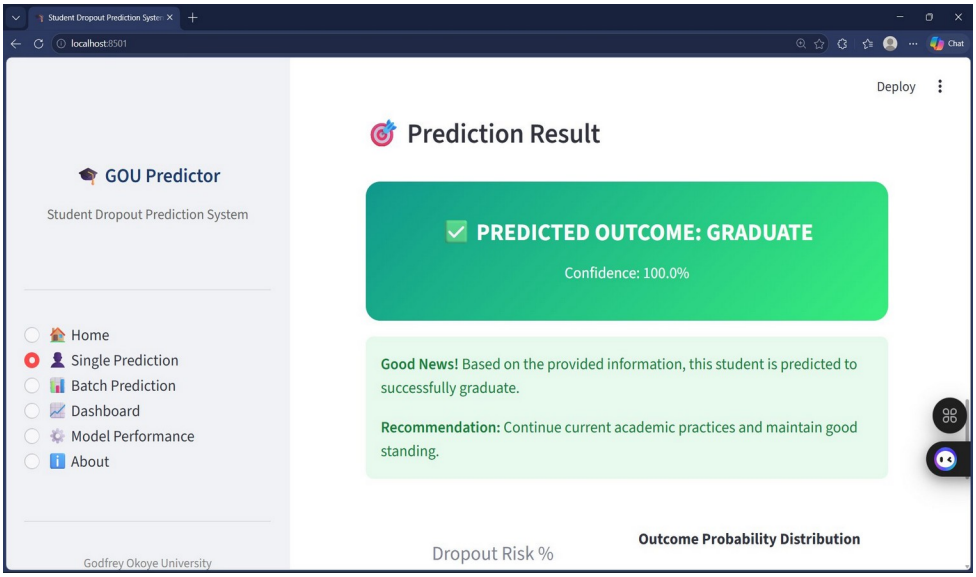

Figure 4.5: Single Prediction Result

The prediction output shows a Graduate outcome with 100% confidence, along with a recommendation to continue current academic practices

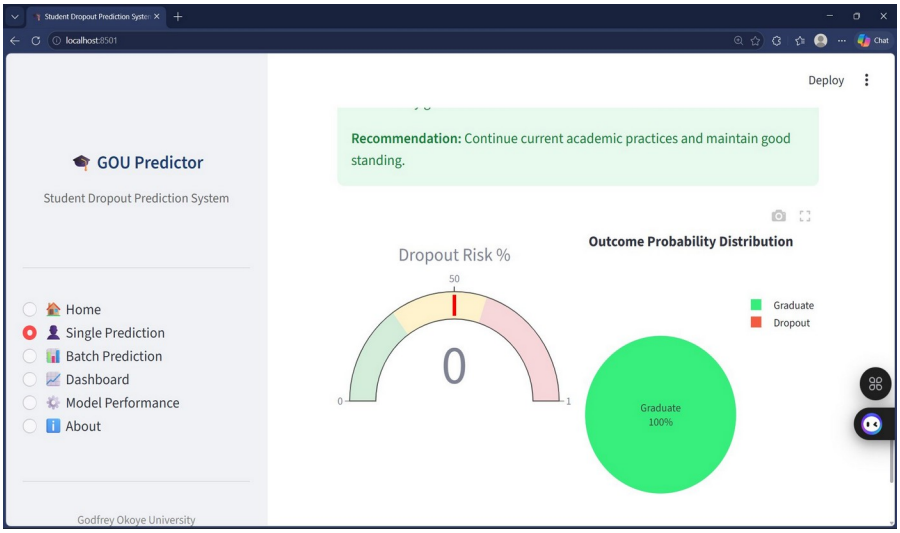

Figure 4.6: Single Prediction Result

The dropout risk gauge at 0 (low risk) and the pie chart confirms 100% Graduate probability for this student input

4.5.3 Batch Prediction Page

This page allows administrators to upload a csv file containing multiple students, and predict multiple students in one go. It contains a template CSV to ensure the correct format of the file, an area to upload the CSV file (CSV files up to 200MB), a data preview after uploading, and a progress bar while the batch processing occurs.

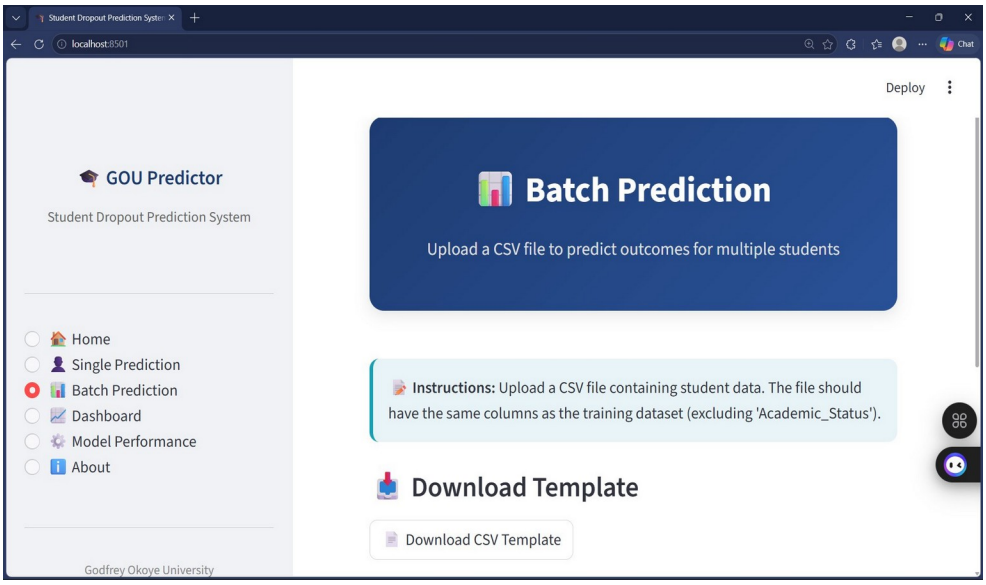

Figure 4.7: Batch Prediction Page – Instructions and CSV template download

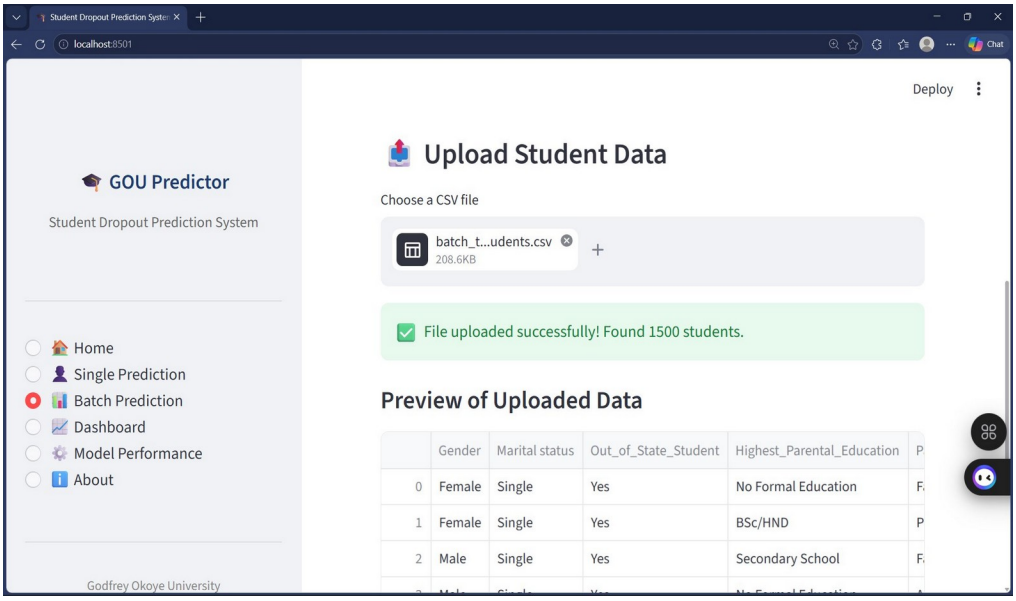

Figure 4.8: Batch Prediction – CSV file uploaded with 1,500 student records

After processing, the system displays summary metrics (total students, predicted graduates, at-risk students), a pie chart showing the batch prediction distribution, a histogram of dropout risk scores, and a detailed results table with individual predictions.

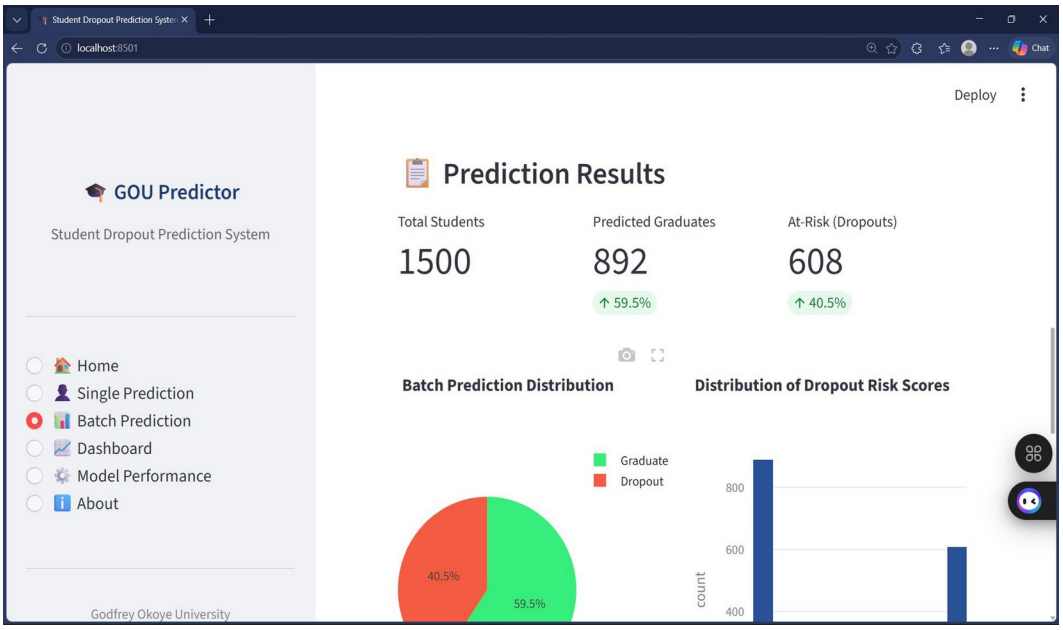

Figure 4.9: Batch Prediction Results – Summary metrics and distribution charts

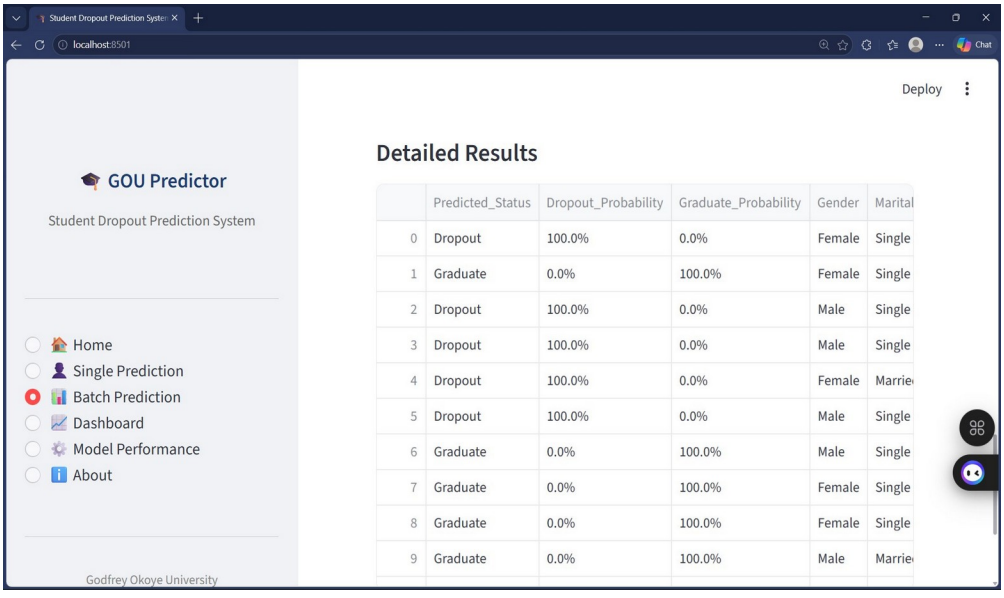

Figure 4.10: Batch Prediction Results – Detailed results table with probabilities

4.5.4 Analytics Dashboard

The Analytics dashboard offers visualizations where users can interact and examine the trends in the student dataset. The student data for all 3,630 students is plotted in the dashboard. The charts available in the dashboard include: distribution of CGPA, CGPA by academic status comparison, monthly allowance and parents' education comparison, analysis of lifestyle, hours spent on studying, playing games, doing side hustle, features correlation heatmap and the top 15 features which correlates to dropout risk.

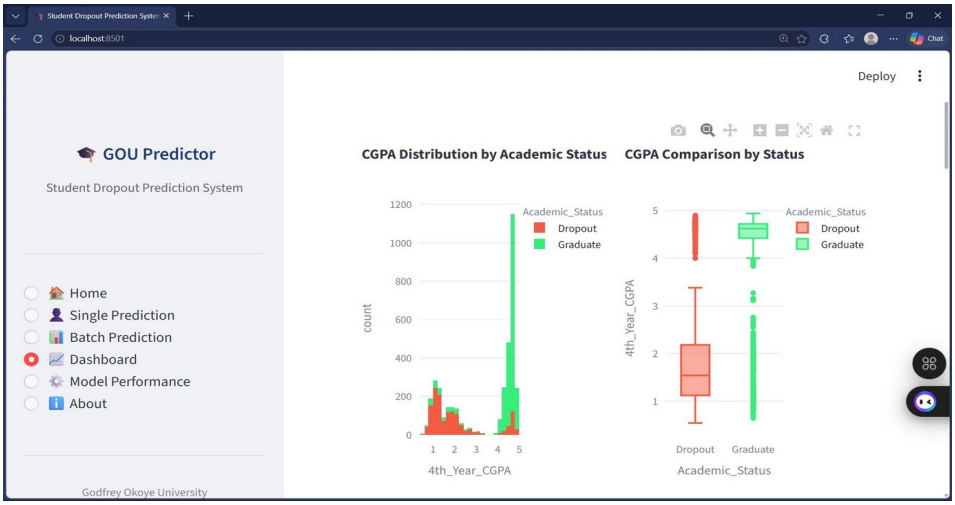

Figure 4.11: Analytics Dashboard – CGPA Distribution and Comparison by Academic Status

Figure 4.12: Analytics Dashboard – Monthly Allowance and Parental Education vs Academic Outcome

Figure 4.13: Analytics Dashboard – Lifestyle Factors Analysis (Study, Gaming, Side Hustle Hours)

Figure 4.14: Analytics Dashboard – Top 15 Features Correlated with Dropout Risk

4.5.5 Model Performance Page

The Model Performance page offers a thorough comparison of all five machine learning algorithms. It spotlights the top-performing model (Decision Tree), includes the algorithm comparison table covering all four metrics, and presents both a grouped bar chart and a radar chart to make visual comparisons easier. In addition, it provides confusion matrices for each algorithm.

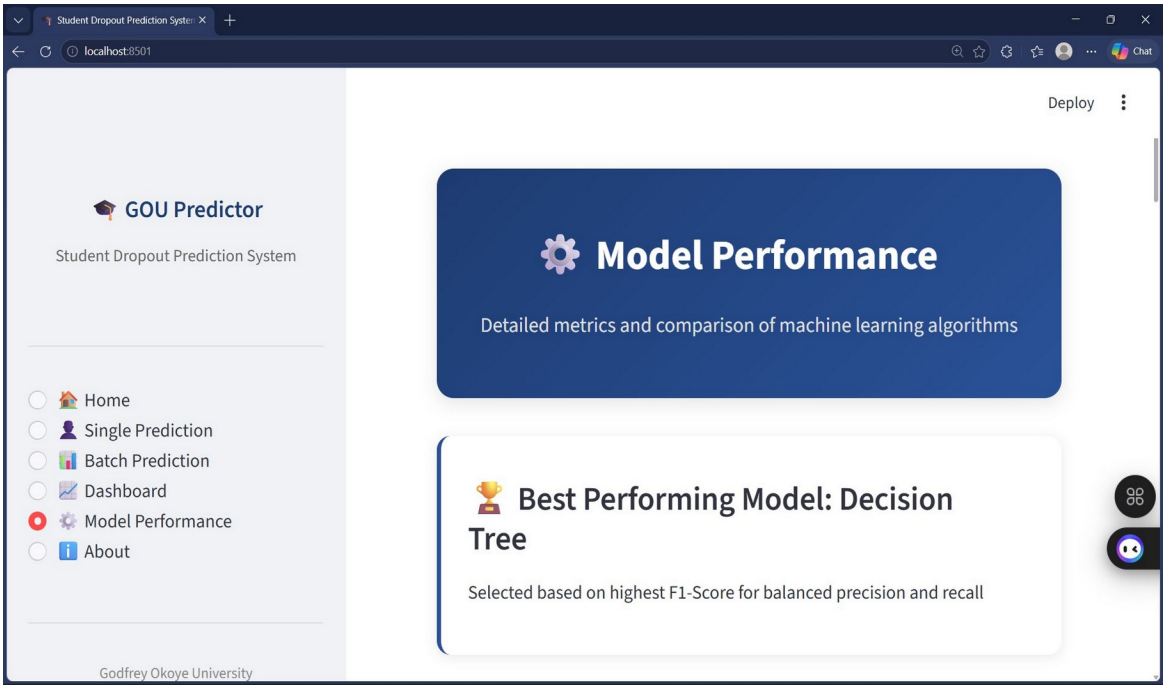

Figure 4.15: Model Performance Page – Best Performing Model: Decision Tree

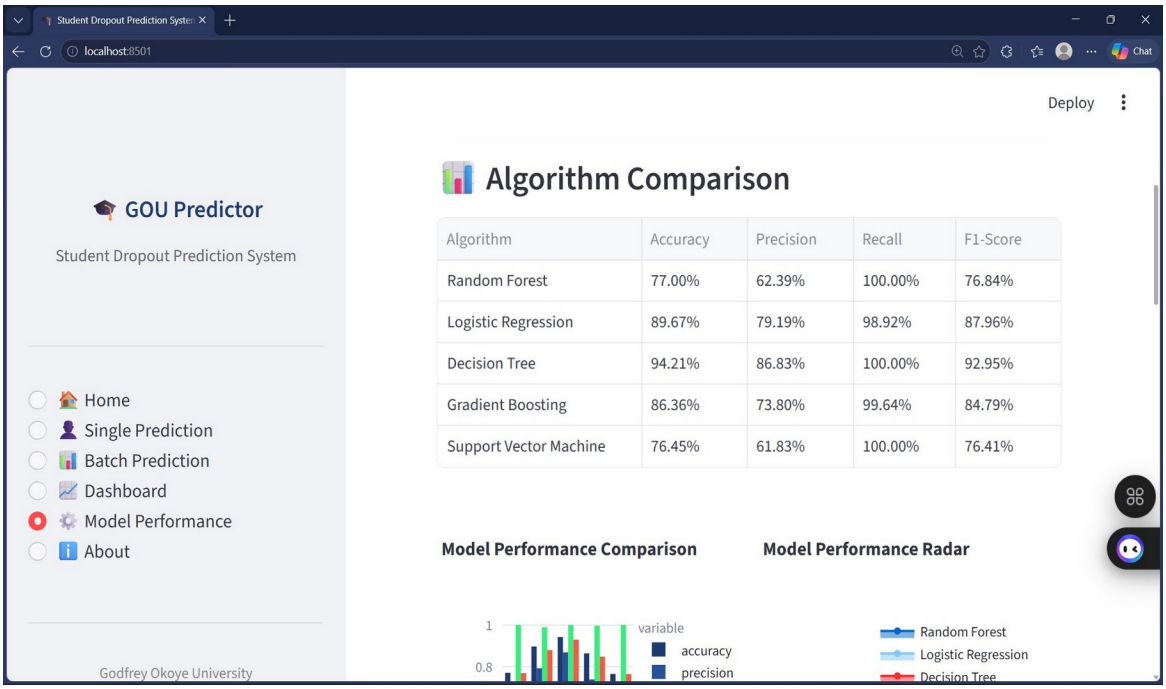

Figure 4.16: Model Performance Page – Algorithm Comparison Table

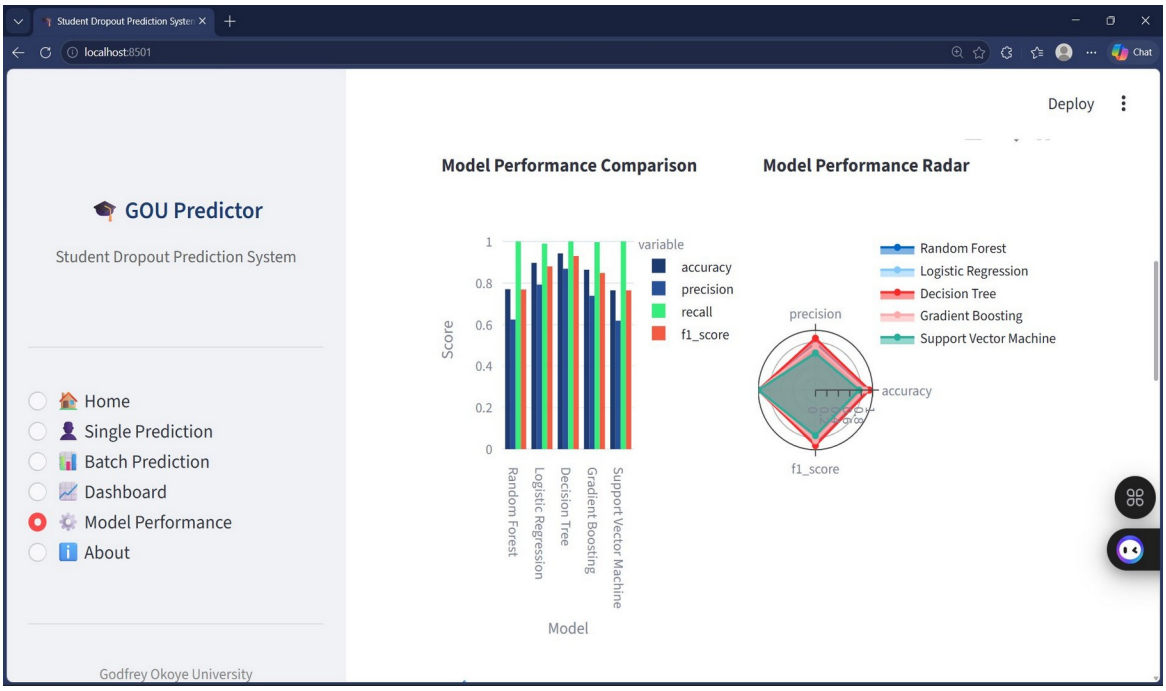

Figure 4.17: Model Performance Page – Bar Chart and Radar Chart Comparison

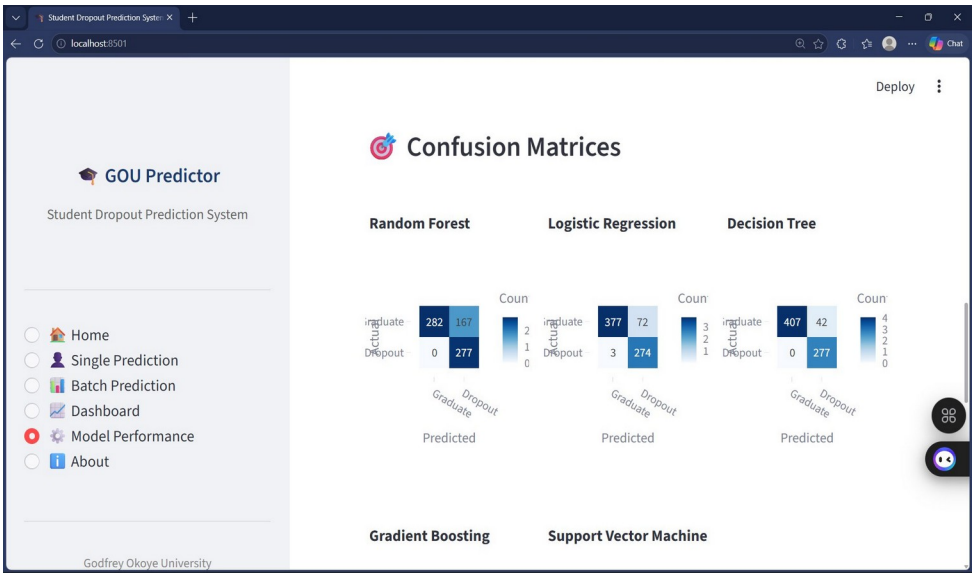

Figure 4.18: Model Performance Page – Confusion Matrices (Random Forest, Logistic Regression, Decision Tree)

4.5.6 About Page

The About page provides project documentation, including the project title, institution, department, and academic session; the system's purpose and the factors it analyses; the technology stack used; the CRISP-DM methodology followed; and dataset information (3,630 records, 109 features, 2 target classes).

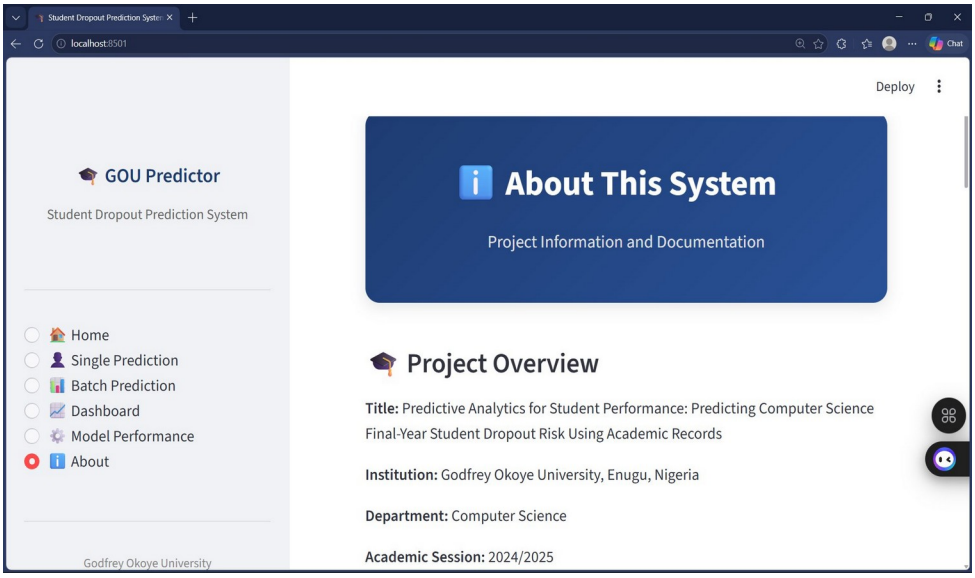

Figure 4.19: About Page – Project Overview and Institution Details

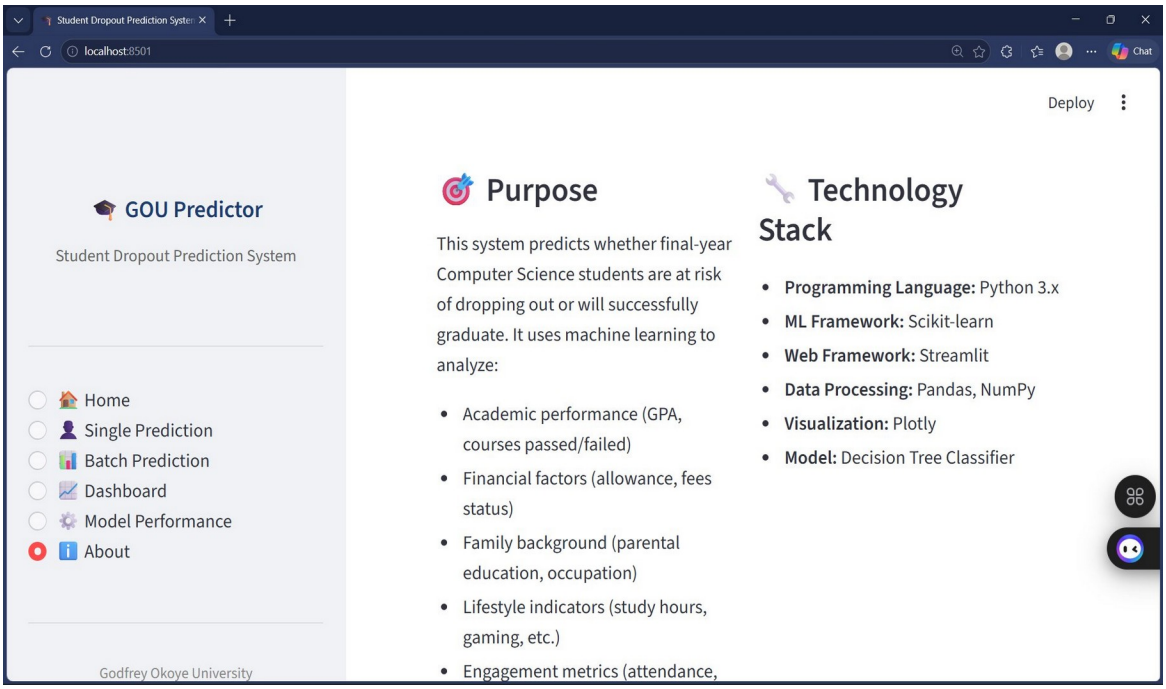

Figure 4.20: About Page – System Purpose and Technology Stack

4.6 Discussion of Results

The experimental results show that the Decision Tree Classifier outperformed the other four algorithms across all evaluation metrics. This section discusses the key findings and their implications.

4.6.1 Why the Decision Tree Performed Best

There are several reasons why Decision Tree might be best performing. Firstly, a mixture of both numerical and categorical data with obvious decision boundaries would fit Decision Trees well (Quinlan, 1986). Secondly, most of the features (1st_Sem_Courses_Failed, 2nd_Sem_Courses_Passed and Daily_Study_Hours) offer obvious enough splits where a decision tree will perform well. Thirdly, the dataset size is of a reasonable size (3,630 records) allowing sufficient samples at each split, enabling Decision Trees, generally susceptible to over-fitting, overcome this issue.

4.6.2 Feature Importance Analysis

In this correlation study, the highest positive correlates (risk factors) with student dropout were 1st_Sem_Courses_Failed, 2nd_Sem_Courses_Failed, Side_Hustle_Hours, Off_Campus_Nights, Weekly_Clubbing_Frequency, and Weekly_Gaming_Hours. The highest negatively correlated correlates (protection factors) with student dropout were 2nd_Sem_Courses_Passed, 1st_Sem_1st_Midsemester, 2nd_Sem_1st_Midsemester, Daily_Study_Hours and 4th_Year_CGPA. The implication of this finding is that there exist academic and social/lifestyle factors contributing to student dropout.

4.6.3 Practical Implications

Overall accuracy of 94.21% with perfect recall for predicting dropout implies that it is very good at predicting which students will drop out. Academic administrators at the department can use this information to further categorize the students based on whether or not the system predicted them to be a dropout or graduate. The administrator can then use this to advise on matters ranging from counseling, to referrals to student aid offices, to even life style decisions. The system also allows batch prediction on an entire group of students simultaneously, allowing for general assessment on overall dropout risk.

4.6.4 Comparison with Existing Literature

The F1-Score of 92.95% achieved by the Decision Tree Classifier in this study compares favourably with results reported in related studies. Table 4.2 presents a direct comparison of the performance of this study against four key studies from the literature review. Compared to Realinho *et al*. (2022), whose original UCI dataset yielded accuracies ranging from 75% to 92% across various algorithms, the present study achieved a higher accuracy of 94.21% using the adapted version of the same dataset. Oqaidi *et al*. (2022), working in a Moroccan university

context, reported a best accuracy of 87.5% with Random Forest, which is 6.71 percentage points lower than the result obtained in this study. Osemwegie and Amadin (2023), who conducted one of the few Nigerian studies on dropout prediction at the University of Benin, used six algorithms but relied solely on course grades without incorporating demographic or lifestyle variables. The present study extends their work by including 37 features spanning demographics, finances, academic performance, and lifestyle factors, resulting in a more comprehensive predictive model. Tamada *et al*. (2022), who used LMS log data in Brazil, achieved accuracies above 85%, which is also below the 94.21% achieved in this study. A distinctive contribution of the present study is the incorporation of Nigerian student lifestyle variables such as weekly gaming hours, clubbing frequency, and side hustle hours as predictive features. These context-specific variables, which have not been explored in prior studies, were found to be among the top 15 features correlated with dropout risk, demonstrating the importance of localized feature engineering in educational prediction models.

4.7 Summary

The implementation and result of Student Dropout Prediction System was illustrated in this chapter. The five machine learning models, that is, Decision Tree Classifier, Random Forest Classifier, Logistic Regression, Gradient Boosting Classifier and Support Vector Machine Classifier were applied on the data set consist of 3630 student records and 109 features, and their performance was evaluated. Decision Tree Classifier obtain 94.21%, 86.83%, 100.00% and 92.95% of accuracy, precision, recall and F1-score respectively. A web application was built using Streamlit, consisting of 6 pages: home page, single prediction page, batch prediction page, analytics dashboard page, model performance page, about page. Single and batch prediction are provided and some useful graphs are available to show the trend of student information and model

performance. Through the analysis and experiments it has been shown that the system is able to predict the dropout risk of students accurately and can be a good tool to support academic administrators at Godfrey Okoye University to make a better decision.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

This chapter concludes the study by giving an overview of the whole research, drawing conclusions from the results of the study, outlining the contribution of the study, discussing the weaknesses that have occurred, and recommending what can be done in future. It summarizes what has been learned in previous chapters and what the research may have meant for retention of Nigerian students in higher education.

5.1 Summary of the Study

The purpose of this research was to construct a classification system capable of determining whether final year computer science students of Godfrey Okoye University, Enugu, Nigeria will graduate successfully or are at risk of dropping out. Manual techniques for tracking student progress in Nigerian Universities have proven inadequate in predicting student success or failure in completing their studies, thus motivated this study.

Chapter One set the stage for this research by outlining the context of the study, including the increasing significance of using data to inform decision making in the field of education as well as the shortcomings of current student monitoring practices. The problem statement stated that there is no predictive system designed specifically for the context of higher education in Nigeria. To gather and preprocess student data, to identify the most significant variables that affect dropout-risk, to develop and train 5 machine learning models and assess and deploy the model with the highest prediction accuracy as a web app.

A detailed literature review was provided in Chapter Two that covered the theoretical basis of student retention (Tinto's Model of Student Departure) and educational data mining and learning analytics, as well as machine learning algorithms used for prediction. The related works reviewed showed that predictive analytics has been successfully implemented in the Western context in the educational sector, whereas its use in Nigerian universities is limited, thus creating an educational research gap that this study will address.

In Chapter Three, the system analysis and design process was explained, starting by analysing the existing manual system and its shortcomings. The requirements of the system were identified and shown in the form of UML diagrams such as the use case diagram that illustrates the interactions between Academic Administrators and the System, the activity diagram that depicts the prediction flow and the architecture diagram that depicts the three- tier architecture of the system (Presentation, Application, Data). The system's design was based around the CRISP-DM workflow (Wirth and Hipp, 2000).

System implementation and results were given in Chapter Four. The dataset was adapted from the UCI dataset (Realinho *et al*., 2022) and adapted it to the Nigerian's educational context and contained 3,630 students' records and 37 features (109 features after one-hot encoding). The five machine learning algorithms implemented and evaluated were Random Forest, Logistic Regression, Decision Tree, Gradient Boosting and Support Vector Machine. These results are summarized in Table 5.1.

Table 5.1: Summary of Model Performance Results

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	77.00%	62.39%	100.00%	76.84%
Logistic Regression	89.67%	79.19%	98.92%	87.96%
Decision Tree	94.21%	86.83%	100.00%	92.95%
Gradient Boosting	86.36%	73.80%	99.64%	84.79%
Support Vector Machine	76.45%	61.83%	100.00%	76.41%

Decision Tree Classifier was the top performing model giving us the highest accuracy of 94.21%, precision of 86.83% and the F1 Score of 92.95% along with perfect 100% recall value. The trained model was deployed into a web application which consists of 6 pages: Home, Single Prediction, Batch Prediction, Analytics Dashboard, Model Performance and About as a 6-page Streamlit application.

5.2 Conclusion

The conclusion derived from this study were:

First, a machine learning model can reliably predict student dropout risk in the University setting of Nigeria. Overall, the Decision Tree Classifier proved to be accurate at 94.21% and the F1-Score of 92.95% shows that prediction of student outcomes is possible and reliable with data. The perfect recall of 100% means the classifier is able to correctly identify every student at risk of dropping out, a crucial requirement for a successful prediction system.

Second, academic achievement measures and lifestyle factors are important predictors of student drop-out. The correlation analysis indicated that academic variables like 1st_Sem_Courses_Failed, 2nd_Sem_Courses_Passed and semester GPA are strong predictors. Other lifestyle items specific to the Nigeria student experience, such as Side_Hustle_Hours, Weekly_Gaming_Hours, Weekly_Clubbing_Frequency and Off_Campus_Nights, also had strong correlations with likelihood of leaving. This is a testament to the need to look at the whole student experience and not just grades.

Third, the Decision Tree algorithm was found to outperform more complex ensemble algorithms (Random Forest, Gradient Boosting) and the other algorithms (Logistic Regression, SVM) on this specific dataset. The decision boundaries in the data are well suited to the tree-based partitioning

and simpler models can sometimes outperform more complex models when the relationships between the features are naturally hierarchical.

Fourth, by deploying the model on the web via Streamlit, the interface is accessible and easy to use, enabling academic administrators without extensive technical knowledge of machine learning to use the model. It is simple to use for web forms, as well as for batch production based on uploading CSV files, and flexible enough for a wide variety of applications, including counselling individual students or screening a large population at the start of a semester.

Fifth, the ability to adapt the UCI dataset for use in the Nigerian educational context shows how far international datasets are able to be adapted to be used in developing countries where data collection may be limited by resources. The model was also enriched with Nigeria-specific variables like JAMB_Score, School_Fees_Cleared, and Monthly_Allowance_Bracket, adding Nigeria-specific context to the model.

Overall, the aim of this study to build a predictive analytics system for student dropout prediction has been successfully achieved. The system can serve as a useful data-driven instrument that can help inform student retention policies at Godfrey Okoye University and other universities in Nigeria.

5.3 Contributions of the Study

The following are the key contributions of this study to the fields of educational data mining and learning analytics within the Nigerian context:

1. A functional prediction system: Instead of just a theoretical model, this work developed a fully deployable web-based prediction system. With six fully functional pages (Home, Single Prediction, Batch Prediction, Dashboard, Model Performance, About) the Streamlit

web application covers the entire end-to-end pipeline, right from preprocessing the data, up to presenting user friendly predictions.

2. Comparative evaluation of various algorithms: This work presents an empirical comparison of five different machine learning algorithms (Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, Support Vector Machine) by training each model on the same dataset and using the same evaluation metrics, providing relevant information when it comes to choosing which algorithm might be suitable for similar educational prediction problems.
3. Data contextualization for the Nigerian context: By using the data obtained from the UCI Machine Learning Repository (Realinho *et al*., 2022) and including Nigeria-specific data, features that were added, for example, were the JAMB score, status of school fees clearance, categories of parental occupations in Nigeria, and students' lifestyles. This localized dataset provides a suitable baseline for other educational data mining projects in Nigerian universities.
4. Determination of Nigerian-specific predictors: This work also provided conclusive evidence that the time spent on side hustles, clubbing, and gaming are significant predictors of dropout for Nigerian students. These specific features in the Nigerian context were not explored much in previous works.
5. Contribution to the Local Literature: This research adds to the few existing literature in Nigeria Higher Education that focused on educational data mining, which could serve as guidelines for other scholars and institutions interested in deploying data-driven student support in higher education institutions.

5.4 Limitations of the Study

Although this project fulfills its specified aims the following limitations need to be taken into consideration:

1. Adapted dataset: The dataset used in this project is adapted from the UCI Machine Learning Repository rather than being obtained directly from student records at Godfrey Okoye University. While care has been taken to adapt the dataset to be reflective of the Nigerian educational environment the patterns and distributions of the adapted dataset might differ to those found in real institutional data. Validation on data obtained directly from an institution is necessary.

2. University-specific(Single Institution Focus): The system has been developed and tested based on data derived from Godfrey Okoye University. The patterns predicted by the models might be unique to this university and might not necessarily be generalizable to other Nigerian universities that might have different admission criteria, marking schemes, student populations and course structure. Validation across a number of universities would be necessary.

3. Not tested in a live environment: The system has been built and tested in a local development environment but not implemented or tested on live institutional data or with actual users of the system (academic administrators). User acceptance testing and feedback from end users of the system is crucial.

4. Non-optimization of hyperparameters: The machine learning models were not hyperparameter optimized. Standard or default parameters were employed. Grid search, random search, or bayesian optimization techniques could be employed in the future to achieve higher accuracy.

5. No temporal validation performed: No temporal validation was performed (training on one academic session and testing on a subsequent academic session) to evaluate how the models would perform at different times. Patterns could change from one session to the next.

6. Time and resource constraints: This project was a final-year undergraduate project and therefore resource and time constraints dictated the depth to which the problem could be studied. Further research in the field with more resources would allow the investigation of further algorithms, feature engineering techniques, and deployment.

5.5 Recommendations for Future Work

Based on the findings, conclusions, and limitations of this study, the following recommendations are made for future work:

5.5.1 Recommendations for the University

1. Obtain and Validate with Actual Data: Godfrey Okoye University should strive to validate the system by using real student data from the Computer Science department. This includes fetching past student data of those who graduated or failed out of the university, training the model with real student data and then evaluate its prediction accuracy on the local student population.

2. Integrate with the School Information System: A direct integration should be made to the school's information management system so as to enable automatic data flow and predictions. This way, the system will receive data directly, providing constant monitoring for the students' risks.

3. Create an Intervention plan: A plan should be created specifying interventions such as referrals for academic counseling, tutoring sessions, scholarship evaluations, or lifestyle counseling that should be put in place when a student is deemed to be at-risk.

4. Expand to Other departments: Once the system has been validated for the Computer Science department, Godfrey Okoye University should look into expanding its features and models to the other departments and faculties at the university.

5.5.2 Recommendations for Future Researchers

1. Future studies should use primary data directly from the Nigerian universities using the institutional database, surveys, and learning management systems. This would give more authentic data which would reflect the pattern and distribution that would be peculiar to the Nigerian educational context.

2. Future research should investigate the use of deep learning algorithms for prediction, including Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs), to see if the accuracy of the prediction can be increased beyond that of the Decision Tree in this study.

3. Future studies could take advantage of 'temporal and sequential' data, which is data from multiple semesters in students' performance, allowing for the identification of trends of student performance degradation instead of only snapshots of students' data.

4. Apply Explainable AI: Future research should consider the integration of explainable methods like Explainable AI (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to deliver clear explanations of individual predictions and boost trust and uptake with academic staff.

5. Create a mobile application of the prediction system – This would increase the ease with which academic administrators can access the prediction system and receive alerts on their mobile device, which would allow them to intervene more quickly and effectively.

Multi-Institutional Studies: Collaborative studies among various Nigerian universities would yield more robust and generalizable models which would enable the identification of both institution-specific and universal predictors of student dropouts.

REFERENCES

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon,* 5(2), e01250. https://doi.org/10.1016/j.heliyon.2019.e01250

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel,* 25(4), 297–308.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research,* 55(4), 485–540.

Booch, G., Rumbaugh, J., & Jacobson, I. (2005). *The Unified Modeling Language user guide* (2nd ed.). Addison-Wesley Professional.

Breiman, L. (2001). *Random forests. Machine Learning,* 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Flores, V., Heras, S., & Julian, V. (2022). Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. *Electronics,* 11(3), 457. https://doi.org/10.3390/electronics11030457

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics,* 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature,* 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Kurzweil, M., & Wu, D. (2015). *Building a pathway to student success* at Georgia State University. Ithaka S+R. https://doi.org/10.18665/sr.221043

Lee, S., Kim, J., & Park, H. (2023). Student dropout prediction for university with high precision and recall. *Applied Sciences,* 13(10), 6275. https://doi.org/10.3390/app13106275

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13(1), 5705. https://doi.org/10.1038/s41598-023-32484-w

McKinney, W. (2010). *Data structures for statistical computing in Python*. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 56–61).

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

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

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,* 17(18), 109–125. https://doi.org/10.3991/ijet.v17i18.25567

Osemwegie, E. E., & Amadin, F. I. (2023). Student dropout prediction using machine learning. *FUDMA Journal of Sciences*, 7(6), 347–353. https://doi.org/10.33003/fjs-2023-0706-2103

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12, 2825–2830.

Plotly Technologies Inc. (2015). *Collaborative data science.* Plotly Technologies Inc. https://plot.ly

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning,* 1(1), 81–106. https://doi.org/10.1007/BF00116251

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146. https://doi.org/10.3390/data7110146

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict students' dropout and academic success [Dataset]. *UCI Machine Learning Repository.* https://doi.org/10.24432/C5MC89

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, 2, 100018. https://doi.org/10.1016/j.caeai.2021.100018

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery,* 10(3), e1355. https://doi.org/10.1002/widm.1355

Segura, M., Mello, J., & Hernández, A. (2022). Machine learning prediction of university student dropout: Does preference play a key role? *Mathematics,* 10(18), 3359. https://doi.org/10.3390/math10183359

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. https://doi.org/10.1109/ACCESS.2022.3188767

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

Streamlit Inc. (2019). *Streamlit: The fastest way to build data apps*. https://streamlit.io

Tamada, M. M., Giusti, R., & de Magalhães Netto, J. F. (2022). Predicting students at risk of dropout in technical course using LMS logs. *Electronics,* 11(3), 468. https://doi.org/10.3390/electronics11030468

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). University of Chicago Press.

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society,* 76, 102474. https://doi.org/10.1016/j.techsoc.2024.102474

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

Waheed, H., Hassan, S. U., Nawaz, R., Aljohani, N. R., Chen, G., & Gasevic, D. (2022). Early prediction of learners at risk in self-paced education: A neural network approach. *Expert Systems with Applications,* 213, 118868. https://doi.org/10.1016/j.eswa.2022.118868

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39).

World Bank. (2021). *Education sector analysis: The Federal Republic of Nigeria. World Bank & UNESCO-IIEP.* https://www.iiep.unesco.org/en/publication/education-sector-analysis-federal-republic-nigeria

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments,* 9, 11. https://doi.org/10.1186/s40561-022-00192-z