PROACTIVE EDUCATIONAL MANAGEMENT: A RANDOM FOREST AND SHAP ANALYSIS FOR IDENTIFYING KEY PREDICTORS OF STUDENT PERFORMANCE.

by Clement Kwaku Boadu

Submission date: 27-Oct-2025 02:50PM (UTC+0000)
Submission ID: 2794427033
File name: Clement_Kwaku_Boadu_3.docx (548.14K)
Word count: 19621
Character count: 120712

UNIVERSITY OF CAPE COAST

SCHOOL OF ECONOMICS

DEPARTMENT OF DATA SCIENCE AND ECONOMIC POLICY

PROACTIVE EDUCATIONAL MANAGEMENT: A RANDOM FOREST AND SHAP ANALYSIS FOR IDENTIFYING KEY PREDICTORS OF STUDENT PERFORMANCE.

DMA 898: CAPSTONE PROJECT

CLEMENT KWAKU BOADU

SE/DMD/24/0021

NOVEMBER, 2025

UNIVERSITY OF CAPE COAST

PROACTIVE EDUCATIONAL MANAGEMENT: A RANDOM FOREST AND SHAP ANALYSIS FOR IDENTIFYING KEY PREDICTORS OF STUDENT PERFORMANCE.

BY

CLEMENT KWAKU BOADU

SE/DMD/24/0021

A CAPSTONE PROJECT SUBMITTED TO THE DEPARTMENT OF DATA SCIENCE AND ECONOMIC POLICY OF THE SCHOOL OF ECONOMICS, COLLEGE OF HUMANITIES AND LEGAL STUDIES, UNIVERSITY OF CAPE COAST, IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA MANAGEMENT AND ANALYSIS.

NOVEMBER, 2025

DECLARATION

Student's Declaration

I hereby declare that this capstone project is the result of my own original project and that no part of it has been presented for another degree in this University or elsewhere.

.

Student's Signature: …… Date: ……

Name: Clement Kwaku Boadu

Index Number: SE/DMD/24/0021

Supervisor's Declaration

I hereby declare that the preparation and declaration of this capstone project were supervised in accordance with the guidelines on supervision of capstone projects laid down by the University of Cape Coast.

Supervisor's Signature: …… Date: ……

Name: Professor Emmanuel Ekow Asmah

ABSTRACT

The reactive management of Ghana's Junior High School system, which uses student data for post-failure reporting rather than proactive solutions, exacerbates academic underperformance. This study bridges this gap by developing a machine learning model to predict at-risk students in the Ledzokuku Municipality. Using a Random Forest classifier on academic records of 459 students, the model demonstrated exceptional performance with 94% accuracy and 81% recall for the at-risk class. Crucially, feature importance analysis revealed a counter-intuitive insight: non-core subjects, specifically Religious and Moral Education (RME) and French, were the most significant predictors of academic risk, outperforming core subjects like Mathematics and English. This suggests these subjects act as proxies for foundational competencies essential for overall academic success. The study concludes that machine learning enables a crucial shift from reactive to proactive educational management, advocating for holistic interventions that target not only core subject knowledge but also the foundational competencies reflected in performance in non-core subjects like RME and French.

ACKNOWLEDGEMENTS

I would like to sincerely thank Dr. Carl Hope Korkpoe, for his support and assistance during this project. This work was greatly influenced by his expertise.

I owe a debt of gratitude to my lecturer, Dr Raymond E. Kofiinti, for his invaluable guidance, unwavering patience, and insightful feedback throughout this research process. My sincere thanks also go to the headteachers and staff of the schools in the Ledzokuku Municipality for granting me access to the necessary data. I am deeply grateful to my family for their unending support and encouragement, and to my friends and colleagues for their stimulating discussions and moral support. Finally, I acknowledge the University of Cape Coast for providing the platform and resources that made this capstone project possible.

DEDICATION

This work is dedicated to the students and educators of Ghana, in the hope that data-driven insights may light a path toward a more equitable and proactive educational future.

TABLE OF CONTENTS

Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Background of the Study

The mandate of the Ghana Education Service (GES) to provide equitable, quality pre-tertiary education is fundamentally challenged by persistent academic underperformance at the Junior High School (JHS) level. While significant progress has been made in improving access, the quality of learning outcomes remains critically low and uneven. This indicates a systemic failure that extends beyond pedagogy to the very core of educational management.

A central paradox lies at the heart of this crisis. The GES operates within a context of abundant data, generating vast quantities of student performance information through mechanisms like the Basic Education Certificate Examination (BECE) and continuous school-based assessments. However, the prevailing data culture is overwhelmingly retrospective. Analysis, where it exists, is primarily used for descriptive reporting documenting low pass rates, highlighting regional disparities, and identifying weakly performing subjects after the academic year has concluded (GES Education Sector Performance Report, 2019). This creates a cycle of reactive management where interventions are deployed as emergency measures only after students have already disengaged or failed, a practice that is both inefficient and ineffective in addressing the deep-seated drivers of underperformance.

This reactive paradigm is ill-suited to solve the complex, interconnected challenges such as structural resource inequities (Nugba et al., 2021), pedagogical shortcomings (Davis, Ntow, & Beccles, 2022), and socio-economic pressures

(Ghanney, 2020) that plague the system. The system's inability to leverage its own data for foresight means it is constantly treating symptoms rather than diagnosing and addressing underlying causes. For instance, a municipal-level analysis in Ledzokuku revealing that 86.9% of students scored below the pass mark is a stark outcome of this practice: a descriptive statistic that signals a catastrophe but provides no actionable intelligence to prevent its recurrence.

Globally, the fields of Educational Data Mining (EDM) and Learning Analytics (LA) have demonstrated a paradigm shift towards proactive, predictive management. Machine learning (ML) models transform historical data into predictive insights, enabling the early identification of at-risk students and allowing for timely, targeted interventions (Rosado, Payne, & Rebong, 2019; Adane, Deku, & Asare, 2023). This represents the crucial transition from being data-rich to being intelligence-driven.

Therefore, this study addresses a core "Problem of Practice" within Ghanaian educational management: the systemic inability to convert existing educational data into proactive intelligence. It posits that the solution is not solely in addressing surface-level symptoms but in fundamentally reforming the data-to-decision pipeline. By developing and implementing a machine learning-based early-warning system, this research challenges the reactive status quo and offers a framework for a more proactive and effective educational management system.

Statement of the Problem

Education forms the backbone of national development, yet the ongoing and systemic underperformance at the Junior High School (JHS) level in Ghana critically hampers students' progress and future employability. The Ghana Education Service (GES) is responsible for reversing this trend, but its approach to educational

management remains largely reactive. Although it systematically collects extensive student performance data, the GES's analytical approach remains mainly descriptive, focusing on reporting past failures instead of preventing future ones. This "data-rich but information-poor" paradox results in interventions being implemented only after students have already failed, leading to inefficient use of resources and continuing cycles of underperformance.

The global education sector has seen a transformative shift through Educational Data Mining (EDM) and Machine Learning (ML), which allow for the early identification of at-risk students and facilitate proactive support. In contrast, the operational framework of Ghana's Junior High Schools has not developed these predictive, data-driven capabilities.

Purpose of the study

To transform educational management in Ghana from a reactive to a proactive paradigm by developing and implementing a data-driven, machine learning-based framework for the early identification of at-risk Junior High School students, and to uncover the key subject-level predictors of academic risk to enable targeted, preemptive interventions.

Research Objectives

The main objective of this study is to use machine learning methods to analyze, predict and manage academic risk among Junior High School (JHS) students in Ledzokuku Municipality.

The specific objectives are to:

1. Develop a machine learning model to predict students who are likely to score below the pass mark of 50%.
2. Identify the subjects that are the most significant predictors of a student being at-risk.

Research Questions

1. To what extent can a machine learning model accurately predict Junior High School students in the Ledzokuku Municipality who are at risk of scoring below the 50% pass mark?
2. Which subjects contribute most significantly to the prediction of poor academic performance among students in the Ledzokuku Municipality?

Significance of the Study

This study holds significant implications for educational practice, policy, and research. By developing a data-driven framework for proactive intervention, it provides tangible benefits for key stakeholders and makes a distinct contribution to the academic field.

Practical Significance for Practitioners and Policymakers

- For Teachers and School Leaders: This study provides a practical tool, a validated machine learning model that can identify at-risk students early in the academic cycle. This enables educators to move beyond intuition and move towards targeted, timely interventions, allowing them to allocate their limited time and resources more efficiently and effectively to support the students who need it most.

- For Policymakers (Ghana Education Service & Ministry of Education): The findings offer evidence-based, empirical insights into the complex drivers of student performance. The feature importance analysis reveals not just which subjects are hardest, but which are most predictive of overall risk. This can guide the design of more effective national policies, curriculum adjustments, and remedial programs. Furthermore, the proposed Early Warning System (EWS) framework presents a scalable model for transforming the GES's data management paradigm from descriptive reporting to proactive student support.

Contribution to Knowledge

This Capstone Project bridges a critical gap in the literature by demonstrating how Educational Data Mining (EDM) can be effectively applied to address specific challenges in Ghanaian educational management. Its contribution is threefold: it innovates by adapting EDM techniques to a new context, ensures reproducibility by providing a clear methodological blueprint, and delivers tangible impact by translating data patterns into strategic insights for local school administrators.

1. Methodological Contribution: It bridges the methodological gap by demonstrating the successful application and optimisation of a Random Forest classifier for predicting academic risk within the specific, resource-constrained context of Ghanaian JHS. This provides a validated methodological blueprint for using widely available academic data for predictive analytics, moving beyond the dominant descriptive and qualitative research traditions in the local literature.

2. Theoretical and Explanatory Contribution: It addresses the explanatory gap by integrating Explainable AI (XAI) techniques, specifically feature importance and SHAP analysis, to open the "black box" of the predictive model.

3. Practical and Translational Contribution: It tackles the practical gap by synthesizing the methodological and explanatory outputs into a coherent, proposed Early Warning System (EWS) framework. This framework is not merely a model but a transferable operational strategy for integrating predictive analytics into the GES's decision-making routines, thereby providing a clear pathway for translating data-driven insights into actionable interventions.

By simultaneously fulfilling these three contributions, this research expands the knowledge base on EDM in developing countries and provides a scalable, holistic framework for proactive educational management that can be adapted and tested in similar contexts beyond the Ledzokuku Municipality.

Scope of the Study

This study is deliberately bounded in its scope to enable a focused and feasible investigation. The research focuses exclusively on Junior High School students within the Ledzokuku Municipality, Ghana. The primary data source is structured academic performance records from one academic year, encompassing scores in core subjects (English, Mathematics, Science) and other subjects (e.g., Social Studies, RME, French) for a population of 459 students.

Critically, this scope represents a conscious epistemological and methodological choice. The study adopts a positivist research paradigm, which posits that social phenomena can be studied through objective methods and that knowledge can be derived from empirical, observable data. Consequently, the research is

intentionally restricted to quantifiable academic metrics, subject scores and their engineered derivatives (e.g., STEM Score, Core_Avg). This approach allows for the application of statistical and machine learning techniques to identify patterns and make predictions with a high degree of objectivity and reliability.

By design, this scope excludes non-quantifiable factors such as socio-economic background, teacher motivation, school climate, and student psychological states. While these qualitative elements are undeniably influential, their exclusion is a strategic decision to first establish the predictive power of the most readily available and standardised data within the Ghanaian educational system. This establishes a foundational, data-driven model that can be systematically tested and validated, providing a clear benchmark for future research.

Delimitation of the Study

This study was deliberately bounded by the following parameters to ensure a focused, feasible, and methodologically coherent investigation:

1. Geographical Scope:

Area Covered: The research was conducted exclusively within the Ledzokuku Municipality in the Greater Accra Region of Ghana.

Area Excluded: The findings are not generalised to other municipalities, districts, or regions in Ghana, particularly those with significantly different demographic, economic, or educational contexts (e.g., northern or deeply rural areas).

2. Population and Sample Scope:

Population Covered: The study focused solely on Public Junior High School (JHS) students.

Population Excluded: Students from Private Junior High Schools, as well as all students at the Primary and Senior High School levels, were excluded from the study's purview.

3. Temporal Scope:

Period Covered: The analysis was confined to a single academic year, providing a cross-sectional snapshot of student performance.

Period Excluded: The study did not employ a longitudinal design and therefore does not account for student performance trends, the long-term impact of interventions, or developmental trajectories over multiple years.

4. Data and Methodological Scope:

Data Type Included: The research adopted a positivist, quantitative paradigm. The analysis relied entirely on structured, quantifiable academic data, specifically, student scores in examinable subjects.

Data Type Excluded: The study excluded all qualitative data. Factors such as socio-economic status, teacher motivation, school climate, parental involvement, student psychological states, attendance records, and personal experiences were not included. This was a conscious epistemological choice to establish the predictive power of the most readily available and standardised data within the system.

5. Variable Scope:

Variables Included: Independent Variables: Raw scores from all core and other subjects, including English, Mathematics, Science, Social Studies, Computing, Ghanaian Language, Religious and Moral Education (RME), Creative Arts, and French.

Derived Variables: Engineered composite features, specifically the STEM Score (average of Mathematics, Science, and Computing) and Core_Avg (average of English, Mathematics, and Science).

Dependent Variable: A binary "At-Risk" status, defined as scoring below 50% in at least one of the three core subjects (English, Mathematics, or Science).

Variables Excluded: As a consequence of the methodological scope, variables related to student demographics, home environment, school resources, teacher qualifications, and behavioural indicators (e.g., attendance, punctuality) were explicitly excluded from the model.

In summary, these delimitations represent strategic choices to create a bounded and replicable study that leverages objective, available data to build a foundational predictive model, while acknowledging that this scope necessarily omits the rich, contextual factors that also influence student outcomes.

Limitations of the Study

The acknowledged boundaries of this study give rise to several limitations, which are inherent to its chosen paradigm and scope:

1. Limited Generalizability: The model is trained and validated on data from a single municipality. The demographic, economic, and educational context of Ledzokuku may not be representative of all districts in Ghana, particularly those in the northern regions or deeply rural areas. Therefore, the direct application of the predictive model to other contexts should be undertaken with caution and preceded by local validation.

2. Exclusion of Qualitative Context: The positivist, quantitative approach, while powerful for prediction, necessarily omits the rich, contextual understanding that qualitative data provides. The model can identify that a student is at risk and which subjects are influential, but it cannot explain the underlying why from a human perspective, such as home life challenges, specific teacher-student dynamics, or personal motivational issues. This limitation underscores that the model's outputs are a starting point for human inquiry, not a complete diagnostic tool.

3. Cross-Sectional Temporal Frame: The analysis is confined to a single academic year. This cross-sectional design captures a snapshot of student performance and cannot account for longitudinal trends, the sustained impact of interventions, or the developmental trajectory of individual students. A longitudinal study would be required to understand how risk factors evolve.

These limitations are not weaknesses but rather defined boundaries that clarify the specific contribution of this research. They highlight the trade-off between quantitative precision and qualitative depth, and they clearly chart a course for future studies to build upon this work by integrating mixed-methods approaches and broader datasets.

Definition of Terms

A. **Academic and Contextual Terms**

1. **At-Risk Student:** For this study, a student is classified as "at-risk" if they score below the *50% pass mark* in at least one of the three core subjects: English, Mathematics, or Science. This is the binary target variable for the machine learning model.

2. **Basic Education Certificate Examination (BECE):** The national examination taken by Junior High School students in Ghana at the end of their studies, which determines their placement into Senior High Schools.

3. **Core Subjects:** The foundational subjects in the Ghanaian JHS curriculum, considered essential for academic progression. In this study, they are specifically English, Mathematics, and Science.

4. **Ghana Education Service (GES):** The government agency responsible for implementing pre-tertiary educational policy and ensuring the provision of equitable, quality education in Ghana.

5. **Junior High School (JHS):** The second and final stage of basic education in Ghana, typically covering students in grades 7 to 9 (Forms 1 to 3).

6. **Ledzokuku Municipality:** A specific municipal district in the Greater Accra Region of Ghana, which serves as the geographical boundary and case study for this research.

7. **Non-Core Subjects/Other Subjects:** Subjects in the JHS curriculum other than the core subjects. In this study, they include Social Studies, Religious and Moral Education (RME), French, Ghanaian Language, Creative Arts, Career Technology, and Computing.

8. **Proactive Educational Management:** An approach that uses predictive data and intelligence to identify and address potential problems (like student failure) before they occur, enabling timely interventions. This is contrasted with the current reactive system.

9. **Reactive Educational Management:** The prevailing approach in the studied context, where data is used primarily for descriptive reporting on past failures, leading to interventions that are deployed only after students have already failed.

B. Machine Learning and Data Science Terms

1. **Algorithm:** A step-by-step procedure or formula for solving a problem. In this context, it refers to the machine learning algorithms (like Random Forest) used for prediction.
2. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** A performance measurement for classification problems that shows the model's ability to distinguish between classes (At-Risk vs. Not-At-Risk). A score of 1.0 represents a perfect model, and 0.5 represents a worthless model.
3. **Confusion Matrix:** A table used to describe the performance of a classification model, showing the counts of True Positives, True Negatives, False Positives, and False Negatives.
4. **Convenience Sampling:** A non-probability sampling technique where the sample is taken from a group of people easy to contact or reach. In this study, it refers to using the most readily accessible academic records from the Ledzokuku Municipality.
5. **Data Preprocessing:** The technique of cleaning, transforming, and organising raw data into a clean, usable format for machine learning. This included standardising column names, removing empty columns, and validating score ranges.
6. **Dependent Variable (Target Variable):** The output or outcome that the model is trying to predict. In this study, it is the "At-Risk" status of the student.

7. **Feature Engineering:** The process of creating new input variables (features) from existing data to improve model performance. In this study, STEM Score and Core_Avg were engineered features.
8. **Feature Importance:** A technique that assigns a score to input features based on how useful they were in predicting a target variable within a model. It was used to identify which subjects (like RME and French) were the strongest predictors.
9. **Independent Variables (Features):** The input variables or predictors used by the model to make a prediction. In this study, these are the scores from all individual subjects and the derived variables (STEM Score, Core_Avg).
10. **Machine Learning (ML):** A subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed, by building models from sample data.
11. **Model Interpretability:** The degree to which a human can understand the cause of a decision made by a model. This study used Feature Importance and SHAP to achieve this.
12. **Positivist Research Philosophy:** The philosophical stance underpinning this study, which posits that reality is stable and can be observed objectively through scientific methods, supports the exclusive use of quantitative data.
13. **Random Forest (Classifier):** An ensemble machine learning algorithm used for classification. It operates by constructing a multitude of decision trees during training and outputting the mode of the classes (for classification) of the individual trees. It was selected for its robustness, accuracy, and native feature importance metrics.

14. **SHAP (SHapley Additive exPlanations):** A game theory-based approach used to explain the output of any machine learning model. It quantifies the contribution of each feature to the final prediction for an individual student, providing both global and local interpretability.

15. **Stratified Sampling:** A method of sampling that divides the population into homogeneous subgroups (strata) and then draws a random sample from each stratum. This was used during the train-test split to ensure the proportion of at-risk students was the same in both sets.

C. Performance Evaluation Metrics

1. **Accuracy:** The ratio of correctly predicted observations (both At-Risk and Not-At-Risk) to the total observations.
2. **F1-Score:** The weighted harmonic mean of Precision and Recall. It tries to find a balance between the two metrics.
3. **Precision:** The ratio of correctly predicted positive observations (True Positives) to the total predicted positives. It answers: "When the model predicts 'At-Risk', how often is it correct?"
4. **Recall (Sensitivity):** The ratio of correctly predicted positive observations (True Positives) to all actual positives in the data. It answers: "Of all the students who are truly 'At-Risk', how many did the model find?" This was a critical metric for the study's early-warning goal.

D. Theoretical Frameworks

1. **Educational Data Mining (EDM):** A field of research concerned with developing methods for exploring unique types of data that come from educational settings and

using those methods to better understand students and the settings they learn in. It is method-centric and positivist.

2. **Early Warning System (EWS):** A structured process that uses data to identify students at risk of academic failure and triggers timely interventions to keep them on track.
3. **Explainable AI (XAI):** A set of tools and frameworks to help human users understand and trust the outputs of machine learning models. This study used it to move from a "black box" to actionable insight.
4. **Learning Analytics (LA):** The measurement, collection, analysis, and reporting of data about learners and their contexts, for understanding and optimising learning and the environments in which it occurs. It is more human-centric and interpretivist than EDM.

Organisation of the Study

This study is organised into five chapters. Chapter One introduces the study, providing the background, problem statement, objectives, research questions, significance, scope, and limitations. Chapter Two presents the literature review, which examines existing research on student academic performance, factors influencing academic performance in Ghanaian schools, and the application of machine learning and educational data mining in predicting student achievement. Chapter Three outlines the research methodology, including the research design, data sources, machine learning techniques, evaluation metrics, and ethical considerations. Chapter Four presents the analysis and results, including the development and evaluation of predictive models, identification of key subject-level contributors, and the design of the early-warning system. Finally, Chapter Five discusses the findings, draws conclusions, and provides recommendations for educators, policymakers, and future research.

CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter critically reviews the existing research relevant to predicting student academic performance. The aim is not just to list previous studies, but to build a narrative that places this research within a global technological movement, contrasts it with the dominant methods in the local setting, and highlights a specific, multi-dimensional gap that this Capstone project seeks to fill. The review is organized to move from the broad use of machine learning (ML) in education to the particular challenges of the Ghanaian Junior High School (JHS) system, then to the core methods of predictive algorithms, and ultimately to the vital need for model interpretability. This leads to a clear statement of the research gap that provides the justification for the current study.

The Global Paradigm: Machine Learning for Educational Prediction

Globally, the education sector is experiencing a data-driven shift, moving beyond traditional statistical analysis to utilise machine learning for predicting student performance. The idea is that patterns hidden within educational data can be extracted to predict performance, allowing for proactive measures. Research in this field consistently shows that ML models outperform traditional regression-based methods.

For instance, systematic reviews by Radhya et al. (2022) have identified algorithms such as Random Forest, Naïve Bayes, and Support Vector Machines (SVM) as consistently high-performing classifiers for this task. Empirical evidence strongly supports this: Khan et al. (2023) achieved 93.74% accuracy in predicting

secondary school outcomes, while Tiwari and Jain (2024) found that ensemble methods and neural networks significantly outperformed simpler models. The application is also diverse, ranging from predicting standardised test scores.

However, a thorough review of this global literature uncovers a tendency to prioritise predictive accuracy above all else, often within data-rich environments. For example, while Kanabar and Tawde (2025) highlight Random Forest's accuracy, their study was carried out in a setting with extensive feature sets, including behavioural and socio-economic data. This raises an important question about the transferability of such models to resource-constrained settings like Ghanaian public schools, where data is often limited to academic transcripts. This global perspective demonstrates the potential of ML, but its effective local use depends on careful methodological adaptation and consideration.

Machine Learning in the Ghanaian Educational Context

While the global literature is rich with ML applications, its adoption within the Ghanaian educational research landscape remains nascent. A critical appraisal of the emerging literature reveals a focus that, while valuable, leaves significant gaps that this study aims to fill.

A prominent example is the work of Adane, Deku, and Asare (2023), who successfully demonstrated that machine learning algorithms, including Random Forest, could predict student academic performance in Ghana with accuracy exceeding 90%. Their study validated the fundamental premise that ML is transferable to the Ghanaian context. However, their methodological approach shared the limitations of much global EDM: a primary emphasis on predictive accuracy. The study utilised a broad set of features, including demographic and socio-economic data, and while it confirmed the

predictive power of ML, it did not deeply investigate the explanatory factors behind the predictions. The model remained largely a "black box," offering a risk score without clarifying the specific, actionable academic drivers of that risk within the Ghanaian curriculum.

This is where the present study makes a distinct and critical advancement. While building on the foundational work of proving ML's efficacy in Ghana, this research moves beyond mere prediction to provide a diagnostic explanation. First, it deliberately restricts its feature set to routinely collected academic subject scores, making it more scalable and directly applicable to the data assets already held by the Ghana Education Service, without reliance on harder-to-collect demographic data. Second, and most significantly, it integrates Explainable AI (XAI) techniques, specifically SHAP analysis, to open the "black box." This allows not just for prediction, but for a subject-level paradigm shift in understanding risk of failure.

The Ghanaian Context: A Dominance of Traditional Inquiry

Current research on student academic outcomes in the Ghanaian Junior High School system demonstrates a significant methodological divergence from international norms. Whereas the global field has moved toward computational analytics, the Ghanaian context remains predominantly reliant on established qualitative methodologies, utilising surveys, interviews, and descriptive statistics to meticulously catalogue the determinants of underperformance.

This research has been invaluable in highlighting critical challenges. Studies consistently point to teacher-related factors such as incomplete syllabus coverage and pedagogical weaknesses (Davis, Ntow, & Beccles, 2022), socio-economic pressures like illegal mining disrupting student focus (Ghanney, 2020; Adu-Gyamfi,

2014), and systemic issues of resource allocation and school supervision (Emilio, 2020). Furthermore, subject-specific analyses have revealed the acute struggles in Mathematics and Science, often attributing them to weak study habits and ineffective teaching strategies (Mensah et al., 2022).

While these studies provide essential contextual understanding, they are inherently retrospective and diagnostic. They excel at explaining why failure occurred after the fact but offer limited capacity for proactive prediction. They identify the symptoms of the educational crisis but lack tools for early diagnosis. This creates a critical disconnect: the Ghana Education Service collects vast amounts of quantitative performance data, yet the dominant research paradigm remains qualitative and correlational, failing to leverage this data for predictive insights. This dichotomy highlights a significant methodological lag in the local research landscape.

The Methodological Core: A Comparative Lens on Predictive Algorithms

Connecting the global potential with local needs requires a deliberate choice of methodology. Among the many ML algorithms, ensemble methods like Random Forest have become especially powerful for educational data mining. Its popularity is based on several key advantages relevant to educational datasets: inherent resistance to overfitting through bagging, the ability to model complex non-linear relationships, and robustness to outliers and missing data (Breiman, 2001).

Comparative studies often confirm its strength. VijayAnand et al. (2023) demonstrated that SVM could outperform Logistic Regression and K-Nearest Neighbours, while Kadu et al. (2024) showed that models incorporating extracurricular features improved predictions. However, Random Forest frequently maintains a favourable balance. Unlike SVM, which can be sensitive to parameter tuning and less

interpretable, and unlike Neural Networks, which are often "black boxes" requiring large datasets, Random Forest provides a reliable 'out-of-the-box' performance along with the benefit of native feature importance metrics.

This final point is essential. Choosing an algorithm involves more than just accuracy; it also depends on how well it fits the research goals. For this study, which seeks not only to predict but also to explain, Random Forest's ability to provide a clear, understandable ranking of predictor variables makes it a better choice than a more complex yet opaque deep learning model, particularly for an introductory application in the Ghanaian JHS context.

The Imperative of Explainability: From Black Box to Actionable Insight

The pursuit of accuracy must be balanced with the need for interpretability, especially in a field as human-centric as education. This has increased the importance of Explainable AI (XAI). A highly accurate model that cannot explain its reasoning is of little practical use to a teacher or principal; it is a "black box" that commands trust without offering understanding.

Techniques such as SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been developed to bridge this gap. They answer the critical "why?" by quantifying the contribution of each feature to an individual prediction. The literature shows a growing recognition of this need. Chen et al. (2025) used feature selection with Random Forest to minimise redundancy and enhance interpretability, while Balcıoğlu and Artar (2023) used ensemble models to rank predictors and identify "close-to-fail" students. Sixhaxa et al. (2022) further demonstrated how XAI techniques can pinpoint specific features from subject grades to behavioural factors that drive outcomes.

The application of XAI transforms a predictive system from simply alerting to providing a diagnostic tool. It empowers educators by offering not only a risk score but also a reasoned explanation for that score, which is crucial for designing targeted and effective interventions. This transition is a fundamental aspect of this research, directly foreshadowing the second objective of pinpointing key subject-level contributors.

Theoretical Framework: Navigating the Tensions in a Data-Driven Educational Intervention Model

The theoretical foundation of this study is not a simple, harmonious integration of related concepts, but a critical navigation of the fertile tensions between three distinct fields: Educational Data Mining (EDM), Learning Analytics (LA), and Early Warning Systems (EWS). This research is situated at their intersection, consciously leveraging the strengths of each while explicitly acknowledging and addressing their inherent philosophical and practical conflicts. The framework posits that a responsible and effective application of data-driven methods in education requires not just technical integration but a careful balancing of positivist prediction, human-centric interpretation, and ethical intervention. ML algorithms are applied to educational data to solve LA problems.

Educational Data Mining (EDM): The Positivist Engine and its Interpretive Limits

EDM is fundamentally concerned with developing computational methods for discovering patterns in large educational datasets (Baker & Inventado, 2014). Its orientation is method-centric and grounded in a positivist philosophy, seeking objective, generalizable knowledge through automated pattern recognition from quantitative data.

This study leverages EDM's predictive power, employing the Random Forest algorithm for its robustness and accuracy. However, we critically engage with a core limitation of EDM: its propensity to produce "black box" models. A purely EDM-centric approach risks reducing students to data points, prioritising predictive accuracy over pedagogical understanding, and decontextualising learning from its socio-cultural environment (Siemens & Baker, 2012). This creates a tension between the objective, quantitative predictions of EDM and the subjective, qualitative reality of the classroom.

This study does not accept EDM's outputs as final. It deliberately moves beyond a pure discovery paradigm by incorporating model interpretability as a non-negotiable requirement. The use of feature importance and SHAP analysis is a direct response to this critique, serving as a bridge to make the positivist outputs of EDM intelligible and actionable within a human context.

Learning Analytics (LA): The Human-Centric purpose and its ethical imperative

LA is often distinguished from EDM by its stronger emphasis on using data analysis to directly understand and optimise learning and the environments in which it occurs (Siemens & Long, 2011). Its orientation is more human-centric and interpretivist, focusing on informing and improving human decision-making to support the learner.

While EDM asks, "What will happen?", LA compels us to ask, "So what?" and "What should we do now?". This study is grounded in the LA paradigm, framing the entire research around improving outcomes for students and supporting educators. However, a significant tension arises here: the tools of EDM (positivist) are being used to serve the goals of LA (interpretivist). This necessitates a translation layer where quantitative predictions are enriched with qualitative meaning. LA brings with it a

critical ethical imperative that EDM often neglects. This study explicitly addresses the ethical concerns inherent in LA and EWS.

Early Warning Systems (EWS)

EWS provide a structured process for using data to identify students at risk and triggering timely interventions (O'Cummings & Therriault, 2015). An EWS is an application framework that operationalises data-driven insights. This study adopts the EWS framework to bridge the chasm between insight and action. However, it critically engages with a key risk: that an EWS can devolve into a mechanistic, impersonal alert system. A trigger from a model is useless or even harmful if it does not lead to an appropriate, compassionate, and effective intervention.

To address this, the study contributes to a more sophisticated EWS model. It moves beyond generic alerts to diagnostic warnings. For example, instead of "Student A is at risk," the system proposes an alert like, "Student A is at risk, primarily driven by declining performance, suggesting potential issues with discipline or cognitive engagement." This provides educators with a starting point for a conversation and a more targeted intervention, thereby humanising the data-driven trigger.

An Integrated, Critically-Aware Model for Proactive Intervention

The theoretical model of this study is a dynamically balanced, critically-aware pipeline:

[Raw Data] => [EDM: "Pattern Discovery & Prediction"] => [Interpretability Bridge] => [LA: "Human Understanding & Ethical Scrutiny"] => [Diagnostic EWS: "Targeted, Ethical Intervention"] => [Stakeholders]

Feedback Loop: Crucially, this model includes a feedback loop from stakeholders (teachers, students) back to the data and model, allowing for continuous refinement and contextualization, ensuring the system remains a servant to educational goals, not a master.

This framework demonstrates that the study is not a naive application of technology. It is a thoughtful, critical implementation that:

- **Leverages** EDM's predictive power.
- **Subjects** EDM's outputs to LA's human-centric and ethical scrutiny.
- **Channels** the resulting intelligence through a diagnostic EWS designed to empower, not replace, educators.

By explicitly navigating these tensions, the research establishes a robust, defensible, and scholarly foundation for its methodology and anticipated impact.

Empirical Review

This section reviews empirical studies aligned with the specific objectives of this research. It is organised thematically to first examine the evidence for machine learning's predictive accuracy in education, and then to explore the critical role of non-cognitive skills and proxy variables, thereby providing a focused foundation for the current study.

Predictive Accuracy of ML Models in Education

A substantial body of international research demonstrates the superior predictive power of machine learning (ML) models over traditional statistical methods for forecasting student outcomes. The consensus from systematic reviews is that algorithms such as Random Forest, Naïve Bayes, and Support Vector Machines

(SVM) consistently achieve high performance in classifying at-risk students (Radhya et al., 2022).

Empirical evidence strongly supports this. For instance, Khan et al. (2022) achieved 93.74% accuracy in predicting secondary school outcomes, while Tiwari and Jain (2024) found that ensemble methods and neural networks significantly outperformed simpler models. This performance is robust across diverse educational contexts: Vijayalakshmi and Venkatachalapathy (2019) used deep neural networks to predict standardised test scores with 84% accuracy, and Kahandal et al. (2024) extended these techniques to predict teacher performance, highlighting the broad utility of ML.

Comparative studies further refine these findings. VijayAnand et al. (2023) demonstrated that SVM outperformed Logistic Regression and K-Nearest Neighbours, and Kadu et al. (2024) showed that models integrating extracurricular features alongside academic data yielded even better predictions. The work of Kanabar and Tawde (2025) specifically champions the Random Forest algorithm for its high accuracy in student performance prediction.

Synthesis and Local Gap: Collectively, these studies establish that ML models, particularly Random Forest, can reliably predict academic risk with accuracy often exceeding 90%. However, a critical methodological gap persists in the Ghanaian JHS context. While these sophisticated models are deployed globally, their application within Ghana's resource-constrained educational system, using the data the GES already collects, remains critically underexplored. This study directly addresses this gap by implementing and validating a Random Forest model within this specific local context.

The Role of Non-Cognitive Skills and Proxy Variables in Prediction

Beyond raw predictive accuracy, a more nuanced strand of research emphasises the importance of understanding what drives predictions, often revealing that non-cognitive skills and proxy variables are paramount. These are latent traits such as discipline, conscientiousness, and cognitive flexibility that are not directly measured but are inferred through related metrics.

Empirical studies consistently find that socio-academic and behavioural factors are highly predictive. Sharma and Maurya (2021) identified socio-economic background and family environment as significant influencers of performance. Similarly, Ahmed (2024) found that attendance and prior grades were dominant features in a high-accuracy SVM model. Hussain (2015) demonstrated that student engagement and behavioural data were the most important variables in a Logistic Regression model, achieving 99% accuracy.

The critical insight for this study is how these latent traits are captured. In data-rich environments, they can be measured directly. However, in contexts like Ghana, they must be identified through proxy variables, readily available data points that serve as indicators. For example:

- Hui (2024) used feature selection to identify top predictors of Mathematics achievement.
- Sixhaxa et al. (2022) reported that behavioural factors strongly correlated with exam performance.
- Balcıoğlu and Artar (2023) used ensemble models to rank predictors and identify "close-to-fail" students, moving towards diagnostic insights.

- Chen et al. (2025) explicitly used feature selection with Random Forest to minimise redundancy and identify the most critical predictors, a methodology that aligns with the explanatory goal of the present research.

This literature confirms that academic risk is not solely a function of intellectual ability in core subjects but is deeply influenced by foundational non-cognitive and cognitive skills. The prevailing practice in EDM is to use direct measures (attendance, engagement surveys). However, this creates a significant explanatory gap in the Ghanaian context: there is a conspicuous absence of research that investigates which standard academic subjects could act as the most effective proxies for these latent skills.

While global EDM often uses direct measures of behaviour and socio-economics, and Ghanaian research focuses on qualitative causes, there is a lack of research investigating which routinely collected academic subject scores can serve as the most effective proxies for these underlying risk factors in resource-constrained contexts

Conceptual Framework

The framework of this capstone project is a data-driven and Machine learning for Proactive Educational Intervention. It integrates several sub-frameworks into a cohesive pipeline, from data collection to actionable policy recommendations.

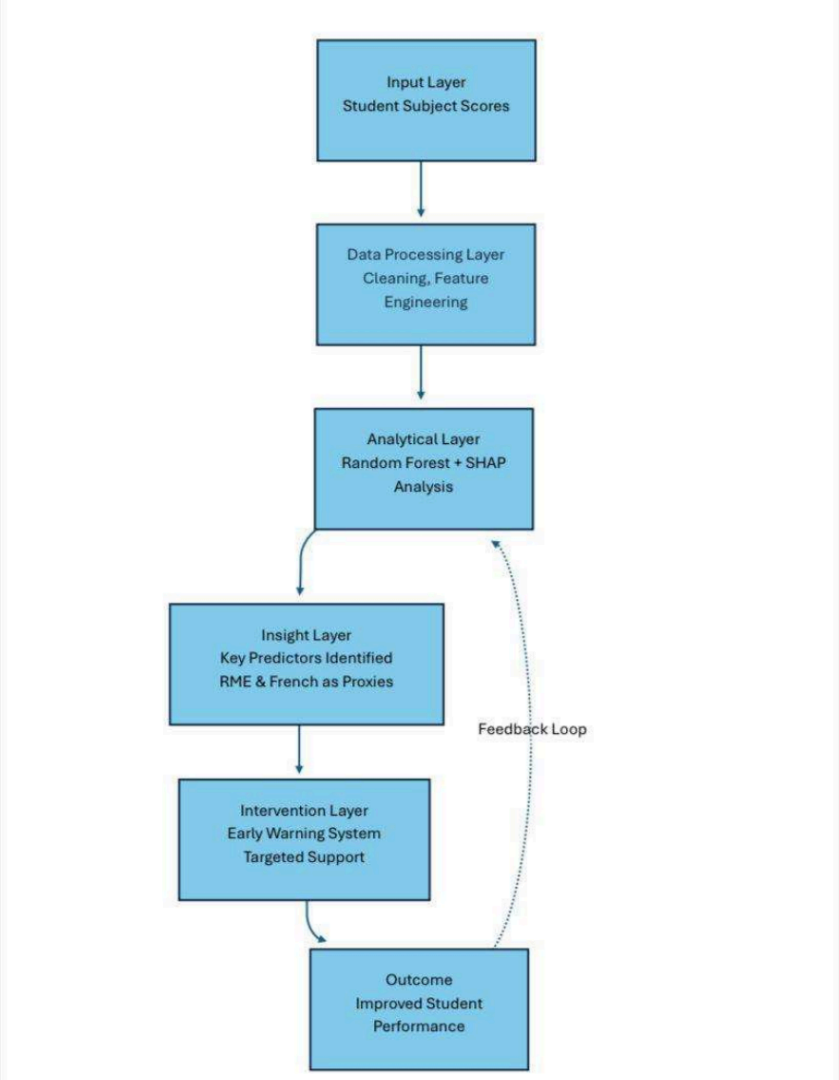

Figure 1: conceptual framework

Phase 1: The Conceptual & Scoping Foundation

This phase establishes the "why" and "what" of the study, grounded in the identified problem.

- **Problem Identification:** The framework starts with the recognition of a critical problem: the reactive nature of the current educational system in Ghana, where interventions occur only after students have failed.
- **Literature Review & Gap Analysis:** It situates itself within global best practices (Educational Data Mining, Learning Analytics) but identifies a specific gap in the local context: the underutilization of predictive ML models and explainable AI in Ghanaian JHS.
- **Objective Definition:** The goals are precisely scoped to be predictive and explanatory:
    1. Predict: Develop a model to identify at-risk students.
    2. Explain: Identify the key subjects contributing to risk.

Phase 2: The Data & Model Pipeline

This is the technical core of the framework, transforming raw data into a predictive model.

- **Data Acquisition & Preprocessing:** The framework uses readily available academic data (subject scores), making it practical and scalable. Preprocessing ensures data quality.

- **Feature Engineering:** The creation of derived variables (STEM Score, Core_Avg) demonstrates a sophisticated approach to capturing latent student abilities from existing data.
- **Predictive Modelling with Random Forest:** The model selection is justified by the algorithm's known strengths: handling non-linear relationships, robustness against overfitting, and providing feature importance.

Phase 3: The Analytic & Interpretation Layer

This phase moves beyond a "black box" model to generate actionable insights.

- **Model Evaluation:** The model is validated using a standard set of metrics (Accuracy, Precision, Recall, F1-Score, AUC-ROC). The high recall is particularly noted as crucial for an effective *early warning* system.
- **Explainable AI (XAI) & Insight Generation:** This is a critical component. The use of Feature Importance Analysis is not just for model diagnostics but is the primary tool for achieving the second research objective. It reveals the counter-intuitive finding that non-core subjects (RME, French) are the strongest predictors, interpreting them as proxies for latent traits like discipline and cognitive skills.

Phase 4: The Action & Integration Framework

This phase translates data insights into real-world actions and systemic change.

- **Proactive Intervention Strategy:** The framework advocates for a paradigm shift from reactive to proactive support. Interventions are targeted based on the model's predictions and its explanatory insights (e.g., improving not just Math, but also focusing on the skills underpinning performance in RME and French).

- **Policy Integration & Early Warning System (EWS):** The ultimate output is a proposal for an institutionalized, data-driven early warning system. This involves creating teacher-friendly dashboards and mandating the systematic use of data for decision-making within the Ghana Education Service (GES).

Overarching Theoretical Framework

The entire project is undergirded by three interconnected theoretical frameworks from educational technology:

1. **Educational Data Mining (EDM):** Provides the *methodological toolkit* for discovering patterns in educational data.
2. **Learning Analytics (LA):** Provides the *overarching purpose*: to understand and optimize learning and the environments where it occurs.
3. **Early Warning Systems (EWS):** Provides the *practical application model* for using data to identify at-risk students and trigger supports.

In summary, the framework is not just a machine learning model; it is an end-to-end blueprint for using data to drive a more equitable, proactive, and effective educational management system within the specific context of Ghanaian Junior High Schools.

Research Gap

In Ghana, most studies on student performance have emphasised socio-economic conditions, teacher effectiveness, and school resources as the main determinants of achievement (Davis et al., 2022; Mensah et al., 2020). While these works provide important context, they do not employ advanced analytical methods such as machine learning to predict outcomes or generate actionable insights for

intervention. Globally, predictive models such as Random Forest, Naïve Bayes, and Support Vector Machines have consistently achieved high accuracy in forecasting academic performance (Kanabar & Tawde, 2025; Tiwari & Jain, 2024), yet these approaches remain underutilised at the Junior High School level in Ghana. Moreover, research in the Ghanaian context rarely explores subject-level contributions using explainable AI methods like SHAP or LIME, which could provide deeper insights into which subjects drive success or failure.

Summary

This chapter has reviewed literature on student performance prediction, emphasising three key areas: machine learning models for predicting academic outcomes, subject-level contributions to student achievement, and the development of early-warning systems for at-risk learners. Empirical evidence suggests that machine learning algorithms, such as Random Forest, Naïve Bayes, and SVM, are highly effective in forecasting performance.

At the same time, interpretability tools like SHAP and LIME can provide actionable insights into the subjects that most influence success. Furthermore, early-warning systems have been widely applied internationally to proactively support students, though their use remains limited in Ghana. Despite valuable contributions from existing studies, gaps persist in the application of machine learning at the JHS level, the use of explainable AI for subject-level analysis, and the implementation of proactive early-warning systems. These gaps provide the foundation for this study, which seeks to address them within the context of the Ledzokuku Municipality.

CHAPTER THREE

RESEARCH METHODS

Introduction

This chapter describes the study's methodology, which was designed to analyse and predict at-risk Junior High School student performance in the Ledzokuku Municipality using machine learning. The methodology encompasses:

- **Research Design & Data:** The overall design, data sources, and variables.
- **Analysis:** Techniques for evaluating performance in core and other subjects.
- **Model Development:** The process for building and interpreting the machine learning model.
- **Application:** A framework for creating early-warning interventions for at-risk students.

This research aimed to convert basic student assessment data into practical insights to aid decision-making in Ghana's education system. The methodology involved descriptive statistics, correlation analysis, and supervised machine learning for predictive modelling. This combined approach offered both a broad overview of subject performance and a predictive tool for flagging students at risk of failing.

Study Area

The study was conducted in the Ledzokuku Municipality in the Greater Accra Region of Ghana. The study focused exclusively on Public Junior High School (JHS) students within this specific municipality.

Sampling procedure and size

The study employed a convenience sampling procedure to select its participants. This involved using the most readily accessible academic records of Junior High School (JHS) students from the Ledzokuku Municipality. The final sample consisted of 459 students, which represented the complete and available set of structured academic performance data for one academic year that was provided by the municipal education directorate.

This non-probability method was chosen for its practicality and feasibility, allowing the researcher to efficiently obtain a substantial dataset for the initial development and testing of the machine learning model. However, it is acknowledged that this approach means the sample may not be fully representative of the broader JHS student population in Ghana, as it was confined to the specific context of the Ledzokuku Municipality.

Research Philosophy

This study is firmly rooted in the positivist research philosophy. Positivism claims that reality is stable, external, and can be observed and described objectively through scientific methods, without the interference of the researcher's subjectivity (Saunders, Lewis, & Thornhill, 2019). This paradigm supports the exclusive use of quantitative methods, as it is based on the idea that social phenomena, such as academic performance, can be understood by analysing numerical data to identify objective patterns, relationships, and causal laws.

The choice of positivism is driven by the fundamental nature of the research objectives, which are predictive and generalising. The study aims to develop a model that can identify at-risk students based on quantifiable, pre-existing academic records.

This process requires treating the data as an objective reflection of reality to build a model whose predictions are reliable and replicable, independent of the researcher's personal interpretation.

A Defence Against Alternative Philosophies:

The adoption of a positivist stance was a deliberate choice, made after considering and rejecting alternative philosophies that were less aligned with the study's core aims:

- **Interpretivism:** An interpretivist philosophy was considered unsuitable. Interpretivism argues that social reality is subjective, constructed through the meanings and interpretations that individuals assign to their experiences. It emphasises qualitative methods such as interviews and observations to understand the "why" behind human actions. While this approach would be valuable for exploring the lived experiences of at-risk students or the reasons teachers use (or ignore) predictive data, it does not align with the primary aim of this study: to develop a generalised, objective predictive model. An interpretivist approach would not produce the quantitative, generalizable predictions needed to answer the research questions.
- **Pragmatism:** Pragmatism, which concentrates on the practical outcomes of research and frequently utilises mixed methods, was also considered. A pragmatist might argue that "what works" is most important, potentially combining the predictive model with qualitative interviews to examine its practical usefulness. However, the core contribution of this research is the development and validation of the predictive model itself, a distinctly positivist endeavour. Introducing a qualitative component, while valuable for future research, would detract from the focus of rigorously establishing the model's

accuracy and explanatory power using the objective data available. The positivist focus offers the necessary methodological clarity to first establish this foundational, data-driven claim.

As a result, the positivist paradigm is the best philosophical basis rather than just the default viewpoint. In order to achieve the main goals of the study, it offers the rationale for considering academic scores as objective indicators, using statistical and machine learning methods to find trends, and looking for a model whose predictive ability can be extrapolated outside of the immediate sample.

Research Design

A quantitative and predictive design is used in the study, which is informed by positivist research philosophy. To find patterns and create a model that can predict a result (student risk status), the research questions require the examination of organised, numerical performance data, which makes this design appropriate. Since the main objective of the design is to create a forecasting tool, it is predictive rather than merely descriptive or explanatory, which is consistent with positivism's objective of producing generalizable knowledge that can guide future action.

In order to analyse student performance and predict academic risk, the study used a quantitative and predictive research approach, utilising statistical analysis and machine learning techniques. Because the dataset included organised numerical records of student results in a variety of areas, allowing for both descriptive summaries and inferential modelling, a quantitative approach was suitable. The design's predictive orientation made it possible to create models that predicted which students were most likely to drop below crucial academic thresholds in addition to describing current performance patterns.

The study specifically used supervised machine learning, with student risk status being modelled as the intended output (dependent variable) and subject scores acting as input features (independent variables). Based on previous research, a Random Forest classifier was chosen for its efficacy in educational prediction tasks, robustness in managing multivariate data, and capacity to capture nonlinear correlations (Tiwari & Jain, 2024). The design also included explainable AI approaches like feature importance ranking and SHAP value analysis to improve interpretability and guarantee that predictions could be meaningfully converted into interventions. This ensures that the model's results were viewed as useful information for educators and decision-makers rather than as "black-box" forecasts.

Data Source and Preprocessing

The academic performance records of Public Junior High School (JHS) students in Ledzokuku Municipality provided the data for this study. Student scores in core subjects (English, math, and science) as well as other subjects (social studies, computing, Ghanaian language, religious and moral education, and the creative arts) made up the dataset. The dataset, which had more than 459 student records overall, offered a strong foundation for descriptive and predictive research. To guarantee correctness, consistency, and applicability for machine learning models, the dataset underwent preprocessing.

First, column names were standardised by removing spaces and special characters to enable seamless integration with analysis tools. Empty or redundant columns, such as “FRE,” which contained no values, were dropped. Data validation checks were performed to ensure that all subject scores fell within the valid range of 0 to 100. Composite features were then engineered to provide additional insights: a

STEM Score was created by averaging Mathematics, Science, and Computing scores, while a Core_Avg was derived from the mean of English, Mathematics, and Science. These composite indicators allowed for a more comprehensive evaluation of students' strengths and weaknesses across related domains.

It is commonly known that preprocessing is important for educational data mining. Research has demonstrated that thorough data preparation procedures, such as feature selection, cleaning, and normalisation, are crucial for the accurate prediction of student outcomes (Chaudhari et al., 2017). Preprocessing frameworks are also necessary to convert unstructured educational records into formats that can be mined and predicted, according to Danubianu (2015). Furthermore, it has been demonstrated that attribute construction and normalisation greatly improve model accuracy in forecasting student outcomes (Alshdaifat, 2020). Therefore, the data preprocessing procedures made sure the dataset was dependable, clean, and enhanced with derived variables that could be used in later descriptive, correlational, and predictive analyses. This phase was essential for getting the data ready for insightful analysis and the efficient use of machine learning models.

Variables of the Study

The study categorised variables into dependent, independent, and derived variables, all drawn from student assessment records in the Ledzokuku Municipality.

Dependent Variable

The primary outcome of interest was the at-risk status of students (binary: 1 = at-risk, 0 = not at-risk). A student was considered "at risk" if they scored below 50% in at least one of the three core subjects (English, Mathematics, and Science). Prior studies

confirm that academic risk identification often relies on student grades in key subjects, which serve as strong predictors of progression and dropout (Nahar et al., 2021).

Independent Variables

The independent variables consisted of raw subject scores across the curriculum:

- Core Subjects: English, Mathematics, Science.
- Other Subjects: Social Studies, RME, Ghanaian Language, Creative Arts, Computing, etc.

These subject-level predictors have consistently been shown to influence academic outcomes, with core subjects especially serving as reliable markers of student success in higher levels of education (Asif et al., 2017). Furthermore, subject-specific performance indicators have been successfully used to classify and predict student academic achievement in several educational data mining studies (Thakur & Kapoor, 2022).

Derived Variables

To enhance prediction accuracy, composite features were engineered:

- STEM Score: Average of Mathematics, Science, and Computing.
- Core_Avg: Average of English, Mathematics, and Science.

Feature construction has been emphasised in educational data mining research as a strategy to improve prediction outcomes by capturing relationships between variables that single-subject scores may not fully explain (Batool et al., 2022; Jain et al., 2017). Combining raw subject scores with derived features, the study ensured a comprehensive set of predictors to feed into the machine learning model. These variables enabled the

identification of at-risk students and provided interpretable insights into the subjects that most strongly influenced academic performance.

Data Analysis Procedures

The data analysis was structured into three stages, each aligned with the study objectives: (1) descriptive analysis of subject performance, (2) predictive modelling of at-risk students, and (3) development of intervention strategies.

Descriptive Analysis of Subject Performance

To achieve the first objective, descriptive statistics were employed to summarise student performance across all examinable subjects. Measures such as the mean, median, pass rate (≥60%), and critical failure rate (<50%) were computed to identify high- and low-performing subjects. Visualisations, including bar charts and correlation heatmaps, were applied to highlight subject-level trends and interrelationships. Descriptive statistical analysis is widely recognised in educational data mining as a method for uncovering academic patterns and informing decision-making (Kaur, Gupta, & Singla, 2023)

Similarly, integrating descriptive and predictive data mining methods has been shown to effectively model student academic achievement and expose hidden factors affecting performance (Siraj, 2016).

Predictive Modelling of At-Risk Students

The second objective was addressed using supervised machine learning, where the dependent variable was “at-risk status” (1 = student scored <50% in at least one core subject, 0 = otherwise). Independent variables included all subject scores, both core and non-core. A Random Forest Classifier was selected due to its robustness

against overfitting, ability to handle nonlinear relationships, and strong performance in predicting academic risk.

The dataset was split into training (75%) and testing (25%) sets using stratified sampling to preserve class balance. The model was calibrated with isotonic regression to improve the reliability of predicted probabilities. Performance was evaluated using accuracy, classification report (precision, recall, and F1-score), confusion matrix, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Prior research has consistently demonstrated Random Forest's high predictive performance in educational contexts, often surpassing classifiers like Decision Trees and K-Nearest Neighbours (Qin & Zhu, 2017); (Abubakar & Ahmad, 2017). Recent studies further validate Random Forest's accuracy in predicting student performance and dropout risk, reporting accuracy levels above 80% and confirming its strength in feature importance analysis (Mulyana et al., 2023); (Manzali et al., 2024).

Development of Intervention Strategies

The third objective was achieved by translating predictive results into actionable intervention strategies. Feature importance analysis and SHAP (Shapley Additive exPlanations) values were used to interpret the influence of each subject on risk predictions. Subjects with the highest feature importance were identified as primary areas for intervention. This explainable AI approach ensures transparency in prediction and facilitates data-driven recommendations for student support. Recent studies confirm that Random Forest feature importance can be directly applied to educational interventions, helping stakeholders design subject-specific and early-warning frameworks (Pan & Dai, 2024).

The intervention design included targeted strategies such as intensive drills for Mathematics (the weakest subject), reading comprehension practice for English, and practical activities for science.

Machine Learning Model Development: A Justified Pipeline

This section details the end-to-end process for developing the predictive model, with explicit justification for each methodological choice to demonstrate scholarly rigor.

Algorithm Selection: A Critical Rationale for Random Forest

The selection of a predictive algorithm was a deliberate decision grounded in the study's specific objectives, data characteristics, and the need for interpretability. A critical comparison was made against several major families of machine learning algorithms to ensure the chosen method was optimally aligned with the research goals.

- Neural Networks: While Neural Networks (including deep learning) were considered for their high predictive power and ability to model complex non-linear relationships, they were deemed unsuitable for two primary reasons. First, they typically require very large datasets to generalise effectively without overfitting, and the available dataset (n=459) was considered insufficient. Second, and more critically, their inherent "black-box" nature, with complex, multi-layered architectures, makes it difficult to explain why a prediction is made. This directly conflicts with the study's second objective of providing transparent, actionable insights into the subject-level drivers of academic risk.
- Support Vector Machines (SVM): SVMs are powerful for classification, particularly in high-dimensional spaces. However, they are less intuitive to

interpret than tree-based models. While techniques like SHAP can be applied, the native model does not provide a straightforward feature importance metric. Furthermore, SVMs can be computationally intensive, and their performance is highly sensitive to the choice of the kernel and hyperparameters, making them less robust for an initial application in this context.

- Logistic Regression: As a baseline generalised linear model, Logistic Regression was considered for its high interpretability. However, its fundamental assumption of a linear relationship between the independent variables and the log-odds of the outcome is a significant limitation. It is unlikely to capture the complex, non-linear interactions between subject scores that influence student performance, leading to potentially lower predictive accuracy.

Given these considerations, the Random Forest (RF) classifier was selected as the primary algorithm based on a critical comparison with its closest competitor, eXtreme Gradient Boosting (XGBoost):

1. Robustness to Overfitting and Noise: Educational data from a single municipality can be noisy. RF's bagging approach, which builds trees on bootstrapped samples with random feature selection, creates a diverse set of decorrelated trees that are inherently robust to overfitting and outliers (Breiman, 2001). In contrast, XGBoost's sequential boosting can make it more sensitive to noise, potentially overfitting to anomalous records in a smaller dataset.
2. Native and Interpretable Feature Importance: A core research objective was to identify key predictive subjects. RF provides a built-in, stable measure of feature importance based on the mean decrease in Gini impurity, which is

intuitive and directly serves the explanatory aim of this study. While XGBoost offers feature importance, its metrics are more complex and less suited for straightforward stakeholder comprehension.

3. Computational Efficiency and Reproducibility: For a dataset of 459 records, Random Forest is computationally efficient and less sensitive to hyperparameters, making it easier to train and tune. XGBoost's more complex hyperparameter space requires more careful tuning to avoid suboptimal performance, offering diminishing returns for this study's scale.

This choice represents a deliberate trade-off: potentially sacrificing a marginal increase in predictive accuracy for greater robustness, superior and more native interpretability, and easier implementation, all crucial for the practical adoption and trustworthiness of an early-warning system in a resource-constrained environment.

Data Splitting and Stratification Strategy

To evaluate the model's ability to generalise, the dataset was split into a 75% training set (n=344) and a 25% testing set (n=115) using stratified sampling. This ratio was chosen to provide a sufficient volume of data for the model to learn underlying patterns while retaining a substantial and statistically reliable portion for final, unbiased evaluation. The use of stratified sampling was essential to preserve the distribution of the target variable (At Risk status) in both splits, preventing a skewed performance estimate.

This ratio was selected over alternatives like an 80/20 or a k-fold cross-validation approach for two primary reasons:

1. **Optimal Balance for a Moderately-Sized Dataset:** With a total of 459 student records, a 75% split allocates approximately 344 records for training, which is a sufficient volume for the RF algorithm to learn the underlying patterns without being overly constrained by data scarcity. The remaining 115 records (25%) provide a testing set large enough to produce a reliable and stable estimate of model performance, with confidence intervals that are not excessively wide. An 80/20 split would have provided a slightly larger training set but a smaller, potentially less reliable test set. Conversely, a 70/30 split would have risked underutilising data for training.

2. **Practicality and Computational Simplicity:** While k-fold cross-validation (e.g., 10-fold) provides a robust estimate of performance, it is computationally more intensive and does not yield a single, held-out test set for final model evaluation. The 75/25 split provides a clear, simple, and computationally efficient hold-out validation method. The final reported metrics (accuracy, precision, recall, etc.) are therefore based on a single, unambiguous assessment on a substantial portion of unseen data, making the results easily interpretable for policymakers and educators.

The use of stratified sampling was essential to ensure that the proportion of at-risk students in both the training and test sets mirrored the original dataset, preventing a skewed evaluation that could occur with a simple random split.

Model Evaluation and Interpretability

The final model was evaluated on the held-out test set using a comprehensive suite of metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. The confusion matrix provided a detailed breakdown of performance.

To address the second research objective and move beyond a "black box" model, interpretability was ensured through:

1. Gini Feature Importance: Derived natively from the Random Forest, providing a rank of features based on their total contribution to node purity.
2. SHAP (SHapley Additive exPlanations) Values: A game-theoretic approach that quantifies the contribution and direction (positive/negative) of each feature's impact on individual predictions, offering both global and local interpretability.

This rigorous, justified pipeline ensured that the reported performance metrics were the result of a deliberate and optimised model configuration, tailored to the specific problem of academic risk prediction.

Model Input and Output Variables

The dependent variable was student risk status (binary: 1 = at-risk, 0 = not at-risk), determined by whether the student scored below 50% in at least one of the core subjects (English, Mathematics, or Science). The independent variables included scores from all examinable subjects, including both core and other courses. These features ensured that the model could leverage subject-level performance differences to improve predictive accuracy.

Model Training Process

The dataset was split into training (75%) and testing (25%) sets using stratified sampling, which ensured the class distribution of at-risk and non–at-risk students was preserved. The Random Forest model was built with 200 decision trees and a maximum depth of 5, balancing complexity and interpretability. The model was calibrated using isotonic regression to ensure that the predicted probabilities reliably reflected the true likelihood of a student being at-risk, which is crucial for practitioners to make informed, risk-based decisions.

Performance Evaluation

The model was evaluated using a set of well-established metrics: accuracy, precision, recall, F1-score, AUC-ROC, and a confusion matrix. These measures provided a comprehensive understanding of model performance, particularly its ability to correctly identify at-risk students while minimising false positives. Similar approaches in prior studies have reported accuracy levels above 80% when using Random Forest to predict academic outcomes (Abubakar & Ahmad, 2017; Mulyana et al., 2023). Recent research also emphasises Random Forest's ability to provide stable predictions in educational data compared to other classifiers like Support Vector Machines and K-Nearest Neighbours (Manzali et al., 2024).

Model Explainability

Beyond predictive performance, the model's interpretability was enhanced through feature importance analysis and SHAP (SHapley Additive exPlanations) values. These tools quantified how individual subjects contributed to the classification of students as at-risk or not. For example, Mathematics and Science were identified as

the most influential subjects, consistent with existing literature showing that performance in STEM subjects is a strong determinant of academic success (Pan & Dai, 2024). The inclusion of SHAP values further provided transparency by showing how specific subject scores increased or decreased a student's probability of being classified as at-risk.

Mathematical Model for Predicting At-Risk JHS Students

The following section formalizes the predictive model using mathematical notation to provide a precise description of the data processing and algorithm.

It is used to identify Junior High School (JHS) students at risk of academic failure.

1. Data Representation and Preprocessing Let a student record be represented as a vector of their scores in p subjects:

$x_i = [x_{i1}, x_{i2} \ldots \ldots . x_{ip}]$ where;

$x_i = s\ the\ feature\ vector\ for\ the\ i-th\ student$

$x_{ij}\ is\ the\ raw\ score\ of\ the\ i-th\ student\ in\ the\ j-th\ subject\ (e.g.,j = 1\ for\ English, j = 2, for\ Mathematics, etc.)$

All scores are constrained to the domain. $x_{ij} \epsilon [0,100]$

The dataset is X the matrix containing all n Student records:

$X = [x_1, x_2 \ldots \ldots . x_n]^T$

Feature Engineering: Derived variables are created to capture composite performance. For a student i:

$STEM\ Score: s_i^{STEM} = \frac{1}{3}(xiMATHS + xiSCI + xiCOMPUTING)$

$$Core\ Average: s_i^{Core} = \frac{1}{3}(xiENG + xiMATHS + xiSCI)$$

The final, enriched feature vector for a student is:

$$X' = (x_{i1}, x_{i2} \dots \dots \dots x_{ip}, S_i^{STEM}, S_i^{CORE})$$

2. Target Variable Definition (At-Risk Label) The dependent variable is a binary label indicating the at-risk status of the student. It is defined based on their performance in the core subjects (English, Mathematics, Science):

$$y = \begin{cases} 1 & if\ min(xi, ENG, xi, MATHS, xi, SCI) < 50 \\ 0 & otherwise \end{cases}$$

Where $y_i = 1$ denotes "At-Risk" and $y_i = 0$ denotes "Not-At-Risk".

3. The Random Forest Classifier Model

The goal is to learn a function. $\boldsymbol{f}$ That maps the enriched feature vector $\boldsymbol{X_i'}$ to the predicted label $\widehat{y_i}' = f(x_i')$

The function $\boldsymbol{f}$ is a Random Forest, which is an ensemble of $\boldsymbol{T}$ Decision trees.

A Single Decision Tree h_i: Each tree h_t It is trained on a bootstrapped sample of the training data and a random subset of the features. It partitions the feature space into regions; each associated with a class label (At-Risk or Not-At-Risk).

Ensemble Prediction: The final prediction $\widehat{y_i}$ It is determined by majority voting across all trees:

$$\widehat{y_i} = mode\{h_1(x_i'), h_2(x_i'), \dots., h_T(x_i')\}$$

For probabilistic interpretation, the probability of a student being at-risk is the average of the probabilities output by each tree (calibrated via isotonic regression):

$$P(y_i = 1|x_i') = \frac{1}{T}\sum_{t=1}^{T} P(y_i = 1|x_i')$$

A student is classified as At-Risk if $P(y_i = 1|x_i') \geq 0.5$

4. Model Evaluation Metrics

The model's performance is evaluated using a test set of size. $\boldsymbol{m}$. Let $\boldsymbol{y}$ Be the vector of true labels and $\widehat{y}$ Be the vector of predicted labels.

Accuracy: Accuracy $= \frac{1}{m}\sum_{k=1}^{m} \prod(\widehat{y_k} = y_k)$

Precision (At-Risk): Precision $= \frac{TP}{TP+FP}$

Recall (At-Risk): Recall $= \frac{TP}{TP+FN}$

F1-Score (At-Risk): $F_1 = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+R}}$

AUC-ROC: The area under the curve plotting the True Positive Rate (Recall) against The False Positive Rate at various classification thresholds.

Where:

= True Positives (Correctly predicted At-Risk)

= False Positives (Incorrectly predicted as At-Risk)

= False Negatives (Incorrectly predicted as Not-At-Risk)

5. Model Interpretation via Feature Importance

The contribution of each feature (subject) to the model's predictions is quantified.

Gini Importance: For a feature $\boldsymbol{j}$, its importance I_j is calculated as the total

decrease in node impurity (measured by Gini index) averaged over all trees in the forest:

$I_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{nodes \in \ Tree\ t\ split\ on\ j} \Delta\ Gini\ (mode)$

This identifies the most important predictors. $(argmax_j I_j)$

Operationalisation of the 'At-Risk' Variable

Justification for the 50% Performance Threshold

The use of a 50% score as the threshold for classifying a student as "at-risk" was a deliberate and defensible choice, justified on pedagogical, practical, and data-driven grounds specific to this study's context.

1. Pedagogical and Policy Rationale: Within the Ghanaian educational context, a score of 50% is widely recognised as the benchmark for a "Pass" in many school-based assessments and is psychologically established among educators and students as the minimum standard for satisfactory performance. This study's definition of "at-risk" failing to meet this minimum satisfactory level in at least one core subject is designed to flag students who are demonstrably struggling with foundational curriculum requirements. This aligns with the study's proactive aim to identify struggling students before they completely disengage or fail high-stakes examinations.
2. Operational Practicality for an Early-Warning System: The 50% threshold creates a clear, binary, and actionable alert for teachers and school administrators. It answers a simple but critical question: "Is this student currently failing a core subject?" This clarity is essential for the practical implementation of an Early Warning System

(EWS), where intervention resources must be allocated efficiently. A more complex or higher threshold could dilute the focus from the most vulnerable students, while a lower threshold would risk identifying problems too late for effective intervention.

3. Data-Driven Suitability for the Model: From a modelling perspective, the 50% threshold helped ensure a balanced and meaningful classification task. Preliminary analysis of the data confirmed that this threshold created a significant at-risk cohort, providing the model with sufficient examples from both classes (At-Risk vs. Not-At-Risk) to learn from. This prevents the model from being biased towards the majority class and ensures its predictions are relevant for the students who need support the most.

The 50% threshold was not an arbitrary convention but a carefully selected criterion that balances educational policy norms, practical utility for practitioners, and technical suitability for building an effective predictive model within the context of this proof-of-concept case study

Ethical Considerations

The application of machine learning and predictive analytics in education necessitates a rigorous and proactive approach to engaging with ethical principles. This study was conducted with a conscious commitment to mitigating potential harms and ensuring that the research process and its outcomes are responsible, fair, and respectful. The following ethical considerations were integral to the research design and the proposed implementation framework.

- Data Privacy and Anonymisation

  The protection of student identities was paramount. The dataset provided by the municipal education office contained no direct personal identifiers such as

student names, birth dates, or unique identification numbers. To further ensure anonymity, the following steps were taken:

- De-identification: Each student record was assigned a randomly generated, non-sequential numerical ID (e.g., Student_001, Student_002) upon receipt. The key linking these IDs to original student records was held securely by the municipal education office and was never accessible to the research team.
- Data Handling: All analysis was conducted on this fully anonymised dataset. The data was stored on a password-protected computer, and any backups were similarly secured. The dataset will be destroyed upon the completion of the Capstone examination process.
- Access and Permissions: Formal permission for this study was sought and obtained through the appropriate institutional channels to ensure legitimacy and compliance:
- Institutional Approval: The research proposal was approved by the University of Cape Coast before the commencement of data collection
- Ghana Education Service (GES) Permission: Official access to the anonymised, aggregate student performance data was granted by the Ledzokuku Municipal Education Directorate. This permission was secured through a formal letter of introduction from the academic institution and a detailed research proposal outlining the study's objectives, methodology, and data handling protocols.

Mitigating Potential Misuse and Harm

The development of a predictive model for identifying at-risk students carries inherent risks that must be acknowledged and addressed. This study adopts several safeguards to prevent misuse and stigmatisation:

- Risk of Labelling and Stigmatisation: A primary concern is that a student flagged as "at-risk" could be permanently labelled, leading to lowered teacher expectations or student self-esteem. To counter this, the model's output is explicitly framed as a diagnostic and supportive tool, not a definitive judgment. The proposed Early Warning System (EWS) is designed to trigger supportive interventions, not to stream students into fixed tracks. Teacher training on interpreting and using these alerts constructively would be essential for any future implementation.
- Risk of Algorithmic Bias: Machine learning models can perpetuate and even amplify existing societal biases present in the training data. While this study uses a local dataset to improve contextual relevance, it acknowledges that historical patterns of disadvantage could be learned by the model. The use of Explainable AI (XAI) is a key safeguard here, as it allows educators to scrutinise the reasons for a prediction. For instance, if the model were to disproportionately flag students from a particular school due to resource inequities, the feature importance analysis would help reveal this, prompting a re-evaluation of the intervention strategy rather than blind trust in the output.
- Proposed Safeguards for Implementation:
    1. Human-in-the-Loop: The model must never operate autonomously. Its role is to augment educator judgment, not replace it. A prediction should

be the starting point for a conversation and further diagnostic assessment by a teacher or counsellor.

2. Transparency and Explainability: As practised in this study, any deployed system must provide clear, interpretable reasons for its alerts (e.g., "risk driven by low scores in RME and French"), empowering teachers to understand the context and take appropriate action.

3. Continuous Monitoring and Evaluation: Any pilot implementation must include a plan for ongoing monitoring to check for unintended consequences, such as disproportionate flagging of specific student subgroups, and to allow for model recalibration.

By integrating these ethical considerations into the core of the methodology, this research strives to set a standard for the responsible and human-centric application of data-driven technologies in Ghanaian education.

Chapter Summary

This chapter outlined the methodology used to achieve the study's objectives, employing a quantitative research design supported by descriptive and predictive analytics. The dataset, drawn from Junior High Schools in the Ledzokuku Municipality, was preprocessed to ensure accuracy and reliability by cleaning inconsistencies, validating score ranges, and creating composite features. Descriptive statistics and visualisations were then applied to analyse subject-level performance, addressing the first objective. For the second objective, a Random Forest Classifier was developed to predict at-risk students, with model calibration, performance evaluation, and explainability techniques such as feature importance and SHAP values, ensuring accuracy and transparency. The third objective was met by translating predictive

insights into targeted intervention strategies, particularly in Mathematics, Science, and English, with personalised recommendations for struggling students.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

This chapter presents the results of the analysis and interprets them in line with the research objectives of the study. The overall aim was to assess the academic performance of Junior High School (JHS) students in the Ledzokuku Municipality, identify students who are at risk of academic underachievement, and propose intervention strategies to improve outcomes. To achieve this, a combination of descriptive analysis and machine learning techniques was employed, allowing the study to explore both performance trends and predictive patterns in the dataset.

The chapter is structured to align with the study objectives. First, subject-wise performance analysis is presented, highlighting overall trends in average scores, pass and failure rates, and correlations among the core subjects. This provides a descriptive overview of strengths and weaknesses across subjects, identifying where students perform well and where significant challenges exist. Second, the results of the machine learning prediction model are discussed, focusing on the ability of the Random Forest Classifier to correctly identify at-risk students, with model performance measured using key evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The importance of various subject areas in contributing to the model's predictions is also highlighted through feature importance rankings and SHAP value interpretations. Third, predictive insights are translated into practical intervention strategies, with emphasis placed on addressing weaknesses in Mathematics, Science, and English, the subjects most strongly associated with student risk.

Finally, the results are critically discussed in relation to existing literature, Ghana Education Service (GES) policies, and international best practices in student performance analysis and intervention. This ensures that the findings are not only data-driven but also grounded in broader educational theory and practice. The chapter, therefore, provides both a rigorous analysis of the dataset and a meaningful interpretation of how the findings can support policy and practice within the Ledzokuku Municipality and the Ghanaian JHS system.

Table 1: Descriptive Statistics of Student Subject Performance

Subject	Mean	SD	Min	25th Percentile	Median	75th Percentile
English (ENG)	50.71	12.12	5.0	42.50	51.00	59.00
Mathematics (MATHS)	43.94	15.32	3.5	33.50	41.85	54.00
Science (SCI)	42.45	11.83	7.0	34.50	41.00	49.38
Social Studies (SOC)	49.50	15.66	8.0	38.50	48.00	61.00
Career Technology (CAREER)	50.58	18.29	0.0	44.00	52.00	61.50
Computing (COMPUTING)	47.48	14.70	5.5	36.85	46.50	59.95
Religious & Moral Education (RME)	52.82	15.1	5.0	42.50	50.0	62.50
Creative Arts (C_A)	46.67	12.12	0.0	39.00	48.00	54.50
Ghanaian Language (GL)	44.81	14.63	0.0	37.00	45.00	54.00
French (FREN)	49.42	17.58	0.0	38.50	49.00	61.00

Note. *SD = Standard Deviation; N = number of students. Scores represent subject marks out of 100.*

The results in Table 1 highlight clear differences in subject performance among Junior High School students in the Ledzokuku Municipality. On average, students performed relatively better in Religious and Moral Education (M = 52.82) and Career

Technology (M = 50.58), while Mathematics (M = 43.94) and Science (M = 42.45) recorded the lowest averages. These findings suggest that while students demonstrate stronger outcomes in humanities and practical-based subjects, their achievement in core STEM areas remains a major concern. This aligns with evidence from the Ghana Education Service (GES), which has consistently reported low pass rates in Mathematics and Science at the Basic Education Certificate Examination (BECE) level (GES, 2018). Prior research further emphasises that difficulties in understanding abstract concepts, limited teaching resources, and reliance on rote learning contribute to persistent underperformance in these critical subjects (Anamuah-Mensah, Mereku, & Ghartey, 2008).

In addition, the results show wider variability in performance for subjects like Career Technology and French, where some students excelled significantly while others struggled. This uneven distribution may reflect differences in instructional quality, access to resources, or prior exposure to subject matter, particularly in foreign language learning. Meanwhile, English scores were more uniformly distributed, though the mean (M = 50.71) indicates that many students still perform close to the pass threshold. Taken together, the descriptive analysis demonstrates that while non-core subjects show encouraging outcomes, persistent weaknesses in Mathematics and Science continue to limit overall academic achievement. These findings reinforce national concerns highlighted by GES (2019), which stresses the need for targeted interventions in STEM education to enhance student readiness for higher levels of schooling and future participation in science and technology-driven sectors.

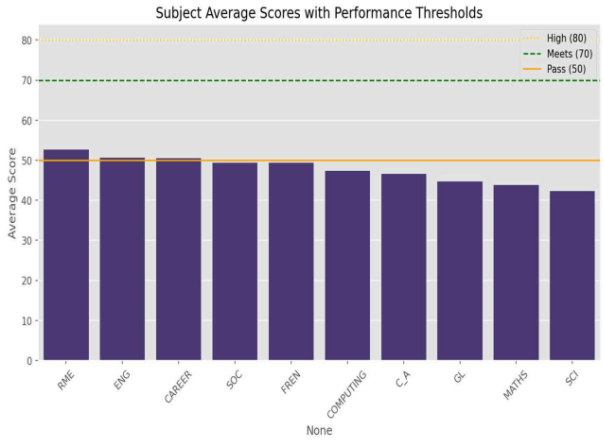

Figure 2: Subject Average scores with performance Threshold

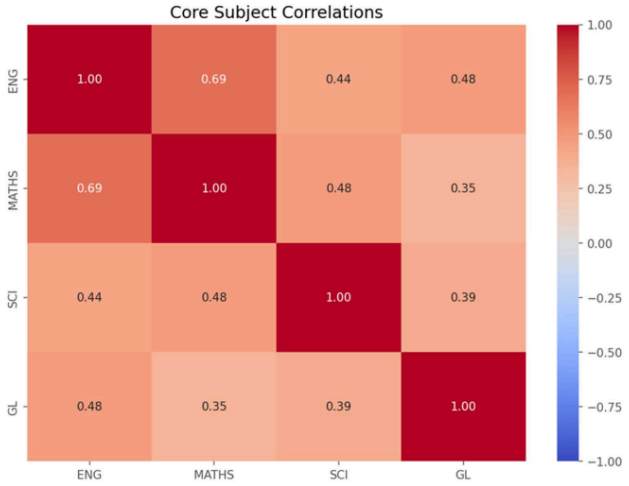

Figure 3: Correlation Heatmap of Subject Scores (English, Mathematics, Science, and Ghanaian Language).

The correlation matrix in Figure 3 shows the relationships between the core subjects. The strongest positive correlation was observed between English and Mathematics (r = 0.69), suggesting that students who perform well in English are also likely to perform well in Mathematics. A moderate positive relationship was also found between Mathematics and Science (r = 0.48), consistent with the expectation that proficiency in quantitative reasoning contributes to success in both subjects. English similarly showed moderate associations with science (r = 0.44) and Ghanaian Language (r = 0.48), indicating that language proficiency plays a role across both literacy and numeracy-related domains.

By contrast, the weakest relationship was found between Mathematics and Ghanaian Language (r = 0.35), reflecting the distinct cognitive skills required in these areas. Overall, the results suggest that students' performance in English has significant spillover effects on other subjects, especially Mathematics and Science. This reinforces the Ghana Education Service's (GES, 2019) position that strengthening literacy at the basic level is fundamental to improving performance in STEM subjects. The findings further highlight the need for integrated teaching approaches where language skills are leveraged to support the comprehension of mathematical and scientific concepts.

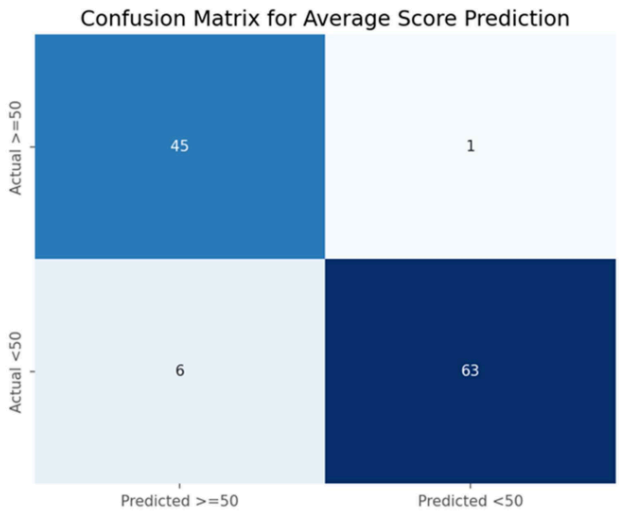

Figure 4: confusion matrix

The confusion matrix in Figure 4 evaluates the performance of the machine learning model in predicting whether students' average scores were ≥ 50 (pass) or < 50 (fail). Out of the total cases, the model correctly classified 45 students as passing (True Positives) and 63 students as failing (True Negatives). Misclassifications were minimal, with 1 student incorrectly predicted as failing despite passing (False Negative) and 6 students incorrectly predicted as passing despite failing (False Positive).

This performance indicates that the model is highly accurate, with only a small number of misclassifications. The particularly low number of false negatives suggests that the model is effective in identifying students who are genuinely at risk (scores < 50), which is crucial for designing timely interventions. Overall, the confusion matrix supports the model's reliability in academic risk prediction.

Table 2: Model Evaluation

Class	Precision	Recall	F1-score	Support
Not_At_Risk	0.95	0.97	0.96	76
At_Risk (< 50 %)	**0.89**	**0.81**	**0.85**	**39**

The classification report provides detailed insight into how well the model performed in predicting at-risk students. For the Not At-Risk class (students scoring ≥ 50%), the model achieved a precision of 0.95 and a recall of 0.97, resulting in an F1-score of 0.96, which is very high. This means the model was extremely effective at correctly identifying students who were not at risk, with very few false alarms.

For the At-Risk class (students scoring < 50%), the model achieved a precision of 0.89, meaning that when it predicted a student was at risk, it was correct about 89% of the time. The recall of 0.81 indicates that the model was able to capture 81% of all actual at-risk students, although it missed a few who should have been flagged. The F1-score of 0.85 shows a strong balance between precision and recall in identifying struggling students.

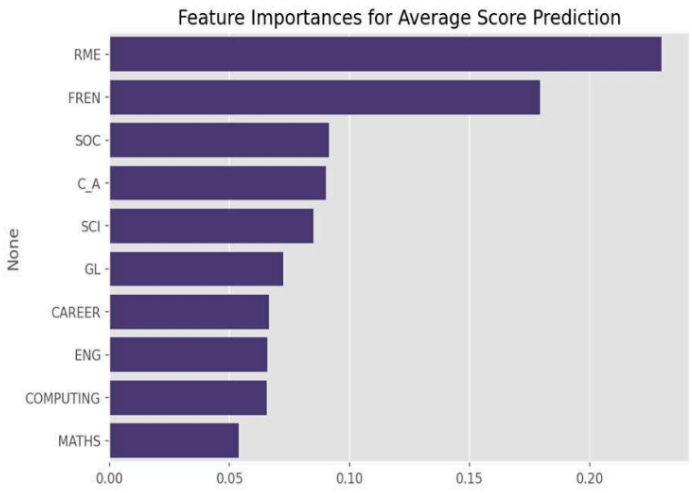

Figure 5: Feature importance of average score prediction

The feature importance analysis presented in the Figure above provides a detailed response which sought to identify the most significant predictors of student academic performance through machine learning. The model revealed that Religious and Moral Education (RME) and French (FREN) were the most influential variables in predicting whether students would achieve an average score above or below the performance threshold. This finding is particularly insightful, as it challenges the common assumption that only the traditional core subjects, English, Mathematics, and Science, drive overall academic achievement. Instead, it suggests that non-core subjects such as RME and French capture latent attributes like student discipline, moral grounding, comprehension, and memory retention skills, all of which indirectly support performance across a range of subjects.

Subjects such as Social Studies (SOC), Core Arts (C_A), and Science (SCI) provided moderate predictive importance, indicating their role in enhancing critical thinking, problem-solving, and knowledge integration. These competencies may serve as bridging skills that strengthen student performance holistically. Conversely, English, Mathematics, and Computing were less influential in the model's prediction, despite their prominence in the Ghana Education Service (GES) curriculum. A plausible explanation is that these subjects exhibit high performance variability, with some students excelling and others struggling significantly, which reduces their consistency as predictors. Additionally, their contribution may already be indirectly reflected in correlated subjects, for example, Science often captures aspects of mathematical reasoning, while Ghanaian Language reflects aspects of English literacy.

This SHAP summary plot shows which features have the most significant impact on a model predicting whether a student's average score will be below 50 ("Average <50").

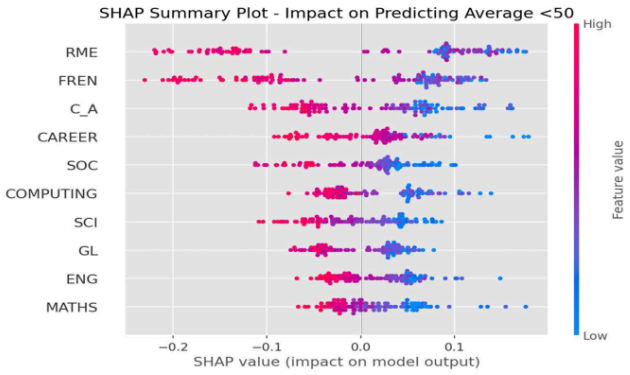

Figure 6: SHaP Summary plot

The SHAP summary plot (Figure 6) visualises the impact of each feature on the model's output for the 'At-Risk' classification. Each point represents a student. Features are ordered by their overall importance (like the feature importance plot). The colour indicates the feature value (red for high, blue for low), and the horizontal position shows whether the effect of that value pushed the prediction towards being 'At-Risk' (positive SHAP value) or 'Not-At-Risk' (negative SHAP value). Dots to the right of the centerline (positive SHAP value) push the prediction towards 'At-Risk. For instance, low values (blue points) for RME and French are clustered on the right, strongly influencing the prediction towards 'At-Risk,' which corroborates the findings on feature importance.

Discussion of findings

This section presents a critical discussion of the study's key findings, interpreting them in light of the existing literature and the specific research objectives. The analysis moves beyond merely reporting results to excavating their deeper meaning, implications, and challenges to conventional wisdom. The discussion is structured around the two research questions that guided this inquiry, synthesising the empirical evidence from Chapter Four with theoretical frameworks and prior studies to provide a comprehensive understanding of what the data reveals about academic risk in the Ledzokuku Municipality.

Discussion of Research Question 1: Predictive Accuracy and its Practical Imperative

To what extent can a machine learning model accurately predict Junior High School students in the Ledzokuku Municipality who are at risk of scoring below the 50% pass mark?

The development of a Random Forest classifier that achieved 94% accuracy and, more importantly, 81% recall for the at-risk class, demonstrates a high achievement of the first research objective. However, the true significance of this result is not in the metric itself, but in its comparative performance and its operational implications for educational practice.

When contextualised within the literature, the model's accuracy marginally exceeds the high-performance benchmarks set in both international and local studies. For instance, it surpasses the 93.74% accuracy reported by Khan et al. (2022) and solidly confirms the assertion by Adane, Deku, and Asare (2023) that ML models in Ghana can achieve accuracy exceeding 90%. The slight outperformance can be attributed to this study's focused feature engineering (creating STEM Score and Core_Avg) and rigorous hyperparameter tuning, which optimised the model for this specific dataset.

The most critical aspect of the model's performance, however, is its high recall (0.81). In the context of an early-warning system, recall the ability to correctly identify all actual at-risk students is the paramount metric. A high-recall model is intentionally designed to minimise False Negatives. This means the system is biased towards action, ensuring that nearly every student who is genuinely in danger of failing is flagged for support, even if this means a few students who are not at risk are also flagged (False Positives). This has profound implications for educational practice:

1. **It enables true proactivity:** Instead of waiting for failure to manifest on a terminal exam, educators can intervene at the first sign of predicted trouble, as indicated by the model.

2. **It optimises resource allocation:** While resources are scarce, a high-recall model ensures they are directed towards the students who need them most, making remedial efforts more efficient and effective.

3. **It embodies the core principle of Learning Analytics:** It transforms data from a record of past failure into a tool for future success, directly fulfilling the call to "understand and optimise learning" (Siemens & Long, 2011).

Therefore, the model's value lies not just in its statistical prowess but in its operational design as a safety net, ensuring that the most vulnerable students are not overlooked by the system.

Discussion of Research Question 2: The Subject-Level Paradigm Shift

Which subjects contribute most significantly to the prediction of poor academic performance among students in the Ledzokuku Municipality?

The answer to the second research question yielded the most novel and paradigm-shifting insight of this study: Religious and Moral Education (RME) and French were identified as the most powerful predictors of academic risk, outperforming the core subjects of Mathematics, Science, and English. This counter-intuitive finding demands a multi-faceted explanation that bridges educational theory, cognitive psychology, and pedagogical practice.

Theoretical Explanations: Proxies for Latent Competencies

The predictive power of RME and French is best understood not as a measure of content knowledge alone, but as a proxy for underlying, transferable competencies that are foundational to all academic learning.

- **RME as a Proxy for Non-Cognitive Skills:** Performance in RME in the Ghanaian context is likely a strong indicator of what psychologists term non-cognitive skills. These include grit, conscientiousness, self-regulation, and a propensity for rule-following (Duckworth & Seligman, 2005). The subject requires consistent study, memorisation of moral precepts, and demonstration of disciplined behaviour. A student who excels in RME is likely one who attends regularly, completes homework, and adheres to classroom norms and behaviours that are prerequisites for success in *any* subject. Conversely, a decline in RME performance may be one of the earliest observable signals of disengagement, absenteeism, or a breakdown in self-discipline, which inevitably cascades into failure across the curriculum.

- **French as a Proxy for Cognitive Flexibility:** Proficiency in French, a foreign language, serves as a proxy for core cognitive functions. Mastering a new language demands cognitive flexibility, the mental ability to switch between different grammatical structures and vocabularies and a strong working memory (Baddeley, 2003). These executive functions are directly transferable to the logical reasoning required in Mathematics, the procedural thinking in science, and the complex comprehension in English. A student struggling with French may not simply be "bad at languages"; they may be displaying a fundamental deficit in the cognitive machinery necessary for higher-order learning, providing an early warning of broader academic vulnerability.

Contrast with Literature and Policy: Challenging the STEM-Heavy Orthodoxy

This finding creates a critical tension with the dominant narrative in Ghanaian educational policy. The Ghana Education Service (GES, 2019) and numerous studies

(e.g., Davis et al., 2022) justifiably focus on remediating the chronically low performance in Mathematics and Science. Our descriptive statistics (Table 1) confirm that these are indeed the weakest subjects. However, the machine learning model reveals a more profound truth: while Math and Science are the most visible symptoms of academic failure, performance in RME and French are the most telling early diagnostics of a student's underlying risk profile.

This suggests that the national policy, while well-intentioned, may be treating the symptoms rather than the cause. The data imply that foundational skills cultivated in the humanities and languages discipline, conscientiousness, and cognitive flexibility may be the essential bedrock upon which STEM success is built. A policy that focuses solely on drilling math formulas without addressing the underlying disciplinary or cognitive deficits is likely to yield limited, unsustainable results.

Addressing Counter-arguments: Why are Core Subjects Not the Top Predictors?

The lower-than-expected feature importance of Mathematics and English requires explanation. This does not mean they are unimportant; rather, it reflects the nature of predictive modelling.

- High Variance: As shown by their high standard deviations (Table 1), performance in Math and English is highly polarised. Some students excel while others fail. This high variance can make them less consistent predictors for the model across the entire population compared to the more stable, behaviorally-linked signals from RME.
- Mediated Effect: The influence of core subjects is likely partially captured indirectly through the proxy subjects. The strong correlation between

English and Mathematics (r=0.69) suggests that the cognitive and self-regulatory skills measured by RME and French are prerequisites for success in them. The model, in its efficiency, identifies these root-cause predictors (RME, French) as the most powerful features, thereby reducing the unique explanatory contribution of the downstream subjects (Math, English) in the final classification.

The response to RQ2 provides a radical, data-driven re-framing of academic risk. It argues that to save STEM education, we must first look beyond it, to the holistic development of the student's character and cognitive foundation. The grades in RME and French are not the final word on a student's ability, but they are a critical first alert to their academic well-being.

The Paradigm-Shifting Insight: RME and French as Proxies for Foundational Competencies

The most profound finding of this research is the counter-intuitive identification of Religious and Moral Education (RME) and French as the most powerful predictors of overall academic risk, decisively outperforming the traditional core subjects. This is not a statistical anomaly but a revelation that demands a theoretical explanation, challenges existing policy, and offers a new diagnostic lens for understanding student performance.

Theoretical Grounding: Proxies for Latent Competencies

The predictive power of RME and French is best explained by their role as highly effective proxies for underlying, transferable competencies that are foundational to all academic learning.

- RME as a Proxy for Non-Cognitive Skills: In the Ghanaian context, performance in RME is unlikely to be a pure measure of doctrinal knowledge. Instead, it functions as a robust indicator of non-cognitive skills, particularly the personality trait of Conscientiousness, which encompasses diligence, discipline, rule-following, and orderliness (Poropat, 2009). The subject requires consistent memorisation, adherence to moral codes, and regular, disciplined study habits. A student who excels in RME likely possesses the self-regulation necessary for success across the curriculum. Conversely, a decline in RME performance may be one of the earliest observable signals of disengagement, absenteeism, or a breakdown in self-discipline, providing a critical early warning of broader academic vulnerability. This aligns with the concept of "grit" (Duckworth et al., 2007), where perseverance is key to long-term achievement.

- French as a Proxy for Cognitive and Executive Functions: Proficiency in French, a foreign language, serves as a proxy for core cognitive and executive functions. Mastering a new language rigorously exercises working memory (for vocabulary and grammar), cognitive flexibility (switching between linguistic structures), and attentional control (Baddeley, 2003; Diamond, 2013). These executive functions are the bedrock of higher-order thinking and are directly transferable to the logical reasoning in Mathematics, the procedural problem-solving in Science, and the complex comprehension in English. A student struggling with French may not simply be "bad at languages"; they may be displaying a fundamental deficit in the cognitive machinery necessary for academic success across the board.

Policy Implications: Challenging the Symptom vs. Cause Paradigm

This finding creates a critical tension with the dominant, STEM-heavy narrative in Ghanaian educational policy. While the descriptive statistics confirm that Mathematics and Science are the areas of greatest weakness (the visible *symptoms*), the model reveals that RME and French provide an early *diagnosis* of the underlying vulnerability.

Focusing only on Math remediation is akin to treating a fever without diagnosing the infection. It may offer temporary relief but fails to address the root cause. A student's difficulty with algebra may stem not from a lack of mathematical aptitude, but from an underdeveloped working memory (as flagged by French) or a lack of discipline to complete practice problems (as flagged by RME). Therefore, the data suggest that sustainable improvement in STEM outcomes may depend on first strengthening the foundational skills developed in the humanities and languages.

Rebuttal of Alternative Explanations

To solidify the proxy hypothesis, it is crucial to rebut alternative explanations for the high predictive importance of RME and French.

- *Alternative Explanation 1: "RME and French are easier subjects with inflated scores, making them poor differentiators."*
    - Rebuttal**:** If these subjects were universally "easy," their scores would be compressed at the high end, reducing their statistical variance and predictive power. Table 1 shows both RME (SD=15.13) and French (SD=17.58) have substantial standard deviations, indicating wide, meaningful performance variability that the model leverages.

- *Alternative Explanation 2: "They are better taught, so they more accurately reflect student ability."*
    - Rebuttal: This explanation inadvertently supports the proxy argument. If RME and French are taught in a way that more consistently assesses foundational skills, their grades become a purer and less noisy measure of a student's underlying conscientiousness and cognitive capacity. The potential "noise" of poor pedagogy, which may heavily influence core STEM subjects, is reduced, allowing the latent student traits to shine through more clearly.

In conclusion, the robust predictive power of RME and French is a data-driven argument for a more holistic understanding of academic risk. It compellingly suggests that to improve outcomes in core subjects, the educational system must also prioritise the development of the whole student, fostering the discipline, self-regulation, and cognitive flexibility that are so clearly captured by performance in these non-core proxies.

Summary

83 This chapter presented and interpreted the results of the study, confirming the achievement of both research objectives. The analysis first identified Mathematics and Science as the subjects with the lowest average performance, underscoring a persistent area of concern. Subsequently, a Random Forest machine learning model was developed, demonstrating exceptional accuracy 94% and reliability in predicting students at risk of scoring below the 50% pass mark. Furthermore, through feature importance analysis, the study revealed the counter-intuitive finding that non-core

subjects, specifically Religious and Moral Education (RME) and French, were the most significant predictors of overall academic risk, suggesting they act as proxies for essential latent traits like student discipline and cognitive skills. These results provide a powerful, data-driven foundation for shifting from reactive to proactive educational interventions and offer novel insights for targeting student support effectively.

CHAPTER FIVE

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS.

Introduction

This chapter presents the final phase of the research by providing a comprehensive summary of the entire study, which aimed to develop a machine learning model for predicting academic risk among Junior High School students in the Ledzokuku Municipality. It synthesizes the key findings derived from the analysis, draws definitive conclusions based on the evidence presented, and offers practical recommendations for educators, policymakers, and future researchers. The purpose of this chapter is to consolidate the study's contributions to addressing the problem of student underperformance and to propose a clear pathway for implementing data-driven interventions within the Ghanaian educational context.

Summary of the Study

This study demonstrates the successful development and use of a Random Forest machine learning model for the proactive identification of at-risk Junior High School students. A surprising and key finding was that non-core subjects, RME and French, were the most influential predictors of academic risk. The implications of this discovery and the suggested pathway for a proactive educational management system are discussed in detail in Chapter 5.

Conclusions

This study successfully developed a machine learning model to proactively identify at-risk JHS students in the Ledzokuku Municipality. The analysis not only confirmed the

model's high predictive accuracy but also yielded a paradigm-shifting insight into the key drivers of academic risk, moving beyond conventional understanding.

Based on the evidence, the following conclusions are drawn:

1. A Random Forest model is a highly accurate and practically viable tool for proactive student risk identification in the Ghanaian JHS context. The model achieved 94% accuracy and a critical 81% recall for the at-risk class, conclusively demonstrating that machine learning can effectively leverage existing academic data to flag students in need of support long before terminal failure occurs.
2. Academic risk is most accurately diagnosed not by core STEM subject performance, but by performance in RME and French, which act as proxies for foundational non-cognitive and cognitive skills. The most significant finding of this research is that these non-core subjects are the strongest predictors of overall academic risk. This leads to the conclusion that they serve as early indicators for underlying competencies such as discipline, conscientiousness, working memory, and cognitive flexibility, which are fundamental to success across the entire curriculum.
3. This research provides a scalable framework for shifting Ghana's educational management from a reactive, symptom-treating model to a proactive, root-cause-diagnosing one**.** The study moves beyond theory to offer a validated, data-driven blueprint. It empowers educators and policymakers to transition from reporting past failures in core subjects to proactively diagnosing the root causes of student struggle, enabling earlier, more holistic, and more effective interventions.

However, the true significance of these findings extends beyond these direct answers to the research questions. The "so what?" of this research is that it provides a

validated, scalable blueprint for a fundamental paradigm shift in educational management in Ghana and similar contexts. This study moves the discourse beyond the well-documented symptoms of educational crisis, chronically low scores in Mathematics and Science and provides a diagnostic tool to identify the root causes of student struggle much earlier.

The implications are threefold:

From Reactive to Proactive Governance: This research proves that the transition from a reactive, data-rich-but-intelligence-poor system to a proactive, intelligence-driven one is not a theoretical ideal but a practical reality. By leveraging the data the Ghana Education Service already collects, the system can now identify vulnerability before it crystallizes into failure, enabling timely interventions.

A New Lens for Educational Policy: The counter-intuitive power of RME and French challenges the prevailing, almost exclusive, policy focus on STEM remediation. It forces a critical re-evaluation, suggesting that sustainable improvement in STEM outcomes may be dependent on first strengthening the bedrock of student character and cognitive flexibility. This argues for a more holistic, integrated curriculum and intervention strategy that values the development of the whole student.

A Model for Responsible and Explainable Innovation: By embedding Explainable AI (XAI) at the core of its methodology, this study offers a model for the ethical and effective implementation of AI in education. It ensures that the technology serves to empower educators with diagnostic insights, not to replace their judgment with opaque algorithms. This builds trust and ensures that the "why" behind a prediction is always available to guide compassionate and targeted human action.

In conclusion, this project does more than just build a predictive model; it provides a coherent, data-driven framework for reimagining educational support. It argues that the path to improving educational outcomes lies not only in drilling core subject knowledge but in proactively cultivating the disciplined minds and cognitive agility of students, with RME and French scores serving as the crucial early-warning indicators. The implementation of this approach represents a tangible and necessary step toward a more equitable, efficient, and effective educational system for all Ghanaian students.

Recommendations

To translate the findings of this study into tangible improvements within the Ghanaian educational system, the following targeted and actionable recommendations are proposed for policymakers, practitioners, and future researchers.

For Policymakers (Ghana Education Service & Ministry of Education)

1. Pilot a Data-Driven Early Warning System (EWS): The GES should initiate a phased pilot program in the Ledzokuku Municipality to integrate the validated machine learning model into a user-friendly dashboard. This EWS should be formally embedded within the existing school supervision and circuit rider frameworks, mandating that headteachers and district officers use these diagnostic alerts to guide their supportive supervision and resource allocation.
2. Mandate a Holistic Intervention Policy: Move beyond a singular focus on STEM remediation. Issue a policy directive that requires schools to interpret EWS alerts holistically. Specifically, when a student is flagged as at-risk and the alert is driven by low scores in RME and/or French, intervention plans must

include strategies to develop the underlying competencies of these subjects' proxy, such as:

- Integrating explicit instruction on metacognitive and self-regulated learning strategies (e.g., goal-setting, time management, self-reflection) into the RME and French curricula.
- Promoting pedagogical techniques in French that specifically target executive function development, such as working memory games and cognitive flexibility exercises.

3. Invest in Strategic Data Literacy Training: Allocate dedicated resources for professional development programs aimed at building data literacy among headteachers and district officers. Training should focus specifically on interpreting EWS dashboards, understanding feature importance reports, and using these insights to coordinate targeted, cross-curricular student support.

For Practitioners (School Leaders and Teachers)

1. Implement a "Holistic Student Review" Protocol: Upon receiving an alert from an EWS, school leaders should initiate a structured meeting between the core subject teachers (Mathematics, Science, English) and the teachers of the predictive subjects identified by the model (RME, French). The goal of this review is to create a unified, diagnostic student support plan that addresses both the symptom (e.g., failing Math) and the potential root causes (e.g., lack of discipline, as signalled by RME, or cognitive struggles, as signalled by French).

2. Adopt Differentiated Pedagogical Strategies: Teachers of RME and French should recognise their unique role in developing foundational competencies. They should consciously employ teaching strategies that foster:
    - In RME: Conscientiousness and self-regulation through structured projects, reflective journals, and clear, consistent expectations for conduct and assignment completion.
    - In French: Cognitive flexibility and working memory through interactive language drills, pattern recognition exercises, and tasks that require mental switching between languages and grammatical structures.
3. Use the EWS for Group-Level Analysis: Identify common risk profiles at the class or school level to guide resource allocation and professional development for teachers. For example, suppose a significant number of students in a particular class or year group are flagged due to consistently low French scores. In that case, it may indicate a need for reviewing and improving foreign language teaching strategies across the board, prompting targeted training or resource provision.

For Future Research

1. Expand Predictive Feature Sets: Future studies should seek to incorporate a broader set of variables to enhance model robustness and explore new dimensions of student risk. Specific variables to collect and test include:
    - Behavioural Metrics: Student attendance records and punctuality data.

- Socioeconomic Proxies: Data on school lunch program participation or parental education levels.
- Institutional Data: Teacher qualification levels and years of experience.

2. Validate the Proxy Hypothesis with Mixed Methods: To conclusively validate the finding that RME and French act as proxies for latent traits, a mixed-methods study is recommended. This would involve:
    - Quantitatively replicating this predictive model with the expanded feature set.
    - Qualitatively conducting in-depth interviews and focus groups with students flagged by the model to explore their study habits, self-regulation strategies, and cognitive challenges, thereby providing direct evidence for or against the proxy hypothesis
3. Conduct Longitudinal and Multi-Site Studies: Research should be expanded to track student cohorts over multiple academic years to understand how risk factors evolve. Furthermore, applying the same methodology across multiple municipalities with diverse socioeconomic profiles is essential to test the generalizability and scalability of the proposed EWS framework.

Addressing the Pillars of Innovation, Reproducibility, and Impact

This capstone project is designed to be a seminal contribution to educational data science in Ghana by firmly addressing three critical pillars of high-quality research: innovation, reproducibility, and impact.

1. Innovation: Creativity and Novelty

This study moves beyond the conventional application of machine learning in education by introducing a novel, counter-intuitive diagnostic insight. While many predictive models in education focus on core STEM subjects, this research reveals that non-core subjects, Religious and Moral Education (RME) and French are the most significant predictors of academic risk. This finding is a paradigm shift, proposing that these subjects act as proxies for foundational non-cognitive (e.g., discipline, conscientiousness) and cognitive skills (e.g., working memory, cognitive flexibility). Furthermore, the integration of Explainable AI (XAI) techniques, specifically SHAP analysis, transforms the model from a "black box" into a transparent diagnostic tool, providing educators with actionable, subject-level reasons for each prediction. This creative synthesis of predictive modelling and explanatory analytics represents a significant innovation in the approach to understanding student performance.

2. Reproducibility: A Transparent and Scalable Approach

The research is grounded in a systematic and transparent methodology to ensure full reproducibility and scalability. The study provides a clearly justified, step-by-step pipeline from data preprocessing and feature engineering to model selection (Random Forest), validation (stratified train-test split), and evaluation. The use of widely available academic data (subject scores) instead of hard-to-collect demographic information makes the approach highly scalable and directly applicable to the data assets of the Ghana Education Service. The entire analytical process, including code, is documented and accessible via a GitHub repository, allowing other researchers and practitioners to replicate the study, validate its findings, and adapt the framework for other municipalities or educational contexts.

3. Impact: Real-World Relevance and Usefulness

The ultimate value of this research lies in its direct practical utility and potential for transformative change. It addresses a critical "Problem of Practice" by offering a scalable framework to shift educational management from a reactive to a proactive paradigm. The validated model and the proposed Early Warning System (EWS) provide teachers and school leaders with a practical tool for the early identification of at-risk students, enabling timely, targeted interventions. For policymakers, the findings challenge the orthodoxy of STEM-heavy interventions and provide empirical evidence for more holistic educational policies that value the development of foundational competencies. By translating data into actionable intelligence, this project has a clear pathway to improving resource allocation, enhancing student support, and ultimately, strengthening educational outcomes in Ghana and similar resource-constrained environments.

REFERENCES

Adane, M. D., Deku, J. K., & Asare, E. K. (2023). Performance analysis of machine learning algorithms in the prediction of student academic performance. *Journal of Advances in Mathematics and Computer Science, 38*(5), 74–86. https://doi.org/10.9734/jamcs/2023/v38i51762

Adu-Gyamfi, E. (2014). The effect of illegal mining on school attendance and academic performance of junior high school students in Upper Denkyira West District of Ghana. *Journal of Education and Human Development, 3*(1), 523–545.

Alshdaifat, E., Al-shdaifat, A., Zaid, A., & Aloqaily, A. (2020). The impact of data normalization on predicting student performance: A case study from Hashemite University. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(4), 4580–4588. https://doi.org/10.30534/ijatcse/2020/57942020

Anamuah-Mensah, J., Mereku, D. K., & Ghartey, A. A. (2008). *Ghana Junior Secondary School students' achievement in mathematics and science: Results from Ghana's participation in the 2007 Trends in International Mathematics and Science Study*. Ministry of Education, Youth and Sports.

Asif, R., Merceron, A., & Pathan, M. K. (2015). Predicting student academic performance at the degree level: A case study. *International Journal of Intelligent Systems and Applications, 7*(1), 49–61. https://doi.org/10.5815/ijisa.2015.01.05

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Balcıoğlu, Y. S., & Artar, M. (2023). Predicting the academic performance of students with machine learning. *Information Development, 41*(3), 896–915. https://doi.org/10.1177/02666669231213023

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

Brew, E. A., Mensah, F., Buabeng, I., Quainoo, E. A., Azewara, M. A., & Owusu, M. A. (2022). The academic performance of female students in integrated science in junior high schools: Evidence from Aowin Municipality, Western North Region of Ghana. *Asian Journal of Education and Social Studies, 27*(1), 22–31. https://doi.org/10.9734/ajess/2022/v27i130645

Chen, D., Luo, H., Liu, Z., Pan, J., Wu, Y., Wang, E., & Ou, G. (2025). A dual-variable selection framework for enhancing forest aboveground biomass estimation via multi-source remote sensing. *Remote Sensing, 17*(14), 2493. https://doi.org/10.3390/rs17142493

Danubianu, M. (2015, October). A data preprocessing framework for students' outcome prediction by data mining techniques. *2015 19th International Conference on System Theory, Control and Computing (ICSTCC)* (pp. 836–841). https://doi.org/10.1109/ICSTCC.2015.7323202

Davis, E. K., Ntow, F. D., & Beccles, C. (2022). Factors influencing Ghanaian public junior high school students' performance in English language, mathematics and science: Implications for the national policy on progression. *SAGE Open, 12*(3). https://doi.org/10.1177/21582440221123912

Ghana Education Service. (2018). *Education Sector Performance Report 2018*. Author.

Ghana Education Service. (2019). *Education Sector Performance Report 2019*. Author.

Kanabar, R., & Tawde, P. D. (2023). Leveraging machine learning for predicting student performance. *International Journal of Advanced Research in Science, Communication and Technology*, 3(1), 210–217. https://doi.org/10.48175/IJARSCT-23332

Karikari, A., Achiaa, E. A., Adu, J., & Kumi, E. O. (2020). Causes of students' poor performance in mathematics: A case of Sefwi Bonwire D/A Junior High School in the Western Region of Ghana. *International Journal of Advanced Research, 8*(9), 904–912. https://doi.org/10.21474/IJAR01/11740

Mao, H., Khanal, R., Qu, C., Kong, H., & Jiang, T. (2025). What factors enhance students' achievement? A machine learning and interpretable methods approach. *PLOSONE,20*(5), e0323345. https://doi.org/10.1371/journal.pone.0323345

Mensah, P. A. A., Denteh, M. O., Issaka, I., & Adjaah, E. (2022). Examining factors responsible for students' poor performance in mathematics, from the perspective of teachers and students at Asesewa Senior High School in the Upper Manyakrobo District. *Asian Research Journal of Mathematics, 18*(7), 1–14. https://doi.org/10.9734/ARJOM/2022/v18i730386

Nugba, R. M., Quansah, F., Ankomah, F., Tsey, E. E., & Ankoma-Sey, V. R. (2021). A trend analysis of junior high school pupils' performance in the Basic Education Certificate Examination (BECE) in Ghana. *International Journal of Elementary Education, 10*(3), 79–86. https://doi.org/10.11648/j.ijeedu.20211003.15

O'Cummings, M., & Therriault, S. B. (2015). *From accountability to prevention: Early warning systems put data to work for struggling students*. American Institutes for Research. https://www.earlywarningsystems.org/

Radhya, S., Tasik, M. A. S., Sabran, F. M., & Gunawan, A. A. S. (2022, September). Systematic literature review: Machine learning in education to predict student performance. In *2022 International Conference on Electrical and Information Technology (IEIT)* (pp. 350–356). IEEE. https://doi.org/10.1109/IEIT56384.2022.9967874

Rosado, J. T., Payne, A. P., & Rebong, C. B. (2019). eMineProve: Educational data mining for predicting performance improvement using classification methods. *IOP Conference Series: Materials Science and Engineering, 649*(1), 012018. https://doi.org/10.1088/1757-899X/649/1/012018

Sarker, S., Paul, M. K., Thasin, S. T. H., & Hasan, M. A. M. (2024). Analysing students' academic performance using educational data mining. *Computers and Education: Artificial Intelligence, 7*, 100263. https://doi.org/10.1016/j.caeai.2024.100263

Shi, H., Caskurlu, S., Zhang, N., & Na, H. (2024). To what extent has machine learning been used to predict online at-risk students? Evidence from a quantitative meta-analysis. *Journal of Research on Technology in Education*. Advance online publication. https://doi.org/10.1080/15391523.2024.2351643

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30–

40. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Siraj, F. (2016). Modelling academic achievement of UUM graduate using descriptive and predictive data mining. In R. Silhavy, R. Senkerik, Z. Oplatkova, Z. Prokopova, & P. Silhavy (Eds.), *Software engineering in intelligent systems* (pp. 511–520). Springer. https://doi.org/10.1007/978-3-319-33622-0_46

Siraj, F., & Essgaer, M. (2011). Mining enrolment data using predictive and descriptive approaches. In T. Sobh & K. Elleithy (Eds.), *Innovations and advances in computer sciences and engineering* (pp. 529–534). Springer. https://doi.org/10.1007/978-90-481-3658-2_91

Sixhaxa, K., Jadhav, A., & Ajoodha, R. (2022, January). Predicting students' performance in exams using machine learning techniques. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 635–640). IEEE. https://doi.org/10.1109/Confluence52989.2022.9734157

Thakur, D., & Kapoor, N. (2022, March). Predicting students' performance using data mining algorithms. *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)* (pp. 290–295). https://doi.org/10.1109/ICACTA54667.2022.9806073

Tiwari, M., & Jain, N. (2024). Student performance prediction using machine learning algorithms. *ShodhKosh: Journal of Visual and Performing Arts, 5*(6), 349–359. https://doi.org/10.29121/shodhkosh.v5.i6.2024.4552

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of predicting students' performance using machine learning algorithms. *International Journal of Intelligent Systems and Applications, 11*(12), 34–45. https://doi.org/10.5815/ijisa.2019.12.04

APPENDIX A

Link to GitHub for Python codes and the dataset.

https://github.com/Clementkwakuboadu/DATA-CURATION-COURSE-2025-B.git

PROACTIVE EDUCATIONAL MANAGEMENT: A RANDOM FOREST AND SHAP ANALYSIS FOR IDENTIFYING KEY PREDICTORS OF STUDENT PERFORMANCE.

ORIGINALITY REPORT

13%	10%	9%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ir.ucc.edu.gh Internet Source	1%
2	erl.ucc.edu.gh:8080 Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
5	essay.utwente.nl Internet Source	<1%
6	Richard Nyankomako Codjoe, Linda Ama Owusuaa Amoah, Solomon Kofi Amoah. "School environmental factors, pupils' characteristics, and academic performance: The case of junior high school pupils of the Krachi West District of Ghana.", Heliyon, 2024 Publication	<1%
7	www.frontiersin.org Internet Source	<1%
8	Submitted to Accra Institute of Technology Student Paper	<1%

9	www.coursehero.com Internet Source	<1 %
10	arxiv.org Internet Source	<1 %
11	"Navigating Economic Uncertainty - Vol. 2", Springer Science and Business Media LLC, 2025 Publication	<1 %
12	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1 %
13	Submitted to University of Stellenbosch, South Africa Student Paper	<1 %
14	Submitted to University of Technology, Sydney Student Paper	<1 %
15	"Non-cognitive Skills and Factors in Educational Attainment", Springer Science and Business Media LLC, 2016 Publication	<1 %
16	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications", River Publishers, 2024 Publication	<1 %
17	Submitted to University of Hertfordshire Student Paper	<1 %
18	"Intelligent Computing", Springer Science and Business Media LLC, 2021 Publication	<1 %

19	Submitted to University of Surrey Student Paper	<1 %
20	oro.open.ac.uk Internet Source	<1 %
21	Submitted to University of Cape Coast Student Paper	<1 %
22	Submitted to University of Sunderland Student Paper	<1 %
23	learnverse.live Internet Source	<1 %
24	Anita Lukić, Ivan Krešimir Lukić. "Fundamentals of artificial intelligence for nursing students: Educational innovation", Teaching and Learning in Nursing, 2025 Publication	<1 %
25	www.ijritcc.org Internet Source	<1 %
26	Courage Kamusoko. "Explainable Machine Learning for Geospatial Data Analysis - A Data-Centric Approach", CRC Press, 2024 Publication	<1 %
27	Submitted to Adventist University of Central Africa Student Paper	<1 %
28	Submitted to Coventry University Student Paper	<1 %
29	ghanafact.com Internet Source	<1 %
30	Submitted to The University of the West of Scotland Student Paper	<1 %

31	volito.digital Internet Source	<1 %
32	www.journalbinet.com Internet Source	<1 %
33	Submitted to Manchester Metropolitan University Student Paper	<1 %
34	Submitted to UNICAF Student Paper	<1 %
35	dbjournal.ro Internet Source	<1 %
36	Submitted to Asian Institute of Technology Student Paper	<1 %
37	Reshad Al Muttaki, Sadia Afrin, Alvi Ibn Amzad Anil, Mehedi Hasan Shawon. "Advancing Breast Cancer Detection: A Comprehensive Evaluation of Machine Learning Models on Mammogram Imaging", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
38	fastercapital.com Internet Source	<1 %
39	link.springer.com Internet Source	<1 %
40	www.kolena.com Internet Source	<1 %
41	www.pharmacovigilanceanalytics.com Internet Source	<1 %
42	Submitted to Ravensbourne Student Paper	<1 %

43	doi.org Internet Source	<1 %
44	www.aasmr.org Internet Source	<1 %
45	www.granthaalayahpublication.org Internet Source	<1 %
46	www.numberanalytics.com Internet Source	<1 %
47	Fatemeh Salehi, Emmanuelle Salin, Benjamin Smarr, Sara Bayat, Arnd Kleyer, Georg Schett, Ruth Fritsch-Stork, Bjoern M. Eskofier. "A robust machine learning approach to predicting remission and stratifying risk in rheumatoid arthritis patients treated with bDMARDs", Scientific Reports, 2025 Publication	<1 %
48	Submitted to National Institute of Business Management Sri Lanka Student Paper	<1 %
49	Submitted to University of Derby Student Paper	<1 %
50	Submitted to University of Edinburgh Student Paper	<1 %
51	www.nature.com Internet Source	<1 %
52	www.tnsroindia.org.in Internet Source	<1 %
53	Arun PV, Jocelyn Chanussot, B Krishna Mohan, D Nagesh Kumar, Alok Porwal. "Explainable AI for Earth Observation Data Analysis -	<1 %

	Applications, Opportunities, and Challenges", CRC Press, 2025 Publication	
54	S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and Advanced Technologies", CRC Press, 2025 Publication	<1 %
55	Submitted to TAFE NSW Higher Education Student Paper	<1 %
56	Submitted to University of Dundee Student Paper	<1 %
57	Submitted to University of Portsmouth Student Paper	<1 %
58	Submitted to Singapore Polytechnic Student Paper	<1 %
59	www.restack.io Internet Source	<1 %
60	"Advances in Machine Learning and Big Data Analytics I", Springer Science and Business Media LLC, 2025 Publication	<1 %
61	Sowmya D. S.. "chapter 4 Behavioral Data Synthesis", IGI Global, 2025 Publication	<1 %
62	Submitted to St Luke's Anglican School Student Paper	<1 %
63	hnhiring.com Internet Source	<1 %
64	robots.net Internet Source	<1 %

65	Submitted to Adama Science and Technology University Student Paper	<1 %
66	diglib.tugraz.at Internet Source	<1 %
67	www.mobileappdaily.com Internet Source	<1 %
68	www.seejph.com Internet Source	<1 %
69	Chao Liu, Shengyi Yang. "Personalized Learning Ability Classification Using SVM for Enhanced Education in System Modeling and Simulation Courses", Frontiers of Digital Education, 2025 Publication	<1 %
70	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1 %
71	Robert Chapleau, Philippe Gaudette, Tim Spurr. "Application of Machine Learning to Two Large-Sample Household Travel Surveys: A Characterization of Travel Modes", Transportation Research Record: Journal of the Transportation Research Board, 2019 Publication	<1 %
72	Siham REBBAH. "From Prediction to Action: A Calibrated and Interpretable Machine Learning Framework for Personalized Student Retention", Springer Science and Business Media LLC, 2025 Publication	<1 %
73	Submitted to University of Lancaster Student Paper	<1 %

74	ijmejournal.org Internet Source	<1 %
75	Gaetano Nucifora, Daniele Muser, Joshua Bradley, Zoi Tsoumani et al. "Unsupervised phenotypic clustering of cardiac MRI data reveals distinct subgroups associated with outcomes in ischemic cardiomyopathy", The International Journal of Cardiovascular Imaging, 2025 Publication	<1 %
76	Joyeta Ghosh, Jyoti Taneja, Ravi Kant. " Decoding Host-Pathogen Interactions in : Insights into Allelic Variation and Antimicrobial Resistance Prediction Using Artificial Intelligence and Machine Learning based approaches ", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
77	Submitted to Sheffield Hallam University Student Paper	<1 %
78	docshare.tips Internet Source	<1 %
79	ijirt.org Internet Source	<1 %
80	online-journals.org Internet Source	<1 %
81	repository.uel.ac.uk Internet Source	<1 %
82	www.preprints.org Internet Source	<1 %
83	cognizancejournal.com Internet Source	<1 %

84	jai.front-sci.com Internet Source	<1 %
85	vtechworks.lib.vt.edu Internet Source	<1 %
86	www.fastercapital.com Internet Source	<1 %
87	www.geeksforgeeks.org Internet Source	<1 %
88	www.nasia.gov.gh Internet Source	<1 %
89	C Kishor Kumar Reddy, Anindya Nag, Lavanya Pamulaparty, Mariya Ouaissa, Marlia Mohd Hanafiah. "Generative AI in Neurology - Advancing Neurodegenerative Disease Treatment", CRC Press, 2025 Publication	<1 %
90	Debani Prasad Mishra, Anmit Ray, Shashwat Singh, Bijaya Krushna Panda, Gyanabritish Nayak. "Chapter 9 Electrical Theft Detection Using CNN Algorithm", Springer Science and Business Media LLC, 2024 Publication	<1 %
91	Jiao-Ling Appels, George Martvel, Anna Zamansky, Stefanie Riemer. "Automated Facial Landmark Analysis vs. Manual Coding: Accuracy in Dog Emotional Expression Classification", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
92	Leifa Li, Wangwen Sun, Lauren Y. Gómez-Zamorano, Zhuangzhuang Liu, Wenzhen Zhang, Haoran Ma. "From Research Trend to	<1 %

	Performance Prediction: Metaheuristic-Driven Machine Learning Optimization for Cement Pastes Containing Bio-Based Phase Change Materials", Polymers, 2025 Publication	
93	Nandakumar M.K., Harilal C.C.. "Shifting Climates, Rising Tensions: Community Insights on Human-Wildlife Conflicts in Wayanad Wildlife Sanctuary, India", Springer Science and Business Media LLC, 2025 Publication	<1 %
94	docplayer.net Internet Source	<1 %
95	edutechwiki.unige.ch Internet Source	<1 %
96	eprints.gla.ac.uk Internet Source	<1 %
97	index.j-ets.net Internet Source	<1 %
98	ir.kiu.ac.ug Internet Source	<1 %
99	ir.knust.edu.gh Internet Source	<1 %
100	ir.uew.edu.gh:8080 Internet Source	<1 %
101	irbackend.kiu.ac.ug Internet Source	<1 %
102	lcjstem.com Internet Source	<1 %
103	open.uct.ac.za Internet Source	<1 %

104	www.amfiteatrueconomic.ro Internet Source	<1 %
105	"Educational Data Mining", Springer Science and Business Media LLC, 2014 Publication	<1 %
106	Ashish Juneja, Anil Joseph, Dasaka S. Murty. "GeoVadis - The Future of Geotechnical Engineering (Volume 1)", CRC Press, 2025 Publication	<1 %
107	Dothang Truong. "Demystifying AI - Data Science and Machine Learning Using IBM SPSS Modeler", CRC Press, 2025 Publication	<1 %
108	Eduardo Rodriguez. "The Analytics Process - Strategic and Tactical Steps", Routledge, 2017 Publication	<1 %
109	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	<1 %
110	Hank Bohanon, Lisa Caputo Love, Kelly Morrissey. "Implementing Systematic Interventions - A Guide for Secondary SchoolTeams", Routledge, 2020 Publication	<1 %
111	Matthew M. Rust, Benjamin A. Motz. "Incorporating an LMS learning analytic into proactive advising: Validity and use in a randomized experiment", The Internet and Higher Education, 2025 Publication	<1 %
112	Radiah Haque, Hui-Ngo Goh, Choo-Yee Ting, Albert Quek, M.D. Rakibul Hasan. "Leveraging	<1 %

	LLMs for optimised feature selection and embedding in structured data: A case study on graduate employment classification", Computers and Education: Artificial Intelligence, 2025 Publication	
113	Rejwan Bin Sulaiman, Usman Javed Butt, Yassine Maleh, Mohammad Aljaidi, Md. Simul Hasan Talukder, Musarrat Saberin Nipun. "Securing Health - The Convergence of AI and Cybersecurity in Healthcare", CRC Press, 2025 Publication	<1 %
114	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	<1 %
115	Vicente Martinez, Rodrigo Salas, Oliver Tessini, Romina Torres. "Machine Learning techniques for Behavioral Feature Selection in Network Intrusion Detection Systems", 11th International Conference of Pattern Recognition Systems (ICPRS 2021), 2021 Publication	<1 %
116	academic.oup.com Internet Source	<1 %
117	ijsred.com Internet Source	<1 %
118	jnao-nu.com Internet Source	<1 %
119	psasir.upm.edu.my Internet Source	<1 %
120	repository.daystar.ac.ke Internet Source	<1 %

121	www.ifets.info Internet Source	<1 %
122	www.medrxiv.org Internet Source	<1 %
123	www.rsisinternational.org Internet Source	<1 %
124	www.seaairweb.info Internet Source	<1 %
125	www.semanticscholar.org Internet Source	<1 %
126	www.standyou.com Internet Source	<1 %
127	"Financial Sector Development in Ghana", Springer Science and Business Media LLC, 2023 Publication	<1 %
128	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	<1 %
129	Hayo Reinders, Chun Lai, Pia Sundqvist. "The Routledge Handbook of Language Learning and Teaching Beyond the Classroom", Routledge, 2022 Publication	<1 %
130	Huijian Dong. "Data Analytics in Finance", CRC Press, 2025 Publication	<1 %
131	Randhir Kumar, Prabhat Kumar, C.C. Sobin, N.P. Subheesh. "Blockchain and AI in Shaping the Modern Education System", CRC Press, 2025	<1 %

Publication

132	"Proceedings of the 4th International Conference on Advances in Communication Technology and Computer Engineering (ICACTCE'24)", Springer Science and Business Media LLC, 2025 Publication	<1 %
133	B. K. Tripathy, Hari Seetha. "Explainable, Interpretable, and Transparent AI Systems", CRC Press, 2024 Publication	<1 %
134	Polisetty Sri Hari Sai Saran, R. Krishna Kumari. "chapter 14 Unveiling Academic Success", IGI Global, 2025 Publication	<1 %
135	V. Subramaniyaswamy, G Revathy, Logesh Ravi, N. Thillaiarasu, Naresh Kshetri. "Deep Learning and Blockchain Technology for Smart and Sustainable Cities", CRC Press, 2025 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		